Algebraic & Geometric Topology

Volume 25 (2025)

Issue 8 (pages 4437–5174)

msp

ALGEBRAIC & GEOMETRIC TOPOLOGY

msp.org/agt

See inside back cover or msp.org/agt for submission instructions.

The subscription price for 2025 is US $760/year for the electronic version, and $1110/year (+$75, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP. Algebraic & Geometric Topology is indexed by Mathematical Reviews, Zentralblatt MATH, Current Mathematical Publications and the Science Citation Index.

Algebraic & Geometric Topology (ISSN 1472-2747 printed, 1472-2739 electronic) is published 9 times per year and continuously online, by Mathematical Sciences Publishers, 2000 Allston Way # 59, Berkeley, CA 94701-4004. Periodical rate postage paid at Oakland, CA 94615-9651, and additional mailing offices. POSTMASTER: send address changes to Mathematical Sciences Publishers, 2000 Allston Way # 59, Berkeley, CA 94701-4004.

AGT peer review and production are managed by EditFlow® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

https://msp.org/

msp

Algebraic & Geometric Topology **25**:8 (2025) 4437–4497
DOI: 10.2140/agt.2025.25.4437
Published: 20 November 2025

Hierarchies for relatively hyperbolic virtually special groups

EDUARD EINSTEIN

Wise's quasiconvex hierarchy theorem classifying hyperbolic virtually compact special groups in terms of quasiconvex hierarchies played an essential role in Agol's proof of the virtual Haken conjecture. Answering a question of Wise, we construct a new virtual quasiconvex hierarchy for relatively hyperbolic virtually compact special groups. We use this hierarchy to prove a generalization of Wise's malnormal special quotient theorem for relatively hyperbolic virtually compact special groups with arbitrary peripheral subgroups.

20F65, 20F67

1 Introduction

1.1 Background, history and motivation

One of the main goals of cube complex theory is to use the geometry and combinatorial structure of cube complexes to better understand groups. The study of cubical groups has played an important role in recent developments in the theory of hyperbolic 3-manifold groups, particularly in Agol's proof of the virtual Haken conjecture [1].

Virtually special cube complexes, developed by Wise and his collaborators, are central to the theory of cubical groups. A group is called *compact virtually special* if it is the fundamental group of a compact virtually special cube complex whose hyperplanes satisfy certain combinatorial conditions. Virtually special cube complexes have desirable separability properties that allow certain immersions to be promoted to embeddings using Scott's criterion [27].

A construction in [24] due to Sageev provides a method for constructing a group action on a CAT(0) cube complex using "*codimension*-1 *subgroups*"; however, in general, this action may not be proper, cocompact, or have a virtually special quotient. For hyperbolic groups, the situation is much clearer: Bergeron and Wise [5] proved that hyperbolic groups with an ample supply of quasiconvex codimension-1 subgroups have a proper and cocompact action on a CAT(0) cube complex. The key to Agol's proof of the virtual Haken conjecture is that any geometric action of a hyperbolic group on a CAT(0) cube complex has virtually special quotient [1, Theorem 1.1]. In the case of closed 3-manifolds, the ample supply of codimension-1 subgroups comes from immersed surfaces constructed by Kahn and Markovic in [20].

Two key ingredients in Agol's theorem are Wise's quasiconvex hierarchy theorem and malnormal special quotient theorem (MSQT). Wise's quasiconvex hierarchy theorem [30, Theorem 13.3] characterizes the virtually special hyperbolic groups in terms of virtual quasiconvex hierarchies.

Definition 1.1 [30, Definition 11.5] Let $\mathcal{QVH}$ be the smallest class of hyperbolic groups closed under the following operations.

1. $\{1\} \in \mathcal{QVH}$.
2. If $G = A *_C B$ and $A, B \in \mathcal{QVH}$ and C is finitely generated and quasi-isometrically embedded in G then $G \in \mathcal{QVH}$.
3. If $G = A_{*C}$, $A \in \mathcal{QVH}$ and C is finitely generated and quasi-isometrically embedded in G, then $G \in \mathcal{QVH}$.
4. If $H \leqslant G$ with $|G : H| < \infty$ and $H \in \mathcal{QVH}$, then $G \in \mathcal{QVH}$.

In other words, groups in $\mathcal{QVH}$ are hyperbolic groups that can be built from the trivial group by taking finite index subgroups or taking amalgamations and HNN extensions over quasiconvex subgroups.

Theorem 1.2 ([30, Theorem 13.3], Wise's quasiconvex hierarchy theorem) *Let G be a hyperbolic group. Then $G \in \mathcal{QVH}$ if and only if G is virtually compact special.*

As Wise notes in [30, Section 12], the MSQT is an essential ingredient in the proof of the quasiconvex hierarchy theorem.

Theorem 1.3 (Wise's malnormal special quotient theorem [30, Theorem 12.2]) *Let G be a hyperbolic and virtually special group with G hyperbolic relative to a collection of subgroups $\{P_1, \dots, P_m\}$. Then there exist finite index subgroups $\dot{P}_i \leqslant P_i$ such that if $\overline{G} = G(N_1, \dots, N_m)$ is any peripherally finite Dehn filling with $N_i \leqslant \dot{P}_i$, then $\overline{G}$ is hyperbolic and virtually special.*

The MSQT together with virtually special amalgamation criteria from [13; 19] are used to prove Theorem 1.2.

For relatively hyperbolic groups, much less is known. Wise's methods from [30] extend to more general situations than hyperbolic groups. In particular, many of the methods for hyperbolic groups extend to finite volume hyperbolic 3-manifolds. Hsu and Wise [19] also proved a special combination result for relatively hyperbolic groups albeit with much more restrictive hypotheses.

The main goal of this paper is to prove relatively hyperbolic analogs of important ingredients in the proof of Theorem 1.2. The first result answers a question posed by Wise:

Theorem 1 *Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair and let G be a virtually compact special group. Then there exists a finite index subgroup $G_0 \leqslant G$ and an induced relatively hyperbolic group pair $(G_0, \mathcal{P}_0)$ so that G_0 has a quasiconvex, malnormal and fully $\mathcal{P}_0$-elliptic hierarchy terminating in groups isomorphic to elements of $\mathcal{P}_0$.*

Proving that the hierarchy is not only quasiconvex and *malnormal* but also *fully $\mathcal{P}_0$-elliptic* is a way of ensuring that the hierarchy is compatible with the relatively hyperbolic structure on G and allows for the use of relatively hyperbolic Dehn filling arguments. See Sections 3.2 and 3.3 for definitions of quasiconvex, malnormal and fully $\mathcal{P}_0$-elliptic hierarchies.

Theorem 1 will be used to prove a relatively hyperbolic generalization of the MSQT using relatively hyperbolic Dehn filling techniques similar to those used in [3]:

Theorem 2 *Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair with $\mathcal{P} = \{P_1, \dots, P_m\}$. If G is virtually compact special, then there exist subgroups $\{\dot{P}_i \lhd P_i\}$ where $\dot{P}_i$ is finite index in P_i such that if $\overline{G} = G(N_1, \dots, N_m)$ is any peripherally finite filling with $N_i \lhd \dot{P}_i$, then $\overline{G}$ is hyperbolic and virtually special.*

Peripherally finite fillings are defined formally in Definition 8.2. While Wise proved a generalized relatively hyperbolic version of the MSQT in [30, Theorem 15.6] for relatively hyperbolic groups with virtually abelian peripherals, Theorem 2 holds for arbitrary peripheral subgroups.

1.2 Outline

Section 2 contains a brief overview of the geometry of relatively hyperbolic groups. Section 3 covers preliminaries about graphs of groups and quasiconvex hierarchies.

Section 4 is devoted to proving a relative fellow traveling result for a CAT(0) space with a geometric action by a relatively hyperbolic group, a generalized version of quasigeodesic stability in hyperbolic spaces. The main result is Theorem 4.7. Similar results were proved by Hruska [14] and Hruska–Kleiner in [17] for CAT(0) spaces with isolated flats, and this result was previously known to experts in the field. However, it was difficult to find an exact formulation of Theorem 4.7 in the literature, so a proof is produced here.

Section 5 contains a combination lemma for certain subspaces of CAT(0) spaces with a geometric action by a relatively hyperbolic group. The main result, Theorem 5.6 shows that subspaces of such a CAT(0) space that are unions of convex cores for peripheral coset orbits and convex subspaces that obey a separation property are quasiconvex. The proof technique is inspired partly by the proof of the combination lemma in [19].

Section 6 reviews the properties of special cube complexes. In particular, Section 6.3 will introduce separability and explain how to pass to a finite cover so that each hyperplane's elevations to the universal cover obey a separation property. Section 6.4 recalls a result of Sageev and Wise [26] used to represent peripheral subgroups of a relatively hyperbolic compact special group G as immersed complexes in an NPC cube complex X with $\pi_1 X = G$.

Section 7 follows the outline of [3, Section 5] and uses Wise's double dot hierarchy construction to prove Theorem 1. While the general strategy is the same, the hyperbolic geometry used in [3] to prove the edge groups of the hierarchy are π_1-injective and quasi-isometrically embedded needs to be replaced by relatively hyperbolic geometric results from the preceding sections.

Section 8 uses Theorem 1 along with a relatively hyperbolic Dehn filling argument similar to the one used in a new proof of Wise's MSQT from [3] to prove Theorem 2, a relatively hyperbolic analog of Wise's MSQT.

Acknowledgements

The author would like to thank Jason Manning and Daniel Groves for their invaluable guidance and suggestions. Specifically, the author would like to thank Groves for explaining the proof of Proposition 7.29. The author also thanks Lucien Clavier, Yen Duong, Chris Hruska, Michael Hull and Daniel Wise for useful conversations that helped shape this work. Finally, the author thanks the anonymous referee for careful readings and suggestions that greatly helped improve the structure of this paper.

2 Relatively hyperbolic geometry

2.1 The geometry of CAT(0) spaces being acted on by relatively hyperbolic groups

In the situation where a relatively hyperbolic group acts properly and cocompactly on a CAT(0) space, it is reasonable to hope to partially recover the geometric features of a hyperbolic space. There are many equivalent definitions of a relatively hyperbolic group, see [16] for several examples; one definition, originally due to Farb [10], is produced here:

Definition 2.1 [16, Definition 3.6] Let G be finitely generated relative to $\mathcal{P}$ with each $P \in \mathcal{P}$ finitely generated. The pair $(G, \mathcal{P})$ is a *relatively hyperbolic group pair* if for some finite relative generating set S, the coned-off Cayley graph $\widehat{\Gamma}(G, \mathcal{P}, S)$ is hyperbolic and $(G, \mathcal{P}, S)$ has Farb's bounded coset penetration property (see [10, Section 3.3]).

The elements of $\mathcal{P}$ and their conjugates are called *peripheral subgroups* and the cosets $\{gP : g \in G,\ P \in \mathcal{P}\}$ are called peripheral cosets.

Definition 2.1 establishes useful notation to refer to a relatively hyperbolic group pair, but the technical details will be less useful. Instead, most of the arguments involving relatively hyperbolic groups will be made using two key properties: that coarse intersections of peripheral cosets are uniformly bounded and that triangles are *relatively thin* in a sense defined in Section 2.2.

The following fact is well known:

Proposition 2.2 *Let $(G,\mathcal{P})$ be a relatively hyperbolic group pair. Let S be a finite generating set for G. For all $R \geqslant 0$, there exists $M_R \geqslant 0$ such if gP, $g'P'$ is a pair of distinct peripheral cosets, then* $\operatorname{diam} \mathcal{N}_R(gP) \cap \mathcal{N}_R(g'P') \leqslant M_R$ *in the word metric on $\Gamma(G,S)$.*

The uniform bounds on coarse intersections of peripheral cosets transfers nicely to the case where a relatively hyperbolic group acts properly and cocompactly on a geodesic space by isometries:

Corollary 2.3 *Let G be a finitely generated group acting properly and cocompactly by isometries on a geodesic metric space X, and let $x \in X$ be a base point. If $(G,\mathcal{P})$ is a relatively hyperbolic group pair, then for all $R \geqslant 0$, there exists $M_{R,X,x} \geqslant 0$ such that if $P, P' \in \mathcal{P}$, $g, g' \in G$ with $gP \neq g'P'$, then* $\operatorname{diam} \mathcal{N}_R(gPx) \cap \mathcal{N}_R(g'P'x) \leqslant M_{R,X,x}$.

2.2 Relatively thin triangles

Comparison tripods help compare geodesic triangles in X with tripods:

Definition 2.4 Let $a,b,c \in X$ and let $\triangle abc$ be a geodesic triangle. There exists a map $h\colon \triangle abc \to T(a,b,c)$ where $T(a,b,c)$ is a unique tripod (up to isometry) with center point x such that h is isometric on each side of the triangle and the three legs of the tripod are $[h(a),x]$, $[h(b),x]$ and $[h(c),x]$. The tripod $T(a,b,c)$ is called a *comparison tripod* for $\triangle abc$. The map h is the *comparison map*.

A geodesic metric space X is *hyperbolic* if there exists a $\delta > 0$ so that for every geodesic triangle in X, the preimage of every point in the comparison map has diameter less than δ.

Definition 2.5 Let X be a geodesic metric space, and let $F \subseteq X$ be a subset of X.

Let $\triangle abc$ be a geodesic triangle in X and let $\delta > 0$. Let $T(a,b,c)$ be the comparison tripod, and let $h\colon \triangle abc \to T(a,b,c)$ be the comparison map. If, for all $p \in T(a,b,c)$,

(1) $\operatorname{diam} h^{-1}(p) < \delta$ or

(2) $h^{-1}(p) \subseteq \mathcal{N}_\delta(F)$,

then $\triangle abc$ is *δ-thin relative to F*.

Definition 2.6 Let X be a geodesic metric space, $\delta \geqslant 0$ and let $\mathcal{B}$ be a collection of subspaces. The space X has the *δ-relatively thin triangle property relative to $\mathcal{B}$* if each geodesic triangle Δ is δ-thin relative to some $F \in \mathcal{B}$.

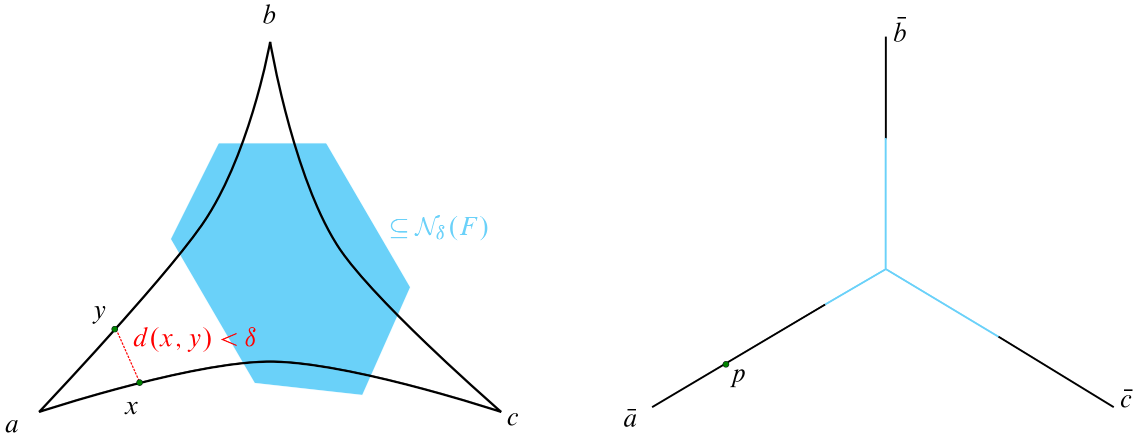


Figure 1: An example of a triangle which is δ-thin relative to some F with its comparison tripod. Points in the blue part of the tripod have preimages in the triangle which lie in the blue shaded region. All other points have preimages in the triangle with diameter δ like the point p whose preimages x, y have $d(x,y) < \delta$. The fat part (see Definition 2.10) of each side is the subsegment that intersects the blue shaded region.

See Figure 1 for an illustration of Definition 2.6.

The space X may contain triangles that are δ-thin. By definition, these triangles are δ-thin relative to every element of $\mathcal{B}$. In the applications, X will usually be a CAT(0) space with a geometric action by a relatively hyperbolic group G where the elements of $\mathcal{B}$ are convex subspaces of X that lie in uniformly bounded neighborhoods of peripheral coset orbits. If $(G,\mathcal{P})$ is a relatively hyperbolic group pair, a CAT(0) space with a geometric action by G has the relatively thin triangle property relative to $\mathcal{B} = \{gPx \mid g \in G,\ P \in \mathcal{P}\}$:

Proposition 2.7 ([26, Theorem 4.1, Proposition 4.2], see also [8, Section 8.1.3]) *Let $(G,\mathcal{P})$ be a relatively hyperbolic group pair and let G act properly and cocompactly on a* CAT(0) *space X by isometries. Let $x \in X$ be a base point and set*

$$\mathcal{B} = \{gPx \mid g \in G,\ P \in \mathcal{P}\}.$$

Then for some $\delta > 0$, X has the δ-relatively thin triangle property relative to $\mathcal{B}$.

When X has the relatively thin triangle property relative to $\mathcal{B}$, $R \geqslant 0$ and $\mathcal{B}' = \{\mathcal{N}_R(F) : F \in \mathcal{B}\}$, then X still has the relatively thin triangle property relative to $\mathcal{B}'$.

The notion of fellow traveling will be useful for describing behavior of geodesics that issue from the same point. Definitions of fellow traveling may vary, so the one that will be used is recorded here:

Definition 2.8 Let $\alpha : [a_1,a_2] \to X$ and $\beta : [b_1,b_2] \to X$ be geodesics, and let $k \geqslant 0$. The geodesics α and β *k-fellow travel for distance D* if $d(\alpha(a_1+t), \beta(b_1+t)) \leqslant k$ for all $0 \leqslant t \leqslant D$. If $x := \alpha(a_1) = \beta(b_1)$ and α and β k-fellow travel for distance D, then *α and β k-fellow travel distance D from x*.

We also introduce tails of a geodesic to help us make geometric arguments:

Definition 2.9 Let γ be a geodesic in $\widetilde{X}$, let p be an endpoint of γ, and let $k \geqslant 0$. The *k-tail of γ at p* is the geodesic subsegment of T consisting of all $x \in \gamma$ so that $d(x, p) \leqslant k$.

Definition 2.10 Let X be a CAT(0) geodesic metric space with triangles that are δ-thin relative to $\mathcal{B}$. Let $\triangle \subseteq X$ with vertices a, b, c with comparison map $h\colon \triangle abc \to T(a,b,c)$. Let L_a be the closure of the leg of the tripod $T(a,b,c)$ that contains $h(a)$. Let $\mathbf{Thin}_a := \{x \in h^{-1}(L_a) : \operatorname{diam} h^{-1}(h(x)) < \delta\}$. The *corner segments* of $\triangle$ at a are the two closures of the parts of $\mathbf{Thin}_a$ in each side and the *corner length* is the length of a corner segment at a.

The *fat part* of the side $ab \subseteq \triangle$ in $\triangle$ is $ab \setminus (\mathbf{Thin}_a \cup \mathbf{Thin}_b)$.

The corner segments at a are subsegments of the sides issuing from a that δ-fellow travel. Each of these segments have the same length, which is defined to be the corner length. If $\triangle$ is δ-thin relative to $B_\triangle \in \mathcal{B}$, the fat part of each side of $\triangle$ is the maximal subsegment that does not lie in any of the corner segments and hence lies in $\mathcal{N}_\delta(B_\triangle)$. Note that the fat part of a side may be empty. Since X is CAT(0), each corner segment or fat part of a side is connected.

A *(λ, ϵ)-quasigeodesic* in X is a (λ, ϵ)-quasi-isometric embedding of a (possibly unbounded) interval in the real line in X, see [7, Definition I.8.22] for details.

Quasigeodesic triangles in the Cayley graph of a relatively hyperbolic group also satisfy a thinness condition which is used to obtain Proposition 2.7:

Theorem 2.11 ([26, Theorem 4.1], originally due to [8]) *Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair with Cayley graph Γ. For all $\lambda \geqslant 1, \epsilon > 0$ there exists a $\delta > 0$ such that if $\triangle$ is a (λ, ϵ)-quasigeodesic triangle in Γ with sides c_0, c_1, c_2, either*

(1) *there exists a point p that lies within $\frac{\delta}{2}$ of each side or*

(2) *there is a peripheral coset gP so that each side c_i of $\triangle$ has a subpath c_i' where $c_i' \subseteq \mathcal{N}_\delta(gP)$ and the terminal endpoint of c_i' and the initial point of c_{i+1}' (indices mod 3) are within distance δ of each other.*

Lemma 2.12 is simple but is instrumental for working with relatively thin triangles.

Lemma 2.12 *Let $\widetilde{X}$ be a CAT(0) space. Let Δabc be a geodesic triangle in $\widetilde{X}$ that is δ-thin relative to F. Let ab, bc, ac denote the sides of Δabc. If the length of the fat part of ac in Δabc is bounded above by $k_{\mathrm{fat}} \geqslant 0$, then the length of the fat part of bc and the length of the fat part of ab differ by at most $k_{\mathrm{fat}} + 3\delta$.*

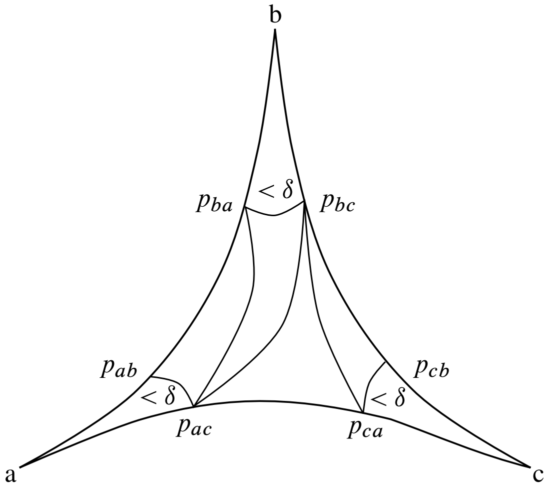


Figure 2: Applying the triangle inequality four times gives a bound on the difference between the length of $[p_{ab}, p_{ba}]$ and the length of $[p_{bc}, p_{cb}]$ in terms of $|[p_{ac}, p_{ca}]|, \delta$.

The proof involves four applications of the triangle inequality. See Figure 2 for a schematic. With Lemma 2.12, a bound on the fat part of one side of a relatively thin triangle helps control the lengths of the fat parts of the other two sides. This technique will be used repeatedly, particularly in Section 5.

Relatively hyperbolic groups interact nicely with passing to finite index subgroups:

Proposition 2.13 [3, Notation 2.9] *Let G be a group and let $\mathcal{P}$ be a finite collection of subgroups of G. Let $H \lhd G$ be a finite index normal subgroup. For each $P \in \mathcal{P}$, let $\mathcal{E}_0(P) = \{gPg^{-1} \cap H \mid g \in G\}$ and let $\mathcal{E}(P)$ be a set of representatives of H-conjugacy classes in $\mathcal{E}_0(P)$. Let $\mathcal{P}' = \bigsqcup_{P \in \mathcal{P}} \mathcal{E}(P)$.*

The pair $(G, \mathcal{P})$ is relatively hyperbolic if and only if $(H, \mathcal{P}')$ is relatively hyperbolic.

There is also a generalized version of quasiconvexity for relatively hyperbolic groups.

Definition 2.14 [16, Definition 6.10] Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair. Let $H \leqslant G$. Let S be any finite set such that $S \cup \mathcal{P}$ generates G. Suppose there exists $\kappa(S, d_S)$ such that for any $\widehat{\Gamma}(G, \mathcal{P}, S)$-geodesic γ with endpoints in H, $\gamma \cap G$ lies in $\mathcal{N}_\kappa(H)$ with respect to d_S. Then H is *relatively quasiconvex in* $(G, \mathcal{P})$.

There are other equivalent definitions which are discussed in [16]. The definition is also independent of the choice of finite relative generating set (see [16, Theorem 7.10]). Relative quasiconvexity will only be needed for the peripheral subgroups:

Proposition 2.15 *Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair. Then every element of $\mathcal{P}$ is relatively quasiconvex in G.*

Proof In $\widehat{\Gamma}(G, \mathcal{P}, S)$ every $P \in \mathcal{P}$ has diameter 1. □

3 Graphs of groups and hierarchies

3.1 Graphs of groups

A graph of groups (together with an isomorphism from the fundamental group) is a way of decomposing a group along a finite number of splittings and HNN extensions. Further decomposing the vertex groups as graphs of groups, decomposing the resulting vertex groups as a graph of groups again and continuing this process a finite number of times yields a kind of "multilevel graph of groups" called a *hierarchy* which will be defined in Definition 3.6.

Definition 3.1 A *graph of groups* (Γ, χ) consists of the following data:

1. a connected finite graph $\Gamma = \Gamma(V, E)$ where V is the vertex set of Γ and E is the oriented edge set of Γ with an involution $e \mapsto \bar{e}$ that switches the orientation of each edge,
2. an *assignment map* $\chi\colon V \sqcup E \to \mathbf{Grp}$ that assigns a group to each vertex and edge,
3. for all $e \in E$, $\chi(e) = \chi(\bar{e})$,
4. *attachment homomorphisms* $\psi_e\colon \chi(e) \to \chi(t(e))$ where $t(e)$ is the terminal vertex of the edge e.

Γ is a *faithful* graph of groups if the attachment homomorphisms ψ_e are injective.

A *graph of spaces* is constructed like a graph of groups, except that the assignment map χ assigns a (path connected) topological space instead of a group to each edge and vertex. The attachment homomorphisms are replaced by continuous *attachment maps*, and a *faithful graph of spaces* has π_1-injective attachment maps. A *graph of spaces realization of a space* X for a graph of spaces (Γ, χ) is a triple (Γ, χ, q) where q is a homotopy equivalence from X to the mapping cylinders of the attachment maps glued along vertex spaces.

Some authors, for example Wise and Serre, take faithfulness to be a part of the definition of a graph of groups. Not requiring faithfulness makes it easier to define graphs of groups in terms of graphs of spaces. For the applications in Section 7, graphs of groups will be constructed first without showing that they are faithful, but these graphs of groups will turn out to be faithful.

If (Γ, χ) is a graph of groups, and T is a maximal tree in Γ, then $\pi_1(\Gamma, T)$ will denote the *fundamental group of the graph of groups* Γ *with respect to the tree* T. See [28] for further details about graphs of groups.

A *graph of groups structure* is the group-theoretic analog of a graph of spaces realization:

Definition 3.2 Let G be a group, let (Γ, χ) be a graph of groups where T is a maximal tree and let $\phi\colon G \to \pi_1(\Gamma, T)$ be an isomorphism. The triple (Γ, ϕ, T) is a *graph of groups structure on* G.

The structure (Γ, ϕ, T) is *degenerate* if Γ is a single vertex labeled with G and ϕ is the identity.

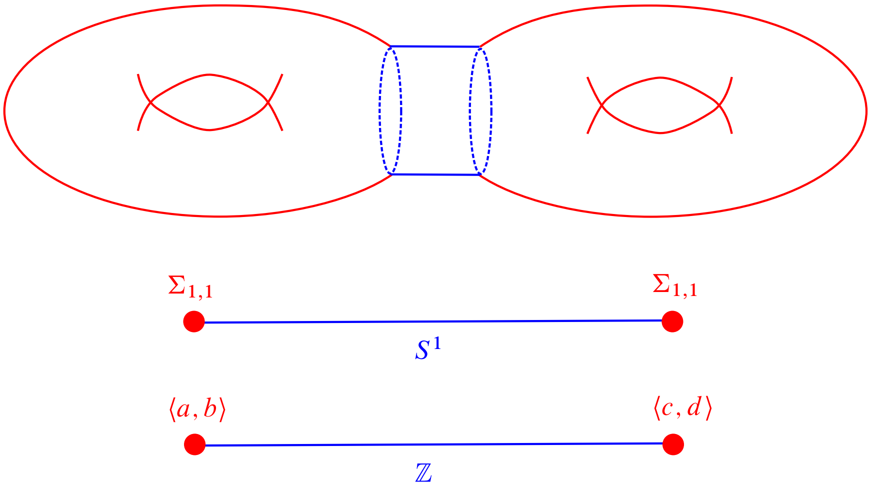


Figure 3: A graph of spaces realization of a genus-2 surface where $\Sigma_{1,1}$ is a punctured torus, together with the corresponding graph of groups obtained by applying the π_1 functor.

While a graph of groups structure determines a splitting of G, the choice of isomorphism and maximal tree affects the precise splitting. In many cases, it suffices to give a splitting of G up to conjugacy which will be the case in the examples below. When the splitting is given up to conjugacy, the choice of maximal tree also becomes unnecessary.

Example 3.3 Figure 3 shows a graph of spaces decomposition of a genus-2 surface and a graph of groups splitting of the fundamental group induced by the graph of spaces decomposition.

Example 3.4 If Σ_g is a closed surface of genus g, then a pants decomposition of Σ_g induces a splitting of $\pi_1\Sigma_g$ as a graph of groups where the vertex groups are isomorphic to a free group of rank 2 and the edge groups are infinite cyclic groups.

Graph of groups structures interact naturally with finite index normal subgroups. The following is [3, Proposition 3.18] but is originally due to Bass [4].

Proposition 3.5 *Suppose G has a graph of groups structure (Γ, ϕ, T), $H \lhd G$ and H is finite index in G. Then H has an induced graph of groups structure $(\widetilde{\Gamma}, \widetilde{\phi}, T')$ so that:*

(1) *Every vertex group of $(\widetilde{\Gamma}, T')$ has the form $(K^g \cap H) \lhd K^g$ and is finite index in K^g for some vertex group K of (Γ, T) and some $g \in G$.*

(2) *Every edge group of $(\widetilde{\Gamma}, T')$ has the form $(K^g \cap H) \lhd K^g$ and is finite index in K^g for some edge group K of (Γ, T) and some $g \in G$.*

3.2 Hierarchies

Hierarchies of groups are inductively defined multilevel graphs of groups:

Definition 3.6 A *hierarchy of groups of length* 0 is a single vertex labeled by a group.

A *hierarchy of groups of length* n is a graph of groups (Γ_n, χ_n) together with hierarchies of length $n-1$ on each vertex of Γ_n.

If $\mathcal{H}$ is a length-n hierarchy of groups, the n^{th} *level* of $\mathcal{H}$ is the graph of groups Γ_n. For $1 \leqslant k \leqslant n$, the $(n-k)^{\text{th}}$ level of $\mathcal{H}$ is the disjoint union of the $(n-k)^{\text{th}}$ levels of the hierarchies on the vertices of Γ_n.

The *terminal groups* are the groups labeling the vertices at level 0.

It will be useful to think of graphs of groups as length-1 hierarchies. Realizing a group as a hierarchy is similar to finding a graph of groups structure for that group:

Definition 3.7 Let G be a group, $\mathcal{H}$ be a hierarchy of length n. Let (Γ_n, χ_n) be the level-n graph of groups. When $n = 0$, a *hierarchy for* G is a single vertex labeled by G. If $n \geqslant 1$, a *hierarchy for* G is $\mathcal{H}$ together with a graph of groups structure (Γ_n, ϕ, T) for G so that for every vertex v of Γ_n, the hierarchy on length $n-1$ on v is a hierarchy for the vertex group $\chi_n(v)$. Let $\mathcal{P}$ be a collection of subgroups of G. The hierarchy structure *terminates in* $\mathcal{P}$ if every terminal group of $\mathcal{H}$ is conjugate to $\phi(\mathcal{P})$ for some $P \in \mathcal{P}$.

It will often be convenient to forget the choice of maximal tree and only give a hierarchy structure for a group up to conjugacy. In general, hierarchies will be allowed to contain degenerate splittings, but in order to obtain nontrivial results, it will be necessary to ensure that at least one of the splittings in the hierarchy is nondegenerate.

Wise's hierarchies in [30] permit only one-edge splittings rather than allowing a graph of groups splitting for each vertex group in the hierarchy. The hierarchies in Definition 3.7 can be converted to hierarchies with one-edge splittings for each vertex group at the expense of increasing the length of the hierarchy. Wise's hierarchies also terminate in the trivial group while Definition 3.7 allows arbitrary terminal groups. In practice, the goal in Section 7 will be to (virtually) find a hierarchy for a relatively hyperbolic group $(G, \mathcal{P})$ that terminates in groups isomorphic to those in the induced peripheral structure. Section 8 will explore what happens to the hierarchy after quotienting out finite index subgroups of the peripheral subgroups.

A *hierarchy of spaces* and a *hierarchy realization for a space* X can be defined analogously by replacing groups in Definition 3.6 with topological spaces and replacing graph of groups structures by realizations in Definition 3.7.

Malnormality is an important group property which will play a role in Section 8 and is useful for amalgamating virtually special groups to make new virtually special groups (see [19]).

Definition 3.8 Let G be a group and let $H \leqslant G$. The subgroup H is *malnormal in* G if for all $g \in G \setminus H$, $g^{-1}Hg \cap H = \{1\}$. Similarly, H is *almost malnormal in* G if for all $g \in G \setminus H$, $|g^{-1}Hg \cap H| < \infty$.

Malnormality also extends to collections of subgroups. Let $\mathcal{P}$ be a collection of subgroups of G. The collection $\mathcal{P}$ is (almost) malnormal in G if for all $g \in G$ and $P, P' \in \mathcal{P}$ either $g^{-1}Pg \cap P'$ is trivial (finite) or $P = P'$ and $g \in P$.

For example, if $(G, \mathcal{P})$ is a relatively hyperbolic group pair and G is finitely generated, then the collection $\mathcal{P}$ is almost malnormal in G by Proposition 2.2.

Definition 3.1 (graphs of groups) and Definition 3.6 (hierarchies) are very flexible, but in practice, some further restrictions will be needed to ensure that graphs of groups and hierarchies produce useful splittings:

Definition 3.9 Let (Γ, χ) be a faithful graph of groups and let (Γ, ϕ) be a graph of groups structure (up to conjugacy) for a group G.

1. Γ is *quasiconvex* if every edge attachment map is a quasi-isometric embedding into $\pi_1(\Gamma)$.
2. Γ is (*almost*) *malnormal* if for every $e \in E$, the image of the attachment homomorphism ψ_e in $\pi_1(\Gamma)$ is (almost) malnormal in $\pi_1(\Gamma)$.

Let $\mathcal{H}$ be a hierarchy for G.

1. $\mathcal{H}$ is *faithful* if every graph of groups at every level of $\mathcal{H}$ is faithful.
2. $\mathcal{H}$ is *quasiconvex* if every edge group of every graph of groups at every level of $\mathcal{H}$ quasi-isometrically embeds in G.
3. $\mathcal{H}$ is (*almost*) *malnormal* if every edge group of every graph of groups at every level of $\mathcal{H}$ is (almost) malnormal in G.

It may be possible to give a reasonable weaker definition of quasiconvex (or malnormal) hierarchy by only requiring an edge group G_e of a graph of groups H in $\mathcal{H}$ to be quasi-isometrically embedded (malnormal) in each adjacent vertex group, but the stronger definition given here will be needed in Section 8.

Here are some examples to help illustrate the definition of a hierarchy:

Example 3.10 A splitting of the fundamental group of a hyperbolic surface group can be realized along quasiconvex infinite cyclic subgroups by using a pants decomposition. The splitting can be achieved either as a sequence of 1-edge splittings to create a hierarchy or can be achieved a single multiedge graph of groups splitting.

There are iterated hierarchy splittings that cannot be realized by a single graph of groups splitting:

Example 3.11 Figure 4 shows a length-2 hierarchy for the fundamental group of a genus-2 surface, Σ_2. Cuts are made along the both the blue and green simple closed curves which intersect, so the iterated splitting of the fundamental group cannot be accomplished by a graph of groups (length-1 hierarchy).

Other notable examples of hierarchies are the Haken hierarchy for Haken 3-manifolds, see [22, Section 9.4], and the Magnus–Moldvanskii hierarchy for one-relator groups, see [30, Chapter 19].

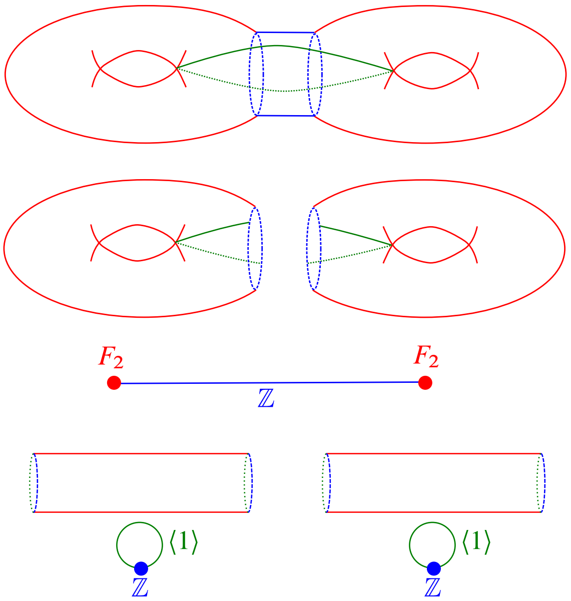


Figure 4: A hierarchy for $\pi_1(\Sigma_2)$, the fundamental group of a genus-2 surface Σ_2, where the iterated splitting of $\pi_1(\Sigma_2)$ cannot be realized by a graph of groups. The first splitting is over the infinite cyclic subgroup of $\pi_1(\Sigma_2)$ corresponding to one of the blue copies of S^1. The resulting vertex spaces are punctured tori whose fundamental groups are rank-2 free groups. Cutting along the green arc in each punctured torus makes an annulus. Then the fundamental group of a punctured torus splits as an HNN extension of the fundamental group of an annulus ($\mathbb{Z}$) over the trivial group (corresponding to the green arcs in each annulus which are glued together to make a punctured torus).

Proposition 3.5 extends to hierarchies by induction on the length of the hierarchy.

Corollary 3.12 *Suppose G has a hierarchy $\mathcal{H}$ and H is a finite index normal subgroup of G. Then $\mathcal{H}$ has an induced hierarchy $\mathcal{H}'$ such that the length of $\mathcal{H}$ is the length of $\mathcal{H}'$ and*:

1. *Every vertex group at level i of the hierarchy $\mathcal{H}'$ is of the form $K^g \cap H$ which is finite index and normal in K^g for some vertex group K of $\mathcal{H}$ at level i and some $g \in G$.*
2. *Every edge group at level i of the hierarchy $\mathcal{H}'$ is of the form $K^g \cap H$ which is finite index and normal in K^g for some edge group K of $\mathcal{H}$ at level i and some $g \in G$.*

Lemma 3.13 follows from Corollary 3.12:

Lemma 3.13 *If $\mathcal{H}$ is a quasiconvex hierarchy for G and G_0 is a finite index normal subgroup of G, then the induced hierarchy on $\mathcal{H}_0$ on G_0 is quasiconvex.*

The definition of a quasiconvex hierarchy for a group G only requires that the edge groups are quasi-isometrically embedded in G; when a graph of groups (Γ, ϕ, T) structure for G is quasiconvex, the vertex groups are quasi-isometrically embedded as well.

Lemma 3.14 *Let (Γ, T) be a graph of groups structure for G. If the edge groups of Γ are quasi-isometrically embedded in G, then the vertex groups of Γ are quasi-isometrically embedded in G.*

Here is a rough sketch of the proof of Lemma 3.14. A Cayley graph $\Lambda(G,S)$ of G coarsely looks like a "tree of spaces" whose underlying (infinite) graph is the covering tree of (Γ, T) where the edge spaces are Cayley graphs of edge groups and the vertex spaces are Cayley graphs of vertex groups. If $\Lambda_v := \Lambda(G_v, S_v)$ is one of the vertex spaces, the coarse tree structure ensures that if a $\Lambda(G,S)$-geodesic shortcut γ between two points in Λ_v exits Λ_v through an edge space Λ_e, it must return through Λ_e. If γ enters and exits Λ_v at points $p_{e_1}, p'_{e_1}, \dots, p_{e_m}, p'_{e_m}$, let γ_i be the image (in $\Lambda(G,S)$) of a Λ_e-geodesic between p_{e_i} and p'_{e_i}. There exist $\lambda \geqslant 1$ and $\epsilon > 0$ so that every γ_i is (λ, ϵ)-quasigeodesic in $\Lambda(G,S)$. We can build a new path ρ from γ by replacing the subsegment of γ from p_{e_i} to p'_{e_i} with γ_i. Then ρ lies entirely in the image of Λ_v and hence ρ is at least as long as the Λ_v-distance between its endpoints. Now the length of ρ is at most $\lambda|\gamma| + \epsilon$, or equivalently, $|\gamma| \geqslant \frac{1}{\lambda}|\rho| - \epsilon$. Thus γ cannot be much shorter than the shortest path in Λ_v between the endpoints of γ.

3.3 Fully $\mathcal{P}$-elliptic hierarchies

Given a relatively hyperbolic group pair $(G, \mathcal{P})$ and a hierarchy $\mathcal{H}$ for G, the goal in Section 8 will be to strategically find a quotient of G that has a hierarchy induced by $\mathcal{H}$ and inherits a relatively hyperbolic structure from $(G, \mathcal{P})$ that is also compatible with the induced hierarchy structure. Theorem 1.2 can then be used to show the resulting quotient is virtually special. To ensure that this happens, some additional restrictions must be imposed on the interactions between the edge and vertex groups of the hierarchy and the peripheral subgroups of G.

Definition 3.15 Let $\mathcal{H}$ be a hierarchy for a group G and let $\mathcal{P}$ be a collection of subgroups of G. Let $\mathcal{V}$ be the vertex groups of $\mathcal{H}$. For each $H \in \mathcal{V}$, let $\pi_1(\Gamma_H, \phi_H, T_H)$ be the graph of groups structure for H induced by the hierarchy $\mathcal{H}$. The hierarchy $\mathcal{H}$ is *$\mathcal{P}$-elliptic* if whenever there exists a $g \in G$ such that $P^g := gPg^{-1} \subseteq H \in \mathcal{V}$, then there exists an $h \in H$ such that $hP^g h^{-1}$ is contained in some vertex group of Γ_H. A $\mathcal{P}$-elliptic hierarchy is *fully $\mathcal{P}$ elliptic* if whenever E is an edge group in $\mathcal{H}$, then for all $g \in G$, either $P^g \cap E$ is finite or $P^g \leqslant E$.

When $\mathcal{H}$ is a fully $\mathcal{P}$-elliptic hierarchy for G and G_0 is a finite index normal subgroup of G, the induced hierarchy from Corollary 3.12 for H is also fully $\mathcal{P}$-elliptic in the induced peripheral structure provided by Proposition 2.13:

Proposition 3.16 *Suppose that G_0 is finite index normal in G and let $(G_0, \mathcal{P}_0)$ be the peripheral structure induced on G_0 by Proposition 2.13. If G has a fully $\mathcal{P}$-elliptic hierarchy, then the induced hierarchy $\mathcal{H}_0$ of G_0 is fully $\mathcal{P}_0$-elliptic.*

Proposition 3.16 follows immediately from the explicit characterizations of the edge and vertex groups of the induced hierarchies in Corollary 3.12 and from the explicit description of the induced peripheral structure.

4 The relative fellow traveling property

4.1 CAT(0) relatively hyperbolic pairs

The main result of the section is Theorem 4.7. In [14], Hruska proved that piecewise Euclidean 2-complexes satisfy a relative form of quasigeodesic stability called the *relative fellow traveling property*. In [17, Proposition 4.1.6], Hruska and Kleiner showed that CAT(0) spaces with isolated flats have the relative fellow traveling property relative to the isolated flats. Earlier, Epstein proved a version of relative fellow traveling for truncated hyperbolic spaces associated to finite volume cusped hyperbolic manifolds [9, Theorem 11.3.1]. Theorem 4.7 is a version of relative fellow traveling for CAT(0) spaces with a proper cocompact action by a relatively hyperbolic group. Theorem 4.7 is presumed to be known to experts based on the works of [8; 14; 15; 17] and others, but the exact formulation used here proved difficult to find in the literature. Therefore, a proof is provided here.

Definition 4.1 Let $\widetilde{X}$ be a CAT(0) space, let $\delta \geqslant 0$, let $f\colon \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$ be a function and let $\mathcal{B}$ be a collection of subsets of $\widetilde{X}$. The pair $(\widetilde{X}, \mathcal{B})$ is a (δ, f)-CAT(0) *relatively hyperbolic pair* if

(1) every geodesic triangle in $\widetilde{X}$ is δ-thin relative to some $F \in \mathcal{B}$,

(2) for all $r \geqslant 0$ and $F_1, F_2 \in \mathcal{B}$ with $F_1 \neq F_2$, $\operatorname{diam} \mathcal{N}_r(F_1) \cap \mathcal{N}_r(F_2) \leqslant f(r)$.

We say that a (δ, f)-CAT(0) relatively hyperbolic pair has *the L-quasiconvexity property* if there exists $L \geqslant 0$ so that each $F \in \mathcal{B}$ is *L-quasiconvex* in the sense that any $\widetilde{X}$-geodesic with endpoints in F lies in $\mathcal{N}_L(F)$. The subspaces $\mathcal{B}$ are called *peripheral spaces*.

An immediate consequence of CAT(0) geometry is the following useful fact that we will use repeatedly:

Observation 4.2 If $\widetilde{Y}$ is an L-quasiconvex subspace of a CAT(0) space $\widetilde{X}$, then for any $R \geqslant 0$, $\mathcal{N}_R(\widetilde{Y})$ is also L-quasiconvex. In other words, if $x, y \in \mathcal{N}_R(\widetilde{Y})$, then any geodesic between x, y lies in $\mathcal{N}_{R+L}(\widetilde{Y})$.

Definition 4.3 Let $(\widetilde{X}, \mathcal{B}_0)$ be a (δ, f_0)-CAT(0) relatively hyperbolic pair, and let $R \geqslant 0$. An *R-thickening of $\mathcal{B}_0$* is a collection, $\mathcal{B}$, of subspaces of $\widetilde{X}$ so that there exists a bijection $B_0 \in \mathcal{B}_0 \leftrightarrow B \in \mathcal{B}$ where $B_0 \subseteq B$, and $B \subseteq \mathcal{N}_R(B_0)$.

Proposition 4.4 *Let $(\widetilde{X}, \mathcal{B}_0)$ be a (δ, f_0)-CAT(0) relatively hyperbolic pair, and let $\mathcal{B}$ be an R-thickening of $\mathcal{B}_0$. Let $f(r) = f_0(r + R)$. Then $(\widetilde{X}, \mathcal{B})$ is a (δ, f)-CAT(0) relatively hyperbolic pair.*

Proof Let $F_1, F_2 \in \mathcal{B}$ with $F_1 \neq F_2$. Then there exist $F_{1,0}, F_{2,0} \in \mathcal{B}_0$ so that $F_1 \subseteq \mathcal{N}_R(F_{1,0})$ and $F_2 \subseteq \mathcal{N}_R(F_{2,0})$. Then

$$\operatorname{diam} \mathcal{N}_r(F_1) \cap \mathcal{N}_r(F_2) \leqslant f(r).$$

A geodesic triangle $\triangle$ in $\widetilde{X}$ is δ-relatively thin relative to some F_0 in $\mathcal{B}_0$. Since F_0 is contained in some $F \in \mathcal{B}$ element, $\triangle$ is δ-relatively thin relative to F. □

Definition 4.5 (similar to [17, Definition 4.1.4]) Let $(\widetilde{X}, \mathcal{B})$ be a (δ, f)-CAT(0) relatively hyperbolic pair. The pair $(\widetilde{X}, \mathcal{B})$ has the *relative fellow traveling property* if for all $\lambda \geqslant 1$ and $\epsilon \geqslant 0$, there exist $U, V \geqslant 0$ depending on λ, ϵ such that for any (λ, ϵ)-quasigeodesics $\sigma \colon [0, t_\sigma] \to \widetilde{X}$ and $\gamma \colon [0, s_\gamma] \to \widetilde{X}$ with the same endpoints, there exist partitions

$$0 = s_0 \leqslant s_1 \leqslant \cdots \leqslant s_{2n+1} = s_\gamma \quad \text{and} \quad 0 = t_0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_{2n+1} = t_\sigma$$

such that

(1) for all i, $d(\gamma(s_i), \sigma(t_i)) \leqslant U$,

(2) if i is even, then $d_{\mathrm{Haus}}\big(\gamma([s_i, s_{i+1}]), \sigma([t_i, t_{i+1}])\big) \leqslant U$ or

(3) if i is odd, $\gamma([s_i, s_{i+1}]), \sigma([t_i, t_{i+1}]) \subseteq \mathcal{N}_V(F_i)$ for some $F_i \in \mathcal{B}$.

For a fixed (λ, ϵ), we say that (λ, ϵ)-quasigeodesics *(U, V)-fellow travel relative to $\mathcal{B}$*.

All the CAT(0) relatively hyperbolic pairs we consider in later sections are of the form considered in the next proposition:

Proposition 4.6 *Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair so that G acts geometrically on a* CAT(0) *cube complex $\widetilde{X}$. Let $x \in \widetilde{X}$ be a basepoint. Let $\mathcal{B}_\mathcal{P} = \{gPx : g \in G,\ P \in \mathcal{P}\}$, and let $\mathcal{B}$ be any R-thickening of $\mathcal{B}_\mathcal{P}$. There exist $\delta, L(R) \geqslant 0$ and $f \colon \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$ so that $(\widetilde{X}, \mathcal{B})$ is a (δ, f)-*CAT(0) *relatively hyperbolic pair that has the $L(R)$-quasiconvexity property.*

Proof By [26, Theorem 1.1], for each $P \in \mathcal{P}$, the convex hull of Px lies in a bounded neighborhood of Px. Since $\mathcal{P}$ is finite, there is an $L \geqslant 0$ so that the convex hull of gPx lies in $\mathcal{N}_L(gPx)$. Thus any geodesic between points in gPx lies in $\mathcal{N}_L(gPx)$. By Observation 4.2, any R-thickening will have the $(L+R)$-quasiconvexity property because the R-neighborhood of each $B \in \mathcal{B}$ is L-quasiconvex. Let B_{gP} be the convex hull of $gPx \in \mathcal{B}_\mathcal{P}$. Since $\mathcal{P}$ is finite, there is an R (independent of g, P) so that each $B_{gP} \subseteq \mathcal{N}_R(gPx)$. Hence $\mathcal{B} = \{B_{gP} : g \in G,\ P \in \mathcal{P}\}$ is an R-thickening of $\mathcal{B}_\mathcal{P}$. By Proposition 4.4, it suffices to show that there exist $\delta \geqslant 0$ and $f_\mathcal{P} \colon \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$ so that $(\widetilde{X}, \mathcal{B}_\mathcal{P})$ is a $(\delta, f_\mathcal{P})$-CAT(0) relatively hyperbolic pair. Proposition 2.7 implies Definition 4.1(1) holds. Corollary 2.3 ensures that Definition 4.1(2) holds. □

Theorem 4.7 *Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair where G acts geometrically on a* CAT(0) *space $\widetilde{X}$ with basepoint $x \in \widetilde{X}$. If $\mathcal{B}$ is any R-thickening of $\{gPx \mid g \in G,\ P \in \mathcal{P}\}$ then $(\widetilde{X}, \mathcal{B})$ has the relative fellow traveling property.*

The remainder of this section is devoted to the proof of Theorem 4.7. The proof of Theorem 4.7 is completely self-contained, so a reader who is not interested in the technical details may wish to skip to the next section. We now set the following standing hypotheses for the remainder of Section 4:

Hypotheses 4.8 Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair where G acts geometrically on a CAT(0) cube complex $\widetilde{X}$. Fix a basepoint x and let $\mathcal{B}$ be an R-thickening of $\{gPx \mid g \in G,\ P \in \mathcal{P}\}$. Fix $\delta \geqslant 0$, $L \geqslant 0$ and $f : \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$ so that $(\widetilde{X}, \mathcal{B})$ is a (δ, f)-CAT(0) relatively hyperbolic pair with the L-quasiconvexity property.

4.2 Some geometric features of $(\widetilde{X}, \mathcal{B})$ under Hypotheses 4.8.

In this section, we establish some geometric facts about the (δ, f)-CAT(0) relatively hyperbolic pair $(\widetilde{X}, \mathcal{B})$.

Definition 4.9 Let $(\widetilde{X}, \mathcal{B})$ be a (δ, f)-CAT(0) relatively hyperbolic pair. Let $\gamma \subseteq \widetilde{X}$ and let $\mu \geqslant 0$. The *μ-saturation of γ* (with respect to $\mathcal{B}$) is

$$\mathrm{Sat}_\mu(\gamma) = \bigcup \{B \in \mathcal{B} : \gamma \cap \mathcal{N}_\mu(B) \neq \varnothing\}.$$

In the following, γ will usually be a quasigeodesic.

The following is a consequence of [8, Lemma 8.10] and the Milnor–Švarc lemma:

Proposition 4.10 *Under Hypotheses 4.8, for every $\lambda \geqslant 1$ and $\epsilon \geqslant 0$, there exists $u_{\lambda,\epsilon}$ so that if γ, σ are (λ, ϵ)-quasigeodesics with the same endpoints, then*

$$\sigma \subseteq \mathcal{N}_{u_{\lambda,\epsilon}}(\gamma) \cup \Bigl(\bigcup_{F \in \mathrm{Sat}_{u_{\lambda,\epsilon}}(\gamma)} \mathcal{N}_{u_{\lambda,\epsilon}}(F) \Bigr).$$

Definition 4.11 Let $\widetilde{X}$ be a geodesic metric space and let $\mathcal{B}$ be a collection of subspaces of $\widetilde{X}$. Let $B \in \mathcal{B}$, $\lambda \geqslant 1$ and $\epsilon > 0$. Let $\triangle$ be a (λ, ϵ)-quasigeodesic triangle. Let $\gamma_1, \gamma_2, \gamma_3$ be the sides of $\triangle$. We say that $\triangle$ is *coarsely ξ-thin relative to $F \in \mathcal{B}$* if

(1) there exists a point $p \in \widetilde{X}$ so that $d(p, \gamma_1), d(p, \gamma_2), d(p, \gamma_3) < \frac{\xi}{2}$ or

(2) there exist subpaths $c_i \subseteq \gamma_i$ so that $c_i \subseteq \mathcal{N}_\xi(F)$ and the distance between the terminal point of c_i and the initial point of c_{i+1} (where indices are taken mod 3) is less than ξ.

Theorem 2.11 and the Milnor–Švarc lemma imply:

Proposition 4.12 *With Hypotheses 4.8, for all $\lambda \geqslant 1$ and $\epsilon \geqslant 0$, there exist $\delta_{\lambda,\epsilon}$ so that if $\triangle$ is a (λ, ϵ)-quasigeodesic triangle, then there is an $F_\triangle \in \mathcal{B}$ so that $\triangle$ is coarsely $\delta_{\lambda,\epsilon}$-thin relative to $F_\triangle$.*

To simplify the proof of relative fellow traveling, we can make the following reduction:

Proposition 4.13 *Assume Hypotheses 4.8. To show that $(\widetilde{X}, \mathcal{B})$ has the relative fellow traveling property, it suffices to prove Definition 4.5 holds in the special case that γ is geodesic.*

The proof of Proposition 4.13 is essentially identical to the reduction step in [14, proof of Theorem 13.1].

Proposition 4.10 suggests it might be possible for a quasigeodesic to remain far from a geodesic with the same endpoints by passing from one peripheral space to another. However, Lemma 4.14 shows that such a quasigeodesic must always come close to the geodesic with the same endpoints when transitioning from one peripheral space to another:

Lemma 4.14 *Given $\mu \geqslant 0$, $\lambda \geqslant 1$ and $\epsilon > 0$, there exists $D_\cap(\mu, \lambda, \epsilon) \geqslant \mu$ so that if σ is a (λ, ϵ)-quasigeodesic, γ is a geodesic with the same endpoints as σ, and $\sigma(t) \in \mathcal{N}_\mu(F_1) \cap \mathcal{N}_\mu(F_2)$ for some distinct $F_1, F_2 \in \operatorname{Sat}_\mu(\gamma)$, then $\sigma(t) \in \mathcal{N}_{D_\cap(\mu,\lambda,\epsilon)}(\gamma)$.*

Proof There exist $p_1, p_2 \in \gamma$ so that $p_i \in \mathcal{N}_\mu(F_i)$. Let τ_1, τ_2 be geodesics so that τ_i joins $\sigma(t)$ to p_i. By Observation 4.2 and the L-quasiconvexity of F_i, $\tau_i \subseteq \mathcal{N}_{\mu+L}(F_i)$. Let $\triangle$ be the geodesic triangle with sides τ_1, τ_2 and the subpath of γ joining p_1 to p_2. Then $\triangle$ is δ-thin relative to some $F \in \mathcal{B}$.

Recall corner segments and fat parts of relatively thin triangles from Definition 2.10. Let τ_1' and τ_2' be the corner segments of $\triangle$ at $\sigma(t)$. Observe that $\tau_1' \subseteq \mathcal{N}_{\mu+L}(F_1) \cap \mathcal{N}_{\mu+L+\delta}(F_2)$, so $|\tau_1'| = |\tau_2'| \leqslant f(\mu+L+\delta)$.

Up to exchanging the indices of F_1, F_2, we may assume that $F \neq F_1$.

The fat part of τ_1 in $\triangle$ lies in $\mathcal{N}_\delta(F) \cap \mathcal{N}_{\mu+L}(F_1)$, so it has length at most $f(\mu+L+\delta)$. The fat part of τ_1 also intersects $\mathcal{N}_\delta(\gamma)$. Therefore, $d(\sigma(t), \gamma) \leqslant 2f(\mu+L+\delta)+\delta$.

If necessary, we may enlarge $D_\cap(\mu, \lambda, \epsilon)$ to ensure $D_\cap(\mu, \lambda, \epsilon) \geqslant \mu$. □

4.3 Relative fellow traveling

Hypotheses 4.15 For the following subsection, we adopt the following baseline hypotheses in addition to Hypotheses 4.8:

(1) Fix $\lambda \geqslant 1$ and $\epsilon \geqslant 0$.

(2) Let $\sigma\colon [0, t_\sigma] \to \widetilde{X}$ be a (λ, ϵ)-quasigeodesic triangle and let $\gamma\colon [0, s_\gamma] \to \widetilde{X}$ be a geodesic that has the same endpoints as σ.

(3) Enlarge δ from Hypotheses 4.8 so that all (λ, ϵ)-quasigeodesic triangles are coarsely δ-relatively thin relative to some $F \in \mathcal{B}$ (recall Definition 4.11 and Proposition 4.12) and all geodesic triangles are δ-relatively thin relative to some $F \in \mathcal{B}$.

(4) Let $u = u_{\lambda,\epsilon}$ as in Proposition 4.10.

(5) We abuse notation slightly and use $D_\cap = D_\cap(u+\epsilon+1, \lambda, \epsilon)$ (see Lemma 4.14). Note that $D_\cap \geqslant u+\epsilon+1 \geqslant u$.

(6) Let $\epsilon' = \epsilon + 2D_\cap$.

(7) Choose $D \gg \delta_{\lambda,\epsilon'} + \epsilon'$ where $\delta_{\lambda,\epsilon'}$ is a constant such that all (λ, ϵ')-quasigeodesic triangles are coarsely $\delta_{\lambda,\epsilon'}$-thin relative to some $F \in \mathcal{B}$ (recall Proposition 4.12).

(8) Let $\ell \geqslant f(D)$.

We first obtain a stability result for (λ,ϵ)-quasigeodesics with endpoints in $\mathcal{N}_q(F)$ for some $F \in \mathcal{B}$:

Proposition 4.16 *Let $q \geqslant 0$. There exists $K(q) \geqslant 0$ so that if $\alpha\colon [a_1,a_2] \to \widetilde{X}$ is a (λ,ϵ)-quasigeodesic with $\alpha(a_1),\alpha(a_2) \in \mathcal{N}_q(F)$ for some $F \in \mathcal{B}$, then $\alpha([a_1,a_2]) \subseteq \mathcal{N}_{K(q)}(F)$.*

Proof Let $\beta\colon [b_1,b_2] \to \widetilde{X}$ be a geodesic with $\beta(b_1)=\alpha(a_1)$ and $\beta(b_2)=\alpha(a_2)$. Since $\mathcal{N}_q(F)$ is L-quasiconvex by Observation 4.2, $\beta \subseteq \mathcal{N}_{L+q}(F)$. Let $y=\alpha(x)$ for some $a_1 \leqslant x \leqslant a_2$. Let $\alpha_l=\alpha([a_1,x])$ and let $\alpha_r=\alpha([x,a_2])$. The sides α_l, α_r, β define a (λ,ϵ)-quasigeodesic triangle that is coarsely thin relative to some $F' \in \mathcal{B}$.

If there exist p, $\alpha(a_l)$, $\alpha(a_r)$, and $\beta(x_b) \in \beta$ so that $d(p,\alpha(a_l)), d(p,\alpha(a_r)), d(p,\beta(x_b)) \leqslant \frac{\delta}{2}$, then $|x-a_l| \leqslant |a_l-a_r| \leqslant \lambda(\delta+\epsilon)$. Then

$$d(\beta(b),y) \leqslant d(\alpha(a_l),y)+d(\alpha(a_l),\beta(x_b)) \leqslant \lambda(|x-a_l|)+\epsilon+\delta \leqslant \lambda^2\delta+\lambda\epsilon+\epsilon+\delta.$$

If $F=F'$, then there exist $a_l \leqslant x \leqslant a_r$ so that $\alpha(a_l),\alpha(a_r) \in \mathcal{N}_\delta(F)$ and $d(\alpha(a_l),\alpha(a_r)) \leqslant \delta$. Hence $|a_l-x| \leqslant |a_l-a_r| \leqslant \lambda(\delta+\epsilon)$. Then $d(\alpha(a_l),y) \leqslant \lambda^2\delta+\lambda^2\epsilon+\epsilon$, so $y \in \mathcal{N}_{\delta+\lambda^2\delta+\lambda^2\epsilon+\epsilon}(F)$.

Finally, if $F \neq F'$, then there exist a_l, a_r, b_l, b_r with $a_l \leqslant x \leqslant a_r$ so that $d(\alpha(a_l),\beta(b_l)) \leqslant \delta$, $(\alpha(a_r),\beta(b_r)) \leqslant \delta$ and $\beta([b_l,b_r]) \subseteq \mathcal{N}_{q+L}(F) \cap \mathcal{N}_\delta(F')$. Therefore

$$d(\alpha(a_l),\alpha(a_r)) \leqslant d(\beta(b_l),\beta(b_r))+2\delta \leqslant f(q+L+\delta)+2\delta.$$

Following computations similar to those in the previous cases,

$$\begin{aligned}|a_l-x| \leqslant |a_l-a_r| &\leqslant \lambda(f(q+L+\delta)+2\delta)+\epsilon,\\ d(\alpha(a_l),y) &\leqslant \lambda^2(f(q+L+\delta)+2\delta)+\lambda^2\epsilon+\epsilon,\\ d(\beta(b_l),y) &\leqslant \lambda^2(f(q+L+\delta)+2\delta)+\lambda^2\epsilon+\epsilon+\delta.\end{aligned}$$

Therefore, $y \in \mathcal{N}_{q+L+\lambda^2(f(q+L+\delta)+2\delta)+\lambda^2\epsilon+\epsilon+\delta}(F)$. Taking $K(q)$ to be the maximum of the constants generated in the three cases yields an appropriate constant. □

Here is a brief overview of our strategy for the rest of this section:

(1) We will partition $[0,t_\sigma]$ into subintervals so that on each subinterval either σ is near an element of $\mathcal{B}$ or σ does not stay close to any element of $\mathcal{B}$ for long (Proposition 4.17).

(2) In Lemma 4.18, we alter our partition of $[0,t_\sigma]$ by widening the intervals where σ remains near some element of F so that σ is near γ at the endpoints of these intervals. In exchange, we need to calculate looser upper bounds (Proposition 4.19) on how close σ is to an element of $\mathcal{B}$ on these intervals.

(3) On what remains of the subintervals where σ is not near an element of $\mathcal{B}$, we prove that σ lies within bounded Hausdorff distance of a part of γ (Proposition 4.21).

(4) We use this information to find subintervals of $[0, s_\gamma]$ that cover $[0, s_\gamma]$ where γ is either close to an element of $\mathcal{B}$ or within bounded Hausdorff distance of σ. However, these subintervals may overlap. In Propositions 4.22 and 4.24, we show that overlapping can be controlled.

(5) In Propositions 4.25 and 4.26, we rearrange the interval endpoints and delete some subintervals of $[0, t_\sigma]$ and $[0, s_\gamma]$ to eliminate any overlap and use the bounds found in Propositions 4.22, 4.24 and 4.25 to ultimately construct a partition that witnesses relative fellow traveling.

In the following, we will use superscripts to help track the stages of partitioning and repartitioning $[0, t_\sigma]$ and covering $[0, s_\gamma]$ by subintervals.

Proposition 4.17 *There exists a partition* $0 = t_0^0 \leqslant t_1^0 \leqslant t_2^0 \leqslant \cdots \leqslant t_{2n+1}^0 = t_\sigma$ *and* $F_0, F_1, \ldots, F_{n-1} \in \mathcal{B}$ *with the following properties*:

(1) $\operatorname{diam}\{t \in [t_{2i}^0, t_{2i+1}^0] : \sigma(t) \in \mathcal{N}_D(F)\} \leqslant \ell$ *for all* $F \in \mathcal{B}$.

(2) $\sigma(t_{2i+1}^0), \sigma(t_{2i+2}^0) \in \mathcal{N}_{D+\epsilon}(F_i)$.

(3) *For all* $F \in \mathcal{B}$, *there do not exist* $t_F^- < t_{2i+1}^0 \leqslant t_{2i+2}^0 < t_F^+$ *so that* $\sigma(t_F^-), \sigma(t_F^+) \in \mathcal{N}_{u+\epsilon}(F)$.

(4) $F_j \neq F_k$ *for* $j \neq k$.

It turns out the choice of ℓ is somewhat arbitrary, but it does affect how much the partition produced by Proposition 4.17 will need to be altered to give partitions of $[0, t_\sigma]$ and $[0, s_\gamma]$ that witness relative fellow traveling.

Proof Let $m \in \mathbb{N}$ so that $(m-1)\ell \leqslant t_\sigma < m\ell$. We proceed by induction on m.

If $|t_\sigma| < \ell$, then setting $t_0^0 = 0$ and $t_1^0 = t_\sigma$ suffices.

Assume that Proposition 4.17 holds for quasigeodesics parameterized over intervals of length less than $(m-1)\ell$. Find $0 \leqslant t_- \leqslant t_+ \leqslant t_\sigma$ so that $|t_+ - t_-|$ realize $\sup_{F \in \mathcal{B}}\{|a-b| : \sigma(a), \sigma(b) \in \mathcal{N}_D(F)\}$. If $|t_+ - t_-| < \ell$, then $t_0^0 = 0$ and $t_1^0 = t_\sigma$ suffices.

Otherwise, by the inductive hypothesis, we obtain partitions

$$0 = t_0^0 \leqslant t_1^0 \leqslant \cdots \leqslant t_{2j+1}^0 = t_- \quad \text{and} \quad t_+ = t_{2j+2}^0 \leqslant t_{2j+3}^0 \leqslant \cdots \leqslant t_{2n+1}^0 = t_\sigma$$

so that $\operatorname{diam}_{t \in [t_{2i}^0, t_{2i+1}^0]}\{\sigma(t) \in \mathcal{N}_D(F)\} \leqslant \ell$ for all $F \in \mathcal{B}$, $|t_{2i+2} - t_{2i+1}| \geqslant \ell$ and $\sigma(t_{2i+1}^0), \sigma(t_{2i+2}^0) \in \mathcal{N}_{D+\epsilon}(F_i)$ for some $F_i \in \mathcal{B}$. Combining these partitions into a partition of $[0, t_\sigma]$ immediately satisfies the first two requirements. We obtain item (3) because $D \geqslant \epsilon' \geqslant D_\cap \geqslant u + \epsilon$ (recall Hypotheses 4.15), the inductive hypothesis and $|t_+ - t_-|$ is determined by a supremum. However, we need to check that if $k_1 \leqslant j$ and $k_2 \geqslant j$ (with $k_1 \neq k_2$), then $F_{k_1} \neq F_{k_2}$. If $F_{k_1} = F_{k_2}$, then there exist $t_l < t_- < t_+ < t_r$ so that $\sigma(t_l), \sigma(t_r) \in \mathcal{N}_D(F_{k_1})$ with $|t_l - t_r| > |t_- - t_+| \geqslant \ell$, contradicting hypothesis (3). □

In Proposition 4.17, it is not guaranteed that the $\sigma(t_j^0)$ are near γ. To remedy this, we widen the intervals $[t_{2i+1}^0, t_{2i+2}^0]$ as necessary while shrinking $[t_{2i}^0, t_{2i+1}^0]$:

Lemma 4.18 *For* $0 \leqslant j \leqslant 2n+1$, *there exist* t_j^1 *so that*:

1. *For all* $0 \leqslant i \leqslant n$, $\operatorname{diam}\{t \in [t_{2i}^1, t_{2i+1}^1] : \sigma(t) \in \mathcal{N}_D(F)\} \leqslant \ell$ *for all* $F \in \mathcal{B}$.
2. $0 = t_0^1 \leqslant t_1^1 \leqslant \cdots \leqslant t_{2n}^1 \leqslant t_{2n+1}^1 = t_\sigma$.
3. $t_{2i}^0 \leqslant t_{2i}^1 \leqslant t_{2i+1}^1 \leqslant t_{2i+1}^0$.
4. *Either* $t_{2i}^1 = t_{2i+1}^1$ *or* $d(\sigma(t_{2i}^1), \gamma), d(\sigma(t_{2i+1}^1), \gamma) \leqslant D_\cap$.
5. $|t_{2i+1}^1 - t_{2i+1}^0|, |t_{2i+2}^1 - t_{2i+2}^0| \leqslant \ell$.

Proof For each i, we perform the following procedure to set t_{2i}^1. Consider $p_i = \sigma(t_{2i}^0)$. By Proposition 4.10, either $p_i \in \mathcal{N}_u(\gamma)$ or $p_i \in \mathcal{N}_u(F)$ for some $F \in \operatorname{Sat}_u(\gamma)$ (where u is as defined in Hypotheses 4.15). In the first case, we set $t_{2i}^1 = t_{2i}^0$ noting that $u \leqslant D_\cap$.

Suppose we are in the second case: let $t_{\mathrm{ext}}^+ = \sup\{t \in [t_{2i}^0, t_{2i+1}^0] : \sigma(t) \in \mathcal{N}_u(F)\}$. Then $|t_{\mathrm{ext}}^+ - t_{2i}^0| \leqslant \ell$ by Proposition 4.17. One of the following holds:

- $t_{\mathrm{ext}}^+ = t_{2i+1}^0$ and $\sigma(t_{\mathrm{ext}}^+) \in \mathcal{N}_{u+\epsilon}(F)$ because t_{ext}^+ is a supremum.
- $\sigma(t_{\mathrm{ext}}^+) \in \mathcal{N}_{u+\epsilon+1}(\gamma)$.
- $\sigma(t_{\mathrm{ext}}^+) \in \mathcal{N}_{u+\epsilon+1}(F')$ for some $F' \in \mathcal{F}$ with $F' \neq F$.

Indeed, if $t_{\mathrm{ext}}^+ \neq t_{2i+1}^0$, then Proposition 4.10 and the fact that t_{ext}^+ is a supremum ensure either the second or third possibility must hold. In the case that $t_{\mathrm{ext}}^+ = t_{2i+1}^0$, set $t_{2i+1}^1 = t_{2i}^1 = t_{2i+1}^0$. Otherwise, set $t_{2i}^1 = t_{\mathrm{ext}}^+$. In this case, either $\sigma(t_{2i}^1)$ lies in $\mathcal{N}_{D_\cap}(\gamma)$ directly or Lemma 4.14 with $\mu = u + \epsilon + 1$ (recall Hypotheses 4.15(5)) implies that $\sigma(t_{2i}^1) \in \mathcal{N}_{D_\cap}(\gamma)$.

Proceeding similarly, if $\sigma(t_{2i+1}^0) \in \mathcal{N}_{u+\epsilon+1}(\gamma)$, we set $t_{2i+1}^1 = t_{2i+1}^0$. Otherwise, $\sigma(t_{2i+1}^1) \in \mathcal{N}_u(G)$ for some $G \in \operatorname{Sat}_u(\gamma)$. We then set $t_{2i+1}^1 = \inf\{t \in [t_{2i}^1, t_{2i+1}^1] : \gamma(t) \in \mathcal{N}_u(G)\}$ where $G \in \operatorname{Sat}_u(\gamma)$. As in the preceding argument, $|t_{2i+1}^1 - t_{2i+1}^0| \leqslant \ell$ and one of the following holds: $t_{2i+1}^1 = t_{2i}^1$ so that $\sigma(t_{2i+1}^1) \in \mathcal{N}_{D_\cap}(\gamma)$, $\sigma(t_{2i+1}^1)$ immediately lies in $\mathcal{N}_{D_\cap}(\gamma)$ or there exists $G' \in \operatorname{Sat}_u(\gamma)$ so that $\sigma(t_{2i+1}^1) \in \mathcal{N}_{u+\epsilon+1}(G') \cap \mathcal{N}_{u+\epsilon}(G) \subseteq \mathcal{N}_{D_\cap}(\gamma)$. In the third case, the final containment follows from Lemma 4.14 and Hypotheses 4.15(5).

Since $[t_{2i}^1, t_{2i+1}^1] \subseteq [t_{2i}^0, t_{2i+1}^0]$, we automatically retain the property that

$$\operatorname{diam}\{t \in [t_{2i}^1, t_{2i+1}^1] : \sigma(t) \in \mathcal{N}_D(F)\} \leqslant \ell$$

for all $F \in \mathcal{B}$. □

We now show that $\sigma([t_{2i+1}^1, t_{2i+2}^1])$ remains boundedly close to F_i.

Proposition 4.19 *There exists* $D_{\mathrm{depth}} \geqslant 0$ *so that for all* $0 \leqslant i \leqslant n$, $\sigma([t_{2i+1}^1, t_{2i+2}^1]) \subseteq \mathcal{N}_{D_{\mathrm{depth}}}(F_i)$, *and if* $t_{2i}^1 = t_{2i+1}^1$, $d(\sigma(t_{2i}^1), \gamma) \leqslant f(D_{\mathrm{depth}}) + D_\cap$.

Proof Since $\sigma(t^0_{2i+1}) \in \mathcal{N}_{D+\epsilon}(F_i)$ and $|t^0_{2i+1} - t^1_{2i+1}| \leqslant \ell$, $\sigma(t^1_{2i+1}) \in \mathcal{N}_{D+\epsilon+\lambda\ell+\epsilon}(F_i)$. Similarly, $\sigma(t^1_{2i+2}) \in \mathcal{N}_{D+\epsilon+\lambda\ell+\epsilon}(F_i)$. Set $D_{\mathrm{depth}} = K(D+\lambda\ell+2\epsilon)$ where $K(D+\lambda\ell+2\epsilon)$ is determined (as a function of λ, ϵ, ℓ) as in Proposition 4.16.

Now suppose $t^1_{2i} = t^1_{2i+1}$. By Proposition 4.10, if $t^1_{2i} \notin \mathcal{N}_u(\gamma)$, there exists $F \in \mathcal{B}$ so that $\sigma(t^1_{2i}) \in \mathcal{N}_u(F)$.

Suppose first that $F \neq F_i$. Let $t_F = \sup\{t \in [0, t_\sigma] : \sigma([t^1_{2i}, t]) \subseteq \mathcal{N}_u(F)\}$. By Lemma 4.18(3), $t^1_{2i} \leqslant t^0_{2i+1} \leqslant t^0_{2i+2} \leqslant t^1_{2i+2}$. Then Proposition 4.17(3) implies $t_F \leqslant t^1_{2i+2}$. Moreover, $F \neq F_i$ implies that $d(\sigma(t^1_{2i+1}), \sigma(t_F)) \leqslant f(D_{\mathrm{depth}})$. Since t_F is a supremum, there exists a $t > t_F$ with $d(\sigma(t), \sigma(t_F)) \leqslant \epsilon+1$ so that $\sigma(t) \in \mathcal{N}_u(\gamma)$ or $\sigma(t) \in \mathcal{N}_u(F')$ for some $F' \neq F$. Hence by Lemma 4.14, $d(\sigma(t_F), \gamma) \leqslant D_\cap$. Therefore, $d(\sigma(t^1_{2i}), \gamma) \leqslant f(D_{\mathrm{depth}}) + D_\cap$.

For the case $F \neq F_{i-1}$ set $t_F = \inf\{t \in [0, t_\sigma] : \sigma([t, t^1_{2i}]) \subseteq \mathcal{N}_u(F)\}$ and then proceed using a similar argument to the case $F \neq F_i$. □

We apply the bounds from Lemma 4.18 and Proposition 4.19 to obtain the following.

Corollary 4.20 *Let* $D_{\mathrm{endpoints}} = f(D_{\mathrm{depth}}) + D_\cap \geqslant 0$. *Then* $d(\sigma(t^1_j), \gamma) \leqslant D_{\mathrm{endpoints}}$.

We now find s^1_i in $[0, s_\gamma]$ so that $\gamma(s^1_i)$ is close to $\sigma(t^1_i)$. Let $0 \leqslant s^1_j \leqslant s_\gamma$ be such that $d(\gamma(s^1_j), \sigma(t^1_j))$ is at most $D_{\mathrm{endpoints}}$ if $t^1_j = t^1_{j\pm1}$ or $D_\cap$ otherwise. If $t^1_{2i} = t^1_{2i+1}$, ensure that $s^1_{2i} = s^1_{2i+1}$. We may further assume that $s^1_0 = t^1_0 = 0$, $t^1_{2n+1} = t_\sigma$ and $s^1_{2n+1} = s_\gamma$.

Proposition 4.21 *There exists* $D_{\mathrm{hausdorff}}$ *so that* $d_{\mathrm{haus}}(\sigma([s^1_{2i}, s^1_{2i+1}]), \gamma([t^1_{2i}, s^1_{2i+1}])) \leqslant D_{\mathrm{hausdorff}}$ *for all* $0 \leqslant i \leqslant n$.

Proof If $t^1_{2i+1} = t^1_{2i}$, then $D_{\mathrm{hausdorff}} = D_{\mathrm{endpoints}}$ suffices. Otherwise, Lemma 4.18 implies

$$d(\sigma(t^1_{2i}), \gamma(s^1_{2i})), d(\sigma(t^1_{2i+1}), \gamma(s^1_{2i+1})) \leqslant D_\cap.$$

Recall from Hypotheses 4.15 that $\epsilon' = \epsilon + 2D_\cap$. Construct σ_i, a (λ, ϵ')-quasigeodesic from $\sigma([t^1_{2i}, t^1_{2i+1}])$ by adding geodesics of length at most $D_\cap$ connecting $\sigma(t^1_{2i})$ and $\sigma(t^1_{2i+1})$ to $\gamma(s^1_{2i})$ and $\gamma(s^1_{2i+1})$, respectively.

Let $y \in \sigma_i$. Partition σ_i into σ_l and σ_r so that σ_l is from $\gamma(s^1_{2i})$ to y and σ_r is from y to $\sigma(s^1_{2i+1})$. The triangle bounded by $\gamma([s^1_{2i}, s^1_{2i+1}])$, σ_l and σ_r is $\delta_{\lambda,\epsilon'}$-coarsely thin relative to some $F \in \mathcal{B}$.

There are two possibilities:

Case (there exist points $p_l \in \sigma_l$, $p_r \in \sigma_r$ and p_γ in γ so that $d(p_l, p_r), d(p_r, p_\gamma), d(p_l, p_\gamma) \leqslant \delta_{\lambda,\epsilon'}$) Since σ_i is quasigeodesic, $d(y, p_l) \leqslant \lambda(\lambda(\delta_{\lambda,\epsilon'} + \epsilon')) + \epsilon'$ (a similar computation was carried out in more detail in the proof of Proposition 4.16). Then $d(y, \gamma) \leqslant d(y, p_\gamma) \leqslant \delta_{\lambda,\epsilon'} + \lambda(\lambda(\delta_{\lambda,\epsilon'} + \epsilon')) + \epsilon'$.

Case (there exist $p_l, p_{l,\gamma} \in \sigma_l$, $p_r \in \sigma_r$ and $F \in \mathcal{B}$ so that the interval of σ_l between p_l and $p_{l,\gamma}$ lies in $\mathcal{N}_{\delta_{\lambda,\epsilon'}}(F)$, $d(p_l, \gamma([s^1_{2i}, s^1_{2i+1}])) \leqslant \delta_{\lambda,\epsilon'}$ and $d(p_r, p_l) \leqslant \delta_{\lambda,\epsilon'}$) Recall that

$$\operatorname{diam}\{t \in [t^1_{2i}, t^1_{2i+1}] : \sigma(t) \in \mathcal{N}_D(F)\} \leqslant \ell$$

so $d(p_l, p_{l,\gamma}) \leqslant \lambda\ell + 3\epsilon'$ where the additional $2\epsilon'$ is accounting for the length of the segment linking $\gamma(s^1_{2i})$ to $\sigma(t^1_{2i})$ and the segment linking $\gamma(s^1_{2i+1})$ to $\sigma(t^1_{2i+1})$. We have that $d(y, p_l) \leqslant \lambda(\lambda(\delta_{\lambda,\epsilon'} + \epsilon')) + \epsilon'$ following the computation from the previous case. Hence

$$d(y, \gamma) \leqslant \delta_{\lambda,\epsilon'} + \ell + \epsilon' + \lambda(\lambda(\delta_{\lambda,\epsilon'} + \epsilon')) + \epsilon'.$$

From the two previous cases, we determine that $d(y, \gamma)$ is bounded as a function of λ, ϵ'.

Now consider $x \in \gamma$. We will bound $d(x, \sigma)$. Similar to the previous case, divide $\gamma|_{[s^1_{2i}, s^1_{2i+1}]}$ into two segments γ_l from $\gamma(s^1_{2i})$ to x and γ_r from x to $\gamma(s^1_{2i+1})$ and consider the quasigeodesic triangle with sides $\gamma_l, \gamma_r, \sigma_i$ that is $\delta_{\lambda,\epsilon'}$-coarsely thin relative to some $F \in \mathcal{B}$. There are two possibilities:

Case (there exist x_l, x_r, x_σ so that $x_l \in \gamma_l$, $x_r \in \gamma_r$, $x_\sigma \in \sigma_i$ with $d(x_l, \sigma), d(x_r, x_l) \leqslant \delta_{\lambda,\epsilon'}$) Then $d(x_l, x) \leqslant d(x_l, x_r) \leqslant \delta_{\lambda,\epsilon'}$ because γ is geodesic. Hence we have

$$d(x, \sigma_i) \leqslant d(x, x_\sigma) \leqslant d(x, x_l) + d(x_l, x_\sigma) \leqslant 2\delta_{\lambda,\epsilon'}.$$

Thus $d(x, \sigma) \leqslant 2\delta_{\lambda,\epsilon'} + D_\cap$.

Case (there exist x_l, x_r, x_σ and $F \in \mathcal{B}$ so that $x_l \in \gamma_l$, $x_r \in \gamma_r$, $p_{\sigma_l}, p_{\sigma_r} \in \sigma_i$ so that $p_{\sigma_l}, p_{\sigma_r} \in \mathcal{N}_{\delta_{\lambda,\epsilon'}}(F)$ and $d(x_l, p_{\sigma_l}), d(x_r, p_{\sigma_r}) \leqslant \delta_{\lambda,\epsilon'}$) Since $d(p_{\sigma_l}, \sigma), d(p_{\sigma_r}, \sigma) \leqslant \epsilon'$, there exist t_l, t_r so that $p_{\sigma_l} = \sigma(t_l)$, $p_{\sigma_r} = \sigma(t_r) \in \mathcal{N}_{\delta_{\lambda,\epsilon'}+\epsilon'}(F)$. Then by Proposition 4.17 and Lemma 4.18 and the fact that $D \gg \delta_{\lambda,\epsilon'} + \epsilon'$, we have $|t_l - t_r| \leqslant \ell$. It follows that $d(\sigma(t_l), \sigma(t_r)) \leqslant \lambda\ell + \epsilon'$. Hence

$$\begin{aligned} d(x, \sigma(t_l)) &\leqslant d(x_l, x_r) + d(x_l, p_{\sigma_l}) + d(p_{\sigma_l}, \sigma(t_l)) \leqslant d(x_l, x_r) + \delta_{\lambda,\epsilon'} + \epsilon' \\ &\leqslant d(\sigma(t_l), \sigma(t_r)) + 2\epsilon' + 3\delta_{\lambda,\epsilon'} \leqslant \lambda\ell + 3\epsilon' + 3\delta_{\lambda,\epsilon'}. \end{aligned}$$

Taking the largest constant from the four cases above yields an acceptable value for $D_{\mathrm{hausdorff}}$. □

Unfortunately, it is possible that $j < k$ and $s^1_j > s^1_k$, but this behavior can be controlled:

Proposition 4.22 *There exists D_{outorder} so that if $j < k$ and $s^1_j > s^1_k$, then $|t_j - t_k| \leqslant D_{\mathrm{outorder}}$.*

Proof It suffices to consider the case where k is the largest index such that $j < k$ and $s^1_j > s^1_k$.

By construction, $d(\sigma(t^1_j), \gamma(s^1_j)) \leqslant D_{\mathrm{endpoints}}$. Since k is largest, $s^1_j \in [s^1_k, s^1_{k+1}]$ where $s^1_k \leqslant s^1_{k+1}$. Since $s^1_0 = 0$ and $s^1_{2n} = s_\gamma$, there exists $h_- < j$ so that s^1_k lies in $[s^1_{h_-}, s^1_{h_-+1}]$ where $s^1_{h_-} \leqslant s^1_{h_-+1}$.

Case (k is even) Then $d(\gamma(s^1_j), \sigma(t^1_j)) \leqslant D_{\mathrm{endpoints}}$ and there exists $t_+ \in [t^1_k, t^1_{k+1}]$ such that

$$d(\gamma(s^1_j), \sigma(t_+)) \leqslant D_{\mathrm{hausdorff}}.$$

Hence $d(\sigma(t^1_j), \sigma(t_+)) \leqslant D_{\mathrm{hausdorff}} + D_{\mathrm{endpoints}}$. We then obtain

$$|t^1_k - t^1_j| \leqslant |t^1_j - t_+| \leqslant \lambda(D_{\mathrm{hausdorff}} + D_{\mathrm{endpoints}}) + \epsilon.$$

Case (h_- is even) Then $d(\gamma(s_k^1),\sigma(t_k^1)) \leqslant D_{\text{endpoints}}$ and there exists $t_- \in [t_{h_-}^1, t_{h_-+1}^1]$ so that

$$d(\sigma(t_-),\gamma(s_k^1)) \leqslant D_{\text{hausdorff}}.$$

Similar to the previous case, we conclude

$$|t_{h_-}^1 - t_{h_-+1}^1| \leqslant |t_k^1 - t_j^1| \leqslant \lambda(D_{\text{hausdorff}} + D_{\text{endpoints}}) + \epsilon.$$

Case (h_- and k are both odd) Set $h_- = 2i_- + 1$ and $k = 2i_+ + 1$. Observe that $\gamma([s_{h_-}^1, s_{h_-+1}^1]) \subseteq \mathcal{N}_{D_{\text{endpoints}}+D_{\text{depth}}}(F_{i_-})$ and similarly $\gamma([s_k^1, s_{k+1}^1]) \subseteq \mathcal{N}_{D_{\text{endpoints}}+D_{\text{depth}}}(F_{i_+})$.

We have $s_{h_-}^1 \leqslant s_k^1 < s_j^1 \leqslant s_{k+1}^1$. If $s_{h_-}^1 \leqslant s_k^1 \leqslant s_j^1 \leqslant s_{h_-+1}^1, s_{k+1}^1$, then

$$\gamma([s_k^1, s_j^1]) \subseteq \mathcal{N}_{D_{\text{endpoints}}+D_{\text{depth}}}(F_{i_-}) \cap \mathcal{N}_{D_{\text{endpoints}}+D_{\text{depth}}}(F_{i_+}).$$

Therefore,

$$d(\gamma(s_k^1),\gamma(s_j^1)) \leqslant f(D_{\text{endpoints}} + D_{\text{depth}}) \quad \text{and} \quad d(\sigma(t_j^1),\sigma(t_k^1)) \leqslant 2D_{\text{endpoints}} + f(D_{\text{endpoints}} + D_{\text{depth}}).$$

Then

$$|t_j^1 - t_k^1| \leqslant \lambda(2D_{\text{endpoints}} + f(D_{\text{endpoints}} + D_{\text{depth}})) + \epsilon.$$

Otherwise $s_{h_-}^1 \leqslant s_k^1 \leqslant s_{h_-+1}^1 \leqslant s_j^1 \leqslant s_{k+1}^1$ so that

$$\gamma([s_k^1, s_{h_-+1}^1]) \subseteq \mathcal{N}_{D_{\text{endpoints}}+D_{\text{depth}}}(F_{i_-}) \cap \mathcal{N}_{D_{\text{endpoints}}+D_{\text{depth}}}(F_{i_+}).$$

We see $d(\sigma(t_k^1),\sigma(t_{h_-+1}^1)) \leqslant 2D_{\text{endpoints}} + f(D_{\text{endpoints}} + D_{\text{depth}})$. Recalling $h_- < j$, then

$$|t_j^1 - t_k^1| \leqslant |t_{h_-+1}^1 - t_k^1| \leqslant \lambda(2D_{\text{endpoints}} + f(D_{\text{endpoints}} + D_{\text{depth}})) + \epsilon.$$

Taking D_{outorder} to be the maximum of the bounds found in each of the three cases therefore suffices. □

Definition 4.23 An *augmented partition* of $[0, t_\sigma]$ is a partition

$$0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_m = t_\sigma$$

together with choices $0 = s_0, s_1, s_2, \ldots, s_m = s_\gamma$ where $s_i \in [0, s_\gamma]$. We denote such an augmented partition by

$$(t_0, s_0) \leqslant (t_1, s_1) \leqslant \cdots \leqslant (t_{m-1}, s_{m-1}) \leqslant (t_m, s_m). \tag{1}$$

We call $t_j \leqslant t_{j+1} \leqslant \cdots \leqslant t_k$ a *maximal crossover subinterval* of the augmented partition (1) if $s_h < s_j$ for all $h \leqslant j$ and k is the largest index so that $s_k < s_j$.

In Propositions 4.24 and 4.25, we explain how to take an augmented partition like $(t_0^1, s_0^1) \leqslant \cdots \leqslant (t_{2n+1}^1, s_{2n+1}^1)$ and obtain an augmented partition with similar properties that has one fewer maximal crossover interval from an augmented partition. Then, in Proposition 4.26, we work on $(t_0^1, s_0^1) \leqslant \cdots \leqslant (t_{2n+1}^1, s_{2n+1}^1)$ from left to right using Proposition 4.25 to obtain a new augmented partition with similar properties but no maximal crossover intervals.

Proposition 4.24 *Let $t_j^1 \leqslant t_{j+1}^1 \leqslant \cdots \leqslant t_k^1$ be a maximal crossover subinterval of an augmented partition*

$$(t_0, s_0) \leqslant (t_1, s_1) \leqslant \cdots \leqslant (t_{i_j-1}, s_{i_j-1}) \leqslant (t_j^1, s_j^1) \leqslant (t_{j+1}^1, s_{j+1}^1) \leqslant \cdots \leqslant (t_k^1, s_k^1) \leqslant \cdots \leqslant (t_{2n+1}^1, s_{2n+1}^1)$$

of $[0, t_\sigma]$. Then

- $d(\sigma(t_k^1), \gamma(s_j^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon + D_{\text{endpoints}}$,
- $d(\sigma(t_j^1), \gamma(s_k^1) \leqslant \lambda D_{\text{outorder}} + \epsilon + D_{\text{endpoints}}$, *and*
- $d_{\text{haus}}(\sigma([t_j^1, t_k^1]), \gamma([s_k^1, s_j^1])) \leqslant \lambda D_{\text{outorder}} + \epsilon + 3D_{\text{endpoints}}$.

Proof Recall $d(\sigma(t_j^1), \gamma(s_j^1)), d(\sigma(t_{j+1}^1), \gamma(s_{j+1}^1)), \dots, d(\sigma(t_k^1), \gamma(s_k^1)) \leqslant D_{\text{endpoints}}$. By Proposition 4.22, $|t_j^1 - t_k^1| \leqslant D_{\text{outorder}}$. Then $d(\sigma(t_j^1), \sigma(t_k^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon$. We can conclude then that $d(\gamma(s_j^1), \gamma(s_k^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon + 2D_{\text{endpoints}}$. Therefore, for all $s_k^1 \leqslant s \leqslant s_j^1$,

$$d(\gamma(s), \sigma(t_k^1)) \leqslant d(\gamma(s_j^1), \gamma(s_k^1)) + d(\gamma(s_k^1), \sigma(t_k^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon + 2D_{\text{endpoints}} + D_{\text{endpoints}}.$$

Similarly for all $t_j^1 \leqslant t \leqslant t_k^1$, $|t - t_k^1| \leqslant D_{\text{outorder}}$ so

$$d(\sigma(t), \sigma(t_k^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon.$$

Therefore,

$$d(\sigma(t), \gamma(s_k^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon + D_{\text{endpoints}}.$$

A similar argument will also show that $d(\sigma(t_k^1), \gamma(s_j^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon + D_{\text{endpoints}}$. □

Proposition 4.25 *Let $t_j^1 \leqslant t_{j+1}^1 \leqslant \cdots \leqslant t_k^1$ be a maximal crossover subinterval of an augmented partition*

$$(2) \quad (t_0, s_0) \leqslant (t_1, s_1) \leqslant (t_2, s_2) \leqslant \cdots \leqslant (t_{i_j-1}, s_{i_j-1}) \\ \leqslant (t_j^1, s_j^1) \leqslant (t_{j+1}^1, s_{j+1}^1) \leqslant \cdots \leqslant (t_k^1, s_1^k) \leqslant \cdots \leqslant (t_{2n+1}^1, s_{2n+1}^1)$$

of $[0, t_\sigma^1]$ so that $t_0, t_1, t_2, \dots, t_{i_j-1}$ are not contained in any maximal crossover subintervals of (2)*. There is a new augmented partition*

$$(3) \quad 0 = (t_0, s_0) \leqslant \cdots \leqslant (t_{i_j-1}, s_{i_j-1}) \leqslant (t_j^1, s_k^1) \leqslant (t_k^1, s_j^1) \leqslant (t_{k+1}^1, s_{k+1}^1) \leqslant \cdots \leqslant (t_{2n+1}^1, s_{2n+1}^1)$$

that has the properties

- $t_0, t_1, \dots, t_{i_j-1}, t_j^1, t_k^1$ *are not contained in any maximal crossover subinterval of* (3),
- $d(\sigma(t_k^1), \gamma(s_j^1)), d(\sigma(t_j^1), \gamma(s_k^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon + D_{\text{endpoints}}$, *and*
- $d_{\text{haus}}(\sigma([t_j^1, t_k^1]), \gamma([s_k^1, s_j^1])) < \lambda D_{\text{outorder}} + \epsilon + 3D_{\text{endpoints}}$.

Proof Since $t_0, t_1, t_2, \dots, t_{i_j-1}$ are not contained in any maximal crossover subinterval, $s_0 \leqslant s_1 \leqslant s_2 \leqslant \cdots \leqslant s_{i_j-1} \leqslant s_k$ and $s_k \leqslant s_j$ by hypothesis. Moreover, for all $k' > k$, we have $s_{k'}^1 \geqslant s_j^1 \geqslant s_k^1$ because $t_j^1 \leqslant \cdots \leqslant t_k^1$ is a *maximal* crossover subinterval. Therefore, t_j^1 and t_k^1 cannot be contained in a maximal crossover subinterval of the augmented partition (3).

From Proposition 4.24, we immediately obtain $d(\sigma(t_k^1), \gamma(s_j^1)) \leqslant \lambda D_{\text{outorder}} + \epsilon + D_{\text{endpoints}}$ and

$$d_{\text{haus}}(\sigma([t_j^1, t_k^1]), \gamma([s_k^1, s_j^1])) \leqslant \lambda D_{\text{outorder}} + \epsilon + 3D_{\text{endpoints}}.$$ □

Proposition 4.26 *There exist partitions*

$$0 = t_0^2 \leqslant t_1^2 \leqslant t_2^2 \leqslant t_3^2 \leqslant \cdots \leqslant t_{n'}^2 = t_\sigma \quad \text{and} \quad 0 = s_0^2 \leqslant s_1^2 \leqslant s_2^2 \leqslant s_3^2 \leqslant \cdots \leqslant s_{n'}^2 = s_\gamma$$

so that for $0 \leqslant j \leqslant n'$:

(1) $d(\sigma(t_j^2), \gamma(s_j^2)) \leqslant \lambda D_{\text{outorder}} + \epsilon + D_{\text{endpoints}}$.

(2) *For each* j, *one of the following holds*:

- $d_{\text{haus}}(\sigma([t_j^2, t_{j+1}^2]), \gamma([s_j^2, s_{j+1}^2])) \leqslant \lambda D_{\text{outorder}} + \epsilon + 3D_{\text{endpoints}}$.
- $\sigma([t_j^2, t_{j+1}^2]), \gamma([s_j^2, s_{j+1}^2]) \subseteq \mathcal{N}_{K(D_{\text{depth}}+D_{\text{endpoints}})}(F_j^2)$ *for some* $F_j^2 \in \mathcal{B}$.

(3) *If* $j \neq j'$, *then* $F_j^2 \neq F_{j'}^2$.

Proof sketch We can obtain the desired partition by starting with the partition from Lemma 4.18 and then working left to right using Proposition 4.25 to eliminate any maximal crossover subintervals. Immediately, $s_0^1 = 0$, so t_0^1 is not contained in any maximal crossover subintervals. The bound on $d(\sigma(t_j^2), \gamma(s_j^2))$ is implied by Proposition 4.25. One of the following holds:

- $t_j^2 = t_{2i}^1$, $t_{j+1}^2 = t_{2i+1}^1$, $s_j^2 = s_{2i}^1$ and $s_{j+1}^2 = s_{2i+1}^1$ for some i.
- $t_j^2 = t_{2i+1}^1$, $t_{j+1}^2 = t_{2i+2}^1$, $s_j^2 = s_{2i+1}^1$ and $s_{j+1}^2 = s_{2i+2}^1$ for some i.
- Proposition 4.25 implies that $d_{\text{haus}}(\sigma([t_j^2, t_{j+1}^2]), \gamma([s_j^2, s_{j+1}^2])) \leqslant \lambda D_{\text{outorder}} + \epsilon + 3D_{\text{endpoints}}$.

In the first case, Proposition 4.21 implies that $d_{\text{haus}}(\sigma([t_j^2, t_{j+1}^2]), \gamma([s_j^2, s_{j+1}^2]))$ is bounded appropriately. In the second case, Proposition 4.19 implies that $\sigma([t_j^2, t_{j+1}^2]) \subseteq \mathcal{N}_{D_{\text{depth}}}(F_i)$, so set $F_j^2 = F_i$. Since the endpoints of $\gamma([s_j^2, s_{j+1}^2])$ are within $D_{\text{endpoints}}$ of the endpoints of $\sigma([t_j^2, t_{j+1}^2])$ and γ is geodesic, we have

$$\gamma([s_j^2, s_{j+1}^2]) \subseteq \mathcal{N}_{K(D_{\text{depth}}+D_{\text{endpoints}})}(F_j^2)$$

Since the F_i are distinct, if $j \neq j'$, then $F_j^2 \neq F_{j'}^2$. □

In the partition from Proposition 4.26, we call an interval $[t_j^2, t_{j+1}^2]$ a *Hausdorff interval* if

$$d_{\text{haus}}(\sigma([t_j^2, t_{j+1}^2]), \gamma([s_j^2, s_{j+1}^2])) \leqslant \lambda D_{\text{outorder}} + \epsilon + 3D_{\text{endpoints}}.$$

Otherwise, if $\sigma([t_j^2, t_{j+1}^2]), \gamma([s_j^2, s_{j+1}^2]) \subseteq \mathcal{N}_{K(D_{\text{depth}}+D_{\text{endpoints}})}(F_j^2)$, we call $[t_j^2, t_{j+1}^2]$ a *peripheral interval*.

Theorem 4.7 *Let* $(G, \mathcal{P})$ *be a relatively hyperbolic group pair where* G *acts geometrically on a* CAT(0) *space* $\widetilde{X}$ *with basepoint* $x \in \widetilde{X}$. *If* $\mathcal{B}$ *is any* R*-thickening of* $\{gPx \mid g \in G,\ P \in \mathcal{P}\}$ *then* $(\widetilde{X}, \mathcal{B})$ *has the relative fellow traveling property.*

Proof By Proposition 4.6, $(\widetilde{X}, \mathcal{B})$ is a (δ, f)-CAT(0) relatively hyperbolic pair and there exists $L(R)$ so that Hypotheses 4.8 hold.

Given (λ,ϵ)-quasigeodesics γ,σ with the same endpoints, we can reduce to the case where γ is geodesic by Proposition 4.13. Proposition 4.26 nearly provides the partition for relative fellow traveling except that the intervals $[t_j^2, t_{j+1}^2]$ as constructed in Proposition 4.26 do not alternate between Hausdorff intervals and peripheral intervals. This can be easily remedied by turning any two adjacent Hausdorff intervals into a single Hausdorff interval. In other words, if $[t_j^2, t_{j+1}^2]$ and $[t_{j+1}^2, t_{j+2}^2]$ are both Hausdorff intervals, we remove these two intervals from the partition and replace them with the single interval $[t_j^2, t_{j+2}^2]$. Likewise, replace $[s_j^2, s_{j+1}^2]$ and $[s_{j+1}^2, s_{j+2}^2]$ with $[s_j^2, s_{j+2}^2]$. It is easy to check that $d_{\mathrm{haus}}(\sigma([t_j^2, t_{j+2}^2]), \gamma([s_j^2, s_{j+2}^2])) \leqslant \lambda D_{\mathrm{outorder}} + \epsilon + 3D_{\mathrm{endpoints}}$ in this case. Repeat this process until no adjacent Hausdorff intervals remain. □

5 A relatively hyperbolic combination lemma

The construction of hierarchies in Section 7 is quite similar to the hierarchy constructed in [3]. The goal of this section is to prove a combination theorem for the relatively hyperbolic setting that will be used to show the edge groups of the hierarchy are undistorted.

5.1 The attractive property in CAT(0) relatively hyperbolic pairs

The first goal is to improve a CAT(0) relatively hyperbolic pair so that geodesics that stay near a peripheral space intersect the peripheral space.

Definition 5.1 Let $\widetilde{X}$ be a geodesic metric space, let Z be a subspace of $\widetilde{X}$ and let $K_{\mathrm{att}}\colon \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$ be a function. The subspace Z is K_{att}*-attractive* if for all $R \geqslant \delta$ whenever γ is a geodesic with endpoints in $\mathcal{N}_R(Z)$ and $|\gamma| \geqslant K_{\mathrm{att}}(R)$, then $\gamma \cap Z \neq \varnothing$.

We now fix hypotheses for the remainder of the Section 5.1.

Hypotheses 5.2 Suppose that $(\widetilde{X}, \mathcal{B}')$ is a (δ, f')-CAT(0) relatively hyperbolic pair where every $F' \in \mathcal{B}'$ is convex. Let $\mathcal{B} = \{\mathcal{N}_{2\delta}(F') : F' \in \mathcal{B}'\}$ so that for some $f\colon \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$, $(\widetilde{X}, \mathcal{B})$ is a (δ, f)-CAT(0) relatively hyperbolic pair by Proposition 4.4. Fix $M = f(6\delta)$.

Proposition 5.3 *Under Hypotheses 5.2, every* $B \in \mathcal{B}$ *is* $(3M+6R+21\delta)$*-attractive.*

The following result will be used to prove Proposition 5.3:

Proposition 5.4 *Assume Hypotheses 5.2, let* γ *be a geodesic and let* $F \in \mathcal{B}'$. *If* γ *has endpoints in* $\mathcal{N}_R(F)$, *then* $\operatorname{diam} \gamma \cap \mathcal{N}_{2\delta}(F) > |\gamma| - (3M + 6R + 9\delta)$.

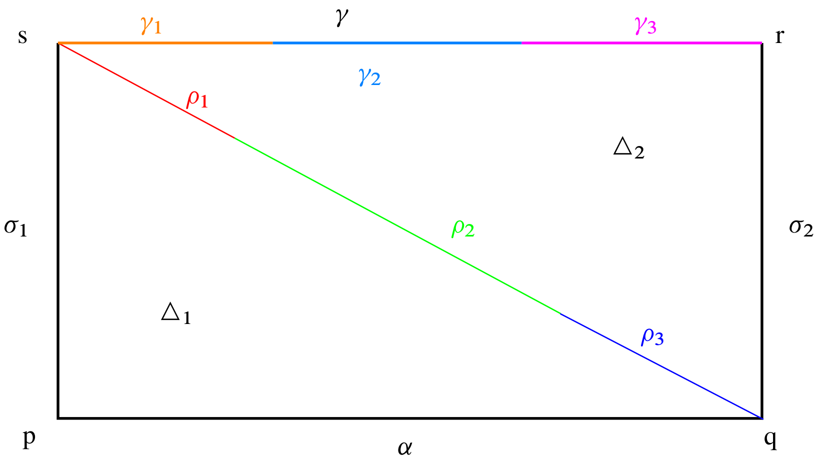


Figure 5: The quadrilateral constructed in the proof of Proposition 5.3.

Proof There is a quadrilateral whose sides are γ, two geodesics σ_1, σ_2 of length at most R connecting the endpoints of γ to points in F and a geodesic α connecting the endpoints of σ_1, σ_2 that are in F. By convexity, $\alpha \subseteq F$. Let ρ be a diagonal so that there are two triangles, $\triangle_1$, $\triangle_2$, so that $\triangle_1$ has sides α, ρ, σ_1 as a side and $\triangle_2$ has sides γ, ρ, σ_2. Designate vertices p, q, r, s so that $\alpha = [p,q]$, $\sigma_2 = [q,r]$, $\gamma = [r,s]$, $\sigma_1 = [p,s]$, and $\rho = [q,s]$ as shown in Figure 5.

Case 1 ($\triangle_1$ is δ-thin relative to some $F' \neq F$) Since $F' \neq F$ and $\alpha \subseteq F$, the length of the fat part of α in $\triangle_1$ is at most M.

Let ρ_1 be the corner segment of ρ in $\triangle_1$ at s. Then $|\rho_1| \leqslant R$. Let ρ_2 be the fat part of ρ in $\triangle_1$. The fat part of σ_1 in $\triangle_1$ has length at most R, so by Lemma 2.12, $|\rho_2| \leqslant M + R + 3\delta$. Let ρ_3 be the corner segment of ρ in $\triangle_1$ at q. By construction, $\rho_3 \subseteq \mathcal{N}_\delta(F)$.

Let γ_1 be the corner segment of γ at s in $\triangle_2$, let γ_2 be the fat part of γ in $\triangle_2$ and let γ_3 be the corner segment of γ in $\triangle_2$ at r. Observe that $\gamma_1 \cap \mathcal{N}_\delta(\rho_3) \subseteq \mathcal{N}_{2\delta}(F)$ and

$$\operatorname{diam} \gamma_1 \cap \mathcal{N}_\delta(\rho_3) \geqslant |\gamma_1| - |\rho_1| - |\rho_2| \geqslant |\gamma_1| - (M + 2R + 3\delta).$$

If $\triangle_2$ is δ-thin relative to F, then $\gamma_2 \subseteq \mathcal{N}_\delta(F)$. If $\triangle_2$ is δ-thin relative to some other element of $\mathcal{B}'$, the fat part of ρ in $\triangle_2$ has length at most $|\rho_1| + |\rho_2| + M \leqslant 2M + 2R + 3\delta$ because $\rho_3 \subseteq \mathcal{N}_\delta(F)$. By Lemma 2.12,

$$|\gamma_2| \leqslant 2M + 2R + 3\delta + R + 3\delta$$

because $|\sigma_2| \leqslant R$. Finally, $|\gamma_3| \leqslant R$.

In summary, at most $M + 2R + 3\delta$ of γ_1 lies outside of $\mathcal{N}_{2\delta}(F)$, at most $2M + 3R + 6\delta$ of γ_2 lies outside of $\mathcal{N}_{2\delta}(F)$, and at most R of γ_3 lies outside of $\mathcal{N}_{2\delta}(F)$, so

$$\operatorname{diam} \gamma \cap N_{2\delta}(F) \geqslant |\gamma| - (3M + 6R + 9\delta).$$

Case 2 ($\triangle_1$ is δ-thin relative to F) Let ρ_1, ρ_2, ρ_3 and γ_1, γ_2, γ_3 be as in the previous case. Here, $|\rho_1| \leqslant R$, $\rho_2 \subseteq \mathcal{N}_\delta(F)$ since $\triangle_1$ is δ-thin relative to F and $\rho_3 \subseteq \mathcal{N}_\delta(\alpha) \subseteq \mathcal{N}_\delta(F)$. Since γ_1 δ-fellow travels a subsegment of ρ at s, $\operatorname{diam} \gamma_1 \cap \mathcal{N}_\delta(\rho_2 \cup \rho_3) \geqslant |\gamma_1| - R$ because $|\rho_1| \leqslant R$. Since $\rho_2 \cup \rho_3 \subseteq \mathcal{N}_\delta(F)$, $\operatorname{diam} \gamma_1 \cap \mathcal{N}_{2\delta}(F) \geqslant |\gamma_1| - R$. If $\triangle_2$ is δ-thin relative to some $F'' \neq F$, the fat part of ρ in $\triangle_2$ has length at most $R + M$ because its intersection with $\rho_2 \cup \rho_3 \subseteq \mathcal{N}_\delta(F)$ has length at most M and $|\rho_1| \leqslant R$. Therefore by Lemma 2.12, $|\gamma_2| < M + 2R + 3\delta$. On the other hand, if $\triangle_2$ is δ-thin relative to F, then $\gamma_2 \subseteq \mathcal{N}_\delta(F)$ so in both cases, all but a less than $M + 2R + 3\delta$ subsegment of γ_2 lies in $\mathcal{N}_{2\delta}(F)$.

In summary, $\operatorname{diam} \gamma_1 \cap \mathcal{N}_{2\delta}(F) \geqslant |\gamma_1| - R$, $\operatorname{diam} \gamma_2 \cap \mathcal{N}_{2\delta}(F) \geqslant |\gamma_2| - (M + 2R + 3\delta)$ and $|\gamma_3| \leqslant R$. Therefore, by the convexity of $\mathcal{N}_{2\delta}(F)$,

$$|\gamma \cap \mathcal{N}_{2\delta}(F)| \geqslant |\gamma| - (M + 4R + 3\delta). \qquad \square$$

Proof of Proposition 5.3 Let γ be a geodesic with endpoints in $\mathcal{N}_R(F)$. Then by convexity, $\gamma \subseteq \mathcal{N}_{R+2\delta}(F')$ for some $F' \in \mathcal{B}'$ where $F = \mathcal{N}_{2\delta}(F')$. By Proposition 5.4, if $|\gamma| > 3M + 6(R + 2\delta) + 9\delta$, then $\gamma \cap \mathcal{N}_{2\delta}(F') \neq \varnothing$. Noting that $F = \mathcal{N}_{2\delta}(F')$ completes the proof. $\square$

5.2 A combination lemma for CAT(0) relatively hyperbolic pairs

Maintain the following baseline hypotheses for Section 5.2:

Hypotheses 5.5 Let $(\widetilde{X}, \mathcal{B})$ be a (δ, f)-CAT(0) relatively hyperbolic pair and let $M = f(6\delta)$ as before. Suppose that every $B \in \mathcal{B}$ is closed, convex and $(3M{+}6R{+}2f(R){+}21\delta)$-attractive.

In Section 7, we will use Proposition 5.3 to obtain attractiveness for a (δ, f)-CAT(0) relatively hyperbolic pair, and then thicken the peripheral spaces to make a new (δ, f)-CAT(0) relatively hyperbolic pair. We will then prove that the new peripheral spaces are $(3M{+}6R{+}2f(R){+}21\delta)$-attractive. For this reason, Hypotheses 5.5 are slightly weaker than what would follow from Hypotheses 5.2 and the conclusions of Proposition 5.3.

Theorem 5.6 *Assume Hypotheses 5.5. Let $\gamma = b_1 a_2 b_2 a_3 b_3 \dots a_n b_n$ be a broken geodesic. Let γ_i be the geodesic connecting the endpoints of the subpath $b_1 a_2 b_2 a_3 b_3 \dots a_i b_i$ of γ. Suppose that:*

1. *For each $1 \leqslant i \leqslant n$, there exists some $F_i \in \mathcal{B}$ so that $b_i \subseteq F_i$.*
2. *If $F_i = F_j$, then $i = j$.*
3. *For $1 \leqslant i \leqslant n-1$, $|b_i| \geqslant 37M + 250\delta$.*
4. *For all $2 \leqslant i \leqslant n$, $\operatorname{diam} a_i \cap \mathcal{N}_{3\delta}(F_i) \leqslant 5M + 39\delta$ and $\operatorname{diam} a_i \cap \mathcal{N}_{3\delta}(F_{i-1}) \leqslant 5M + 39\delta$.*
5. *For all $2 \leqslant i \leqslant n$, $\operatorname{diam} a_i \cap \mathcal{N}_{6\delta}(F_i) \leqslant 5M + 57\delta$ and $\operatorname{diam} a_i \cap \mathcal{N}_{6\delta}(F_{i-1}) \leqslant 5M + 57\delta$.*

Then γ_n has a length at least $|b_n| - (24M + 165\delta)$-tail at the endpoint it shares with b_n (recall Definition 2.9) that lies in $\mathcal{N}_{2\delta}(F_n)$ and for all $2 \leqslant i \leqslant n$, $|\gamma_i| \geqslant |\gamma_{i-1}| + |a_n| + |b_n| - 68M - 628\delta$.

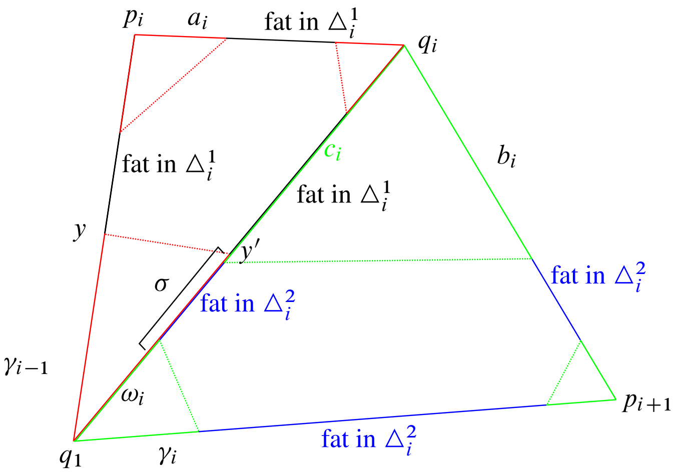


Figure 6: One possible configuration of $\triangle_i^1$ and $\triangle_i^2$ in the proof of Theorem 5.6. Corner segments of triangles at the same point are connected by dotted lines.

Proof In the case $n = 1$, the proof is straightforward. The proof of Theorem 5.6 is by induction on n.

Notation 5.7 We now establish notation that will be used throughout the proof of Theorem 5.6.

(1) For each $2 \leqslant i \leqslant n$, let ω_i be the geodesic connecting the endpoints of the broken geodesic $b_1 a_2 b_2 \dots b_{i-1} a_i$.

(2) For each $2 \leqslant i \leqslant n$, let $\triangle_i^1$ be the triangle with sides γ_{i-1}, ω_i and a_i.

(3) For each $2 \leqslant i \leqslant n$, let $\triangle_i^2$ be the triangle with sides ω_i, b_i and γ_i.

(4) Label vertices so that $a_i = [p_i, q_i]$ and $b_i = [q_i, p_{i+1}]$.

(5) Let c_i be the corner segment of ω_i in $\triangle_i^2$ at q_i.

See Figure 6 for a visual representation.

We make the additional inductive assumption that for $1 \leqslant i < n$, γ_i has a $|b_i| - (24M + 165\delta)$-tail at p_{i+1} in $\mathcal{N}_{2\delta}(F_i)$.

Proposition 5.8 *If $i \geqslant 2$ and we assume the inductive hypotheses for the proof of Theorem 5.6, then there is a point $x_i \in \gamma_i$ so that $d(x_i, F_{i-1}) \leqslant 4\delta$. Further, $|c_i| \leqslant 12M + 81\delta$ (recall Notation 5.7(5)). When $\triangle_i^2$ is δ-thin relative to F_i, then the length of the fat part of ω_i in $\triangle_i^2$ is at most $12M + 81\delta$.*

Proof Since $1 \leqslant i - 1 < n$, γ_{i-1} has a length at least $13M + 85\delta$-tail at p_i in $\mathcal{N}_{2\delta}(F_{i-1})$ by our inductive assumption.

Case ($\triangle_i^1$ is thin relative to $F \neq F_{i-1}$) The corner segments of $\triangle_i^1$ at p_i have length at most $5M + 39\delta$ to avoid violating Theorem 5.6(4) because a more than $5M + 39\delta$-tail of γ_{i-1} at p_i lies in $\mathcal{N}_{2\delta}(F_{i-1})$.

Since $\triangle_i^1$ is thin relative to $F \neq F_{i-1}$, the length of the fat part of γ_{i-1} in $\triangle_i^1$ is at most M. Therefore, there is a point $y \in \gamma_{i-1}$ and a point $y' \in \omega_i$ so that $d(y, p_i) \leqslant 6M + 39\delta$ and $d(y, y') \leqslant \delta$ so that y, y' are endpoints of the corner segments of $\triangle_i^1$ at q_1 and further, there exists a subsegment σ (see Figure 6) of the corner segment $[q_1, y'] \subseteq \omega_i$ with endpoint y' so that $|\sigma| > 2M$ and $\sigma \subseteq \mathcal{N}_{3\delta}(F_{i-1})$.

The intersection of σ with the corner segment of ω_i in $\triangle_i^2$ at q_i lies in $\mathcal{N}_{3\delta}(F_{i-1}) \cap \mathcal{N}_{3\delta}(F_i)$ and therefore has length at most M. The fat part of ω_i in $\triangle_i^2$ is either contained in $\mathcal{N}_\delta(F_{i-1})$ or intersects σ in a segment of length at most M. Therefore, either there is a point in γ_i that is at most δ from the fat part of ω_i in $\triangle_i^2$ and the fat part of ω_i in $\triangle_i^2$ is contained in $\mathcal{N}_\delta(F_{i-1})$ or σ intersects the corner segment of $\triangle_i^2$ at q_1. In the first case, there is a point $x_i \in \gamma_i$ that lies in $\mathcal{N}_{2\delta}(F_{i-1})$ and in the second case, there is a point $x_i \in \gamma_i$ so that $d(x_i, \sigma) < \delta$, so $x_i \in \mathcal{N}_{4\delta}(F_{i-1})$.

The next tasks are to bound $|c_i|$ from above and to prove that when $\triangle_i^2$ is δ-thin relative to F_i, the fat part of ω_i in $\triangle_i^2$ has length at most M. Note that $c_i \subseteq \mathcal{N}_{2\delta}(F_i)$. The intersection of c_i with the corner segment of ω_i in $\triangle_i^1$ at q_i has length at most $5M + 39\delta$ because $\operatorname{diam} a_i \cap \mathcal{N}_{3\delta}(F_i) \leqslant 5M + 39\delta$. If $F \neq F_i$, the intersection of c_i with the fat part of ω_i in $\triangle_i^1$ is a segment of length at most M. Since $|c_i \cap \sigma| \leqslant M$ and $|\sigma| > 2M$, $|c_i| \leqslant 7M + 39\delta$. Further, if $\triangle_i^2$ is δ-thin relative to F_i, then the fat part of ω_i in $\triangle_i^2$ intersects σ in a segment of length at most M, intersects the fat part of ω_i in $\triangle_i^1$ in a length at most M segment and intersects the corner segment of ω_i in $\triangle_i^1$ at q_i in a segment of length at most $5M + 39\delta$. Hence the fat part of ω_i in $\triangle_i^2$ has length at most $7M + 39\delta$ when $\triangle_i^2$ is thin relative to F_i.

If $F = F_i$, then the fat parts of a_i and γ_{i-1} in $\triangle_i^1$, which are contained in $\mathcal{N}_\delta(F_i)$, have length at most $5M + 39\delta$ and M, respectively. Therefore, the length of the fat part of ω_i in $\triangle_i^1$ is at most $5M + 39\delta + M + 3\delta$. Then $|c_i| \leqslant 12M + 81\delta$ by a computation similar to the one in the previous case.

When $F = F_i$, the fat part of ω_i in $\triangle_i^2$ intersects σ in a segment of length at most M, intersects the fat part of ω_i in $\triangle_i^1$ in a segment of length at most $6M + 42\delta$ and intersects the corner segment of ω_i in $\triangle_i^1$ at q_i in a segment of length at most $5M + 39\delta$. Therefore, if $\triangle_i^2$ is thin relative to F_i, then the length of the fat part of ω_i in $\triangle_i^2$ is at most $12M + 81\delta$.

Case ($\triangle_i^1$ is thin relative to F_{i-1}) Recall c_i is the corner segment of ω_i in $\triangle_i^2$ at q_i. The intersection of c_i with the corner segment of ω_i in $\triangle_i^1$ at q_i again lies in $\mathcal{N}_{2\delta}(F_i) \cap \mathcal{N}_\delta(a_i)$ and hence has length at most $5M + 39\delta$. The fat part of ω_i in $\triangle_i^1$ lies in $\mathcal{N}_\delta(F_{i-1})$. Hence, if the length of the fat part of ω_i in $\triangle_i^1$ exceeds M, then its intersection with c_i has length at most M so $|c_i| \leqslant 6M + 39\delta$. Hence for the purposes of bounding $|c_i|$ from above, assume the fat part of ω_i in $\triangle_i^1$ has length at most M. The length of the fat part of a_i in $\triangle_i^1$ is at most $5M + 39\delta$. If the length of the fat part of ω_i in $\triangle_i^1$ is at most M, then by Lemma 2.12, the length of the fat part of γ_{i-1} in $\triangle_i^1$ is at most $6M + 42\delta$. Now, if $y \in \gamma_{i-1}$, $y' \in \gamma_{i-1}$ are the endpoints of the corner segments of $\triangle_i^1$ at q_1, then $d(y, p_i) \leqslant 5M + 39\delta + 6M + 42\delta = 11M + 81\delta$. Therefore there is a tail at y' of the corner segment of ω_i in $\triangle_i^1$ at q_1 called σ so that $|\sigma| > 2M$ and $\sigma \subseteq \mathcal{N}_{3\delta}(F_{i-1})$ because γ_{i-1} has a more than $13M + 84\delta$-tail in $\mathcal{N}_{2\delta}(F_{i-1})$. Therefore, c_i intersects

$[y', q_1]$ in a segment of length at most M because $c_i \subseteq \mathcal{N}_{2\delta}(F_i)$. Hence $|c_i| \leqslant 7M + 39\delta$ because the union of the two corner segments of ω_i in $\triangle_i^1$ and the fat part of ω_i in $\triangle_i^1$ is ω_i.

In all cases, $|c_i| \leqslant 12M + 81\delta$.

If $\triangle_i^2$ is δ-thin relative to F_i, the fat part of ω_i in $\triangle_i^2$ has length at most $6M + 39\delta$ because the corner segment of ω_i in $\triangle_i^1$ at q_i lies in $\mathcal{N}_\delta(a_i)$, and both σ and the fat part of $\triangle_i^1$ lie in $\mathcal{N}_\delta(F_{i-1})$. In particular, the fat part of ω_i in $\triangle_i^2$ may only intersect $[q_1, y']$ in σ because otherwise its intersection with σ has length more than M and lies in $\mathcal{N}_{3\delta}(F_{i-1}) \cap \mathcal{N}_\delta(F_i)$.

The only remaining thing to prove is that there is a point $x_i \in \gamma_i$ so that $d(x_i, F_{i-1}) \leqslant 4\delta$. If $\triangle_i^2$ is δ-thin relative to F_{i-1} and is not δ-thin relative to any other $F \in \mathcal{B}$, then there is a point on γ_i in $\mathcal{N}_{2\delta}(F_{i-1})$. Hence assume $\triangle_i^2$ is thin relative to some $G \in \mathcal{B}$ with $G \neq F_{i-1}$.

Let $\omega^1 \subseteq \mathcal{N}_\delta(F_{i-1})$ be the fat part of ω_i in $\triangle_i^1$ and let ω^2 be the corner segment of ω_i in $\triangle_i^2$ at q_1. If there exists $r \in \omega^1 \cap \omega^2$, then $d(r, \gamma_i) < \delta$, so there exists an $x_i \in \gamma_i$ such that $\gamma_i \in \mathcal{N}_{2\delta}(F_{i-1})$.

Otherwise, ω^1 intersects c_i in a segment of length at most M because c_i lies in $\mathcal{N}_{2\delta}(F_i)$ and intersects the fat part of ω_i in $\triangle_i^2$ in a segment of length at most M (the fat part of ω_i in $\triangle_i^2$ lies in $\mathcal{N}_\delta(G)$). Hence $|\omega^1| \leqslant 2M$. Let ω^3 be the corner segment of ω_i in $\triangle_i^1$ at q_1. Let $z \in \omega_i$ be the point where ω^1 intersects ω^3. By Lemma 2.12, the fat part of γ_{i-1} in $\triangle_i^1$ has length at most $2M + 5M + 39\delta + 3\delta = 7M + 42\delta$ because $\operatorname{diam} a_i \cap \mathcal{N}_{2\delta}(F_{i-1}) \leqslant 5M + 39\delta$. The corner length of $\triangle_i^1$ at p_i is at most $5M + 39\delta$ because any subsegment of a_i in $\mathcal{N}_{3\delta}(F_i)$ has length at most $5M + 39\delta$. Then at least a $13M + 84\delta - (5M + 39\delta + 7M + 42\delta) > M$-tail of ω^3 at z, which will be called ω', lies in $\mathcal{N}_{3\delta}(F_{i-1})$ because it δ-fellow travels a subsegment of the tail of γ_{i-1} at p_i contained in $\mathcal{N}_{2\delta}(F_{i-1})$. The union of c_i and the fat part of $\triangle_i^2$ lie in $\mathcal{N}_{2\delta}(F_i)$, so they collectively cannot extend past ω' in the direction of q_1 because otherwise ω'contains a length more than M subsegment in $\mathcal{N}_{3\delta}(F_i) \cap \mathcal{N}_{3\delta}(F_{i-1})$. Therefore, ω^2, the corner segment of $\triangle_i^2$ at q_1, must intersect ω'. Since ω' lies in $\mathcal{N}_{3\delta}(F_{i-1})$ and ω^2 is a corner segment of $\triangle_i^2$ at q_1, there is a point $x_i \in \gamma_i$ so that $x \in \mathcal{N}_{4\delta}(F_{i-1})$. □

Proposition 5.9 *If $b_i \subseteq \mathcal{N}_\delta(F_i)$, then the geodesic γ_i has a $|b_i| - (24M + 165\delta)$-tail at p_{i+1} that is contained in $\mathcal{N}_{2\delta}(F_i)$.*

Proof There are two cases:

Case 1 ($\triangle_i^2$ is δ-thin relative to some $F \neq F_i$) The corner length of $\triangle_i^2$ at q_i is at most $12M + 81\delta$ by Proposition 5.8. The length of the fat part of b_i in $\triangle_i^2$ is at most M because $b_i \subseteq \mathcal{N}_\delta(F)$. Therefore, the corner length of $\triangle_i^2$ at p_{i+1} is at least $|b_i| - (13M + 81\delta)$. Thus the corner segment of γ_i at p_{i+1} has length at least $|b_i| - (13M + 81\delta)$ and lies in $\mathcal{N}_\delta(b_i) \subseteq \mathcal{N}_{2\delta}(F_i)$.

Case 2 ($\triangle_i^2$ is δ-thin relative to F_i) The corner length of $\triangle_i^2$ at q_i is at most $12M + 81\delta$. Let s be the length of the fat part of b_i in $\triangle_i^2$. Then the corner length of $\triangle_i^2$ at p_{i+1} is at least $|b_i| - s - (12M + 81\delta)$.

By Proposition 5.8, the length of the fat part of ω_i in $\triangle_i^2$ is at most $12M+81\delta$. By Lemma 2.12, the fat part of γ_i in $\triangle_i^2$ has length at least $s-(12M+81\delta+3\delta)$. The corner segment of γ_i at p_{i+1} in $\triangle_i^2$ and the fat part of γ_i in $\triangle_i^2$ both lie in $\mathcal{N}_{2\delta}(F_i)$ and their combined length is at least $s-(12M+84\delta)+|b_i|-s-(12M+81\delta)=|b_i|-(24M+165\delta)$. □

Lemma 5.10 *Let* $\eta:=[p_i,p_{i+1}]$. *Then* $\operatorname{diam}\eta\cap\mathcal{N}_{5\delta}(F_{i-1})\leqslant 12M+117\delta$.

Further, $d(q_i,\eta)\leqslant 10M+79\delta$.

Proof Let $\triangle$ be the geodesic triangle with sides a_i, b_i, η. If the corner segment of η in $\triangle$ at p_i lies in $\mathcal{N}_{5\delta}(F_{i-1})$, then the corner length of $\triangle$ at p_i is at most $5M+57\delta$ because $a_i\cap\mathcal{N}_{6\delta}(F_{i-1})$ has diameter at most $5M+57\delta$.

Suppose $\triangle$ is δ-thin relative to F_{i-1}. The fat part of b_i in $\triangle$ then lies in F_i and therefore has length at most M. The fat part of a_i in $\triangle$ has length at most $5M+57\delta$ because $a_i\cap\mathcal{N}_{6\delta}(F_{i-1})$ has diameter at most $5M+57\delta$. Hence by Lemma 2.12, the length of the fat part of η in $\triangle$ is at most $6M+60\delta$. On the other hand, if η is δ-thin relative to some $F\neq F_{i-1}$, then the intersection of the fat part of η with $\mathcal{N}_{5\delta}(F_{i-1})$ has length at most M. In all cases, the fat part of η in $\triangle$ intersects $\mathcal{N}_{5\delta}(F_{i-1})$ in a segment of length at most $6M+60\delta$.

Finally, the corner segment of η in $\triangle$ at p_{i+1} lies in $\mathcal{N}_{2\delta}(F_i)$ and can hence intersect $\mathcal{N}_{5\delta}(F_{i-1})$ in a segment of length at most M.

Since η is the union of its two corner segments and its fat part in $\triangle$, its intersection with $\mathcal{N}_{5\delta}(F_{i-1})$ has diameter at most $12M+117\delta$.

The corner length of $\triangle$ at q_i is at most $5M+39\delta$, because the corner segment of a_i in $\triangle$ at q_i lies in $a_i\cap\mathcal{N}_{2\delta}(F_i)$. If $\triangle$ is δ-thin relative to F_i, then the length of the fat part of a_i in $\triangle$ is at most $5M+39\delta$. Otherwise, if $\triangle$ is δ-thin relative to $F\neq F_i$, then the length of the fat part b_i in $\triangle$ is at most M. Since $\triangle$ is relatively δ-thin, in both cases, there exists a point on η that is at most $5M+39\delta+5M+39\delta+\delta=10M+79\delta$ from q_i. □

Lemma 5.11 *Let x_i be a point on γ_i so that $x_i\in\mathcal{N}_{4\delta}(F_{i-1})$ and x_i is the point closest to p_{i+1} with this property. Let $\eta'=[p_i,x_i]$ and let $\eta''=[x_i,p_{i+1}]\subseteq\gamma_i$. Let $\triangle'$ be the triangle with sides η, η', η''. Then at least one of the following holds*:

(1) *The length of the fat part of η in $\triangle'$ is at most* $12M+117\delta$.

(2) *The length of the fat part of η' in $\triangle'$ is at most* $M\leqslant 12M+117\delta$.

Proof Suppose $\triangle'$ is δ-thin relative to F_{i-1}. Then by Lemma 5.10, the fat part of η has length at most $12M+117\delta$. On the other hand if $\triangle'$ is δ-thin relative to some $F\neq F_{i-1}$, then the fat part of η' in $\triangle'$ lies in $\mathcal{N}_{4\delta}(F_{i-1})$ by convexity, so the length of the fat part of η' in $\triangle'$ is at most M. □

Lemma 5.12 *There exists $y_i \in \gamma_i$ so that $d(p_i, y_i) \leqslant 24M + 235\delta$.*

Proof The corner segment of η in $\triangle'$ at p_i lies in $\mathcal{N}_{5\delta}(F_{i-1}) \cap \eta$, so by Lemma 5.10, the corner length of $\triangle'$ at p_i is at most $12M + 117\delta$. By Lemma 5.11, the length of fat part of η in $\triangle'$ or the length of the fat part of η' in $\triangle'$ is at most $12M + 117\delta$, so there is a point y_i in $\eta'' \subseteq \gamma_i$ so that $d(p_i, y_i) \leqslant 24M + 235\delta$ because $\triangle'$ is relatively δ-thin. □

The next lemma follows immediately from the triangle inequality, but is convenient to have recorded:

Lemma 5.13 *Let $\triangle_0$ be a geodesic triangle in $\widetilde{X}$ with sides abc and suppose that a and b meet at the vertex p and $d(p, c) \leqslant J$. Then $|c| \geqslant |a| + |b| - 2J$.*

Proposition 5.14 *We have*

$$\begin{aligned} |\gamma_n| &\geqslant |\gamma_{n-1}| + |a_n| + |b_n| - 2(24M + 235\delta) - 2(10M + 79\delta) \\ &= |\gamma_{n-1}| + |a_n| + |b_n| - 68M - 628\delta. \end{aligned}$$

Proof By Lemmas 5.12 and 5.13,

$$|\gamma_n| \geqslant |\gamma_{n-1}| + |\eta| - 2(24M + 235\delta).$$

Then by Lemmas 5.10 and 5.13,

$$|\eta| \geqslant |a_n| + |b_n| - 2(10M + 79\delta).$$

Putting the two preceding inequalities together yields the desired inequality. □

Propositions 5.9 and 5.14 complete the inductive proof of Theorem 5.6. □

Definition 5.15 Let $\mathcal{A}$ be a collection of subsets of a geodesic metric space and let $K \geqslant 0$. Suppose that for all $A_1, A_2 \in \mathcal{A}$ with $A_1 \neq A_2$, $d(A_1, A_2) \geqslant K$. Then the collection $\mathcal{A}$ is *K-separated.*

The paths in Theorem 5.6 are of a special type to facilitate the inductive proof. Proposition 5.17 generalizes Theorem 5.6 to apply to all geodesic paths coming from certain subspaces of $\widetilde{X}$ with some additional assumptions:

Hypotheses 5.16 Assume Hypotheses 5.5 and assume the following:

(1) Let $\Lambda := 500M + 10000\delta$.

(2) Let $\mathcal{A}$ be a Λ-separated collection of convex subspaces of $\widetilde{X}$.

(3) Let $\mathcal{B}_0 \subseteq \mathcal{B}$.

(4) Let $T = \big(\bigsqcup_{A\in\mathcal{A}} A\big) \sqcup \big(\bigsqcup_{B\in\mathcal{B}_0} B\big)$. Define an equivalence relation $\sim$ on T by $x \sim y$ if and only if $x = y$ or for some $A \in \mathcal{A}$ and $B \in \mathcal{B}_0$, the images of x and y in $\widetilde{X}$ agree and lie in the images of both A and B.

Proposition 5.17 *Under Hypotheses 5.16, if $T/\!\sim$ is path connected, then the natural inclusion of $T/\!\sim\ \hookrightarrow \widetilde{X}$ is a $(2, 114M+1592\delta)$-quasi-isometric embedding (where the metric on $T/\!\sim$ is the induced path metric).*

Proof Let γ be the image in $\widetilde{X}$ of a geodesic in $T/\!\sim$ and let γ' be the $\widetilde{X}$-geodesic between its endpoints.

Up to reversing the direction of γ, γ can be written as a piecewise geodesic of one of the piecewise geodesic forms

(1) $b_1a_2b_2\dots a_nb_n$ and $|b_1|, |b_n| \geqslant 37M + 250\delta$,

(2) $a_1b_1a_2b_2\dots b_na_{n+1}$ where $|a_1|, |a_{n+1}| \neq 0$,

(3) $a_1b_1\dots a_nb_n$ where $|a_1| \neq 0$ and $|b_n| \geqslant 37M + 250\delta$,

(4) $a_1b_1\dots a_nb_n$ where $|a_1| \neq 0$ and $|b_n| \leqslant 37M + 250\delta$,

(5) $b_1a_2b_2\dots a_nb_n$, where both of $|b_1|$, $|b_n|$ are less than $37M + 250\delta$,

(6) $b_1a_2b_2\dots a_nb_n$, where $|b_1| < 37M + 250\delta$ and $|b_n| \geqslant 37M + 250\delta$,

where for each $1 \leqslant i \leqslant n$, $a_i \subseteq A_i \in \mathcal{A}$, for all $1 \leqslant i \leqslant n$, $b_i \subseteq B_i \in \mathcal{B}$, and for $2 \leqslant i \leqslant n-1$, $|b_i| \geqslant \Lambda$ because $\mathcal{A}$ is a Λ-separated collection. Assume also that n is minimal and γ is subdivided in a way that maximizes the sum of the lengths of the b_i.

If $i \neq j$, then $B_i \neq B_j$ because otherwise the subsegment $b_i\dots b_j$ of γ could be replaced by a single geodesic segment in $B_i \subseteq T/\!\sim$ contradicting minimality of n. By the maximality of the lengths of the b_i and the $(3M+6R+2f(R)+21\delta)$-attractiveness of every $B \in \mathcal{B}$,

$$\operatorname{diam} a_i \cap \mathcal{N}_{3\delta}(B_i), \operatorname{diam} a_i \cap \mathcal{N}_{3\delta}(B_{i-1}) \leqslant 5M + 39\delta,$$
$$\operatorname{diam} a_i \cap \mathcal{N}_{6\delta}(B_{i-1}), \operatorname{diam} a_i \cap \mathcal{N}_{6\delta}(B_i) \leqslant 5M + 57\delta$$

because otherwise the interiors of the a_i intersect either B_i or B_{i-1} so that b_i or b_{i-1}, respectively, could be made longer by convexity.

For the following arguments, recall the earlier convention that the endpoints of the a_i, b_i are labeled so that $a_i = [p_i, q_i]$ and $b_i = [q_i, p_{i+1}]$.

Case (1) ($\gamma = b_1a_2b_2\dots a_nb_n$ and $|b_1|, |b_n| \geqslant 37M + 250\delta$) By Theorem 5.6,

$$|\gamma'| \geqslant |b_1| + \Big(\sum_{i=2}^{n} |a_i| + |b_i|\Big) - (n-1)\cdot(68M + 628\delta).$$

Since $|b_i| \geqslant 136M + 1256\delta$, for $2 \leqslant i \leqslant n-1$ then

$$\begin{aligned}
|\gamma'| &\geqslant |b_1| + \Bigl(\sum_{i=2}^{n} |a_i| + |b_i|\Bigr) - (n-1)(68M + 628\delta) \\
&\geqslant \tfrac{1}{2}\Bigl(\sum_{i=2}^{n} |a_i|\Bigr) + |b_1| + \Bigl(\sum_{i=2}^{n-1}(|b_i| - (68M + 628\delta))\Bigr) + |b_n| - (68M + 628\delta) \\
&\geqslant \tfrac{1}{2}\Bigl(\sum_{i=2}^{n} |a_i|\Bigr) + 2(37M + 250\delta) + \Bigl(\sum_{i=2}^{n-1}(|b_i| - (68M + 628\delta))\Bigr) + (68M + 628\delta) \\
&\geqslant \tfrac{1}{2}\Bigl(\sum_{i=2}^{n} |a_i|\Bigr) + \tfrac{1}{2}\Bigl(\sum_{i=1}^{n} |b_i|\Bigr) - 128\delta \\
&\geqslant \tfrac{1}{2}|\gamma| - 128\delta,
\end{aligned}$$

and hence γ is a $(2, 128\delta)$-quasigeodesic in $\widetilde{X}$ in this case.

Case (2) ($\gamma = a_1b_1a_2b_2 \dots b_n a_{n+1}$ where $|a_1|, |a_{n+1}| \neq 0$) Since $\mathcal{A}$ is a Λ-separated collection, the path $\gamma_0 = b_1a_2b_2 \dots b_n$ satisfies the hypotheses of Theorem 5.6. Let γ_0' be the geodesic connecting the endpoints of γ_0. Then $|\gamma_0'| \geqslant |\gamma_0| - n(68M + 628\delta)$ by Theorem 5.6. By Theorem 5.6, γ_0' has a length at least $100M + 2000\delta$-tail in $\mathcal{N}_{2\delta}(B_n)$ at p_{n+1} and a $100M + 2000\delta$-tail at q_1 in $\mathcal{N}_{2\delta}(B_1)$.

Let γ_1 be the geodesic $[p_1, p_{n+1}]$. Let $\triangle_1$ be the geodesic triangle with sides a_1, γ_0' and γ_1. The corner length of $\triangle_1$ at q_1 is at most $5M + 57\delta$ because $\operatorname{diam} a_1 \cap \mathcal{N}_{5\delta}(B_1) \leqslant 5M + 57\delta$ and γ_0' has a long tail at q_1 in $\mathcal{N}_{2\delta}(B_1)$. Either $\triangle_1$ is δ-thin relative to $B \neq B_1$ so that the length of the fat part of γ_0' in $\triangle_1$ has length at most M because a long tail of γ_0' at q_1 is contained in $\mathcal{N}_{2\delta}(B_1)$, or $\triangle_1$ is δ-thin relative to B_1 in which case the length of the fat part of a_1 in $\triangle_1$ has length at most $5M + 57\delta$. Hence there is a point z_1 on γ_1 so that $d(z_1, q_1) \leqslant 10M + 116\delta$ because $\triangle_1$ is δ-relatively thin. Therefore by Lemma 5.13, $|\gamma_1| \geqslant |\gamma_0'| + |a_1| - (20M + 232\delta)$.

Next we want to show that γ_1 has a long tail at p_{i+1} in $\mathcal{N}_{2\delta}(B_n)$. If $\triangle_1$ is δ-thin relative to B_1, the corner length at q_1 is at most $5M + 57\delta$, and the fat part of γ_0' in $\triangle_1$ can have an at most length-M intersection with the at least $100M + 2000\delta$-tail of γ_0' at p_{i+1} that lies in $\mathcal{N}_{2\delta}(B_n)$. On the other hand, if $\triangle_1$ is δ-thin relative to $B \neq B_1$, then the corner length of $\triangle_1$ at q_1 is still at most $5M + 57\delta$ and the long tail of γ_0' at q_1 that lies in $\mathcal{N}_{2\delta}(B_1)$ forces the length of the fat part of γ_0' in $\triangle_1$ to be at most M. In both cases, all but $6M + 57\delta$ of the $100M + 2000\delta$-tail of γ_0' at p_{n+1} that lies in $\mathcal{N}_{2\delta}(B_n)$ must lie in the corner segment of γ_0' at p_{n+1}. Hence an at least $94M + 1000\delta$-tail of γ_1 at p_{n+1} must lie in $\mathcal{N}_{3\delta}(B_n)$.

Let $\triangle_2$ be the triangle with sides γ_1, a_n, γ'. By imitating the argument for $\triangle_1$, there is a point $z_2 \in \gamma'$ so that $d(z_2, p_{n+1}) \leqslant 10M + 116\delta$. Hence by Lemma 5.13,

$$|\gamma'| \geqslant |\gamma_1| + |a_n| - (20M + 232\delta)$$

so that

$$|\gamma'| \geqslant |a_0| + |\gamma_0'| + |a_n| - (40M + 464\delta)$$

and, by the computation from the previous case,

$$|\gamma'| \geqslant |a_0| + \tfrac{1}{2}|\gamma_0| - 128\delta + |a_n| - (40M + 464\delta) \geqslant \tfrac{1}{2}|\gamma| - 128\delta - (40M + 464\delta)$$

so that γ is a $(2, 40M + 592\delta)$-quasigeodesic in $\widetilde{X}$.

Case (3) ($\gamma = a_1 b_1 \dots a_n b_n$, $|a_1| \neq 0$ and $|b_n| \geqslant 37M + 250\delta$) Since $\mathcal{A}$ is a Λ-separated collection, the path $\gamma_0 = b_1 a_2 b_2 \dots b_n$ satisfies the hypotheses of Theorem 5.6. Let γ_0' be the geodesic connecting the endpoints of γ_0. Then $|\gamma_0'| \geqslant |\gamma_0| - (n-1)(68M + 628\delta)$ by Theorem 5.6. By an argument similar to the one in the previous case,

$$|\gamma'| \geqslant |\gamma_0'| + |a_1| - (20M + 232\delta)$$

and by arguments similar to the ones above,

$$|\gamma'| \geqslant \tfrac{1}{2}|\gamma| - (20M + 360\delta)$$

so in this case, γ is a $(2, 20M+360\delta)$-quasigeodesic in $\widetilde{X}$.

Case (4) ($\gamma = a_1 b_1 \dots a_n b_n$ where $|a_1| \neq 0$, $|b_n| \leqslant 37M + 250\delta$) By a previous case, the path $a_1 b_1 \dots a_n$ is a $(2, 40M+592\delta)$-quasigeodesic in $\widetilde{X}$. Hence γ is a $(2, 77M+1000\delta)$-quasigeodesic in $\widetilde{X}$.

Case (5) ($\gamma = b_1 \dots a_n b_n$ where $|b_1|, |b_n| < 37M + 250\delta$) Applying the immediately preceding case to $a_2 b_1 \dots a_n b_n$ and the fact that $|b_1| \leqslant 37M+250\delta$ implies that γ is a $(2, 114M+1250\delta)$-quasigeodesic in $\widetilde{X}$.

Case (6) ($\gamma = b_1 a_2 b_2 \dots a_n b_n$, where $|b_1| < 37M + 250\delta$ and $|b_n| \geqslant 37M + 250\delta$) By case (3), $a_2 b_2 \dots a_n b_n$ is a $(2, 20M+360\delta)$-quasigeodesic. Thus γ is a $(2, 57M+510\delta)$-quasigeodesic because $|b_1| < 37M + 250\delta$.

Now, assume $T/\sim$ is path connected. Let T_0 be the image of $T/\sim$ in $\widetilde{X}$. Let $x, y \in T/\sim$. Let ρ_T, ρ_{T_0}, ρ be the geodesics connecting x and y in $T/\sim$, T_0 and $\widetilde{X}$, respectively. Since $T/\sim$ is path connected, ρ_T maps to a path in T_0, $|\rho_{T_0}| \leqslant |\rho_T|$. From the preceding, $\frac{1}{2}|\rho_T| - (114M + 1592\delta) \leqslant |\rho|$. Combining these inequalities,

$$\tfrac{1}{2}|\rho_{T_0}| - (114M + 1592\delta) \leqslant |\rho| \leqslant |\rho_{T_0}|,$$

making ρ_{T_0} a $(2, 114M+1592\delta)$-quasigeodesic. □

Proposition 5.18 *Under Hypotheses 5.16, any geodesic in $T/\sim$ is not mapped to a loop in $\widetilde{X}$.*

Proof Let γ be a $T/\sim$-geodesic that maps to a loop in $\widetilde{X}$. If $\gamma \subseteq A \in \mathcal{A}$ or $\gamma \subseteq B \in \mathcal{B}$, then γ cannot map to a loop in A or a loop in B. Then γ can be written as a piecewise geodesic of the form

$$b_1 a_2 b_2 \dots a_n b_n,$$

where $b_i \subseteq B_i \in \mathcal{B}$ and $a_i \subseteq A_i \subseteq A \in \mathcal{A}$, $|b_1|, |b_n| \geqslant \frac{1}{2}\Lambda$ and $|b_i| \geqslant \Lambda$ for all $1 \leqslant i \leqslant n$. Since $\Lambda > 4(114M + 1592\delta)$, $|\gamma| > 2(114M + 1592\delta)$. Since γ maps to a $(2, 114M+1592\delta)$-quasigeodesic in $\widetilde{X}$, the distance between the endpoints of γ must be positive, so γ cannot map to a loop. □

6 The geometry of special cube complexes

6.1 Nonpositively curved cube complexes

A *cube complex* is a union of Euclidean cubes $[0,1]^n$ of possibly varying dimensions glued isometrically along faces. A *nonpositively curved* (*NPC*) cube complex is a cube complex such that the link of every vertex is a flag simplicial complex. See [29] Section 2.1 for details.

In each cube $[0,1]^n$, fixing one coordinate at $\frac{1}{2}$ makes a *codimension*-1 midcube. A *hyperplane* H is a connected union of midcubes glued isometrically along faces so that the intersection of H with any cube is either a codimension-1 midcube or empty. See Figure 7 for an example of an NPC cube complex and the link of a vertex.

6.2 Special cube complexes and separability

A *special cube complex* is a type of NPC cube complex developed by Wise and others whose hyperplanes are embedded, are 2-sided and avoid two other pathologies, see [29, Definition 4.2]. The important properties of special cube complexes that will be used in the following are the embeddedness and 2-sidedness of the hyperplanes and the fact that hyperplane subgroups of special cube complexes are separable (see Proposition 6.3).

A group is *special* if it is the fundamental group of a special cube complex. By work of Haglund and Wise [12], compact special groups embed into right angled Artin groups and are hence residually finite. Recall that if G is a group and H is a subgroup, H is *separable in* G if it is the intersection of the finite index subgroups containing H.

Passing to finite index subgroups is compatible with separability:

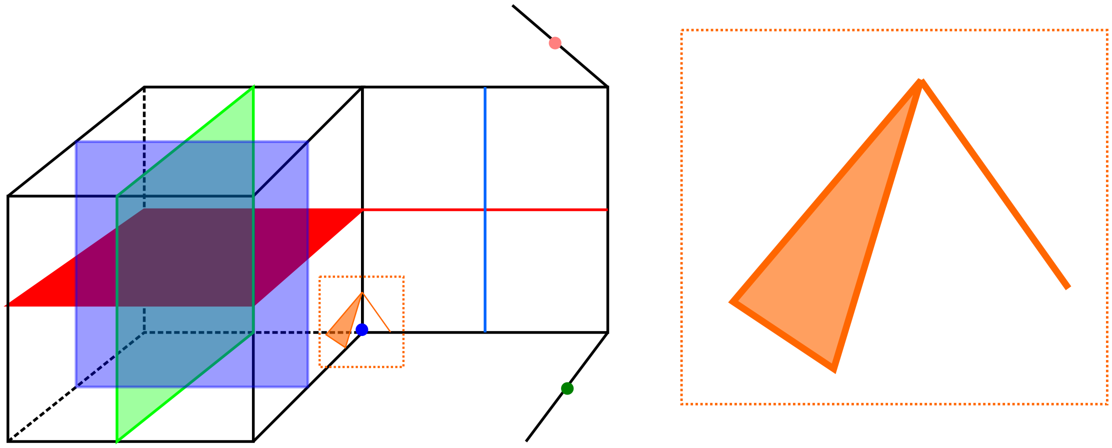


Figure 7: An example of an NPC cube complex (including a 3-cube) with its hyperplanes as well as the link of the blue vertex shown in orange and enlarged on the right.

Lemma 6.1 *Let G be a group, let G_0 be a finite index subgroup of G and let $H \leqslant G$. Then H is separable in G if and only if $H \cap G_0$ is separable in G_0.*

Theorem 6.2 (Scott's criterion, [27]) *Let X be a connected complex, $G = \pi_1 X$ and $H \leqslant G$. Let $p\colon X^H \to X$ be the cover corresponding to H. The subgroup H is separable in G if and only if for every compact subcomplex $Y \subseteq X^H$, there exists an intermediate finite cover $X^H \to \widehat{X} \to X$ such that $Y \hookrightarrow \widehat{X}$.*

Every finitely generated subgroup of a free group is separable. Likewise, special groups have an ample supply of separable subgroups. For example, the hyperplane subgroups of a special cube complex are separable:

Proposition 6.3 *Let X be a virtually special compact and nonpositively curved cube complex. Let W be a hyperplane of X. Then $\pi_1(W)$ is separable in $\pi_1(X)$.*

Proposition 6.3 follows from Haglund and Wise's canonical completion and retraction (see [29, Construction 4.12] or [12, Corollary 6.7]).

6.3 Elevations and R-embeddings

This subsection builds up the technical tools and terminology used to obtain finite covers whose hyperplanes elevate to sufficiently separated images in the universal cover.

The first step is to formalize the notion of an elevation:

Definition 6.4 Let W be a connected topological space and let $\phi\colon W \to Z$ be a continuous map. Let $p\colon \widehat{Z} \to Z$ be a covering map. There is a minimal covering $\widehat{p}\colon \widehat{W} \to W$ such that $\phi \circ \widehat{p}$ lifts to a map $\widehat{\phi}\colon \widehat{W} \to \widehat{Z}$. The map $\widehat{\phi}$ is an *elevation of W to $\widehat{Z}$*.

Often, the map $\widehat{W} \to \widehat{Z}$ will be implied and an elevation of ϕ will instead refer to the image of some elevation.

Elevations may not be unique: two elevations of the same map are *distinct* if they have different images.

When $\phi\colon W \to Z$ is an inclusion map, then the distinct elevations of ϕ are precisely the components of $p^{-1}(W)$.

Definition 6.5 Let X be a metric space, $R \geqslant 0$ and let $Y \subseteq X$ be connected. Let $p\colon X^Y \to X$ be the covering space associated to $\pi_1(Y)$ so that the inclusion $Y \hookrightarrow X$ lifts canonically to X^Y. The subspace Y is *R-embedded in X* if p is injective on $\mathcal{N}_R(Y) \subseteq X^Y$.

The following lemma is straightforward but will be important:

Lemma 6.6 *Let $p\colon \widehat{X} \to X$ be a finite regular cover. If A is R-embedded in X, then each component of $p^{-1}(A)$ is R-embedded in $\widehat{X}$.*

The main application of hyperplane separability is to show that every compact virtually special cube complex has a finite cover where every hyperplane is R-embedded.

Proposition 6.7 *Let X be a compact nonpositively curved cube complex, and let $V_1, V_2, \dots, V_n$ be hyperplanes of X so that $\pi_1 V_i$ is separable in $\pi_1 X$. Given $R \geqslant 0$, then there exists a finite regular cover C such that $V_1, \dots, V_n \subseteq C$ are R-embedded in C.*

If $\widetilde{W}_1, \widetilde{W}_2$ are distinct elevations of a hyperplane V of C to the universal cover $\widetilde{X}$, then $d_{\widetilde{X}}(\widetilde{W_1}, \widetilde{W_2}) \geqslant 2R$.

Proof For each hyperplane W of X, $\pi_1(W)$ is separable by Proposition 6.3. By Theorem 6.2, there exists a finite covering $\widehat{p}\colon \widehat{X} \to X$ such that there is an embedding $i_W\colon \mathcal{N}_R(W) \hookrightarrow \widehat{X}$.

Let $\widetilde{p}\colon \widetilde{X} \to X$, $p^W\colon \widetilde{X}^W \to X$ and $p\colon \widetilde{X} \to X^W$ be canonical covering maps so that $\widetilde{p} = p^W \circ p$. Let $\widetilde{W} \to \widetilde{W}_1, \widetilde{W} \to \widetilde{W}_2$ be distinct elevations of W to $\widetilde{X}$, and let $\widetilde{w}_1 \in \widetilde{W}_1$ and $\widetilde{w}_2 \in \widetilde{W}_2$.

Suppose toward a contradiction there exists a path $\gamma \subseteq \widetilde{X}$ with $|\gamma| \leqslant 2R$ between $\widetilde{W}_1$ and $\widetilde{W}_2$. Let $\widetilde{x} \in \gamma$ such that $d(\widetilde{x}, \widetilde{W}_1) < R$ and $d(\widetilde{x}, \widetilde{W}_2) < R$.

There exists $g \in \pi_1(X)$ such that $g \cdot \widetilde{w}_1 \in \widetilde{W}_2$, and $g \notin \pi_1(W)$ because otherwise $g \cdot \widetilde{w}_1 \in W_1 \cap W_2$ in which case $\widetilde{w}_1 \in \widetilde{W}_2$, but $\widetilde{w}_1 \notin \widetilde{W}_2$. Now $d(g \cdot \widetilde{x}, \widetilde{W}_2) \leqslant R$. Since $g \notin \pi_1(W)$, $p(\widetilde{x}) \neq p(g \cdot \widetilde{x})$. By definition of an elevation, $p(\widetilde{W}_2)$ is contained in the image of an inclusion of W into X^W. Also $p(\widetilde{x})$, $p(g \cdot \widetilde{x})$ lie in an R-neighborhood of the image of W in X^W. However,

$$p^W \circ p(\widetilde{x}) = \widetilde{p}(\widetilde{x}) = \widetilde{p}(g \cdot \widetilde{x}) = p^W \circ p(g \cdot \widetilde{x})$$

contradicting the fact that $i_W\colon \mathcal{N}_R(W) \hookrightarrow \widehat{X}$ is an embedding.

Suppose X has n hyperplanes. By passing to a finite cover if necessary, assume X^W is regular. The number of hyperplane orbits under deck transformations of X^W is at most n, and every hyperplane in the orbit of an elevation of W to X^W is R-embedded. Therefore, performing this procedure at most n times, will produce a finite cover $C \to X$ where every hyperplane is R-embedded. □

Proposition 6.7 will be used later in Section 7 to make the elevations of a hyperplane a $2R$-separated family in the sense of Definition 5.15.

6.4 Convex cores

Specialness also plays a role in building a geometric representation of the peripheral structure. In the hyperbolic case, Wise and others [11; 25] (see also [12, Proposition 7.2]) proved that quasiconvex subgroups of virtually special groups have "convex cores" in the CAT(0) universal cover. This fact and canonical completion and retraction can be used to show that hyperbolic special groups are *QCERF* or *quasiconvex extended residually finite* [12, Theorem 1.3] meaning that if G is hyperbolic and special, then every quasiconvex subgroup of G is separable.

A similar result exists in the relatively hyperbolic case. One might imagine that replacing the quasiconvex subgroup H by a relatively quasiconvex subgroup might yield a generalization; however, some care is required. In particular, a subgroup may stabilize a quasiconvex subset of a CAT(0) cube complex but may fail to stabilize a convex proper subcomplex, see Example 6.9.

Definition 6.8 If $\widetilde{X}$ is a CAT(0) cube complex and $\widetilde{Y} \subseteq \widetilde{X}$, the *cubical convex hull* of $\widetilde{Y}$ is the smallest convex sub*complex* of $\widetilde{X}$ containing $\widetilde{Y}$.

Example 6.9 Take the standard action of $\mathbb{Z}^2 = \langle (1,0), (0,1) \rangle$ on $\mathbb{R}^2$ by translation. The diagonal $D := \{(r,r) : r \in \mathbb{R}\}$ is a subspace stabilized by $L := \langle (1,1) \rangle \leqslant \mathbb{Z}^2$. The subgroup L is $(2,0)$-quasi-isometrically embedded in the given presentation of $\mathbb{Z}^2$, but the cubical convex hull of D is all of $\mathbb{R}^2$.

Full relatively quasiconvex subgroups eliminate these pathologies:

Definition 6.10 [26, Section 4] Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair and let H be a relatively quasiconvex subgroup of G. The subgroup H is a *full relatively quasiconvex subgroup* of G if for each $g \in G$ and $P \in \mathcal{P}$, either $gPg^{-1} \cap H$ is finite or $gPg^{-1} \cap H$ is finite index in gPg^{-1}.

Theorem 6.11 [26, Theorem 1.1] *Let X be a compact nonpositively curved cube complex with $G = \pi_1(X)$ hyperbolic relative to subgroups $P_1, \dots, P_n$. Let $\widetilde{X}$ be the* CAT(0) *universal cover of X. If H is a full relatively quasiconvex subgroup of G, then for any compact $U \subseteq \widetilde{X}$, then there exists an H-cocompact convex subcomplex $\widetilde{Y} \subseteq \widetilde{X}$ with $U \subseteq \widetilde{Y}$.*

By Proposition 2.15, if $(G, \mathcal{P})$ is a relatively hyperbolic group pair, the elements of $\mathcal{P}$ and their conjugates are relatively quasiconvex. By Proposition 2.2, the elements of $\mathcal{P}$ and their conjugates are full relatively quasiconvex. Therefore:

Lemma 6.12 *Let X be a nonpositively curved cube complex with* CAT(0) *universal cover $\widetilde{X}$ and $G := \pi_1(X)$. Let $(G, \mathcal{P})$ be a relatively hyperbolic pair. Let $x \in \widetilde{X}$ be a base point in the universal cover. For each $P \in \mathcal{P}$, there exists a $Z'(P, x)$ such that $Z'(P, x)$ is a P-cocompact convex subcomplex of $\widetilde{X}$ containing x.*

It follows immediately that there exists a $Q \geqslant 0$ such that the cubical convex hull of Px is contained in $\mathcal{N}_Q(Px)$.

7 A malnormal quasiconvex fully $\mathcal{P}$-elliptic hierarchy

For the following section, let X be a compact nonpositively curved cube complex with CAT(0) universal cover $\widetilde{X}$ and $G = \pi_1(X)$ hyperbolic relative to subgroups $\mathcal{P} := \{P_1, \dots, P_n\}$. Fix a base point $x \in \widetilde{X}$. By Lemma 6.12, there is a convex subcomplex $\widetilde{Z}'_{P,x}$ that is a P-cocompact convex subcomplex of $\widetilde{X}$ containing Px.

Let $\mathcal{B}_0 := \{g\widetilde{Z}'_{P,x} : g \in G,\ P \in \mathcal{P}\}$. By Proposition 4.6, there exists $f_0 \colon \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$ and $\delta \geqslant 2$ so that $(\widetilde{X}, \mathcal{B}_0)$ is a $(\delta{-}2, f_0)$-relatively hyperbolic pair.

Let $\widetilde{Z}_{P,x} = \mathcal{N}_{2\delta}(\widetilde{Z}'_{P,x})$. Theorem 6.11 implies that the collection $\mathcal{B}' = \{g\widetilde{Z}_{P,x} : g \in G,\ P \in \mathcal{P}\}$ is a thickening of $\mathcal{B}_0$. Proposition 4.4 implies there exists $f' \colon \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$ so that $(\widetilde{X}, \mathcal{B}')$ is a $(\delta{-}2, f')$-CAT(0) relatively hyperbolic pair. We also define $f \colon \mathbb{R}^{\geqslant 0} \to \mathbb{R}^{\geqslant 0}$ where $f(r) = f'(r+2)$. The function f will be useful later when we carry out the augmentation construction defined in Section 7.1.

To maintain consistency with previous notation, we will use the notation $M = f(6\delta)$ throughout Section 7. Proposition 5.3 implies:

Proposition 7.1 *For every $g \in G$, $g\widetilde{Z}_{P,x}$ is $(3M{+}6R{+}21\delta)$-attractive in the sense of Definition 5.1.*

7.1 Superconvexity, peripheral complexes and augmented complexes

Here we will prove that bi-infinite geodesics contained in a bounded neighborhood of $\widetilde{Z}_{P,x}$ actually lie in $\widetilde{Z}_{P,x}$.

Definition 7.2 Let X be a nonpositively curved cube complex and let $\phi \colon Z \to X$ be a local isometry. The map ϕ is *superconvex* if for any elevation $\widetilde{\phi} \colon \widetilde{Z} \hookrightarrow \widetilde{X}$ of Z to the universal cover $\widetilde{X}$ of X and any bi-infinite geodesic γ in $\widetilde{X}$ such γ lies in a bounded neighborhood of (the $\widetilde{\phi}$ image of) $\widetilde{Z}$ in $\widetilde{X}$, then γ is contained (in the $\widetilde{\phi}$ image of) $\widetilde{Z}$.

If the immersion $\phi \colon Z \to X$ is superconvex, then Z is said to be superconvex in X (with respect to ϕ).

Since $\widetilde{Z}_{P,x}$ is a P-cocompact convex subcomplex of $\widetilde{X}$, the quotient $\overline{Z}_{P,x} := P\backslash\widetilde{Z}_{P,x}$ is a cube complex and there is a natural local isometry $\phi_{P,x} \colon \overline{Z}_{P,x} \to X$ that carries $\overline{Z}_{P,x}$ to the image of $G\backslash\widetilde{Z}_{P,x}$ in X.

Proposition 7.3 *$\phi_{P,x}$ is superconvex.*

Proof Suppose γ is a bi-infinite geodesic contained in $\mathcal{N}_R(\widetilde{Z}_{P,x})$ and $p \in \gamma$. There exist $s_1, s_2 \in \gamma$ so that $p \in [s_1, s_2]$ and $d(s_i, p) > 3M + 6R + 21\delta$. Hence by Proposition 7.1 there exist points t_1, t_2 so that $t_1 \in [s_1, p]$ and $t_2 \in [p, s_2]$ so that $t_1, t_2 \in \widetilde{Z}_{P,x}$. Therefore by convexity $p \in \widetilde{Z}_{P,x}$. Hence $\gamma \subseteq \widetilde{Z}_{P,x}$. □

The complexes $\overline{Z}_{P,x}$ are called *peripheral complexes*. There is a convenient way to upgrade the immersion to an embedding:

Definition 7.4 Let X be a nonpositively curved cube complex with CAT(0) universal cover $\widetilde{X}$ and $G := \pi_1(X)$. Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair. Let $\mathcal{Z} := \bigsqcup_{P \in \mathcal{P}} \bar{Z}_{P,x}$, and let $\Phi\colon \mathcal{Z} \to X$ be the map so that $\Phi|_{\bar{Z}_{P,x}} = \phi_{P,x}$. The *augmented cube complex for the pair* (X, Φ) is the complex

$$C(X, \Phi) := X \cup \Big(\bigsqcup_{P \in \mathcal{P}} \bar{Z}_{P,x} \times [0,1] \Big) / (\bar{Z}_{P,x} \times \{1\}) \sim \phi_{P,x}(\bar{Z}_{P,x}),$$

consisting of the mapping cylinders of the $\phi_{P,x}$ identified along X.

The hyperplanes $\bar{Z}_{P,x} \times \frac{1}{2}$ are called *peripheral hyperplanes* while the remaining hyperplanes of $C(X, \Phi)$ are *nonperipheral*. Note that the nonperipheral hyperplanes of $C(X, \Phi)$ are in one-to-one correspondence with the hyperplanes of X. Since $\pi_1 X \cong \pi_1(C(X, \Phi))$, a (virtual) hierarchy for $\pi_1(C(X, \Phi))$ determines a (virtual) hierarchy of $\pi_1 X$.

Proposition 7.5 *Let $C(X, \Phi)$ be the augmented cube complex for the pair $(X, \mathcal{Z})$ as in Definition 7.4. If X is virtually special and W is a nonperipheral hyperplane of $C(X, \Phi)$, then $\pi_1 W$ is separable in $\pi_1 C(X, \Phi) \cong \pi_1 X$.*

Sketch The natural homotopy equivalence between $C(X, \Phi)$ and X that induces $\pi_1 C(X, \Phi) \cong \pi_1(X)$ brings nonperipheral hyperplanes of $C(X, \Phi)$ to hyperplanes of X. Therefore, W is homotopy equivalent to a hyperplane V of X and $\pi_1 V \cong \pi_1 W$ is separable in $\pi_1 X$ (recall Proposition 6.3). □

Technically, the definition of $C(X, \Phi)$ depends on the base point, but since the following results are given up to conjugacy, there is no need to keep track of base points.

Proposition 7.6 *Let $C(X, \Phi)$ be the augmented cube complex for $(X, \mathcal{Z})$ described in Definition 7.4. Let $\widetilde{C}$ be the universal cover of $C(X, \Phi)$. Let $\mathcal{B}$ be the collection of (images of) elevations of (images of) $\bar{Z}_{P,x} \times [0,1]$ in $C(X, \Phi)$ to $\widetilde{C}$.*

1. *Each $B \in \mathcal{B}$ is closed and convex.*
2. *$(\widetilde{C}, \mathcal{B})$ is a (δ, f)-CAT(0) relatively hyperbolic pair.*
3. *Every $B \in \mathcal{B}$ is $(3M + 6R + 2f(R) + 21\delta)$-attractive (recall Definition 5.1).*

Proof The universal cover $\widetilde{X}$ of X embeds as a closed convex subset of $\widetilde{C}$ so that each $B \in \mathcal{B}$ intersects $\widetilde{X}$ in some $\widetilde{Z}_{P,x}$. Since B intersects $\widetilde{X}$ in a closed convex subspace, B is closed and convex in $\widetilde{C}$.

Every geodesic triangle in $\widetilde{C}$ is Hausdorff distance 1 from a geodesic triangle in $\widetilde{X}$. Since triangles in $\widetilde{X}$ are $(\delta{-}2)$-thin relative to translates of $\widetilde{Z}_{P,x}$, triangles in $\widetilde{C}$ are δ thin relative to $\mathcal{B}$. For every B_1, B_2 in $\mathcal{B}$ with $B_1 \neq B_2$, $\mathcal{N}_t(B_1) \cap \mathcal{N}_t(B_2)$ is distance at most 1 from the intersection of $g_1 \mathcal{N}_t(\widetilde{Z}_{P_1,x})$ and $g_2 \mathcal{N}_t(\widetilde{Z}_{P_2,x})$ in $\widetilde{X}$ for some $g_1, g_2 \in G$ and $P_1, P_2 \in \mathcal{P}$, so the fact that $\widetilde{X}$ is a $(\delta{-}2, f')$-CAT(0) relatively hyperbolic pair implies that $(\widetilde{C}, \mathcal{B})$ is a (δ, f)-CAT(0) relatively hyperbolic pair.

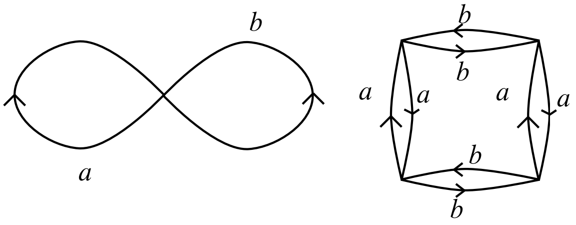


Figure 8: The figure-8 loop on the left whose two hyperplanes are the two edge midpoints and the double dot cover of the figure-8 loop on the right.

Let γ be a geodesic in $\widetilde{C}$ with endpoints in $\mathcal{N}_R(B)$ for some $B \in \mathcal{B}$. Since $\widetilde{X}$ is CAT(0) and B is convex, $\gamma \subseteq \mathcal{N}_R(B)$. Then γ is either contained in B' for some $B' \in \mathcal{B}$ in which case $|\gamma| \leqslant f(R)$ or γ has a subpath σ whose endpoints in $\widetilde{X}$ are at most $f(R)$ from the endpoints of γ. Therefore $|\sigma| \geqslant |\gamma| - 2f(R)$. There is some $g \in G$ and $P \in \mathcal{P}$ so that $g\widetilde{Z}_{P,x} = B \cap \widetilde{X}$. If the length of σ is at least $3M + 6R + 21\delta$, then $\sigma \cap g\widetilde{Z}_{P,x} \neq \varnothing$ by Proposition 7.1. Therefore, if the length of γ is at least $3M + 6R + 2f(R) + 21\delta$, $\varnothing \neq \gamma \cap g\widetilde{Z}_{P,x} \subseteq \gamma \cap B$. □

7.2 The double dot hierarchy

The construction of a hierarchy will use a finite cover called the *double dot cover* whose construction is originally due to Wise [30, Construction 9.1]. This treatment of the double dot cover is similar to the one in [3, Section 5].

Definition 7.7 [30, Construction 9.1] Let X be a cube complex, let $W \subseteq X$ be a hyperplane of X. Let γ be a based loop and let $[\gamma] \in \pi_1 X$. Then $[\gamma]$ has a well-defined (mod 2) intersection number with W. Let $\mathcal{W}$ be the set of embedded, 2-sided, nonseparating hyperplanes of X. For each $W \in \mathcal{W}$ let $i_W \colon \pi_1 X \to \mathbb{Z}/2\mathbb{Z}$ be the algebraic intersection map and define

$$\Psi \colon \pi_1 X \to \bigoplus_{W \in \mathcal{W}} \mathbb{Z}/2\mathbb{Z}, \quad \Psi = \bigoplus_{W \in \mathcal{W}} i_W.$$

The *double dot cover* of X is the cover corresponding to the subgroup $\ker \Psi \leqslant \pi_1 X$.

The double dot cover of a cube complex is usually a high-degree cover. Therefore, constructing examples can be quite difficult. Fortunately, the double dot cover of a rose with 2 petals is easy to construct:

Example 7.8 See Figure 8 for the double dot cover of the figure-8 loop.

An important feature of the double dot cover is that the cover is taken over nonseparating hyperplanes. This serves two purposes: first, making sure that double dot cover is not trivial and second, making sure that the double dot hierarchy constructed later has nontrivial splittings. There is a way to obtain a complex where every hyperplane is nonseparating:

Theorem 7.9 [6, Proposition 2.12] *Let X be a compact special NPC cube complex. Then X is homotopy equivalent to a compact special NPC cube complex whose hyperplanes are all nonseparating.*

Let X be a special cube complex with finitely many hyperplanes $\mathcal{W} := \{W_1, \dots, W_n\}$ where every hyperplane is nonseparating and let $\ddot{p}_X \colon \ddot{X} \to X$ be the double dot cover of X. The hyperplanes of $\ddot{X}$ are elevations of hyperplanes of X, and they divide $\ddot{X}$ in a natural way. Let $x \in \ddot{X} \setminus \bigcup \ddot{p}_X^{-1}(\mathcal{W})$ be a base vertex.

Each component of $\ddot{X} \setminus \bigcup \ddot{p}_X^{-1}(\mathcal{W})$ contains a lift of $\ddot{p}_X(x)$ because the hyperplanes of X are nonseparating. There is only one lift of $\ddot{p}_X(x)$ which lies in each component of $\ddot{X} \setminus \bigcup \ddot{p}_X^{-1}(\mathcal{W})$ because otherwise there is path ν between two points of $\ddot{p}_X^{-1}(x)$ that does not cross $\ddot{p}_X^{-1}(\mathcal{W})$. The path ν must project to a loop that represents a nonidentity element of $\pi_1(X) \setminus \ker \Psi$ but does not cross any $W \in \mathcal{W}$ which is impossible.

Since $\ker \Psi$ is normal, $\pi_1 X / \ker \Psi$ acts by deck transformations on $\ddot{X}$. This action induces a free and transitive action on $\ddot{p}_X^{-1}(x)$. Since each component of $\ddot{X} \setminus \bigcup \ddot{p}_X^{-1}(\mathcal{W})$ contains exactly one element of $\ddot{p}_X^{-1}(x)$, we can label each of the components of $\ddot{X} \setminus \bigcup \ddot{p}_X^{-1}(\mathcal{W})$ by an element of $\pi_1 X / \ker \Psi \cong \bigoplus_{W \in \mathcal{W}} \mathbb{Z}/2\mathbb{Z}$.

With data specified below in Hypotheses 7.10, we will use the labels for components of $\ddot{X} \setminus \bigcup \ddot{p}_X^{-1}(\mathcal{W})$ to construct a *double dot hierarchy* of spaces for the double dot cover $\ddot{C}$ of C. When the data in Hypotheses 7.10 satisfy certain criteria discussed in Section 7.3, the double dot hierarchy gives rise to a quasiconvex and fully $\mathcal{P}$-elliptic hierarchy of groups for $\pi_1(\ddot{C})$ which is isomorphic to a finite index subgroup of $\pi_1 X$. Passing to a particular finite cover will produce an induced hierarchy that is also malnormal. The next several paragraphs outline the construction of the double dot hierarchy as it is presented in [3, Section 5].

We now establish some baseline hypotheses for the remainder of Section 7.2.

Hypotheses 7.10 Let X be a compact special NPC cube complex so that:

- The hyperplanes of X are nonseparating.
- There exist a disjoint union $\mathcal{Z} := \bigsqcup_{i=1}^{n} \bar{Z}_i$ of NPC cube complexes together with a local isometric immersion $\Phi \colon \mathcal{Z} \to X$.
- Let C be the augmented cube complex $C(X, \Phi)$ and let $p \colon \ddot{C} \to C$ be its double dot cover.
- Let $\mathcal{W}$ be the nonperipheral hyperplanes of C and choose an ordering of the elements of $\mathcal{W}$ so that they are $W_1, W_2, \dots, W_n$.
- Additionally, C is a mapping cylinder for the map Φ, so we can view $\mathcal{Z}$ as an embedded subspace of C. In the language of Definition 7.4, $\mathcal{Z}$ is the image of $\bigsqcup_{i=1}^{n} \bar{Z}_i \times \{0\}$ in C.
- Let $\ddot{\mathcal{Z}} = p^{-1}(\mathcal{Z})$ be the preimage of $\mathcal{Z} \subseteq C$ under the double dot covering map.
- Fix a base vertex.

Each component of $\ddot{C} \setminus p^{-1}\big(\bigcup \mathcal{W}\big)$ is labeled (relative to the base vertex) by a vector $\hat{t} \in \bigoplus_{i=1}^{n} \mathbb{Z}/2\mathbb{Z}$. For each $1 \leqslant i \leqslant n$, let $\mathcal{W}_i$ be the first i hyperplanes and let $M_i = \bigoplus_1^i \mathbb{Z}/2\mathbb{Z}$. Then the complementary components of $\bigcup \mathcal{W}_i$ are labeled by elements of M_i. For each $\hat{t} \in M_i$, let $K_{\hat{t}}$ be the closure of the part labeled by $\hat{t}$.

For each $\hat{t} \in M_i$, a *$\hat{t}$-vertex space at level $n-i+1$* is a component of $K_{\hat{t}} \cup \ddot{\mathcal{Z}}$ that intersects $K_{\hat{t}}$. In the construction of the double dot hierarchy, the $\hat{t}$-vertex spaces at level $n-i+1$ specify all of the vertex spaces at each level, but the actual graph of spaces structure at each level must be described.

If A is the closure of a component of $p^{-1}(W_i) \setminus \bigcup_{j<i} p^{-1}(W_j)$, then A is called a *partly cut-up elevation of W_i*. The double dot hierarchy is constructed by cutting along an elevation of a hyperplane W_i to $\ddot{C}$ and any elements of $\ddot{\mathcal{Z}}$ that intersect W_i, but the elevation of the hyperplane W_i may have already been cut by one of the other hyperplane elevations of W_j with $j < i$.

By construction, any two $\hat{t}$-vertex spaces at level $n-i+1$ are either disjoint or intersect in a union of components of $\ddot{\mathcal{Z}}$ and disjoint partly cut-up elevations of W_i.

Now it is time to construct the graph of spaces structures at each level. Let $\hat{t} \in M_i$ and let V be the corresponding $\hat{t}$-vertex space at level $n-i+1$. Consider the canonical projection $\pi\colon M_{i+1} \to M_i$. Let $\hat{t}^+$ and $\hat{t}^-$ be the preimages of $\hat{t}$ under π. Let V^+ and V^- be the collections of complementary components of $V \setminus p^{-1}\big(\bigcup \mathcal{W}_{i+1}\big)$ labeled by $\hat{t}^+$ and $\hat{t}^-$, respectively. Then $V = V^+ \cup V^-$ and the components in V^+, V^- will serve as the vertex spaces in the graph of spaces decomposition of V in this hierarchy.

The edge spaces are components of $V^+ \cap V^-$. The attaching maps are the inclusion maps of edge spaces into vertex spaces while the realization is provided by a homotopy equivalence collapsing the mapping cylinders of the edge spaces onto the images of the edge spaces.

Let $\hat{t} \in M_n$. Then the components of the $\hat{t}$-vertex spaces are the vertex spaces of level 1 of the hierarchy, so the terminal spaces of the hierarchy are precisely these spaces.

Definition 7.11 The hierarchy $\mathcal{H}$ constructed in the preceding paragraphs with vertex spaces is called the *double dot hierarchy for the pair* $(X, \mathcal{Z})$.

The double dot hierarchy actually depends on an ordering on the hyperplanes, but the applications that follow only need an existence of a hierarchy given some local isometric immersion $\mathcal{Z} \to X$, so this complication will be henceforth ignored.

A version of the double dot hierarchy exists for general NPC cube complexes, see [3, Section 5.2]; however, the double dot hierarchy may fail to be faithful and even if it is faithful, the hierarchy may fail to be quasiconvex or malnormal. Also, the terminal spaces may not be useful. However, when hyperplanes are embedded, nonseparating and two-sided, the terminal spaces are easy to understand:

Lemma 7.12 [3, Lemma 5.5] *Assume Hypotheses 7.10. If Y is a terminal space of the double dot hierarchy for $(X, \mathcal{Z})$, then Y has a graph of spaces structure (Γ, χ) such that*

(1) *Γ is bipartite with vertex set $V(Y) = V(Y)^+ \sqcup V(Y)^-$,*

(2) *if $v \in V(Y)^+$, $\chi(v)$ is contractible,*

(3) *if $v \in V(Y)^-$, $\chi(v)$ is a component of $\ddot{\mathcal{Z}}$ and*

(4) *every edge space is contractible.*

Corollary 7.13 *Under Hypotheses 7.10, the fundamental group of a terminal space of the double dot hierarchy is a free product of the form $\big(\mathop{\ast}_{i=1}^{p} G_i\big) * F$ where F is a finitely generated free group and, for all $1 \leqslant i \leqslant p$, $G_i := \pi_1(Z_i)$ where Z_i is a component of $\mathcal{Z}$.*

7.3 A fully $\mathcal{P}$-elliptic malnormal quasiconvex hierarchy

Hypotheses 7.14 We set some basic hypotheses and notation for Section 7.3:

(1) Let X_0 be an NPC compact special cube complex.

(2) Let X be an NPC compact special cube complex that is homotopy equivalent to X so that the hyperplanes of X are all nonseparating (the existence of X follows from Theorem 7.9).

(3) Let $\widetilde{X}$ be the universal cover of X with base point $x \in \widetilde{X}$ that does not lie in any hyperplane.

(4) Let $G := \pi_1 X \cong \pi_1 X_0$ and suppose that $(G, \mathcal{P})$ is a relatively hyperbolic group pair.

(5) For each $P \in \mathcal{P}$, let $\phi_{P,x} : Z_P \to X$ be the superconvex local isometric immersions and let $\mathcal{Z} = \bigsqcup Z_P$ that arise as a consequence of Proposition 7.3. Let $\Phi \colon \mathcal{Z} \to X$ be the map that restricts to $\phi_{P,x}$ on Z_P.

(6) Let $C_1 = C(X, \Phi)$ be the augmented cube complex for (X, Φ) (recall Definition 7.4), and let $\widetilde{C}$ be its universal cover.

(7) Viewing C_1 as a mapping cylinder of Φ, Φ gives rise to a natural embedding $\mathcal{Z} \hookrightarrow C_1$. We call the components $Z_P \times \{0\}$ of the image of Φ *peripheral spaces*.

By strategically passing to finite covers and building the double dot hierarchy, we will produce a faithful, quasiconvex and fully $\mathcal{P}$-elliptic virtual hierarchy for $\pi_1 X$.

Lemma 7.15 (see [3, Lemma 5.18]) *Let C' be a finite regular cover of C_1.*

(1) *There exists a finite cover X' of X with $G' := \pi_1 X'$ and a superconvex local isometric immersion $\Phi' \colon \mathcal{Z}' \to X'$ such that $(G', \mathcal{P}')$ is the induced relatively hyperbolic group pair (see Proposition 2.13) and C' is the augmented cube complex of the pair $(X', \mathcal{Z}')$. The components of $\mathcal{Z}'$ have fundamental group isomorphic to elements of $\mathcal{P}'$ and for each component Z of $\mathcal{Z}'$, the image of $\pi_1 Z$ is conjugate to an element of $\mathcal{P}'$ in G'.*

(2) *Every nonperipheral hyperplane of C' is nonseparating.*

Notation 7.16 (1) Let $\mathcal{B}$ be the collection of elevations of $Z_P \times [0,1]$ (as determined by the mapping $\phi_{P,x}$) to $\widetilde{C}$. Let $\widetilde{\mathcal{Z}}$ be the union of the elements of $\mathcal{B}$ in $\widetilde{C}$.

(2) Recall from Proposition 7.6 that there exist (δ, f) so that $(\widetilde{C}, \mathcal{B})$ is a (δ, f)-CAT(0) relatively hyperbolic pair.

(3) Let $M = f(6\delta)$, let $\lambda = 4$ and $\epsilon = 10000(M + \delta + 1)$.

(4) Proposition 7.6 also implies that every $B \in \mathcal{B}$ is $(3M+6R+2f(R)+21\delta)$-attractive.

(5) Set L_{rftp} so that every pair of (λ, ϵ)-quasigeodesics in $\widetilde{C}$ $(L_{\mathrm{rftp}}, L_{\mathrm{rftp}})$-fellow travel relative to $\mathcal{B}$ (recall Definition 4.5 and Theorem 4.7).

(6) Let $R_{\mathrm{rftp}} = \lambda\big(\lambda(3f(L_{\mathrm{rftp}}) + \epsilon + 2L_{\mathrm{rftp}}) + \epsilon\big) + 2f(L_{\mathrm{rftp}})$.

(7) Let $R_0 > \max\{4, R_{\mathrm{rftp}}, 500M + 10000\delta\}$.

Observation 7.17 The constants established in items (2) and (4) of Notation 7.16 ensure that the pair $(\widetilde{C}, \mathcal{B})$ satisfies Hypotheses 5.5.

Using Propositions 6.7 and 7.5, let C_2 be a finite regular cover of C_1 such that every nonperipheral hyperplane of C_2 is R_0-embedded and nonseparating. Then C_2 is the augmented cube complex of a pair $(X_2, \mathcal{Z}'')$ where X_2 is a finite cover of X by Lemma 7.15. Recall that $\widetilde{X}$ naturally embeds in $\widetilde{C}$, which is also the universal cover of C_2. Let $G_2 = \pi_1(C_2)$ and let $(G_2, \mathcal{P}'')$ be the induced peripheral structure.

Let $c\colon \ddot{C}_2 \to C_2$ be the double dot cover of C_2. Let $(\ddot{G}_2, \ddot{\mathcal{P}}'')$ be the induced peripheral structure on $\ddot{G}_2 := \pi_1\ddot{C}_2$. The next few statements will show that the double dot hierarchy on $\ddot{C}_2$ is faithful, quasiconvex and fully $\ddot{\mathcal{P}}''$-elliptic hierarchy for $\pi_1\ddot{C}_2$. Passing to a finite regular cover will later yield a hierarchy which is also malnormal.

By Lemma 7.15, $\ddot{C}_2$ is an augmented cube complex with respect to a pair $(\ddot{X}_2, \ddot{\mathcal{Z}}_2)$ where $\ddot{\mathcal{Z}}_2$ consists of components of $c^{-1}(\mathcal{Z}'')$. Let E be an edge space of the double dot hierarchy on $\ddot{C}_2$. Then E is a union of partly cut-up elevations of a hyperplane of C_2 and elements of $\ddot{\mathcal{Z}}_2$.

Recall that $(\widetilde{C}, \mathcal{B})$ is a (δ, f)-CAT(0) relatively hyperbolic pair. Let $\widetilde{E}$ be an elevation of E to $\widetilde{C}$. There exist $\mathcal{A}_E$ and $\mathcal{B}_E$ so that $\mathcal{A}_E$ is a collection of elevations to $\widetilde{C}$ of convex partly cut-up hyperplane elevations of W and $\mathcal{B}_E$ is a collection of elevations of the peripheral spaces (recall Hypotheses 7.14(7)) to $\widetilde{C}$ so that $\widetilde{E}$ is the union of the elements of $\mathcal{A}_E$ and $\mathcal{B}_E$.

Each element $B_E \in \mathcal{B}_E$ is an elevation of a peripheral space. While B_E is not an element of $\mathcal{B}$, there is a unique $B'_E \in \mathcal{B}$ containing B_E. In particular, B'_E is the 1-neighborhood of B_E in $\widetilde{C}$. Let $\mathcal{B}'_E = \{B \in \mathcal{B} : B_E \subseteq B \text{ for some } B_E \in \mathcal{B}_E\}$ be the collection of elevations of the $Z_P \times [0,1]$ to $\widetilde{C}$ whose intersection with $\widetilde{X}$ is some $B_E \in \mathcal{B}_E$. See Figure 9. Let $\widetilde{E}'$ be the image of $\big(\bigsqcup \mathcal{A}_E\big) \sqcup \big(\bigsqcup \mathcal{B}'_E\big)$ in $\widetilde{C}$.

By Observation 7.17, the R_0-embeddedness of the hyperplane W and the construction of $\widetilde{E}'$ imply:

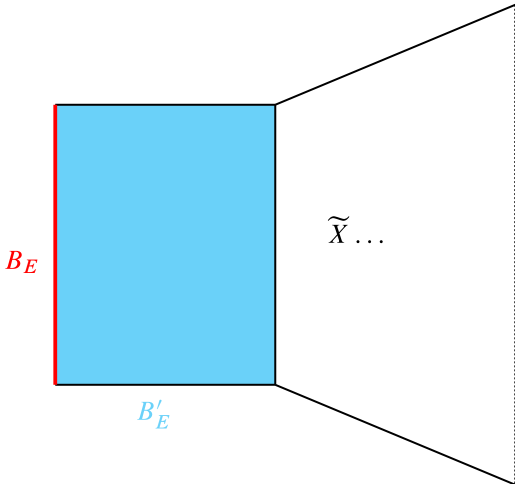


Figure 9: A schematic diagram showing the relationship between B_E, B'_E and their attachment to $\widetilde{X}$. The closure of the shaded region is B'_E.

Proposition 7.18 *The subspace $\widetilde{E}' \subseteq \widetilde{C}$ is a subspace of the form specified by Hypotheses 5.16.*

Proof Observation 7.17 ensures $(\widetilde{X}, \mathcal{B})$ satisfies Hypotheses 5.5.

Recall that C_2 has hyperplanes that are R_0-embedded (recall R_0 from Notation 7.16(7)), and recall that R_0-embeddedness of hyperplanes is preserved by finite covers (Lemma 6.6). Therefore, for all distinct pairs of $A_1, A_2 \in \mathcal{A}_E$, $d(A_1, A_2) \geqslant 2R_0$, and R_0 is large enough to provide the separation between elements of $\mathcal{A}_E$ required by Hypotheses 5.16.

By construction, $\mathcal{B}'_E \subseteq \mathcal{B}$, and $\widetilde{E}'$ is glued together from elements of $\mathcal{A}_E$ and $\mathcal{B}'_E$ as required. □

Proposition 7.19 *Let E be an edge space of the double dot hierarchy on $\ddot{C}_2$. Then the map $E \to \ddot{C}_2$ is π_1-injective.*

Proof Suppose not toward a contradiction. Then there exists a loop γ in E such that γ is essential in E but has trivial image in $\pi_1(\ddot{C}_2)$. Since γ is π_1 trivial in $\pi_1(\ddot{C}_2)$, γ elevates to a loop $\tilde{\gamma} \subseteq \widetilde{E}$ in $\widetilde{C}$. Since $\widetilde{E}$ is homotopy equivalent to $\widetilde{E}'$, there is a loop γ' in $\ddot{C}_2$ that is the image of a geodesic in $\widetilde{E}'$. Since $\widetilde{E}'$ is the image of $\left(\bigsqcup \mathcal{A}_E\right) \sqcup \left(\bigsqcup \mathcal{B}'_E\right)$ in $\widetilde{C}$, $\tilde{\gamma}'$ cannot be a loop by Proposition 5.18. □

The next step is to prove that the double dot hierarchy on $\ddot{C}_2$ is quasiconvex:

Proposition 7.20 *Recall λ, ϵ from Notation 7.16. If E is an edge space of the double dot hierarchy on $\ddot{C}_2$ and $\widetilde{E}$ is the universal cover of E, then any elevation $\widetilde{E} \hookrightarrow \widetilde{C}$ of E to $\widetilde{C}$ is a (λ, ϵ)-quasi-isometric embedding.*

Proof Let γ be a geodesic in $\widetilde{E}$ and let γ' be a geodesic with the same endpoints in $\widetilde{E'}$. Let γ'' be a geodesic in $\widetilde{C}$ with the same endpoints as γ. Proposition 7.18 implies we can use Proposition 5.17, which implies that γ' is a $(2, 114M+1592\delta)$-quasigeodesic in $\widetilde{C}$. Let n be the smallest number so that γ' can be written as $a_1 b_1 \dots b_n a_{n+1}$ where

(1) a_i can be a point if $i = 1$ or $i = n+1$,

(2) otherwise a_i is geodesic in some $A_i \in \mathcal{A}_E$, and

(3) b_i is geodesic in some $B'_i \in \mathcal{B}'_E$.

The endpoints of each b_i lie in $\widetilde{E}$ because every $A_i \subseteq \widetilde{E}$ and γ, γ' have the same endpoints. Thus each b_i can be replaced by a path of length $|b_i| + 2$ that lies entirely in $\widetilde{E}$. It is therefore possible to produce a path in $\widetilde{E}$ between the endpoints of γ whose length is at most $|\gamma'| + 2n$, so $|\gamma| \leqslant |\gamma'| + 2n$. Further, $|b_i| \geqslant R_0 \geqslant 4$ for $1 < i < n$ because the A_i are R_0-separated, so we have that $|\gamma| \geqslant 4n - 8$ which implies

$$2n \leqslant \tfrac{1}{2}|\gamma| + 4. \tag{4}$$

Therefore

$$\begin{aligned} |\gamma''| &\geqslant \tfrac{1}{2}|\gamma'| - (114M + 1592\delta) \\ &\geqslant \tfrac{1}{2}(|\gamma| - 2n) - (114M + 1592\delta) \\ &\geqslant \tfrac{1}{2}\big(\tfrac{1}{2}|\gamma| - 4\big) - (114M + 1592\delta) \\ &\geqslant \tfrac{1}{4}|\gamma| - (114M + 1592\delta + 2), \end{aligned}$$

where the third line follows from the second by the estimate in (4). Hence γ is a $(4, 114M+1592\delta+2)$-quasigeodesic in $\widetilde{C}$. □

Propositions 7.19 and 7.20 together yield the following:

Corollary 7.21 *The double dot hierarchy induced on $\pi_1 \ddot{C}_2$ is faithful and quasiconvex.*

The next step is to prove that the double dot hierarchy on $\ddot{C}_2$ is fully $\ddot{\mathcal{P}}''$-elliptic. Definition 7.22 introduces geometric terminology for the situation where a subgroup of a relatively hyperbolic group pair $(G, \mathcal{P})$ contains an element g conjugate into a peripheral subgroup P such that no positive power of g lies in $E \cap P$.

Definition 7.22 Let Y be a locally convex subspace of $\ddot{C}_2$. Let $E \subseteq \ddot{C}_2$. The subspace E has an *accidental Y-loop* if there exists a homotopically essential loop, γ, which is both freely homotopic to a geodesic loop in Y and has no positive power homotopic in E to a geodesic loop in Y.

The next few statements will show that the edge spaces of the double dot hierarchy for $\ddot{C}_2$ have no accidental $\ddot{\mathcal{Z}}''$-loops. This will imply the hierarchy is fully $\ddot{\mathcal{P}}''$-elliptic. Elevations of partly cut-up hyperplanes do not have accidental $\ddot{\mathcal{Z}}''$-loops:

Lemma 7.23 [3, Lemma 5.15] *Let $(X, \mathcal{Z})$ be a superconvex pair where each component of $\mathcal{Z}$ is embedded and let C be the corresponding augmented cube complex. For $n \geqslant 1$, let $\{W_1, \dots, W_n\}$ be a collection of embedded, 2-sided, nonseparating hyperplanes of C. Let Q be a component of $W_n \setminus \bigcup_{i<n} W_i$. Then Q has no accidental $\mathcal{Z}$-loops.*

Proposition 7.24 *Let E be an edge space of the double dot hierarchy for $\ddot{C}_2$. Then E has no accidental $\ddot{\mathcal{Z}}''$-loops.*

Proof Recall that E is a union of a partly cut-up hyperplane elevations and components of $\ddot{\mathcal{Z}}''$ that intersect these elevations. Let Q be one of the partly cut-up hyperplane elevations. By Lemma 7.23, Q has no accidental $\ddot{\mathcal{Z}}''$-loops.

Suppose there exists a $\ddot{C}_2$-essential loop γ in E such that γ is freely homotopic in $\ddot{C}_2$ into $\ddot{\mathcal{Z}}''$. Then a representative of the homotopy class of γ lifts to a bi-infinite $\widetilde{E}$-geodesic $\widehat{\gamma}$ where $\widetilde{E}$ is an elevation of E to $\widetilde{C}$, and a representative of the homotopy class of γ lifts to a bi-infinite $\widetilde{C}$-geodesic $\rho \subseteq \widetilde{Z}$, an elevation of a component of $\ddot{\mathcal{Z}}''$ and there exists $R \geqslant 0$ so that $\widehat{\gamma} \subseteq \mathcal{N}_R(\rho)$.

Since $\widehat{\gamma}$ is a $\widetilde{E}$-geodesic, $\widehat{\gamma}$ is a (λ, ϵ)-quasigeodesic in $\widetilde{C}$ by Proposition 7.20.

Let $\widehat{\gamma}_0$ be a subsegment of $\widehat{\gamma}$ with $|\widehat{\gamma}_0| = |\gamma|$ (eg take $\widehat{\gamma}_0$ to be the subsegment between two consecutive lifts of a point of γ to $\widehat{\gamma}$). If $\widehat{\gamma}_0 \subseteq \widetilde{Z}'$ where $\widetilde{Z}'$ is an elevation of a component of $\ddot{\mathcal{Z}}''$, then $\widehat{\gamma} \subseteq \widetilde{Z}'$ and $\widehat{\gamma}$ is geodesic in $\widetilde{C}$. Then $\widetilde{Z} = \widetilde{Z}'$ because $\operatorname{diam} \mathcal{N}_R(\widetilde{Z}) \cap \mathcal{N}_R(\widetilde{Z}') = \infty$ in which case γ was not an accidental $\ddot{\mathcal{Z}}''$-loop.

On the other hand, if $\widehat{\gamma}_0 \subseteq \widetilde{Q}$ where $\widetilde{Q}$ is some elevation of Q to $\widetilde{C}$, then Q has an accidental $\ddot{\mathcal{Z}}''$-loop, contradicting the fact that there are no such accidental $\mathcal{Z}$-loops.

Therefore, there exist subsegments of $\widehat{\gamma}$ of the form $\gamma_m = a_{m,1} b_{m,1} a_{m,2} b_{m,2} \dots a_{m,k_m} b_{m,k_m}$ such that $\bigcup_1^\infty \gamma_m = \widehat{\gamma}$, $|\gamma_m| \to \infty$ and $k_m \to \infty$ as $m \to \infty$, $a_{m,i}$ lies in an elevation $\widetilde{Q}_i$ of Q to $\widetilde{C}$, $b_{m,i} \subseteq \widetilde{Z}_{m,i}$ where $\widetilde{Z}_{m,i}$ is an elevation of a component of $\ddot{\mathcal{Z}}''$ to $\widetilde{C}$, and if $i \neq j$, $b_{m,i} \subseteq \widetilde{Z}_i$ and $b_{m,j} \subseteq \widetilde{Z}_j \neq \widetilde{Z}_i$ (otherwise, by convexity of Z_i, γ_m could be written as a concatenation of fewer geodesic segments). Recall that Q is R_0-embedded, so for all m, i, $|b_{m,i}| \geqslant R_0$.

By construction there is a unique $B \in \mathcal{B}$ so that $\widetilde{Z} \subseteq B$. Let τ_m be the $\widetilde{C}$-geodesic connecting the endpoints of γ_m. Since $\tau_m \subseteq \mathcal{N}_R(B)$ and B is $(3M+6R+2f(R)+9\delta)$-attractive, all but $(3M+6R+2f(R)+9\delta)$-tails of the endpoints of τ_m lie in B. Therefore, there exists a subsegment $\tau_m^B \subseteq \tau_m \cap B$ so that $|\tau_m^B| \geqslant |\tau_m| - 2(3M + 6R + 2f(R) + 9\delta)$.

Recall that all (λ, ϵ)-quasigeodesics with the same endpoints $(L_{\mathrm{rftp}}, L_{\mathrm{rftp}})$-fellow travel relative to $\mathcal{B}$. There exists a unique $B_{m,i} \in \mathcal{B}$ containing $\widetilde{Z}_{m,i}$, so $b_{m,i} \subseteq B_{m,i}$. Then for $B \in \mathcal{B}$ with $B \neq B_{m,i}$, $\operatorname{diam} b_{m,i} \cap \mathcal{N}_{L_{\mathrm{rftp}}}(B) \leqslant f(L_{\mathrm{rftp}})$. Since τ_m and γ_m relatively fellow travel, either

- there exist points $p_{m,i}^-$ and $p_{m,i}^+$ on $b_{m,i} \subseteq \gamma_m$ that are at most $f(L_{\mathrm{rftp}})$ from the endpoints of $b_{m,i}$ and are distance L_{rftp} from τ_m or

- there exist $p^-_{m,i}, p^+_{m,i}$ on γ_m so that $p^-_{m,i}, p^+_{m,i}$ are distance at most L_{rftp} from points in τ_m that lie in $\mathcal{N}_{L_{\mathrm{rftp}}}(B_{m,i})$ and the interval of γ_m between $p^-_{m,i}$ and $p^+_{m,i}$ contains all of $b_{m,i}$ except for a length at most $2(f(L_{\mathrm{rftp}}))$ subsegment of $b_{m,i}$.

Indeed, any subsegment $b_{m,i}$ that lies in $\mathcal{N}_{L_{\mathrm{rftp}}}(B)$ for any $B \in \mathcal{B}$ with $B \neq B_{m,i}$ has length at most $f(L_{\mathrm{rftp}})$. In either case since γ_m is (λ, ϵ)-quasigeodesic, there exists a length $\frac{1}{\lambda}\bigl(\frac{1}{\lambda}\bigl(R_0-2(f(L_{\mathrm{rftp}}))\bigr)-\epsilon\bigr)-\epsilon-2L_{\mathrm{rftp}}$ subsegment of τ_m that lies in $\mathcal{N}_{L_{\mathrm{rftp}}}(B_{m,i})$. As $m \to \infty$, $|\tau_m| \to \infty$ while

$$|\tau_m| - |\tau_m^B| \leqslant 2(3M + 6R + 2f(R) + 9\delta),$$

which does not depend on m. Therefore, for $m \gg 0$, there are at least two i such that τ_m^B has a length

$$\frac{1}{\lambda}\left(\frac{1}{\lambda}\bigl(R_0 - 2(f(L_{\mathrm{rftp}}))\bigr) - \epsilon\right) - \epsilon - 2L_{\mathrm{rftp}} > 3f(L_{\mathrm{rftp}})$$

subsegment lying in $\mathcal{N}_{L_{\mathrm{rftp}}}(B_{m,i}) \cap \mathcal{N}_{L_{\mathrm{rftp}}}(B)$ (recall R_0 was chosen in Notation 7.16). Since the $B_{m,i}$ are pairwise distinct, we obtain a contradiction. Therefore, γ cannot be an accidental $\ddot{Z}''$-loop. □

Corollary 7.25 *The double dot hierarchy on $\ddot{C}_2$ is fully $\ddot{\mathcal{P}}''$-elliptic.*

Faithfulness, quasiconvexity and full $\mathcal{P}$-ellipticity are preserved by taking the induced hierarchy of a finite regular cover of $\ddot{C}_2$. The final step is to show that there exists a finite cover of $\ddot{C}_2$ whose induced hierarchy is also a malnormal hierarchy.

The following lemma is straightforward:

Lemma 7.26 *Suppose $H \leqslant G$ and G_0 is a finite index subgroup of G and let $H_0 = H \cap G_0$. If H is malnormal in G, then H_0 is malnormal in H.*

The following is a special case of [26, Corollary 6.4]:

Proposition 7.27 *Let G be the fundamental group of a relatively hyperbolic special compact NPC cube complex, and let $H \leqslant G$ be full relatively quasiconvex. Then H is separable in G.*

Proposition 7.28 (Hruska–Wise [18, Theorem 9.3]) *If G is relatively hyperbolic and $H \leqslant G$ is relatively quasiconvex and separable, then there exists a finite index subgroup $K_0 \leqslant G$ containing H such that for every $g \in K_0 \setminus H$ either $gHg^{-1} \cap H$ is finite or $gHg^{-1} \cap H$ is parabolic in K.*

Proposition 7.29 *If G is relatively hyperbolic and $H \leqslant G$ is full relatively quasiconvex, there is a finite index subgroup $K \leqslant G$ containing H such that H is almost malnormal in G.*

Proof We first prove the following claim: If $H \leqslant G$ is full relatively quasiconvex, then there are only finitely many double cosets of the form HgH so that $H \cap H^g$ is infinite and parabolic.

Let $\mathcal{D}$ be the induced peripheral structure on H. If $H \cap H^g$ is infinite parabolic, then fullness implies there are $Q_1, Q_2 \leqslant H$ that are maximal parabolic in H so that $H \cap H^g$ is finite index in $Q_1 \cap Q_2^g$. Then there exist $D_1, D_2 \in \mathcal{D}$ and $h_1, h_2 \in H$ so that $Q_1 = D_1^{h_1}$ and $Q_2 = D_2^{h_2}$. It is easy to verify that if $g_0 = h_1^{-1} g h_2$, then

(1) $g_0 \in HgH$,

(2) $HgH = Hg_0H$, and

(3) $H \cap H^{g_0} \leqslant D_1 \cap D_2^{g_0}$.

In other words, given a double coset HgH so that $H \cap H^g$ is infinite parabolic, we may assume that g is chosen so that there are maximal parabolic $D_1, D_2 \leqslant H$ so that $H \cap H^g \leqslant D_1 \cap D_2^g$.

Since $\mathcal{D}$ is finite, it suffices to show that for any $D_1, D_2 \in \mathcal{D}$ (D_1, D_2 need not be distinct) there are finitely many double cosets of the form HgH so that $H \cap H^g$ is infinite and $H \cap H^g \subseteq D_1 \cap D_2^g$.

Now suppose Hg_1H is another double coset so that $H \cap H^{g_1}$ is an infinite subgroup of $D_1 \cap D_2^g$. We see that $D_1^{g^{-1}}$ and $D_1^{g_1^{-1}}$ have infinite intersection with D_2 and are therefore finite index in D_2 by fullness, so $D_1^{g^{-1}} \cap D_1^{g_1^{-1}}$ is infinite and hence $D_1 \cap D_1^{gg_1^{-1}}$ is infinite. Let P be the maximal parabolic subgroup of G containing D_1. The fullness of H implies that D_1 is finite index in P. Therefore, $P \cap P^{gg_1^{-1}}$ is infinite, so $gg_1^{-1} \in P$. There are finitely many left cosets $t_1D_1, t_2D_1, \dots, t_\ell D_1$ of D_1 in P. Hence $gg_1^{-1} = t_i d$ for some $d \in D_1 \leqslant H$ and $1 \leqslant i \leqslant \ell$ which means $g_1^{-1} = g^{-1}t_i d$, so $Hg_1^{-1}H = Hg^{-1}t_iH$. There are only finitely many choices for t_i, proving the claim.

Proposition 7.28 implies that if we first pass to a finite index $K_0 \leqslant G$ containing H, we can ensure that if $g \in K_0 \setminus H$ and $H \cap H^g$ is not finite, it is infinite parabolic. By the preceding, there is a finite collection of double cosets $Hk_1H, \dots, Hk_mH$ so that $g \in Hk_iH$ for some $1 \leqslant i \leqslant m$. Note all $k_i \notin H$. The separability of H implies that we can choose a finite index $K \leqslant K_0$ containing H so that $k_1, \dots, k_m \notin K$. Then $Hk_iH \cap K = \varnothing$ because $H \leqslant K$. By the preceding, there exists no $k \in K$ such that $H \cap H^k$ is infinite parabolic, so $H \cap H^k$ is finite for all $k \in K$. □

Corollary 7.30 is based on [3, Corollary 3.29]. Corollary 7.30 follows immediately from the two preceding statements and the fact that when G is virtually special, G is linear and hence virtually torsion free.

Corollary 7.30 *If G is hyperbolic relative to $\mathcal{P}$ and special, and $H \leqslant G$ is full relatively quasiconvex, then H is virtually malnormal.*

Theorem 7.31 *Let G be special, virtually torsion-free and let $(G, \mathcal{P})$ be a relatively hyperbolic group pair. Let $\mathcal{H}$ be a fully $\mathcal{P}$-elliptic quasiconvex hierarchy for G. Then there exists a finite index normal subgroup $G_0 \leqslant G$ with induced fully $\mathcal{P}$-elliptic quasiconvex hierarchy $\mathcal{H}_0$ of G_0 which is malnormal and fully $\mathcal{P}$-elliptic.*

The proof here is nearly the same as in [3, Theorem 3.30].

Proof Because $\mathcal{H}$ is fully $\mathcal{P}$-elliptic, the edge subgroups are full. Since there are finitely many edge groups, by Corollary 7.30, there exists some G_0 such that for every edge group E of $\mathcal{H}$, $E \cap G_0$ is malnormal in G_0. By passing to a deeper finite index subgroup, we may insist that G_0 is normal. Since G_0 is normal, conjugation by $g \in G$ is an automorphism of G_0, so in particular, these edge groups $E \cap G_0$ are malnormal in G. □

At last, it is time to prove Theorem 1.

Theorem 1 *Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair and let G be a virtually compact special group. Then there exists a finite index subgroup $G_0 \leqslant G$ and an induced relatively hyperbolic group pair $(G_0, \mathcal{P}_0)$ so that G_0 has a quasiconvex, malnormal and fully $\mathcal{P}_0$-elliptic hierarchy terminating in groups isomorphic to elements of $\mathcal{P}_0$.*

Proof of Theorem 1 Let X be an NPC compact special cube complex so that $\pi_1(X)$ is finite index in G.

First, pass to a finite index regular cover of X, X_1 that is special. By applying a homotopy equivalence, X_1 is homotopy equivalent to a cube complex where every hyperplane gives a nontrivial splitting of $\pi_1 X_1$ (see [3, Lemma 5.17]).

By Corollary 7.21, there exists a special cube complex X_1' homotopy equivalent to X_1 with a finite regular cover X_2 such that $G_2 := \pi_1 X_2$ with induced peripheral structure $(G_2, \mathcal{P}_2)$ has a faithful, quasiconvex, fully $\mathcal{P}_2$-elliptic hierarchy terminating in $\mathcal{P}_2 * F_k$ where F_k is a free group.

By Theorem 7.31, there exists a finite regular cover X_0 with $G_0 := \pi_1 X_0$ and induced peripheral structure $(G_0, \mathcal{P}_0)$ such that the induced hierarchy on G_0 is malnormal as well and terminates in free products of free groups and elements of $\mathcal{P}_0$ (recall Corollary 7.13). The hierarchy can then be continued to a malnormal, quasiconvex, fully $\mathcal{P}_0$-elliptic one that terminates in $\mathcal{P}_0$. □

8 A relatively hyperbolic version of the malnormal special quotient theorem

Recall Wise's malnormal special quotient theorem (MSQT), see Theorem 1.3 above or [30, Theorem 12.2] mentioned in the introduction. The purpose of this section is to apply Theorem 1 to obtain a relatively hyperbolic version of Wise's MSQT using techniques from [3, Sections 6-9].

Wise's quasiconvex hierarchy theorem [30, Theorem 13.3] has the following useful consequence:

Corollary 8.1 *Let G be a hyperbolic group with a quasiconvex hierarchy terminating in finite groups. Then G is virtually special.*

The technique for proving a relatively hyperbolic analog of Theorem 1.3 will be to start with the hierarchy provided by Theorem 1 and strategically take quotients using group-theoretic Dehn fillings (see Definition 8.2). These quotients can be constructed to be hyperbolic, and with some care, the hierarchy structure can be passed down to the quotient so that Corollary 8.1 can be used. In [3], the authors avoided using Corollary 8.1 because their account aimed to give a new proof of auxiliary results used to prove Corollary 8.1. Consequently, they needed to ensure that the hierarchy structure on the quotient is also a malnormal hierarchy. By using Corollary 8.1, we only need a quasiconvex hierarchy for such a quotient.

8.1 Group-theoretic Dehn filling

For this section, let $(G, \mathcal{P})$ be a relatively hyperbolic group pair where $\mathcal{P} = \{P_1, \dots, P_m\}$ unless stated otherwise. When M is a finite volume hyperbolic 3-manifold with torus cusps, a *Dehn filling* of M is a gluing of solid tori $T_i \cong D \times S^1$ by a diffeomorphism to the boundary components. The result of the gluing depends only on the isotopy class of the curve $\gamma_i \subseteq \partial M$ that each copy of $\partial D \times \{p\} \subseteq T_i$ is glued to (see eg [22, Section 10.1]). In this situation $\pi_1 M$ is hyperbolic relative to a collection of copies of $\mathbb{Z}^2$, one for each boundary component of M.

The next definition is a group-theoretic analog of Dehn filling.

Definition 8.2 Let $\{N_i \lhd P_i : 1 \leqslant i \leqslant m\}$. Then there exists a *group-theoretic Dehn filling* of G with *filling map* π defined by the quotient

$$\pi : G \to G(N_1, \dots, N_m) := G/\langle\langle \bigcup N_i \rangle\rangle.$$

The subgroups N_i are called *filling kernels*.

A filling is called *peripherally finite* if each filling kernel N_i is finite index in P_i.

For a classical filling, if every T_i is filled by gluing along the curves γ_i that are sufficiently long, Thurston's Dehn filling theorem says that the resulting manifold is hyperbolic. The group-theoretic analog of a sufficiently long classical Dehn filling is a group-theoretic Dehn filling where the filling kernels avoid a finite set of elements:

Definition 8.3 A statement $\mathfrak{P}$ holds for all sufficiently long fillings if there exists a finite $B \subseteq G \setminus \{1\}$ such that whenever $B \cap N_i = \varnothing$ for all $1 \leqslant i \leqslant m$, the filling $G(N_1, \dots N_m)$ has $\mathfrak{P}$.

Osin showed that sufficiently long Dehn fillings of relatively hyperbolic groups are relatively hyperbolic, have kernels which intersect each peripheral subgroup P_i precisely in N_i and can be manipulated so that any finite set of elements are not killed by the filling map.

Theorem 8.4 [23, Theorem 1.1] *Let $F \subseteq G$ be any finite subset of G. Then for all sufficiently long Dehn fillings,*

1. $\ker(\phi|_{P_i}) = N_i$ *for* $i = 1, 2, \dots, m$,
2. *the pair* $(G(N_1, \dots, N_m), \{\phi(P_1), \dots, \phi(P_m)\})$ *is a relatively hyperbolic group pair, and*
3. $\phi|_F$ *is injective.*

The edge subgroups of the hierarchy from Theorem 1 will need to be full relatively quasiconvex subgroups of G. The quasiconvexity of the hierarchy will ensure that these subgroups are relatively quasiconvex.

Theorem 8.5 [16, Theorem 1.5] *Let $H \leqslant G$ be a quasi-isometrically embedded subgroup. Then H is relatively quasiconvex in G.*

Theorem 8.6 [16, Theorem 1.2] *Let $H \leqslant G$ be relatively quasiconvex. Then there exists a relatively hyperbolic structure $(H, \mathcal{D})$ where $\mathcal{D}$ is finite and every element of $\mathcal{D}$ is conjugate into an element of $\mathcal{P}$.*

Corollary 8.7 *The collection $\mathcal{D}$ can be chosen so that*

(1) *every element of $\mathcal{D}$ is infinite, and*

(2) *whenever $H \cap P^g$ is infinite, for some $g \in G$, there exists $h \in H$ so that $(H \cap P^g)^h$ is an element of $\mathcal{D}$.*

Proof For the first statement, simply remove all finite elements of $\mathcal{D}$. The second statement follows from [16, Theorem 9.1]. □

When a filling of G interacts nicely with a subgroup H, it is possible to induce a filling on the subgroup H.

Definition 8.8 [21, Definition B.1] Let $H \leqslant G$. A filling $G \to G(N_1, \dots, N_m)$ is an *H-filling* if whenever $gP_ig^{-1} \cap H$ is infinite for some $P_i \in \mathcal{P}$, then $gN_ig^{-1} \subseteq H$.

Definition 8.9 Suppose $H \leqslant G$ is a relatively quasiconvex subgroup and let $(H, \mathcal{D})$ be the relatively hyperbolic structure from Theorem 8.6 and Corollary 8.7. Let $\pi\colon G \to G(N_1, \dots, N_m)$ be an H-filling. Let $D_j \in \mathcal{D}$. Then there exists some $P_i \in \mathcal{P}$ and $g \in G$ with $g^{-1}D_jg \subseteq P_i$. Let $K_j := gN_ig^{-1}$. Since π is an H-filling, $K_j \lhd D_j$, so the groups K_j determine a filling

$$\pi_H\colon H \to H(K_1, \dots, K_N)$$

called the *induced filling of H* with respect to $G(N_1, \dots, N_m)$.

Since N_i is normal in P_i, the groups K_j (and hence the filling) do not depend on the choice of $g \in G$. The following theorem appears as stated in [3] as Theorem 7.11 and collects results about induced Dehn fillings from [2]:

Theorem 8.10 *Let $H \leqslant G$ be a full relatively quasiconvex subgroup and let $F \subseteq G$ be a finite subset. For all sufficiently long H-fillings, $\phi\colon G \to G(N_1, \dots, N_m)$ of G,*

(1) *$\phi(H)$ is a full relatively quasiconvex subgroup of $G(N_1, \dots, N_m)$,*

(2) *$\phi(H)$ is isomorphic to the induced filling in that if $\phi_H : H \to H(K_1, \dots, K_m)$ is the induced filling map, then $\ker \phi_H = \ker \phi \cap H$, and*

(3) *$\phi(F) \cap \phi(H) = \phi(F \cap H)$.*

8.2 The filled hierarchy

Let $\mathcal{H}$ be a quasiconvex fully $\mathcal{P}$-elliptic hierarchy. By Lemma 3.14, Theorem 8.5 and the full $\mathcal{P}$-ellipticity of the hierarchy, the edge and vertex groups of the hierarchy are full relatively quasiconvex. Let $\pi\colon G \to \overline{G}$ be a filling and let $(\overline{G}, \overline{\mathcal{P}})$ be the relatively hyperbolic structure induced on the filling by Theorem 8.4. The goal of this subsection is to build an induced hierarchy $\overline{\mathcal{H}}$ (which may not be faithful) for $\overline{G}$ based on $\mathcal{H}$ where the vertex and edge groups of $\overline{\mathcal{H}}$ are induced fillings of vertex and edge groups of $\mathcal{H}$. The hierarchy $\overline{\mathcal{H}}$ will be called a *filled hierarchy* for $(\overline{G}, \overline{\mathcal{P}})$.

The filled hierarchy is built by starting at the top level and building the hierarchy inductively downward.

At the top level, let $\overline{\mathcal{H}}$ have the degenerate graph of groups decomposition for $\overline{G}$ consisting of a single vertex labeled $\overline{G}$. Let n be the length of $\mathcal{H}$. Suppose the filled hierarchy has been filled down to the $(n-i)^{\text{th}}$ level and let $\overline{A}$ be a vertex group at level $n-i$ so that $\overline{A}$ is the induced filling of a vertex group A at level $n-i$ of $\mathcal{H}$. Let (Γ, χ) be the graph of groups structure for A provided by $\mathcal{H}$. Recall that χ is the assignment map for the graph of groups structure.

If x is a vertex or edge of Γ, let $A_x := \chi(x)$ be the corresponding vertex or edge group. Let $\overline{\chi}(x) := \overline{A}_x$ where $\overline{A}_x$ is the induced filling $\pi_x\colon A_x \to \overline{A}_x$. The problem is that the pair $(\Gamma, \overline{\chi})$ still needs attachment homomorphisms to be a graph of groups.

Let $\phi_e\colon A_e \to A_v$ be an attachment homomorphism of an edge group A_e to a vertex group A_v. Two details need to be checked: first there need to be attachment maps $\overline{\phi}_e\colon \overline{A}_e \to \overline{A}_v$ such that $\overline{\phi}_e \circ \pi_e = \pi_v \circ \phi_e$. Let T be the maximal tree that determines $\pi_1(\Gamma, \chi, T)$. There will also need to be an isomorphism $\overline{\alpha}\colon \pi_1(\Gamma, \overline{\chi}, T) \to \overline{A}$ so that $(\Gamma, \overline{\chi}, T)$ is a graph of groups structure for $\overline{A}$ where $\overline{\alpha} \circ \pi_\Gamma = \pi_A \circ \alpha$.

Completing the square

$$\begin{tikzcd}
A_e \arrow[r, "\pi_e"'] \arrow[d, "\phi_e"] & \overline{A}_e \arrow[d, "\overline{\phi}_e"] \\
A_v \arrow[r, "\pi_v"'] & \overline{A}_v
\end{tikzcd}$$

with a map $\overline{\phi}_e\colon \overline{A}_e \to \overline{A}_v$ is straightforward because π_e is surjective and $\ker \pi_e \subseteq \ker \pi_v \circ \phi_e$.

Constructing the desired isomorphism $\overline{\alpha}\colon \pi_1(\Gamma, \overline{\chi}, T) \to \overline{G}$ amounts to completing the square

$$\begin{tikzcd}
\pi_1(\Gamma, \chi, T) \arrow[r, "\pi_\Gamma"'] \arrow[d, "\alpha"] & \pi_1(\Gamma, \overline{\chi}, T) \arrow[d] \\
A \arrow[r, "\pi_A"'] & \overline{A}
\end{tikzcd}$$

Lemma 8.11 *There exists an isomorphism $\overline{\alpha}\colon \pi_1(\Gamma, \overline{\chi}, T) \to \overline{G}$ that completes the diagram.*

The proof of Lemma 8.11 is essentially identical to [3, Lemma 8.1].

For the following, let $(G,\mathcal{P})$ be a relatively hyperbolic group pair and let $\mathcal{H}$ be a quasiconvex fully $\mathcal{P}$-elliptic hierarchy for G. The next lemma ties together some definitions:

Lemma 8.12 *If $A \leqslant G$ is an edge or vertex group of $\mathcal{H}$, then A is a full relatively quasiconvex subgroup of $(G,\mathcal{P})$ and every filling is an A-filling.*

Proof That A is full relatively quasiconvex follows immediately from the definition of full $\mathcal{P}$-ellipticity and Theorem 8.5.

Whenever $gP_ig^{-1} \cap A$ is infinite, then $gP_ig^{-1} \subseteq A$, so if $N_i \lhd P_i$, then $gN_ig^{-1} \lhd A$. □

Lemma 8.13 *Let A be an edge or vertex group of $\mathcal{H}$. Then for all sufficiently long fillings*

$$\pi\colon (G,\mathcal{P}) \to (\overline{G},\overline{\mathcal{P}})$$

the following hold:

(1) *The subgroup $\overline{A} := \phi(A)$ is full relatively quasiconvex in $(\overline{G},\overline{P})$.*

(2) *If $\overline{G}$ is hyperbolic, then $\overline{A}$ is quasiconvex in $\overline{G}$.*

(3) *The subgroup $\overline{A}$ is isomorphic to the induced filling of A.*

Proof There are only finitely many edge and vertex groups, so the first and third statements follow from Theorem 8.10.

If $\overline{A}$ is full relatively quasiconvex in $(\overline{G},\overline{P})$, then $\overline{A}$ is undistorted in $\overline{G}$ by [16, Theorem 10.5] and by [7, Corollary III.Γ.3.6], $\overline{A}$ is quasiconvex in $\overline{G}$ whenever $\overline{G}$ is hyperbolic. □

The third point also makes the filled hierarchy $\overline{\mathcal{H}}$ faithful:

Corollary 8.14 *For all sufficiently long fillings $\pi\colon (G,\mathcal{P}) \to (\overline{G},\overline{\mathcal{P}})$, the filled hierarchy $\overline{\mathcal{H}}$ for $\overline{G}$ is faithful.*

Proof Let $\phi_e\colon A_e \to A_v$ be an attachment homomorphism mapping an edge group A_e to a vertex group A_v. Since $\pi(A_e)$ and $\pi(A_v)$ are isomorphic to the induced fillings, we can regard the induced filling maps as maps $\pi_v\colon A_v \to \overline{A}_v$ and $\pi_e\colon A_v \to \overline{A}_e$. Let $\overline{\phi}_e\colon \overline{A}_e \to \overline{A}_v$ be the induced edge homomorphism.

We now need to check that given $g_e \in A_e$, $\phi_e \circ \pi_e(g_e) = 1$ implies that $\pi_e(g_e) = 1$. If $\phi_e \circ \pi_e(g_e) = 1$, then $\pi_v \circ \phi_e(g_e) = 1$, so $\phi_e(g_e) \in \ker\pi_v = \ker\pi \cap A_v \subseteq \ker\pi$. Faithfulness of the original hierarchy now implies $g_e \in \ker\pi \cap A_e = \ker\pi_e$, so $\pi_e(g_e) = 1$. □

The preceding results combine to produce a quasiconvex hierarchy:

Theorem 8.15 (see [3, Theorem 2.12]) *Let $(G, \mathcal{P})$ be a relatively hyperbolic group pair and let $\mathcal{H}$ be a quasiconvex fully $\mathcal{P}$-elliptic hierarchy terminating in $\mathcal{P}$. For all sufficiently long peripherally finite fillings $\pi\colon (G,\mathcal{P}) \to (\overline{G}, \overline{\mathcal{P}})$ so that every $\overline{P} \in \overline{\mathcal{P}}$ is hyperbolic, the group $\overline{G}$ is hyperbolic and has a quasiconvex hierarchy terminating in $\overline{\mathcal{P}}$.*

Proof Theorem 8.4 implies that all sufficiently long peripherally finite fillings are hyperbolic.

By Corollary 8.14, the quotient $\overline{G}$ has a faithful hierarchy $\overline{\mathcal{H}}$ where the underlying graphs and every vertex or edge group of $\overline{\mathcal{H}}$ is the image of a vertex or edge group (respectively) of $\mathcal{H}$ under π.

By Lemma 8.13(2), every edge and vertex group of $\overline{\mathcal{H}}$ is quasiconvex in $\overline{G}$ and is hence also quasi-isometrically embedded in $\overline{G}$, so the hierarchy $\overline{\mathcal{H}}$ is quasiconvex.

By construction, the terminal groups are fillings of the terminal groups of $\mathcal{H}$, so the terminal groups of $\overline{\mathcal{H}}$ are in $\overline{\mathcal{P}}$. □

Theorem 8.15 works for a group with a quasiconvex hierarchy, but Theorem 1 only gives a hierarchy for a finite index subgroup. When the filling kernels are chosen carefully, a filling of a finite index subgroup $G' \lhd G$ can be promoted to a filling of G.

Definition 8.16 Let $(G,\mathcal{P})$ be a relatively hyperbolic group pair and let $G' \lhd G$ be a finite index normal subgroup with induced peripheral structure $(G', \mathcal{P}')$. Let $\{N'_j \lhd P'_j \mid P'_j \in \mathcal{P}'_j\}$ be a collection of filling kernels. The collection $\{N'_j\}$ is *equivariantly chosen* if

(1) whenever gP'_jg^{-1} and hP'_kh^{-1} both lie in P_i, then $gN'_jg^{-1} = hN'_kh^{-1}$ and
(2) every such gN'_jg^{-1} is normal in P_i.

An *equivariant filling* of $(G', \mathcal{P}')$ is a filling with equivariantly chosen filling kernels.

An equivariant filling of $(G', \mathcal{P}')$ will induce a nice filling of $(G, \mathcal{P})$:

Proposition 8.17 *An equivariant filling $(G', \mathcal{P}') \to (\overline{G}', \overline{\mathcal{P}}')$ determines a filling $(G,\mathcal{P}) \to (\overline{G}, \overline{\mathcal{P}})$ so that $\overline{G}'$ is finite index normal in $\overline{G}$ and $(\overline{G}', \overline{\mathcal{P}}')$ is the peripheral structure induced by $(\overline{G}, \overline{\mathcal{P}})$.*

For the reader's convenience, here is a restatement of Theorem 2.

Theorem 2 *Let $(G,\mathcal{P})$ be a relatively hyperbolic group pair with $\mathcal{P} = \{P_1, \dots, P_m\}$. If G is virtually compact special, then there exist subgroups $\{\dot{P}_i \lhd P_i\}$ where $\dot{P}_i$ is finite index in P_i such that if $\overline{G} = G(N_1, \dots, N_m)$ is any peripherally finite filling with $N_i \lhd \dot{P}_i$, then $\overline{G}$ is hyperbolic and virtually special.*

Proof By Theorem 1, there exists a finite index $G' \lhd G$ with induced peripheral structure $(G', \mathcal{P}')$ and a quasiconvex, fully $\mathcal{P}'$-elliptic hierarchy terminating in $\mathcal{P}'$. Let $\mathcal{P}' = \{P'_1, \dots, P'_M\}$. Since G is virtually special and hence residually finite, there exist arbitrarily long peripherally finite fillings of $(G', \mathcal{P}')$. In particular, our fillings of $(G', \mathcal{P}')$ will be sufficiently long for Theorem 8.15 to hold.

Let $G'(K_1, \dots, K_M)$ be such a peripherally finite filling. Now pass to subgroups of the filling kernels to obtain an equivariant filling; choose K'_j so that, if $K_j^g \leqslant P_i$,

$$(K'_j)^g = \bigcap \{K_\ell^h \mid h \in G,\ \#(K_\ell^h \cap P_i) = \infty\}, \quad 1 \leqslant j \leqslant M.$$

We set $\dot{P}_i \lhd P_i$ equal to $(K'_j)^g$ for some (any) choice of $g \in G$ where K'_j so that $(K'_j)^g \leqslant P_i$. The new filling $\overline{G}' = G'(K'_1, \dots, K'_M)$ is longer than $G'(K_1, \dots, K_M)$ and remains peripherally finite. By Proposition 8.17, the filling $G'(K'_1, \dots, K'_M)$ determines a filling of G.

Consider any filling $G(N_1, \dots, N_m)$ so that, for each i,

1. $N_i \lhd P_i$,
2. $N_i \leqslant \dot{P}_i$, and
3. P_i / N_i is virtually special and hyperbolic,

with an induced equivariant filling

$$G' \to G'(N'_1, \dots, N'_M)$$

so that $N'_j \leqslant K'_j$ and $N'_j \lhd P'_j$ for each j. Condition (2) ensures the filling is sufficiently long so that Theorem 8.15 implies

1. $\overline{G}'$ is hyperbolic, and
2. $\overline{G}'$ has a quasiconvex hierarchy terminating in $\overline{\mathcal{P}}' = \{P'_j / N'_j\}$.

Then $\overline{G}'$ is a hyperbolic group with a quasiconvex hierarchy that terminates in finite groups (which are hence hyperbolic and virtually special). So by Corollary 8.1 (see [30, Theorem 13.3]), $G'(N'_1, \dots, N'_M)$ is virtually special. By Proposition 8.17, $\overline{G}' = G'(N'_1, \dots, N'_M)$ is finite index normal in $G(N_1, \dots, N_m)$, so the filling $G(N_1, \dots N_m)$ is also virtually special. □

References

[1] **I Agol**, *The virtual Haken conjecture*, Doc. Math. 18 (2013) 1045–1087 MR

[2] **I Agol**, **D Groves**, **J F Manning**, *Residual finiteness, QCERF and fillings of hyperbolic groups*, Geom. Topol. 13 (2009) 1043–1073 MR

[3] **I Agol**, **D Groves**, **J F Manning**, *An alternate proof of Wise's malnormal special quotient theorem*, Forum Math. Pi 4 (2016) art. id. e1 MR

[4] **H Bass**, *Covering theory for graphs of groups*, J. Pure Appl. Algebra 89 (1993) 3–47 MR

[5] **N Bergeron**, **D T Wise**, *A boundary criterion for cubulation*, Amer. J. Math. 134 (2012) 843–859 MR

[6] **C Bregman**, *Automorphisms and homology of non-positively curved cube complexes*, preprint (2016) arXiv 1609.03602

[7] **M R Bridson**, **A Haefliger**, *Metric spaces of non-positive curvature*, Grundl. Math. Wissen. 319, Springer (1999) MR

[8] **C Druţu**, **M Sapir**, *Tree-graded spaces and asymptotic cones of groups*, Topology 44 (2005) 959–1058 MR

[9] **D B A Epstein**, **J W Cannon**, **D F Holt**, **S V F Levy**, **M S Paterson**, **W P Thurston**, *Word processing in groups*, Jones and Bartlett, Boston, MA (1992) MR

[10] **B Farb**, *Relatively hyperbolic groups*, Geom. Funct. Anal. 8 (1998) 810–840 MR

[11] **F Haglund**, *Finite index subgroups of graph products*, Geom. Dedicata 135 (2008) 167–209 MR

[12] **F Haglund**, **D T Wise**, *Special cube complexes*, Geom. Funct. Anal. 17 (2008) 1551–1620 MR

[13] **F Haglund**, **D T Wise**, *A combination theorem for special cube complexes*, Ann. of Math. 176 (2012) 1427–1482 MR

[14] **G C Hruska**, *Nonpositively curved* 2*-complexes with isolated flats*, Geom. Topol. 8 (2004) 205–275 MR

[15] **G C Hruska**, *Geometric invariants of spaces with isolated flats*, Topology 44 (2005) 441–458 MR

[16] **G C Hruska**, *Relative hyperbolicity and relative quasiconvexity for countable groups*, Algebr. Geom. Topol. 10 (2010) 1807–1856 MR

[17] **G C Hruska**, **B Kleiner**, *Hadamard spaces with isolated flats*, Geom. Topol. 9 (2005) 1501–1538 MR

[18] **G C Hruska**, **D T Wise**, *Packing subgroups in relatively hyperbolic groups*, Geom. Topol. 13 (2009) 1945–1988 MR

[19] **T Hsu**, **D T Wise**, *Cubulating malnormal amalgams*, Invent. Math. 199 (2015) 293–331 MR

[20] **J Kahn**, **V Markovic**, *Immersing almost geodesic surfaces in a closed hyperbolic three manifold*, Ann. of Math. 175 (2012) 1127–1190 MR

[21] **J F Manning**, **E Martínez-Pedroza**, *Separation of relatively quasiconvex subgroups*, Pacific J. Math. 244 (2010) 309–334 MR

[22] **B Martelli**, *An introduction to geometric topology*, self published, Pisa, Italy (2016)

[23] **D V Osin**, *Peripheral fillings of relatively hyperbolic groups*, Invent. Math. 167 (2007) 295–326 MR

[24] **M Sageev**, *Ends of group pairs and non-positively curved cube complexes*, Proc. Lond. Math. Soc. 71 (1995) 585–617 MR

[25] **M Sageev**, **D T Wise**, *Periodic flats in* CAT(0) *cube complexes*, Algebr. Geom. Topol. 11 (2011) 1793–1820 MR

[26] **M Sageev**, **D T Wise**, *Cores for quasiconvex actions*, Proc. Amer. Math. Soc. 143 (2015) 2731–2741 MR

[27] **P Scott**, *Subgroups of surface groups are almost geometric*, J. Lond. Math. Soc. 17 (1978) 555–565 MR

[28] **J-P Serre**, *Trees*, Springer (1980) MR

[29] **D T Wise**, *From riches to raags*: 3*-manifolds, right-angled Artin groups, and cubical geometry*, CBMS Region. Conf. Ser. Math. 117, Amer. Math. Soc., Providence, RI (2012) MR

[30] **D T Wise**, *The structure of groups with a quasiconvex hierarchy*, Ann. of Math. Stud. 209, Princeton Univ. Press (2021) MR

Department of Mathematics and Statistics, Swarthmore College
Swarthmore, PA, United States

`eeinste1@swarthmore.edu`

Received: 9 October 2019 Revised: 6 August 2024

Geometry & Topology Publications, an imprint of **mathematical sciences publishers** msp

Algebraic & Geometric Topology **25**:8 (2025) 4499–4545
DOI: 10.2140/agt.2025.25.4499
Published: 20 November 2025

Intersection norms on surfaces and Birkhoff sections for geodesic flows

MARCOS COSSARINI
PIERRE DEHORNOY

Every filling multicurve on a smooth surface determines a norm on the first homology group of the surface. The unit ball of the dual norm is the convex hull of finitely many integer points. We give an interpretation of these points in terms of certain coorientations of the multicurve. Our main result is a classification statement: when the surface is hyperbolic and the filling multicurve is geodesic, integer points in the interior of the unit ball of the dual norm classify isotopy classes of Birkhoff sections for the geodesic flow (on the unit tangent bundle to the surface) whose boundary is the symmetric lift of the multicurve. All results remain true when one replaces the hyperbolic surface by a 2-dimensional orientable hyperbolic orbifold.

37D40; 37D45, 57K30, 57N37

Introduction

This paper deals with the topological study of nonsingular flows on 3-manifolds. With this goal, we study a family of norms on the first homology group of surfaces that may be of independent interest.

Given a smooth 3-manifold M and a smooth, nonvanishing vector field X on M, we denote by $(\varphi_X^t)_{t\in\mathbb{R}}$ the flow induced by X on M. An *embedded Birkhoff section* for $(M, (\varphi_X^t)_{t\in\mathbb{R}})$ is a compact, oriented surface S with boundary, embedded in M, whose interior is positively transverse to X, whose boundary ∂S is tangent to X, and such that every orbit of $(\varphi_X^t)_{t\in\mathbb{R}}$ intersects S in a uniformly bounded time. On

---

the topological side, an embedded Birkhoff section induces an open book decomposition of the underlying 3-manifold, where the binding is the boundary ∂S, and the fibration of the complement over $\mathbb{S}^1$ is given by an appropriate renormalization of the flow. On the dynamical side, when a flow admits an embedded Birkhoff section, its dynamic is encoded by the first-return map on the section — much simpler data. Such a section can be very helpful for understanding some properties of the flow, like the existence or abundance of periodic orbits [5; 18].

There are several existence results on Birkhoff sections for different classes of flows, for example geodesic flows [6; 16], Anosov flows [20], or Reeb flows [29; 30; 10; 9], among others. On the other hand, as far as we know, there are very few situations in which *all* Birkhoff sections are classified. An exception is given by the Hopf flow on $\mathbb{S}^3$, where the Birkhoff sections can be explicitly constructed [15], and the geodesic flow on a flat torus [13]. Our main goal is to provide such a classification, for the geodesic flow on the unit tangent bundle of a hyperbolic surface.

Theorem A *For* Σ *a hyperbolic surface and* γ *a finite collection of closed geodesics that fills* Σ, *denote by* $\overleftrightarrow{\gamma}$ *the symmetric lift of* γ *in* $T^1\Sigma$. *Then there is a one-to-one correspondence between*

- *isotopy classes of embedded Birkhoff sections for the geodesic flow on the unit tangent bundle* $T^1\Sigma$ *bounded by the symmetric lift* $\overleftrightarrow{\gamma}$ *of* γ, *with negative orientation*,
- *points satisfying a certain mod* 2 *condition in the open dual unit ball* $B^*_{\boldsymbol{x}_\gamma} \subset H^1(\Sigma, \mathbb{Z})$ *of the intersection norm* $\boldsymbol{x}_\gamma$ *associated to* γ.

Let us mention that such a statement is not really a surprise: the fact that Birkhoff sections with a given boundary up to isotopy correspond to integral points inside certain polyhedrons follows from theorems of Schwartzman, Thurston and Fried, as we explain later in this introduction. The main contribution of the paper lies in the explicit and combinatorial aspects of all constructions involved.

In the rest of the introduction, we first explain Theorem A by presenting intersection norms, their dual unit balls and the connection with Eulerian coorientations. The one-to-one correspondence in Theorem A is made explicit using a construction we call *Birkhoff–Brunella surfaces* and which is encapsulated in Proposition D. Then we put Theorem A in perspective by connecting it with Thurston and Fried's theory of fibered faces of the Thurston norm ball and with Schwartzman, Fuller, Fried, and Sullivan's theory of global sections for flows.

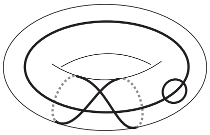
Intersection norms

Let Σ be a smooth surface without boundary. A *multicurve*[1] on Σ is a proper, smoothly immersed 1-submanifold in Σ, without boundary, and in general position (meaning that all multiple points are

---

[1] These are called *divides* by Norbert A'Campo [2] who, along with Sabir M. Gusein-Zade, studied divides on the disc in the context of singularities [1; 26; 27]. They were later generalized to arbitrary surfaces by Masaharu Ishikawa [31]. This terminology is maybe not so common in the worlds of surface topologists or dynamists, so we use the more common term *multicurve*.

double points where the intersection is transverse). On a compact surface, a multicurve consists of finitely many closed curves.



Let γ be a fixed multicurve on a compact surface Σ. We think of γ as the discrete analog of a Riemannian (or Finsler) metric; see [11] for a precise connection. The *length* of a generic path α with respect to γ, denoted by $\mathrm{Len}_\gamma(\alpha)$, is defined as the number of crossings between α and γ, and the length of a homology class $a \in H_1(\Sigma;\mathbb{Z})$, denoted by $\boldsymbol{x}_\gamma(a)$, is the minimal length of a generic integral 1-chain representing the class a; see Section 1.

Our first result was proven by Schrijver on the torus [37], and was stated by Turaev without proof [43, Remark 1.9].

Proposition B *Let Σ be an oriented compact smooth surface and γ a multicurve on Σ. Then the function $\boldsymbol{x}_\gamma\colon H_1(\Sigma;\mathbb{Z}) \to \mathbb{N}$ is a symmetric seminorm, that is, it is*

- *positively homogeneous: $\boldsymbol{x}_\gamma(n\cdot a) = n\,\boldsymbol{x}_\gamma(a)$ for all $a \in H_1(\Sigma;\mathbb{Z})$ and $n \in \mathbb{N}$,*
- *subadditive: $\boldsymbol{x}_\gamma(a+b) \le \boldsymbol{x}_\gamma(a) + \boldsymbol{x}_\gamma(b)$ for all $a, b \in H_1(\Sigma;\mathbb{Z})$,*
- *symmetric: $\boldsymbol{x}_\gamma(-a) = \boldsymbol{x}_\gamma(a)$ for all $a \in H_1(\Sigma;\mathbb{Z})$,*

Furthermore, if the multicurve γ is ***filling*** *(ie it meets every noncontractible closed curve in Σ), then $\boldsymbol{x}_\gamma$ is positive definite: $\boldsymbol{x}_\gamma(a) > 0$ if $a \neq 0$.*

The function $\boldsymbol{x}_\gamma$ is called the *intersection seminorm* (or *intersection norm* if it is positive definite) associated to γ.

Remark 0.1 This seminorm satisfies $\boldsymbol{x}_\gamma(a) \equiv [\gamma]_2(a) \bmod 2$ for each $a \in H_1(\Sigma;\mathbb{Z})$, where $[\gamma]_2$ is the $\mathbb{Z}_2$-cohomology class of the cochain that maps each generic smooth 1-chain α to its modulo 2 number of intersections with γ.

By a theorem of Thurston [42], any integer-valued seminorm N on a *lattice* L (ie an abelian group isomorphic to $\mathbb{Z}^d$ for some $d \in \mathbb{N}$) can be written in the form $N(v) = \max_{\varphi\in F}\varphi(v)$ where F is a finite family of group morphisms $L \to \mathbb{Z}$. In fact, one can take as F the dual unit ball of N, that is, the set B_N^* of homomorphisms $\varphi\colon L \to \mathbb{Z}$ that satisfy $\varphi(v) \le N(v)$ for all $v \in L$. Furthermore, if N coincides modulo m (for a certain integer $m \ge 1$) with a given homomorphism $\mu\colon L \to \mathbb{Z}_m$, then one can restrict F to those functionals in B_N^* that coincide with μ modulo m (see Theorem A.11). A natural question is whether these homomorphisms have a nice interpretation in the case that N is an intersection norm (with $m = 2$ and $\mu = [\gamma]_2$). The answer is positive, as we now explain.

Consider a fixed multicurve γ on a surface Σ. A *coorientation* of γ is a continuous transverse orientation defined on γ except at the double points, where the coorientation is allowed to flip. A coorientation η induces a cochain c_η which maps each generic piecewise-smooth path α in Σ to the signed number of crossings of α with γ, where the sign of each crossing is determined by η. The coorientation η is *Eulerian* if c_η is a closed cochain, that is, if $c_\eta(\alpha) = 0$ whenever α is a contractible closed curve. (Equivalently, η is Eulerian if around each double point p of γ, among the four fragments of γ that meet at p there are exactly two that are positively cooriented and two that are negatively cooriented. See an example in Figure 1, left.) It follows that an Eulerian coorientation η induces an integral cohomology class $[\eta] := [c_\eta] \in H^1(\Sigma; \mathbb{Z})$. Note that different Eulerian coorientations may yield the same cohomology class.

The next result was proven when Σ is a torus by Schrijver, using different methods [36, Theorem 9]. It is illustrated in Figure 1.

Theorem C *Let γ be a multicurve on an orientable closed compact surface Σ. Then the cohomology classes in the closed dual unit ball $\overline{B^*_{\boldsymbol{x}_\gamma}}$ that coincide modulo 2 with $[\gamma]_2$ are precisely the cohomology classes of the Eulerian coorientations of γ. Therefore, for every a in $H_1(\Sigma; \mathbb{Z})$ we have*

$$\boldsymbol{x}_\gamma(a) = \max_{\substack{\eta \text{ Eulerian} \\ \text{coorientation of } \gamma}} [\eta](a).$$

This result also gives an effective way for computing the norm $\boldsymbol{x}_\gamma$, since it reduces the minimization over an infinite number of curves to a maximization over a finite number of coorientations.

Going back to the case where the multicurve γ is a geodesic in a hyperbolic surface, Theorem A states that there is a correspondence between (certain) integral points in the interior of $B^*_{\boldsymbol{x}_\gamma}$ and Birkhoff sections for the geodesic flow, and Theorem C states that (certain) integral points in $\overline{B^*_{\boldsymbol{x}_\gamma}}$ can be represented by Eulerian coorientations. The correspondence of Theorem A is made explicit by associating to every Eulerian coorientation a certain surface in $T^1\Sigma$, as in the following statement. The superscript *BB* stands for Birkhoff–Brunella.

Proposition D *Let Σ be a compact oriented surface with a Riemannian metric and γ a finite collection of closed geodesics on Σ. There is a canonical a map S^{BB} that associates to every Eulerian coorientation η of γ an oriented surface $S^{BB}(\eta)$ in $T^1\Sigma$ whose interior is positively transverse to the geodesic flow and whose oriented boundary is $-\vec{\gamma}$. The Euler characteristic of $S^{BB}(\eta)$ is independent of η and equals minus twice the number of double points of γ.*

If two Eulerian coorientations η_1 and η_2 of γ are cohomologous and their common class lies in the interior of $B^*_{\boldsymbol{x}_\gamma}$, their interiors are isotopic along the flow.

Thurston norm balls, their fibered faces, and suspension flows

We now present Thurston's theory of norms and fibered faces for 3-manifolds. This puts in perspective and explains Theorem A at an abstract level.

Given a compact 3-manifold M with toric boundary, its *Thurston norm* $\boldsymbol{x_M}$ is a function on the space $H_2(M, \partial M; \mathbb{R})$ that encodes the minimal negative part of the Euler characteristic of embedded surfaces in M with boundary in ∂M in the considered homology class [42]. It is a seminorm, and as such it is determined by its unit ball $B_{\boldsymbol{x}_M}$. The latter turns out to be a polyhedron, which is compact when M is atoroidal. It is a topological invariant that is in general hard to compute [22; 3].

Intersection norms can be seen as 2-dimensional siblings of the Thurston norms since they are defined by minimizing a certain complexity measure over homology classes. Their unit balls are also polyhedrons, but, unlike unit balls of Thurston norms, these can be easily computed using Theorem C.

The top-dimensional faces of Thurston norm balls are of two types, namely fibered and nonfibered. A *fibered face* is such that every integral point in the cone generated by the fibered face is the class of the fibers of a fibration of M over the circle.

Fried showed [21] that every pseudo-Anosov flow $(\varphi^t)_{t\in\mathbb{R}}$ (see Section 3.4 for a definition) on M that is tangent to ∂M and that admits a global cross section canonically determines a fibered face of $B_{\boldsymbol{x}_M}$ as follows: denote by D_φ the convex cone generated by the homology classes of the periodic orbits of $(\varphi^t)_{t\in\mathbb{R}}$ in $H_1(M; \mathbb{R})$. This cone actually coincides with the cone over the set of Schwartzman asymptotic cycles [38]. The dual cone C_φ in $H_2(M, \partial M; \mathbb{R})$ is defined as those classes that pair positively with all of D_φ. It turns out that the integral classes in C_φ correspond exactly to the classes of the global sections to $(\varphi^t)_{t\in\mathbb{R}}$. Therefore C_φ is exactly the cone over the interior of a fibered face of the Thurston norm ball. The cones D_φ and C_φ are polyhedral, and Fried also gives an algorithm [19] for computing D_φ and C_φ starting from a Markov partition for $(\varphi^t)_{t\in\mathbb{R}}$.

The connection with Birkhoff sections can be made as follows: Assume that β is a collection of periodic orbits of a flow φ in M. One can blow up the link β and obtain a 3-manifold $\overline{M \setminus \beta}$ with toric boundary $\partial\overline{M \setminus \beta}$. If $(\varphi^t)_{t\in\mathbb{R}}$ is of class C^1, then it extends to a nonsingular flow $(\varphi^t_\beta)_{t\in\mathbb{R}}$ on $\overline{M \setminus \beta}$. If $(\varphi^t)_{t\in\mathbb{R}}$ was of Anosov or pseudo-Anosov type, then $(\varphi^t_\beta)_{t\in\mathbb{R}}$ is pseudo-Anosov. In this context a Birkhoff section for $(\varphi^t)_{t\in\mathbb{R}}$ with boundary in β extends to a global section for the flow $(\varphi^t_\beta)_{t\in\mathbb{R}}$. The discussion of the previous paragraph then implies that, if β bounds a Birkhoff section, isotopy classes of Birkhoff sections whose boundary is in β are classified by integral points in a certain polyhedral cone $C_{\varphi,\beta}$ in $H_2(\overline{M \setminus \beta}, \partial\overline{M \setminus \beta}; \mathbb{R}) \simeq H_2(M, \beta; \mathbb{R})$.

In the context of Theorem A, M is the unit tangent bundle $T^1\Sigma$ to a hyperbolic surface Σ, $(\varphi^t)_{t\in\mathbb{R}}$ is the geodesic flow on $T^1\Sigma$, and β is the symmetric lift $\overleftrightarrow{\gamma}$ of a filling collection γ of geodesics on Σ; see Section 3.1 for the definitions. The set of Birkhoff sections for the geodesic flow bounded by $\overleftrightarrow{\gamma}$ is then the cone over a fibered face of the Thurston norm ball in $H_2(T^1\Sigma, \overleftrightarrow{\gamma}; \mathbb{Z})$ that we denote by $C_{\mathrm{geod},\overleftrightarrow{\gamma}}$.

In Theorem A, the assumption that the oriented boundary is exactly $-\overleftrightarrow{\gamma}$ (that is, every boundary component has multiplicity -1) can be seen as a restriction on the homology class of the section: it has to lie in a certain affine subspace denoted by $\partial^{-1}_{\overleftrightarrow{\gamma}}(-1, \dots, -1)$ of $H_2(T^1\Sigma, \overleftrightarrow{\gamma}; \mathbb{R})$. This means that the Birkhoff

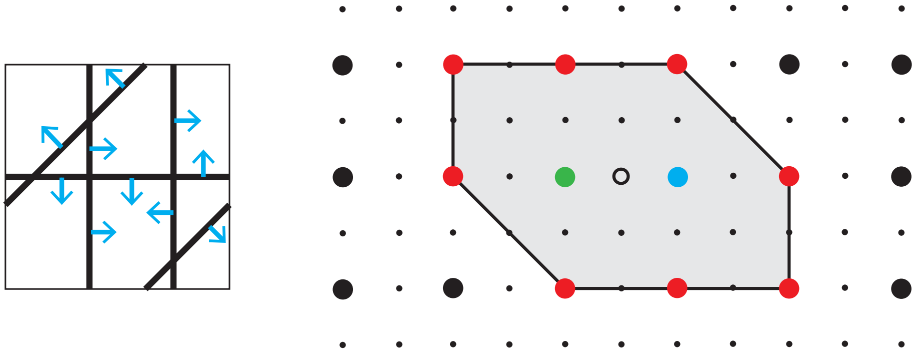


Figure 1: Illustration of Theorems A and C in the case of Σ a torus (with an abuse since Theorem A deals with higher-genus surfaces, whose homology has dimension ≥ 4). On the left, a multicurve γ on Σ consisting of four geodesics, and an Eulerian coorientation (blue arrows). Seen as a graph, γ has five vertices and ten edges. On the right, the dual unit ball $B^*_{\boldsymbol{x}_\gamma}$ of the associated intersection norm. The empty circle denotes the origin. The big dots denote those classes in $H^1(\Sigma;\mathbb{Z})$ congruent to $[\gamma]_2$ mod 2. Among these classes, ten (in blue, green, and red) are in the dual unit ball $B^*_{\boldsymbol{x}_\gamma}$ and correspond to all cohomology classes of Eulerian coorientations of γ (Theorem C). For example, the class corresponding to the blue coorientation is the blue point. The blue and green points lie in the interior of $B^*_{\boldsymbol{x}_\gamma}$, and hence describe two isotopy classes of Birkhoff cross sections for the geodesic flow bounded by $-\vec{\gamma}$. If the genus of Σ was at least 2, there would be no other isotopy class of Birkhoff cross section for the geodesic flow (Theorem A). The eight points are on the boundary of $B^*_{\boldsymbol{x}_\gamma}$ and correspond to classes of surfaces transverse to the geodesic flow, but not intersecting every orbit, and bounded by $-\vec{\gamma}$.

sections we are interested in are enumerated by the intersection of the cone $C_{\mathrm{geod},\vec{\gamma}}$ with the affine subspace $\partial_{\vec{\gamma}}^{-1}(-1,\dots,-1)$. It turns out that a suitable choice of an origin identifies the latter with $H_1(\Sigma;\mathbb{R})$; see Section 3.5. Under this identification, Theorem A can be summarized by the equality

$$C_{\mathrm{geod},\vec{\gamma}} \cap \partial_{\vec{\gamma}}^{-1}(-1,\dots,-1) = \tfrac{1}{2} B^*_{\boldsymbol{x}_\gamma}.$$

Our paper adds to this description the elementary and explicit characters of all the involved constructions. Indeed, as far as we know, there is no other Anosov or pseudo-Anosov flow for which the set of global or Birkhoff cross sections admits such an explicit and combinatorial description.

Remark 0.2 One may wonder how general Theorem A is, namely whether one can hope for an analogous statement for any (transitive) Anosov flow. As explained above, the set of Birkhoff sections up to isotopy fixing the boundary is described by the integral points inside a certain polyhedron. However we do not know how to describe this polyhedron in general. It seems to be related to linking numbers of periodic orbits of the flow [13; 14], but linking numbers are only defined for nullhomologous links. Ghys proved that Gauss linking forms describe all linking numbers between periodic orbits for a vector field in a homology sphere [24]. Moreover he showed how to use these Gauss forms to decide whether all finite collections of periodic orbits bound a Birkhoff section (which he calls *left-* or *right-handed* flows). Probably one should first extend the concept of Gauss linking forms to manifolds that are not rational homology spheres, and see how this helps define linking of periodic orbits and more generally of invariant measures. Then one could hope that these generalized linkings describe exactly the homological information needed to apply Schwartzman's criterion, as we will do in Section 3.

Remark 0.3 It may look strange to deal with Birkhoff cross sections with negative boundary and not with positive ones, ie with surfaces such that the orientation of the boundary inherited from the orientation of the surface (itself inherited from the coorientation of the interior surface by the flow) is opposed to the direction of the flow. The reason is that there is actually no positive Birkhoff cross section for the geodesic flow, as explained in Théo Marty's thesis [33, Chapter 3]. One could then look at *mixed* sections, namely transverse surfaces some of whose boundary components are positively tangent to the geodesic flow and some others are negatively tangent. There are more mixed sections than negative. Alas, we have no analog of Proposition D in this more general case, meaning that we do not have an elementary way to construct all mixed sections.

Remark 0.4 The case of the torus with a flat metric is not covered by Theorem A. In this case, the fact that the unit tangent bundle $T^1\mathbb{T}^2$ is trivial allows us to cut-and-glue horizontal tori to Birkhoff cross sections, so that there are infinitely many isotopy classes with a given boundary. However, modulo this additional operation, there are still only finitely many classes. These have been classified in a previous work by the second author [13, Theorem 3.12]. The statement is similar, namely equivalence classes of Birkhoff sections are classified by points in the interior of a certain polygon with integral vertices. The statement is even more general since, in this restricted case of the torus, there is no assumption that the boundary of the section is symmetric. One could recover this earlier result in the symmetric case by a proof very similar to that of Theorem A.

Extension to 2-dimensional orbifolds

Our results here can be generalized in the following sense. Instead of considering orientable surfaces only, one can consider orientable 2-dimensional orbifolds, as introduced by Thurston [41]. Such a 2-orbifold $\mathbb{O}$ is described by an orientable topological surface $\Sigma_{\mathbb{O}}$ and charts that are local homeomorphisms $\mathbb{R}^2/(\mathbb{Z}/k\mathbb{Z}) \to \Sigma_{\mathbb{O}}$, where $\mathbb{Z}/k\mathbb{Z}$ acts by rotation on $\mathbb{R}^2$.

There are several possible definitions for the homology of an orbifold that yield different spaces. The one that is useful here is the most elementary: we define $H_i(\mathbb{O};\mathbb{R})$ to be the space $H_i(\Sigma_{\mathbb{O}};\mathbb{R})$. In this context the definition of intersection norms extends trivially. Proposition B and Theorem C still hold. Now the unit tangent bundle $T^1\mathbb{O}$ is 3-manifold that is a Seifert fibered space over $\Sigma_{\mathbb{O}}$. The geodesic flow is well defined on $T^1\mathbb{O}$, and when $\mathbb{O}$ is hyperbolic it is still of Anosov type. Proposition D extends directly in this context. Concerning Theorem A, it has to be modified for taking into account orbifolds that are homology spheres — a case that does not occur with hyperbolic surfaces.

Theorem E *Let $\mathbb{O}$ be a hyperbolic orientable 2-dimensional orbifold. Let γ be a finite collection of closed geodesics on $\mathbb{O}$.*

- *If $\Sigma_{\mathbb{O}}$ is a sphere, then $T^1\mathbb{O}$ is a rational homology sphere. In this situation, the link $-\vec{\gamma}$ bounds a Birkhoff section for the geodesic flow in $T^1\mathbb{O}$ if and only if γ is filling in $\Sigma_{\mathbb{O}}$. In that case, the Birkhoff section is unique up to isotopy fixing the boundary.*
- *If $\Sigma_{\mathbb{O}}$ is not a sphere and if γ is filling, then the map $[\eta] \mapsto \{S^{BB}(\eta)\}$ is a one-to-one correspondence between integer points in the open unit ball $\operatorname{int}(B^*_{x_\gamma})$ congruent to $[\gamma]_2 \bmod 2$ and isotopy classes of Birkhoff cross sections for the geodesic flow in $T^1\Sigma$ with boundary $-\vec{\gamma}$.*
- *If $\Sigma_{\mathbb{O}}$ is not a sphere and γ is not filling, then there is no surface bounded by $-\vec{\gamma}$ and transverse to the geodesic flow.*

A particular case is when $\mathbb{O}$ is a hyperbolic triangular orbifold, that is, a sphere with three conic points. In this case every collection γ of closed geodesics is filling, and hence its lift $\vec{\gamma}$ bounds a Birkhoff section. This is a particular case of the main result of [14], which proves that in this case every finite collection of periodic orbits (even nonsymmetric) bounds a Birkhoff section for the geodesic flow.

Acknowledgments

Dehornoy thanks Étienne Ghys and Adrien Boulanger for related discussions, and Elena Kudryavtseva, who initiated this article by asking several questions about Birkhoff sections. The authors thank the referees for useful suggestions, in particular the extension to orbifolds.

1 Intersection norms and proof of Proposition B

In the whole section we fix an oriented compact smooth surface Σ with empty boundary and a multicurve γ on Σ. (Recall that a multicurve in Σ is a compact, closed 1-manifold that is smoothly immersed in Σ, self-transverse, and has no points of multiplicity > 2).

Definition 1.1 A *path* in Σ is a continuous function $\alpha\colon I \to \Sigma$ (where $I \subseteq \mathbb{R}$ is a compact interval), considered up to a uniform shift in the parametrization, so that the concatenation $\alpha\beta$ of two consecutive paths α and β is well defined. The *reverse* of a path α is the path $\alpha^\dagger(t) = \alpha(-t)$. The *trivial path* at a point $p \in \Sigma$ is denoted by 1_p.

Definition 1.2 A smooth path in Σ is *generic* (with respect to γ) if it has no endpoint on γ, it is transverse to γ, and it avoids the double points of γ. We denote by P_γ the set of *generic piecewise-smooth paths*, obtained by concatenating finitely many generic smooth paths. The *length* with respect to γ of a path $\alpha \in P_\gamma$ is the number of times that it meets γ,

$$\operatorname{Len}_\gamma(\alpha) = |\alpha^{-1}(\gamma)|.$$

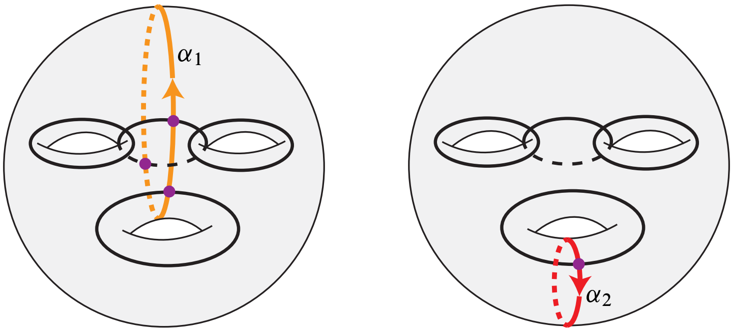


Figure 2: A genus 3 surface with a multicurve γ made of four closed curves (black). On the left the curve α_1 (orange and bold) is transverse to γ and intersects it three times. On the right α_2 (red) is homologous to α_1 since their difference bounds a subsurface, namely the right hemisurface. The curve α_2 intersects γ only once. This number cannot be reduced to 0 in the same homology class, and hence α_2 is $\boldsymbol{x}_\gamma$-realizing and we have $\boldsymbol{x}_\gamma([\alpha_1]) = \boldsymbol{x}_\gamma([\alpha_2]) = |\{\alpha_2^{-1}(\gamma)\}| = 1$.

Definition 1.3 A *generic integral 1-chain* is a linear combination $\alpha = \sum_i c_i \alpha_i$ of paths $\alpha_i \in P_\gamma$ with integer coefficients $c_i \in \mathbb{Z}$. Its *length* is defined as

$$\mathrm{Len}_\gamma(\alpha) = \sum_i |c_i| \, \mathrm{Len}_\gamma(\alpha_i).$$

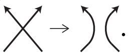
Note that every homology class a in $H_1(\Sigma; \mathbb{Z})$ may be represented by a generic integral 1-chain. The length of the homology class a is defined as

$$\boldsymbol{x}_\gamma(a) = \min_{\substack{\alpha \text{ closed generic integral} \\ \text{1-chain such that } [\alpha]=a}} \mathrm{Len}_\gamma(\alpha).$$

A closed generic 1-chain that minimizes length in its homology class is called an *$\boldsymbol{x}_\gamma$-realizing 1-chain*. The function $\boldsymbol{x}_\gamma : H_1(\Sigma; \mathbb{Z}) \to \mathbb{N}$ is called the *intersection seminorm* (or *intersection norm*, if it is positive definite) associated to γ.

The function $\boldsymbol{x}_\gamma$ has three properties that make it a seminorm, namely it is positively homogeneous, subadditive and symmetric. To prove the first point we need two facts about curves on surfaces. Note first that every homology class $a \in H_1(\Sigma; \mathbb{Z})$ can be represented by an oriented multicurve. A multicurve α is *simple* if it has no double points, and is generic (with respect to γ) if it is transverse to γ and the union $\alpha \cup \gamma$ is a multicurve. (The last condition holds if and only if each of the two multicurves α and γ avoids the double points of the other one.)

Lemma 1.4 (simplification) *Every homology class a in $H_1(\Sigma; \mathbb{Z})$ can be represented by a simple oriented multicurve that is generic with respect to γ and $\boldsymbol{x}_\gamma$-realizing.*

Proof Let α be an oriented multicurve that represents the class a, and is generic with respect to γ and $\boldsymbol{x}_\gamma$-realizing. To make α simple, we eliminate each self-crossing of α by performing a local modification of the form . □

Lemma 1.5 (partitioning) *Every simple oriented multicurve in Σ of homology class $n \cdot a$ (for some $a \in H_1(\Sigma;\mathbb{Z})$ and $n \in \mathbb{N}_{\neq 0}$) is a union of n disjoint simple oriented multicurves, each of class a.*

Proof Let β be a simple oriented multicurve of homology class $n \cdot a$. Since β is of class $n \cdot a$, its algebraic number of crossings with any generic oriented loop is a multiple of n. Therefore we can label the regions (ie connected components) of $\Sigma \setminus \beta$ with integers modulo n in such a way that the label increases by 1 when one crosses β positively (ie from right to left). For every $i \in \mathbb{Z}/n\mathbb{Z}$, denote by α_i the union of those components of β that have regions labeled i on their right, and regions labeled $i+1$ on their left. Every α_i is a simple multicurve, and by construction β is the union of all of them. Any two curves α_i, α_j are homologous since $\alpha_i - \alpha_j$ bounds a subsurface of Σ (namely, the part with labels in $[i, j)$). This implies that $[\beta] = n \cdot [\alpha_i]$ for every i, and since $H_1(\Sigma;\mathbb{Z})$ has no torsion, we conclude that $[\alpha_i] = a$. □

Now let us show that $\boldsymbol{x}_\gamma$ is a seminorm. The symmetry property $\boldsymbol{x}_\gamma(-a) = \boldsymbol{x}_\gamma(a)$ is evident since the number of intersections does not change by reversing the orientation of a curve. We have to prove positive homogeneity and subadditivity.

Lemma 1.6 (positive homogeneity) *For every a in $H_1(\Sigma;\mathbb{Z})$ and for all $n \in \mathbb{N}$ one has*

$$\boldsymbol{x}_\gamma(n \cdot a) = n\, \boldsymbol{x}_\gamma(a).$$

Proof Given $a \in H_1(\Sigma;\mathbb{Z})$ and $n \in \mathbb{N}$, consider a realizing multicurve α in a. Since n parallel copies of α intersect γ at $n\,\boldsymbol{x}_\gamma(a)$ points, we have $\boldsymbol{x}_\gamma(n \cdot a) \le n\,\boldsymbol{x}_\gamma(a)$. For the reverse inequality, consider an $\boldsymbol{x}_\gamma$-realizing multicurve β of homology class $n \cdot a$. By simplification (Lemma 1.4) we can suppose β simple, and then it follows by partitioning (Lemma 1.5) that β is the union of n multicurves α_i of class a. Each multicurve α_i has at least $\boldsymbol{x}_\gamma(a)$ intersections with γ, which implies that β has at least $n\,\boldsymbol{x}_\gamma(a)$ intersections with γ, proving the inequality $\boldsymbol{x}_\gamma(n \cdot a) \ge n\,\boldsymbol{x}_\gamma(a)$. □

Lemma 1.7 (subadditivity) *For every a, b in $H_1(\Sigma;\mathbb{Z})$ one has*

$$\boldsymbol{x}_\gamma(a+b) \le \boldsymbol{x}_\gamma(a) + \boldsymbol{x}_\gamma(b).$$

Proof The union of two multicurves that realize $\boldsymbol{x}_\gamma(a)$ and $\boldsymbol{x}_\gamma(b)$ crosses γ in $\boldsymbol{x}_\gamma(a) + \boldsymbol{x}_\gamma(b)$ points, giving $\boldsymbol{x}_\gamma(a+b) \le \boldsymbol{x}_\gamma(a) + \boldsymbol{x}_\gamma(b)$. □

This finishes the proof of Proposition B which states that the function $\boldsymbol{x}_\gamma$ is a seminorm on $H_1(\Sigma;\mathbb{Z})$.

Remark 1.8 One can easily extend the notion of intersection norm to a surface with boundary Σ, by allowing the multicurves γ to contain arcs with endpoints on $\partial\Sigma$ (as did A'Campo [2; 1]). One then obtains two norms on $H_1(\Sigma;\mathbb{Z})$ and $H_1(\Sigma, \partial\Sigma;\mathbb{Z})$, depending on whether one considers absolute or relative homology classes. Proposition B also holds in the second context.

Remark 1.9 One can wonder how the intersection norms compare with other known norms on the first homology of a surface. For example, the *stable norm* $\boldsymbol{x}_g$ induced by a metric g is defined by $\boldsymbol{x}_g(a) = \liminf_{n\to\infty} \min_{\alpha^{(n)}\in na} g(\alpha^{(n)})/n$. On a surface the stabilization is not necessary, so that one has $\boldsymbol{x}_g(a) = \min_{\alpha\in a} g(\alpha)$. One can check that if $(\gamma_k)_{k\in\mathbb{N}}$ is a sequence of filling geodesics that approximates g, meaning that the sequence of invariant measures on $T^1\Sigma$ that are concentrated on the lift $\vec{\gamma}_k$ tends in the weak-$*$ sense to the Liouville measure defined by g on $T^1\Sigma$, then the rescaled norms $\frac{1}{g(\gamma_k)}\boldsymbol{x}_{\gamma_n}$ tend to the stable norm of g. Equivalently, the rescaled unit balls $g(\gamma_k)B_{\boldsymbol{x}_{\gamma_k}}$ tend to the unit ball of the stable norm.

2 Unit balls and coorientations

The context remains the same as in the previous section: we fix an oriented closed compact smooth surface Σ of genus at least 1 and a multicurve γ on it. We have shown that the intersection norm $\boldsymbol{x}_\gamma$ is an integer-valued seminorm on the lattice $H_1(\Sigma;\mathbb{Z}) \simeq \mathbb{Z}^{2g}$. By Remark 0.1 it coincides modulo 2 with $[\gamma]_2$. Therefore we may apply the following result of Thurston (as extended in the appendix). Recall that a *lattice* L is a finitely generated free abelian group. Its *dual lattice* L^* is the group of homomorphisms $L \to \mathbb{Z}$. Note that $L \simeq L^* \simeq \mathbb{Z}^d$ for some $d \in \mathbb{N}$.

Theorem 2.1 ([42, Theorem 2] and Theorem A.11) *Every integral seminorm N on a lattice L is of the form*

$$N(v) = \max_{\varphi\in B_N^*} \varphi(v),$$

where $B_N^* \subseteq L^*$ is the **dual unit ball** *of N, that is the (finite) set of group homomorphisms $\varphi\colon L \to \mathbb{Z}$ that satisfy $\varphi(v) \le N(v)$ for all $v \in L$. Furthermore, if N coincides modulo a certain integer $m > 1$ with a given homomorphism $\mu\colon L \to \mathbb{Z}_m$, then we have*

$$N(v) = \max_{\substack{\varphi\in B_N^* \\ \varphi_{\operatorname{mod} m}=\mu}} \varphi(v).$$

Our goal in this section is to prove Theorem C, that is, to characterize the points of $\overline{B^*_{\boldsymbol{x}_\gamma}}$ that coincide modulo 2 with $[\gamma]_2$. Specifically, we will show that these cohomology classes are precisely those that can be represented by Eulerian coorientations. We will do so as follows.

Recall, from the introduction, that a *coorientation* of γ is a continuous transverse orientation defined on γ except at the double points, where the coorientation is allowed to flip. A coorientation determines a 1-cochain c_η which maps each generic piecewise-smooth path α in Σ to the signed number of crossings of α with γ, where the sign of each crossing is determined by η. This cochain clearly satisfies

$$c_\eta(\alpha) \le \operatorname{Len}_\gamma(\alpha) \quad \text{and} \quad c_\eta(\alpha) \equiv \operatorname{Len}_\gamma(\alpha) \bmod 2 \quad \text{for each generic path } \alpha.$$

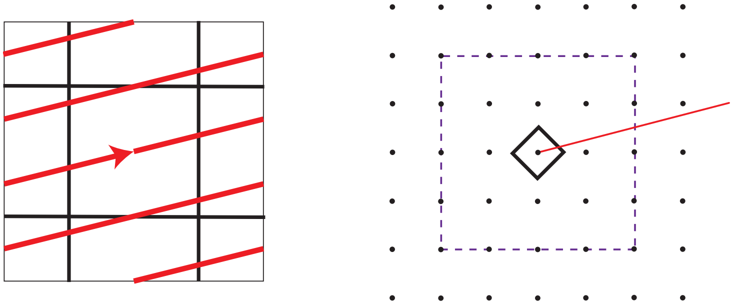


Figure 3: A torus with a collection γ (black) made of four curves, two vertical and two horizontal. The curve α (red and bold) intersects γ in 10 points. It is the smallest number for a curve whose homology class is $(4,1)$, so that $\boldsymbol{x}_\gamma(4,1)=10$. The norm $\boldsymbol{x}_\gamma$ is actually given by $\boldsymbol{x}_\gamma((p,q))=2|p|+2|q|$ in the canonical coordinates.

If we think of the multicurve γ as a discrete metric, then we can see a coorientation η as a discrete field of unit-norm covectors, and the number $c_\eta(\alpha)$ as the value of the integral of η along the path α.

A coorientation η is *Eulerian* if c_η is a closed cochain, that is, if $c_\eta(\alpha)=0$ whenever α is a contractible closed curve. In this case, the coorientation η defines a cohomology class $[\eta]:=[c_\eta]\in H^1(\Sigma;\mathbb{Z})$. This cohomology class $h=[\eta]$ satisfies the properties

$$h(a)\le \boldsymbol{x}_\gamma(a) \quad\text{and}\quad h(a)\equiv \boldsymbol{x}_\gamma(a) \bmod 2 \quad\text{for all } a\in H_1(\Sigma;\mathbb{Z}),$$

and we say then that h is *γ-special*.

To go backwards, from a γ-special cohomology class h to a coorientation η such that $h=[\eta]$, we will rely on an auxiliary object called an *eikonal function*. An *eikonal function* on a surface-with-a-multicurve (Σ,γ) is a function $f\colon \Sigma\setminus\gamma\to\mathbb{Z}$ that satisfies

$$(1)\qquad |f(y)-f(x)|\le d_\gamma(x,y) \quad\text{and}\quad f(y)-f(x)\equiv d(x,y) \bmod 2 \quad\text{for all } x,y\in\Sigma\setminus\gamma.$$

If we think of the multicurve γ as a discrete metric, and we see (Eulerian) coorientations as (closed) unitary 1-forms, then we should see eikonal functions as scalar-valued functions that are nonexpansive (or 1-Lipschitz). We can differentiate an eikonal function to obtain an Eulerian coorientation, and reciprocally, on a simply connected surface, we can integrate an Eulerian coorientation to obtain an eikonal function.

We will use eikonal functions as follows. Let $(\tilde\Sigma,\pi)$ be the universal cover of Σ, and let $x_0\in\Sigma$ be a fixed, arbitrary point. The surface $\tilde\Sigma$ has a multicurve $\tilde\gamma=\pi^*\gamma$ (the pullback of γ by the covering map π). An Eulerian coorientation η determines an eikonal function f_η on $\tilde\Sigma\setminus\tilde\gamma$, called the *primitive* of η, by the formula

$$f_\eta(x)=c_\eta(\pi\circ\alpha_{x_0,x}),$$

where $\alpha_{x_0,x}$ is a generic path in $\tilde\Sigma$ from x_0 to x. We note that f_η is equivariant with respect to the cohomology class $h=[\eta]$, which means that

$$f(T_\beta(x))-f(x)=h[\beta],$$

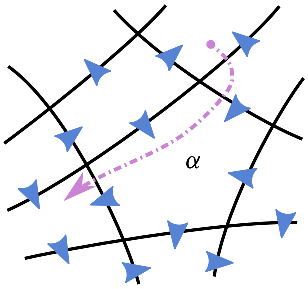


Figure 4: A piece of a multicurve γ (black). A coorientation η of γ is indicated with blue arrows. A path α transverse to γ is shown (purple and dotted). The pairing $\langle\eta,\alpha\rangle$ equals $-1+2=+1$ on this example.

for any points $x \in \widetilde{\Sigma}$, and any loop homotopy class $\{\beta\} \in \Pi_1(\Sigma, \pi(x_0))$, where T_β is the automorphism of $\widetilde{\Sigma}$ induced by the curve β.

Moreover, this process can be reversed: any h-equivariant eikonal function f can be differentiated to obtain a coorientation η of cohomology class h, such that $f_\eta = f$. Therefore, to go backwards, from a γ-special cohomology class h to a coorientation η such that $[\eta]=h$, we do as follows. We define first an h-equivariant function $f\colon \pi^{-1}(p_0) \to \mathbb{Z}$, where $p_0 = \pi(x_0) \in \Sigma$. We show that $\tilde{f}$ is preeikonal (ie it satisfies (1), even thought it is not defined at all points) since h is γ-special. Finally, we show that any preeikonal function can be naturally extended, using a standard formula, to an eikonal function $\bar{f}$ defined on the whole space. Moreover, this extended function $\bar{f}$ is h-equivariant if f is so. Differentiating the eikonal function $\bar{f}$ we obtain the coorientation η such that $f_\eta = \bar{f}$, and therefore $[\eta]=h$.

2.1 Coorientations of multicurves

Recall that γ is a multicurve in Σ. A *cross-vector* on γ is a vector tangent to Σ that is located at a *simple* point of γ, and is transverse to γ. The set of such vectors, considered as a topological subspace of the tangent bundle of Σ, is denoted by C_γ. Note that this space has finitely many connected components.

Definition 2.2 An integral *cross-functional* on γ is a function $\eta\colon C_\gamma \to \mathbb{Z}$ that is locally constant and satisfies the equation $\eta(-v) = -\eta(v)$ for all $v \in C_\gamma$. A *coorientation* of γ is a cross-functional with values ± 1. Note that there are finitely many coorientations of γ.

As mentioned in the introduction, each coorientation η induces a cochain c_η whose cohomology class $[c_\eta] \in H^1(\Sigma;\mathbb{Z})$ is in the dual unit ball $B^*_{\boldsymbol{x}_\gamma}$, as we will see below. To reverse this process and show that each cohomology class $h \in B^*_{\boldsymbol{x}_\gamma}$ (equivalent to $\boldsymbol{x}_\gamma$ mod 2) can be represented by a coorientation, we must understand precisely which cochains are induced by coorientations or, more generally, by cross-functionals of γ. These cochains are called *cross-cochains*, and are characterized as follows.

Recall from Definition 1.2 that P_γ is the space of piecewise-smooth paths on Σ that are generic with respect to γ.

Definition 2.3 An integral *cross-cochain* (with respect to the multicurve γ) is a function $c\colon P_\gamma \to \mathbb{Z}$ with the following properties:

- It is *additive* with respect to concatenation of paths, that is, $c(\delta\varepsilon) = c(\delta) + c(\varepsilon)$ if $\delta, \varepsilon \in P_\gamma$ are consecutive paths.
- It is *alternating* with respect to path reversion, that is, $c(\alpha^\dagger) = -c(\alpha)$ for all paths $\alpha \in P_\gamma$.
- It is *supported on* γ, that is, $c(\alpha) = 0$ if α does not meet γ.
- It is *locally constant*, that is, constant on any continuous family $(\varepsilon_t)_{t\in[0,1]}$ of smooth paths $\varepsilon_t \in P_\gamma$. (Such a family of paths is not called a homotopy of paths because the endpoints may move. However, the endpoints never cross γ, since at the instant of crossing the path would not be in P_γ.)

Definition 2.4 The *integral* of a cross-functional η along a path $\alpha \in P_\gamma$ is the number

$$c_\eta(\alpha) := \sum_{t\in\alpha^{-1}(\gamma)} \eta(\alpha'(t)).$$

Lemma 2.5 *The map $\eta \mapsto c_\eta$ is a bijection from the set of integral cross-functionals to the set of integral cross-cochains on γ.*

The proof is straightforward.

Proof For a cross-functional η, it is clear that c_η is a cross-cochain. Let F be the map from the set of integral cross-functionals to the set of integral cross-cochains given by $F(\eta) = c_\eta$.

To show that F is bijective, we use the following notation. For a cross-vector $v \in C_\gamma$, let $C_\gamma(v)$ be the connected component of C_γ containing v, and let $P_\gamma(v)$ be the set of smooth paths in P_γ that cross γ exactly once, and with velocity v' in $C_\gamma(v)$. Note that any two arcs $\varepsilon_0, \varepsilon_1 \in P_\gamma(v)$ are connected by a continuous family of arcs $(\varepsilon_t)_{t\in[0,1]}$ in $P_\gamma(v)$. This implies that any cross-cochain is constant on $P_\gamma(v)$.

The map F is injective since given two cross-functionals $\eta \neq \eta'$, we see that $c_\eta(v) \neq c_{\eta'}$ by evaluating these two cochains at a path $\gamma \in P_\gamma(v)$, where $v \in C_\gamma$ is a cross-vector such that $\eta(v) \neq \eta'(v)$.

Now let us show that F is surjective. Given a cross-cochain c, we shall produce a cross-functional η such that $c_\eta = c$. We define η as follows: for each cross-vector $v \in C_\gamma$, we set $\eta(v) := c(\alpha)$, for any $\alpha \in P_\gamma(v)$. This value is well defined since c is constant on $P_\gamma(v)$, as noted above. In addition, it is clear that $\eta(-v) = \eta(v)$. This shows that η is a cross-functional.

To finish, let us show that $c_\eta = c$. Given a path $\alpha \in P_\gamma$, we decompose it as a concatenation of smooth paths $\alpha_i \in P_\gamma$, where each α_i meets γ once with certain velocity v_i, or not at all, in which case we say that i is trivial. Then we have

$$c(\alpha) = \sum_i c(\alpha_i) = \sum_{i \text{ nontrivial}} c(\alpha_i) = \sum_{i \text{ nontrivial}} \eta(v_i) = c_\eta(\alpha). \qquad \square$$

Remark 2.6 A cross-functional η is a coorientation if and only if the cross-cochain c_η is *unitary*, that is, it satisfies for each path $\alpha \in P_\gamma$ the condition

$$c_\eta(\alpha) = \pm 1 \quad \text{if } \operatorname{Len}_\gamma(\alpha) = 1,$$

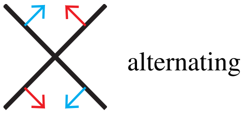
or, equivalently, the conditions

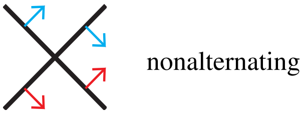
$$c_\eta(\alpha) \leq \operatorname{Len}_\gamma(\alpha) \quad \text{and} \quad c_\eta(\alpha) \equiv \operatorname{Len}_\gamma(\alpha) \mod 2.$$

2.2 Eulerian coorientations

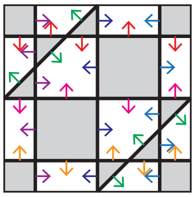
Definition 2.7 A coorientation η of γ is *Eulerian* if the cochain c_η is closed, ie if $c_\eta(\alpha) = 0$ whenever α is a contractible closed curve, or, equivalently, if $c_\eta(\alpha)$ depends only on the homotopy class $\{\alpha\}$. (The homotopies we consider here are with fixed endpoints and disregarding γ, meaning that the intermediate paths may not be in P_γ.) The set of all Eulerian coorientations of γ is denoted by $\operatorname{Eul}(\gamma)$.

Equivalently, η is Eulerian if around each double point p of γ, among the four pieces of η that meet at p there are exactly two with positive coorientation and two with negative coorientation. Hence the local picture of η at p is one of the following two: either when traveling straight along γ and encountering p the coorientation changes — in this case the coorientation is said to be *alternating* at p — or the coorientation does not change when following γ — in which case it is *nonalternating* at p.

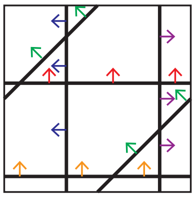




Example 2.8 If $[\gamma]_2 \in H^1(\Sigma; \mathbb{Z}/2\mathbb{Z})$ is zero, then the regions of $\Sigma \setminus \gamma$ can be colored in black and white in such a way that adjacent regions have different colors. In this case we can coorient all edges toward the white regions. The obtained coorientation is Eulerian, all double points being alternating.



Example 2.9 There always exist global Eulerian coorientations, even when $[\gamma]_2 \in H_1(\Sigma; \mathbb{Z}/2\mathbb{Z})$ is not zero. Indeed one can choose a coorientation for every component of γ. This yields an Eulerian coorientation having only nonalternating vertices. If γ consists of c immersed curves, there are 2^c such coorientations.



Remark 2.10 If η is an Eulerian coorientation of γ, then for every generic closed 1-chain α, the number $c_\eta(\alpha)$ depends only of the homology class $[\alpha] \in H_1(\Sigma;\mathbb{Z})$. In consequence, η induces a cohomology class $[\eta] := [c_\eta]$ in $H^1(\Sigma;\mathbb{Z})$.

We denote by $[\mathrm{Eul}(\gamma)]$ the subset of $H^1(\Sigma;\mathbb{Z})$ consisting of the cohomology classes of the Eulerian coorientations on γ. Theorem C states that

$$[\mathrm{Eul}(\gamma)] = \{h \in \overline{B^*_{\boldsymbol{x}_\gamma}} \mid h_{\mathrm{mod}\,2} = [\gamma]_2\}.$$

Let us prove the easy inclusion $\subseteq$, that is, that the cohomology class induced by any Eulerian coorientation is in the dual unit ball $B^*_{\boldsymbol{x}_\gamma}$ (ie it is $\le \boldsymbol{x}_\gamma$) and also coincides with $[\gamma]_2$ modulo 2.

Lemma 2.11 *For every Eulerian coorientation η of γ and every homology class a in $H_1(\Sigma;\mathbb{Z})$, we have $[\eta](a) \le \boldsymbol{x}_\gamma(a)$ and also $[\eta](a) \equiv [\gamma]_2(a) \bmod 2$.*

Proof Let α be an $\boldsymbol{x}_\gamma$-realizing curve of class a. Then $\langle \eta, \alpha\rangle$ counts every intersection point of α and γ with a coefficient ± 1, while $\boldsymbol{x}_\gamma(a)$ counts these same intersection points with a coefficient $+1$ each. Hence we have

$$c_\eta(\alpha) \le \boldsymbol{x}_\gamma(\alpha) \quad \text{and also} \quad c_\eta(\alpha) \equiv \boldsymbol{x}_\gamma(\alpha) \bmod 2. \qquad \square$$

To prove the reverse inclusion we will use eikonal functions.

2.3 Eikonal functions on the universal cover

As before, Σ is a compact closed surface with a multicurve γ on it.

Our task now is to define the eikonal functions on the universal cover $(\tilde{\Sigma}, \pi)$. The space $\tilde{\Sigma}$ has a multicurve $\tilde{\gamma} := \pi^*(\gamma)$ (that is, the pullback of γ by the map π), which induces a length functional $\mathrm{Len}_{\tilde{\gamma}}$ and therefore, a distance function $d_{\tilde{\gamma}}$, which we need to define the notion of eikonal functions. However, we will instead define the distance function directly in terms of the multicurve γ, by taking advantage of the standard explicit construction of the universal cover.

We construct the universal cover $(\tilde{\Sigma}, \pi)$ of the surface Σ as follows. The space $\tilde{\Sigma}$ is the set of homotopy classes of paths in Σ starting at p_0, where $p_0 \in \Sigma \setminus \gamma$ is a fixed, arbitrary point. Thus each point $x \in \tilde{\Sigma}$ is of the form $x = \{\alpha\}$ where α is a path in Σ starting at p_0, and $\{\alpha\}$ denotes its homotopy class (with fixed endpoints). In particular, the space $\tilde{\Sigma}$ has a natural base point $x_0 = \{1_{p_0}\}$, where 1_{p_0} is the trivial path at p_0. The covering map $\pi\colon \tilde{\Sigma} \to \Sigma$ is the function that sends each homotopy class $\{\alpha\}$ to the endpoint of the path α.

The fundamental group $\Pi_1(\Sigma, p_0)$, hereafter denoted by Π_1, acts (on the left) on $\tilde{\Sigma}$ as follows: each loop homotopy class $\{\beta\} \in \Pi_1$ induces on $\tilde{\Sigma}$ a transformation $T_\beta\colon \{\alpha\} \mapsto \{\beta\alpha\}$, where $\beta\alpha$ is the concatenation of the path β followed by the path α. This action commutes with the covering map π (that is, it satisfies $\pi \circ T_\beta = \pi$ for all $\{\beta\} \in \Pi_1$) and is transitive on each fiber of π.

The length with respect to γ of a homotopy class $\{\alpha\}$ is defined as the minimum length of a generic path α' in the class,

$$\mathrm{Len}_\gamma\{\alpha\} = \min_{\alpha' \in \{\alpha\} \cap P_\gamma} \mathrm{Len}_\gamma(\alpha').$$

The *distance* between two points $x = \{\alpha\}$ and $y = \{\beta\} \in \widetilde{\Sigma}$ not located on the multicurve $\tilde{\gamma} := \pi^*(\gamma)$ is

$$d_{\tilde{\gamma}}(x, y) = \mathrm{Len}_\gamma\{\alpha^\dagger \beta\}$$

where $\alpha^\dagger$ is the reverse of the path α. Note that $d_{\tilde{\gamma}}$ satisfies the triangle inequality, therefore it is an integer-valued (but not positive-definite) distance function on $\widetilde{\Sigma} \setminus \tilde{\gamma}$. Moreover, an easy computation shows that the transformations T_β preserve this distance function.

Definition 2.12 An integer-valued function f defined on a subset D of $\widetilde{\Sigma} \setminus \tilde{\gamma}$ is said *preeikonal* if it satisfies

$$|f(y) - f(x)| \le d_{\tilde{\gamma}}(y, x) \quad \text{and} \quad f(y) - f(x) \equiv d_{\tilde{\gamma}}(y, x) \bmod 2 \quad \text{for all } x, y \in D, \tag{2}$$

An *eikonal function* is a preeikonal function defined on the whole set $\widetilde{\Sigma} \setminus \tilde{\gamma}$.

Remark 2.13 A function $f\colon \widetilde{\Sigma} \setminus \tilde{\gamma} \to \mathbb{Z}$ is eikonal if and only if it satisfies the local condition

$$f(y) - f(x) = \begin{cases} 0 & \text{when } d_{\tilde{\gamma}}(x, y) = 0, \\ \pm 1 & \text{when } d_{\tilde{\gamma}}(x, y) = 1. \end{cases}$$

The term "eikonal function" comes from geometric optics, where it describes a (possible singular) real-valued function f that solves the eikonal equation $\|\nabla f\| \equiv 1$. The eikonal functions defined above are discrete analogs of these real-valued functions.

Definition 2.14 An integer-valued function f defined on a subset D of $\widetilde{\Sigma} \setminus \tilde{\gamma}$ is said *equivariant* with respect to a cohomology class $h \in H^1(\Sigma; \mathbb{Z})$, or h-equivariant, if it satisfies

$$f(y) - f(x) = h[\beta]$$

for all pairs of points $x, y \in D$ and all loop homotopy classes $\{\beta\} \in \Pi_1$ such that $T_\beta(x) = y$.

Every Eulerian coorientation η of γ determines a function $f_\eta\colon \widetilde{\Sigma} \setminus \tilde{\gamma} \to \mathbb{Z}$, called the *primitive* of η, by the formula

$$f_\eta\colon \{\alpha\} \mapsto c_\eta(\alpha).$$

The number $c_\eta(\alpha)$ does not depend on how the path α is chosen within its homotopy class since η is Eulerian.

Lemma 2.15 *For each cohomology class $h \in H^1(\Sigma; \mathbb{Z})$, the map $\eta \mapsto f_\eta$ bijects the set of Eulerian coorientations of γ of cohomology class h to the set of h-equivariant eikonal functions on $\widetilde{\Sigma} \setminus \tilde{\gamma}$ that vanish at the base point $x_0 = \{1_{p_0}\}$.*

Proof Let us first see that for each Eulerian coorientation η, the function f_η is eikonal, $[\eta]$-equivariant, and vanishes at the base point. The last claim is clear: since the trivial path 1_{p_0} does not meet γ, we have

$$f_\eta(x_0) = c_\eta(1_{p_0}) = 0.$$

To see that f is eikonal, take two points $\{\alpha\}, \{\beta\} \in \widetilde{\Sigma} \setminus \tilde{\gamma}$ at distance $d_{\tilde{\gamma}}(\{\alpha\}, \{\beta\}) = 1$. This means that there exists a path $\varepsilon \in P_\gamma$ homotopic to $\alpha^\dagger\beta$ with $\mathrm{Len}_\gamma(\varepsilon) = 1$. Hence we can verify that

$$\begin{aligned} f_\eta\{\beta\} - f_\eta\{\alpha\} = c_\eta(\beta) - c_\eta(\alpha) &= c_\eta(\alpha^\dagger\beta) \\ &= c_\eta(\varepsilon) && \text{since } \eta \text{ is Eulerian} \\ &= \pm 1. \end{aligned}$$

Similarly, one can see that $f_\eta\{\beta\} = f_\eta\{\alpha\}$ if $d_{\tilde{\gamma}}(\{\alpha\}, \{\beta\}) = 0$. Finally, to see that f_η is $[\eta]$-equivariant, we take a loop homotopy class $\{\beta\} \in \Pi_1$ and a point $\{\alpha\} \in \widetilde{\Sigma} \setminus \tilde{\gamma}$ and we verify that

$$f_\eta\{\beta\alpha\} = c_\eta(\beta\alpha) = c_\eta(\beta) + c_\eta(\alpha) = [\eta][\beta] + f_\eta\{\alpha\}.$$

Now let us fix a cohomology class $h \in H^1(\Sigma; \mathbb{Z})$. As we have just shown, the map $\eta \mapsto f_\eta$ restricts to a map R_h from the set of Eulerian coorientations of class h to the set of h-equivariant eikonal functions on $\widetilde{\Sigma} \setminus \tilde{\gamma}$ that vanish at the base point $x_0 = \{1_{p_0}\}$. Let us show that R_h is bijective.

To prove that R_h is injective, fix an Eulerian coorientation η and a cross-vector $v \in C_\gamma$. We shall express $\eta(v)$ in terms of the function f_η. Take a path $\varepsilon \in P_\gamma$ which crosses γ just once with velocity v, and let $\alpha \in P_\gamma$ be an auxiliary path from p_0 to the starting point of ε. Then we have

$$f_\eta\{\alpha\varepsilon\} - f\{\alpha\} = c_\eta(\alpha\varepsilon) - c_\eta(\alpha) = c_\eta(\varepsilon) = \eta(v),$$

which shows that η can be recovered from the function f_η, and thus R_h is injective.

Finally, let us show that R_h is surjective. Let $f\colon \widetilde{\Sigma} \setminus \tilde{\gamma} \to \mathbb{Z}$ be an equivariant eikonal function that vanishes at the base point $x_0 = \{1_{p_0}\}$. We have to construct an Eulerian coorientation η such that $f_\eta = f$. To do so, we define first a cross-cochain c as follows. For any generic smooth path $\varepsilon \in P_\gamma$, we let

$$c(\varepsilon) := f\{\alpha\varepsilon\} - f\{\alpha\}$$

where $\alpha \in P_\gamma$ is an auxiliary path from p_0 to the starting point of ε. Let us show first that the value $c(\varepsilon)$ is well defined. Let α' be any other path from p_0 to the starting point of ε. Then we can write $\{\alpha'\} = \{\beta\alpha\}$ where $\beta := \alpha^\dagger\alpha'$, and thus from the fact that f is h-equivariant for some $h\colon \Pi_1 \to \mathbb{Z}$ we get

$$\begin{aligned} f\{\alpha'\varepsilon\} - f\{\alpha'\} &= f\{\beta\alpha\varepsilon\} - f\{\beta\alpha\} \\ &= h[\beta] + f\{\alpha\varepsilon\} - h[\beta] - f\{\alpha\} && \text{since } f \text{ is } h\text{-equivariant} \\ &= f\{\alpha\varepsilon\} - f\{\alpha\}. \end{aligned}$$

We claim that c is a cross-cochain according to Definition 2.3, and in fact, a unitary and closed cross-cochain.

We note first that $c(\varepsilon)$ only depends on the homotopy class $\{\varepsilon\}$. (This will imply that c is closed as a cross-cochain.)

Let us prove that c is additive with respect to concatenation of paths. Let $\delta, \varepsilon \in P_\gamma$ be consecutive paths. To show that $c(\delta\varepsilon) = c(\delta) + c(\varepsilon)$, we take an auxiliary path $\alpha \in P_\gamma$ from p_0 to the starting point of δ. Then we have

$$c(\delta\varepsilon) = f\{\alpha\delta\varepsilon\} - f\{\alpha\} = f\{\alpha\delta\varepsilon\} - f\{\alpha\delta\} + f\{\alpha\delta\} - f\{\alpha\} = c(\varepsilon) + c(\delta).$$

Similarly, let us show that c is alternating with respect to path reversion. Consider a path $\varepsilon \in P_\gamma$ and its reverse $\varepsilon^\dagger$, and let $\alpha \in P_\gamma$ be an auxiliary path from p_0 to the starting point of ε. Note that the path $\alpha\varepsilon$ goes from p_0 to the starting point of $\varepsilon^\dagger$, therefore we have

$$c(\varepsilon^\dagger) = f\{\alpha\varepsilon\varepsilon^\dagger\} - f\{\alpha\varepsilon\} = f\{\alpha\} - f\{\alpha\varepsilon\} = -c(\varepsilon).$$

Next, let us show that c is supported on γ, ie that $c(\varepsilon) = 0$ for any path $\varepsilon \in P_\gamma$ that avoids γ. Let $\varepsilon \in P_\gamma$ be such a path, and let $\alpha \in P_\gamma$ be an auxiliary path from p_0 to the starting point of ε. Then we have

$$\begin{aligned} c(\varepsilon) &= f\{\alpha\varepsilon\} - f\{\alpha\} \\ &\le d_{\tilde{\gamma}}(\{\alpha\}, \{\alpha\varepsilon\}) && \text{since } f \text{ is an eikonal function} \\ &= \mathrm{Len}_\gamma\{\alpha^\dagger\alpha\varepsilon\} = \mathrm{Len}_\gamma\{\varepsilon\} = 0. \end{aligned}$$

Similarly, let us show that c is unitary. For a path $\varepsilon \in P_\gamma$ of length $\mathrm{Len}_\gamma(\varepsilon) = 1$, we have to show that $c(\varepsilon) = \pm 1$. The fact that $\mathrm{Len}_\gamma(\varepsilon) = 1$ implies that $\mathrm{Len}_\gamma\{\varepsilon\} = 1$, since the possibility $\mathrm{Len}_\gamma\{\varepsilon\} = 0$ is excluded because homotopic paths have the same length modulo 2. To compute $c(\varepsilon)$ we take an auxiliary path $\alpha \in P_\gamma$ from p_0 to the starting point of ε and we note that

$$c(\varepsilon) = f\{\alpha\varepsilon\} - f\{\alpha\} = \pm 1$$

since f is an eikonal function and

$$d_{\tilde{\gamma}}(\{\alpha\}, \{\alpha\varepsilon\}) = \mathrm{Len}_\gamma\{\alpha^\dagger\alpha\varepsilon\} = \mathrm{Len}_\gamma\{\varepsilon\} = 1.$$

Finally, let us show that c is constant on any continuous family $(\varepsilon_t)_{t\in[0,1]}$ of smooth paths $\varepsilon_t \in P_\gamma$. It suffices to verify that $c(\varepsilon_0) = c(\varepsilon_1)$. Denote r_t and s_t the starting point and endpoint of ε_t for each $t \in [0,1]$. Note that the curves $r : t \mapsto r_t$ and $s : t \mapsto s_t$ avoid the multicurve γ. These curves r, s may not be in P_γ, but they surely can be approximated by respective curves $\rho, \sigma \in P_\gamma$ that are homotopic to r and s, respectively (with fixed endpoints), and also avoid γ. Then we have $\{\varepsilon_0\} = \{\rho\varepsilon_1\sigma^\dagger\}$, which implies that

$$c(\varepsilon_0) = c(\rho) + c(\varepsilon_1) - c(\sigma) = c(\varepsilon_1)$$

since $c(\rho) = c(\sigma) = 0$ because ρ and σ avoid γ.

This finishes the proof that c is a closed, unitary cross-cochain. Therefore, by Lemma 2.5 (together with Remark 2.6), there exists an Eulerian coorientation η of γ such that $c = c_\eta$. We see that $f_\eta = f$ because

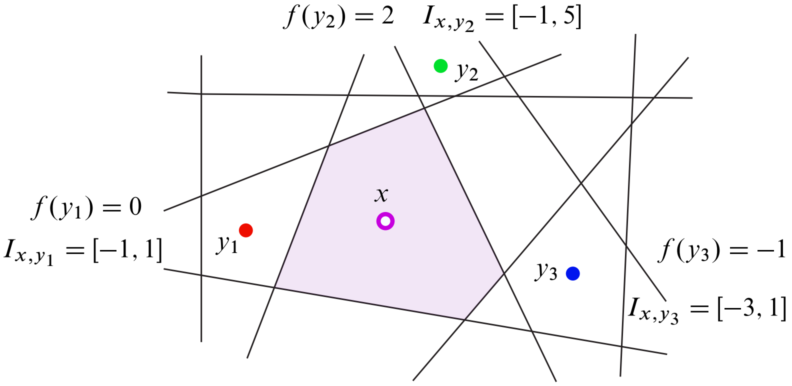


Figure 5: A part of the multicurve $\tilde{\gamma}$ (black and thin). Assume that the set D consists of three points y_1, y_2, y_3 (red, green and blue dots) with prescribed values $f(y_1)=0$, $f(y_2)=2$ and $f(y_3)=-1$. Considering a fourth point x (purple), we see that we have $I_{x,y_1}=[-1,5]$, $I_{x,y_2}=[-1,1]$ and $I_{x,y_3}=[-3,1]$. In particular these three intervals intersect, and one can set $\bar{f}(x)=1$.

for any homotopy class $\{\alpha\} \in \tilde{\Sigma} \setminus \tilde{\gamma}$ (represented by a generic smooth path $\alpha \in P_\gamma$ starting at p_0) we have

$$f_\eta\{\alpha\} = c_\eta(\alpha) = c(\alpha) = f\{1_{p_0}\alpha\} - f\{1_{p_0}\} = f\{\alpha\}$$

since $f\{1_{p_0}\}=0$. This shows that $f_\eta = f$, concluding the proof that R_h is surjective. $\square$

The next result is the key to proving Theorem C.

Lemma 2.16 (extension) *Every preeikonal function f defined on a subset D of $\tilde{\Sigma} \setminus \tilde{\gamma}$ can be extended to an eikonal function $\bar{f}\colon \tilde{\Sigma} \setminus \tilde{\gamma} \to \mathbb{Z}$ given by*

$$\bar{f}(x) = \min_{y \in D} f(y) + d_{\tilde{\gamma}}(x,y).$$

Proof For a point $x \in \tilde{\Sigma} \setminus \tilde{\gamma}$, we want to define $\bar{f}(x)$. We first observe that, for every $y \in D$, the value $\bar{f}(x)$ must lie in the interval $I_{x,y} := [f(y) - d_{\tilde{\gamma}}(x,y), f(y) + d_{\tilde{\gamma}}(x,y)]$. See Figure 5.

We claim that for every y and y' in D, the intervals $I_{x,y}$ and $I_{x,y'}$ intersect. Otherwise there would exist two points y and y' such that $f(y) + d_{\tilde{\gamma}}(x,y) < f(y') - d_{\tilde{\gamma}}(x,y')$, which implies $f(y') - f(y) > d_{\tilde{\gamma}}(x,y) + d_{\tilde{\gamma}}(x,z) \geq d_{\tilde{\gamma}}(y,y')$, contradicting preeikonality of f. Now, any set of intervals in $\mathbb{R}$ that pairwise intersect has a global common point. Therefore the intersection $\cap_{y \in D} I_{x,y}$ is nonempty. So we define $\bar{f}(x)$ as the highest common point $\bar{f}(x) := \min_{y \in D} f(y) + d_{\tilde{\gamma}}(x,y)$ of these intervals.

We claim that the extension $\bar{f}$ is preeikonal (and therefore eikonal, since it is defined at all points of $\tilde{\Sigma} \setminus \tilde{\gamma}$). Indeed, to prove that $|f(x') - f(x)| \leq d_{\tilde{\gamma}}(x,x')$, it is enough to check that

$$|(f(y) + d_{\tilde{\gamma}}(x',y)) - (f(y) + d_{\tilde{\gamma}}(x,y))| \leq d_{\tilde{\gamma}}(x,x')$$

for each y, which follows from the triangle inequality in the form

$$|d_{\tilde\gamma}(x',y)-d_{\tilde\gamma}(x,y)|\le d_{\tilde\gamma}(x,x').$$

To prove that $f(x')-f(x)\equiv d_{\tilde\gamma}(x,x')$ modulo 2, we write

$$\begin{aligned}
f(x')-f(x) &= (f(y')+d_{\tilde\gamma}(x',y'))-(f(y)+d_{\tilde\gamma}(x,y)) && \text{for certain } y,y'\in D\\
&\equiv d_{\tilde\gamma}(y,y')+d_{\tilde\gamma}(x',y')-d_{\tilde\gamma}(x,y) && \text{modulo 2 since } f \text{ is preeikonal}\\
&\equiv d_{\tilde\gamma}(y,y')+d_{\tilde\gamma}(x',y')+d_{\tilde\gamma}(x,y) && \text{since plus and minus coincide mod 2}\\
&\equiv d_{\tilde\gamma}(x,x') && \text{since homotopic paths have equal length mod 2.}
\end{aligned}$$ □

Note that a preeikonal function f generally admits several eikonal extensions. The one we denoted by $\bar f$ is the *highest* one. It has the advantage of being determined by f by an explicit formula.

2.4 Proof of Theorem C

As explained after Lemma 2.11, it remains to be shown that every cohomology class $h\in B^*_{\boldsymbol{x}_\gamma}$ that coincides modulo 2 with $[\gamma]_2$ is the cohomology class of some Eulerian coorientation η. We fix such a cohomology class h.

Recall that we have chosen a point p_0 in $\Sigma\setminus\gamma$ to construct the universal cover $\tilde\Sigma$ and the covering map $\pi\colon\tilde\Sigma\to\Sigma$. Denote $D=\pi^{-1}(p_0)$. We define a function $f\colon D\to\mathbb{Z}$ by the formula $f\{\alpha\}=h[\alpha]$. This function is well defined because homotopic paths are homologous.

Claim 2.17 *The function $f\colon D\to\mathbb{Z}$ is an h-equivariant preeikonal function.*

Proof Let us show that f is h-equivariant. Take a loop β in Σ based at the point p_0. Then for points $y=\{\alpha\}$, $y'=T_\beta(y)=\{\beta\alpha\}\in D$ we have

$$f(y')-f(y)=h[\beta\alpha]-h[\alpha]=h[\beta],$$

as claimed. To show that f is preeikonal we continue as follows. Any two points y, $y'\in D$ can be written as $y=\{\alpha\}$, $y'=T_\beta(y)=\{\beta\cdot\alpha\}$. Therefore we have

$$f(y')-f(y)=h[\beta]\le\boldsymbol{x}_\gamma[\beta]$$

since $h\in B^*_{\boldsymbol{x}_\gamma}$. On the other hand, the distance between y and y' is

$$\begin{aligned}
d_{\tilde\gamma}(y,y') &= \mathrm{Len}_\gamma\{\alpha^\dagger\beta\alpha\}\\
&= \mathrm{Len}_\gamma(\beta') && \text{for some path } \beta'\in\{\alpha^\dagger\beta\alpha\}\\
&\ge \boldsymbol{x}_\gamma[\beta'] && \text{by definition of } \boldsymbol{x}_\gamma\\
&= \boldsymbol{x}_\gamma[\beta] && \text{since } [\beta']=[\alpha^\dagger\beta\alpha]=[\beta]
\end{aligned}$$

which shows that $f(y') - f(y) \le d_{\tilde{\gamma}}(y, y')$. To see that $f(y') - f(y) \equiv d_{\tilde{\gamma}}(y, y')$ modulo 2 we note that

$$\begin{aligned} f(y') - f(y) &= h[\beta] \\ &\equiv [\gamma]_2[\beta] && \text{since } h \equiv [\gamma]_2 \text{ modulo 2} \\ &= [\gamma]_2[\beta'] && \text{since } [\beta'] = [\beta] \text{ with } \beta' \text{ as above} \\ &\equiv \mathrm{Len}_\gamma(\beta') && \text{modulo 2 by definition of } [\gamma]_2 \\ &= d_{\tilde{\gamma}}(y, y'). \end{aligned}$$

This finishes the proof that f is a preeikonal function. □

By the extension lemma (Lemma 2.16), we can extend f to an eikonal function $\bar{f}\colon \tilde{\Sigma} \setminus \tilde{\gamma} \to \mathbb{Z}$ defined by the formula $\bar{f}(x) = \min_{y\in D} f(y) + d_{\tilde{\gamma}}(y, x)$.

Claim 2.18 *The function $\bar{f}$ is h-equivariant.*

Proof This follows from the fact that f is h-equivariant. Indeed, take a loop homotopy class $\{\beta\}$ in Π_1 and a point $x \in \tilde{\Sigma} \setminus \tilde{\gamma}$. Then the value of f at the translate point $x' = T_\beta(x)$ is

$$\begin{aligned} \bar{f}(x') &= \min_{y'\in D} f(y') + d_{\tilde{\gamma}}(y', x') \\ &= \min_{y\in D} f(T_\beta(y)) + d_{\tilde{\gamma}}(T_\beta(y), T_\beta(x)) && \text{since } D = T_\beta(D) \\ &= \min_{y\in D} f(y) + h[\beta] + d_{\tilde{\gamma}}(y, x) && \text{since } f \text{ is } h\text{-equivariant and } T_\beta \text{ preserves } d_{\tilde{\gamma}} \\ &= \bar{f}(x) + h[\beta]. \end{aligned}$$ □

Since f is an h-equivariant eikonal function, by Lemma 2.15 there exists a unique Eulerian coorientation η with cohomology class $[\eta] = h$ such that $f_\eta = \bar{f}$.

Let us put everything together. We have shown in Proposition B that $\boldsymbol{x}_\gamma$ is an integral seminorm on $H_1(\Sigma; \mathbb{Z})$, and this seminorm coincides modulo 2 with the cohomology class $[\gamma]_2$. Therefore we can apply Theorem 2.1 (the extension of Thurston's theorem). We conclude that for each homology class $a \in H_1(\Sigma; \mathbb{Z})$, we have

$$\boldsymbol{x}_\gamma(a) = \max_{\substack{\varphi\in B^*_{\boldsymbol{x}_\gamma} \\ h_{\mathrm{mod}\,2}=[\gamma]_2}} h(a) = \max_{\eta\in \mathrm{Eul}(\gamma)} [\eta](a).$$

This concludes the proof of Theorem C.

3 Birkhoff sections with symmetric boundary for the geodesic flow

We now turn to geodesic flows on unit tangent bundles to hyperbolic surfaces and their Birkhoff sections. Unlike the two previous sections, the surfaces we consider are now equipped with a hyperbolic metric, and all considered multicurves are geodesic. We first recall in Section 3.1 what are the geodesic flow and the symmetric lift of a geodesic. Then in Section 3.2 we associate to every Eulerian coorientation a surface in the unit tangent bundle needed for proving the first part of Proposition D. We recall in Section 3.3 the basic definitions on Birkhoff sections and the elements of Schwartzman–Fried–Sullivan theory we need for our classification. Then in Section 3.4 we recall basic notions on pseudo-Anosov flows and Fried's result on their homology directions. In Section 3.5 we make a bit of elementary algebraic topology for describing homology classes of surfaces with boundary. This allows us to prove in Section 3.6 the second part of Proposition D, as well as Theorem A.

3.1 Geodesic flow and symmetric collections of orbits

Given a hyperbolic surface Σ, its *unit tangent bundle* is the circle bundle $T^1\Sigma$ made of length 1 tangent vectors, that is $T^1\Sigma = \{(p,v) \in T\Sigma \mid \|v\| = 1\}$. The *geodesic flow* $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ on $T^1\Sigma$ is the flow whose orbits are lifts of geodesics. Namely for α a geodesic on Σ parametrized with speed one, the orbit of $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ going through the point $(\alpha(0),\dot\alpha(0)) \in T^1\Sigma$ is described by $\varphi^t_{\mathrm{geod}}(\alpha(0),\dot\alpha(0)) = (\alpha(t),\dot\alpha(t))$. For every oriented periodic geodesic $\underset{\rightarrow}{\alpha}$ on Σ, there is one periodic orbit of $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ corresponding to the oriented lift of $\underset{\rightarrow}{\alpha}$ and denoted by $\vec{\alpha}$. If α now denotes an unoriented geodesic on Σ, there are two associated periodic orbits of $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$, one for each orientation. We denote by $\overset{\leftrightarrows}{\alpha}$ the union of these two periodic orbits, it is an oriented link in $T^1\Sigma$ that is invariant under the involution $(p,v) \mapsto (p,-v)$. A link of the form $\overset{\leftrightarrows}{\alpha}_1 \cup \cdots \cup \overset{\leftrightarrows}{\alpha}_k$ is called a *symmetric link*.[2]

3.2 Birkhoff–Brunella surfaces and the first part of Proposition D

Starting from a hyperbolic surface Σ and a finite collection γ of periodic geodesics[3] on Σ, we now explain how to associate to every Eulerian coorientation of γ a surface in $T^1\Sigma$ bounded by $-\overset{\leftrightarrows}{\gamma}$ and transverse to $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$, thus proving the first part of Proposition D.

Fix a coorientation η (not yet Eulerian) of γ. For every edge e of γ (ie segment between two double points), we consider the set $R^{e,\eta}$ of those tangent vectors based on e and pairing positively with η. It is a subset of in $T^1\Sigma$ of the form $e \times [0,\pi]$ (see Figure 6), and hence we call it an *elementary rectangle*. With the notation of Section 2, it is the closure of a connected component of C_γ. It is bounded by the two lifts of e in $T^1\Sigma$ (called the *horizontal part* of $\partial R^{e,\eta}$) and two halves of the fibers of the extremities of e (called the *vertical part* of $\partial R^{e,\eta}$). Note that the interior of $R^{e,\eta}$ is transverse to the geodesic flow $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ while the horizontal part of $\partial R^{e,\eta}$ is tangent to it. We then orient $R^{e,\eta}$ so that orbits of $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ intersect it positively. One checks that the induced orientation on $\partial R^{e,\eta}$ is opposite to the one given by $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$, as explained in Figure 6. This is the reason why we want to consider negative orientations in Theorem A and Proposition D.

---

[2] The term "antithetic link" was suggested by Bruce Bartlett, but we remarked that symmetric is already used in the literature.

[3] In the sequel we always assume γ to be in general position, meaning in particular that no point belong to three different arcs. This is a restriction as there exists collection of geodesics on surfaces that exhibit triple points for all constantly curved metrics. One way to deal with this situation is to perturb the metric, allowing the curvature to slightly change so that the position of the collection becomes general. Indeed the arguments we use do not require constant curvature, only negative.

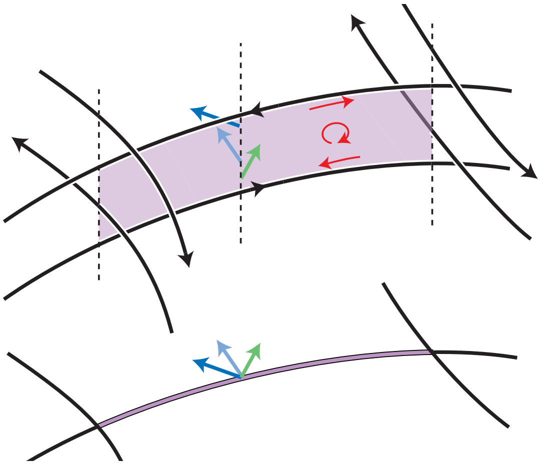


Figure 6: Bottom: an edge e of γ and a coorientation η on it. Top: the corresponding rectangle $R^{e,\eta}$ in $T^1\Sigma$. The dotted lines represent the fibers of some points of Σ, that is, each point on these lines represent a unit tangent vector to Σ. Since the fibers are actually circles, the top and bottom extremities of the dotted lines should be glued. The rectangle $R^{e,\eta}$ is transverse to the orbits of $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ and the induced orientation is shown in red. The induced orientation of the horizontal boundary of $R^{e,\eta}$ (red) is opposed to the orientation of the flow (black). Thus the surfaces we construct are transverse surfaces whose boundary components have multiplicity -1.

Consider now the 2-dimensional CW-complex $S^\times(\eta)$ that is the union of the rectangles $R^{e,\eta}$ over all edges e of γ; see the left parts of Figures 7 and 8.

Lemma 3.1 *The 2-complex $S^\times(\eta)$ described above has boundary $-\vec{\gamma}$ if and only if the coorientation η is Eulerian.*

Proof Since $S^\times(\eta)$ is the union of one rectangle per edge of γ, the horizontal boundary of $S^\times(\eta)$ is always in $\vec{\gamma}$. Since the orientation is opposite to the geodesic flow (see Figure 6), it is actually $-\vec{\gamma}$.

What we have to check is that the vertical boundary is empty if and only if η is Eulerian. At every double point v of γ there are four incident rectangles, corresponding to the four adjacent edges. Now the vertical boundary of a rectangle $R^{e,\eta}$ is oriented upwards at the right extremity of e (when cooriented by η) and downwards at the left extremity. Then the vertical boundary in a vertex of γ is empty if only if all vertical contributions cancel. This is the case exactly when two edges are cooriented in a direction, and two others in the opposite direction: this means that η is Eulerian around v. Conversely, if η is Eulerian, then up to rotation there are two local configurations around v (that we called alternating and nonalternating), and one checks that in both cases, the vertical boundary is empty (see the left parts of Figures 7 and 8). □

When η is Eulerian, the complex $S^\times(\eta)$ is not a topological surface if η has some nonalternating points: as depicted in Figure 8, there are edges in the vertical boundary of four adjacent rectangles, instead of two for obtaining a topological surface. But it is the only obstruction and we can desingularize such

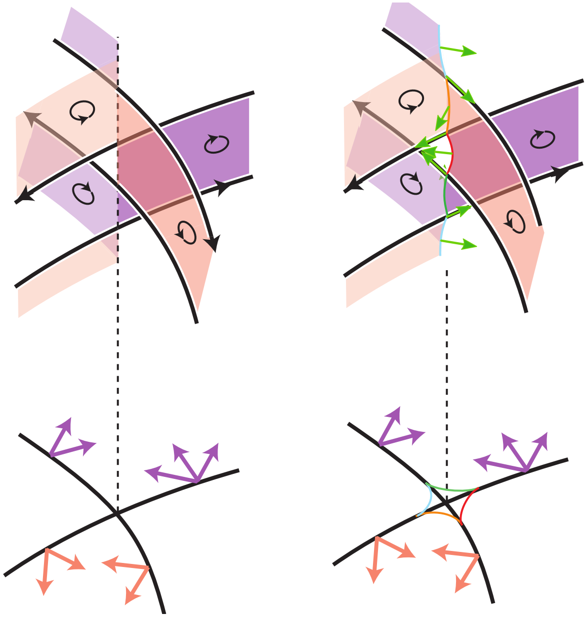


Figure 7: On the left, the complex $S^\times(\eta)$ around the fiber of an alternating double point of γ. Every point of the fiber of v is adjacent to exactly two rectangles. On the right the surface $S^{BB}(\eta)$ is obtained by smoothing $S^\times(\eta)$ in a neighborhood of the fiber of the double point. Its interior is transverse to the vector field generating the geodesic flow (green).

segments as shown on the right of Figure 8. More precisely, label by $1, 2, 3, 4$ the quadrants around the considered nonalternating point so that two edges point toward 1 under the coorientation η. Then the set s of those tangent vectors based on the double point and pointing toward quadrant number 1 is the singular segment to which four rectangles are adjacent. We thus split s into two segments s_1 and s_3, so that the extremities of both segments (in $T^1\Sigma$) coincide with the extremities of s, but s_1 is pushed a bit into quadrant number 1, and s_3 is pushed a bit into quadrant number 3. Then we distort a bit the two rectangles adjacent to quadrant 1 so that their vertical boundary is s_1, and we distort a bit the two rectangles adjacent to quadrant 3 so that their vertical boundary is s_3. These gluings are made in a smooth way.

The main tool connecting Eulerian coorientations to Birkhoff sections is the following.

Definition 3.2 For η an Eulerian coorientation, the associated *BB-surface* is the surface $S^{BB}(\eta)$ obtained from $S^\times(\eta)$ by desingularizing and smoothing the fibers of the double points of γ, as on the right parts of Figures 7 and 8.

The term BB stands for Birkhoff–Brunella, as this construction generalizes previous constructions by these two authors. Indeed, the BB-surface associated to a Birkhoff coorientation (Example 2.8) is isotopic to the construction suggested by Birkhoff and popularized by Fried [6; 20]. Also the BB-surface associated to a Brunella coorientation (Example 2.9) was introduced by Brunella [8, Description 2]. This construction already yields the first part of Proposition D:

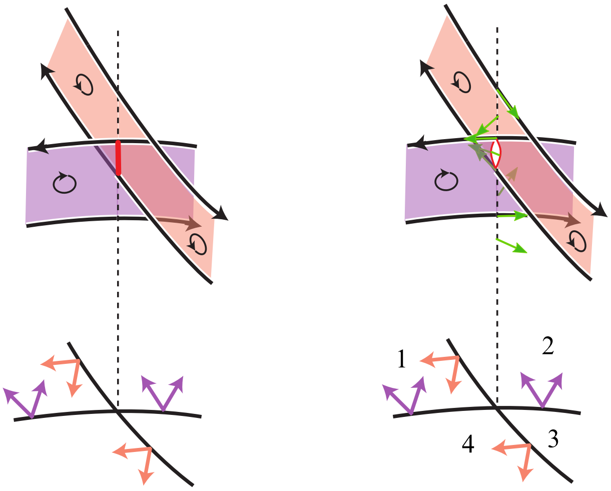


Figure 8: On the left, the complex $S^{\times}(\eta)$ around the fiber of a nonalternating double point of γ. Every point of the fiber of v is adjacent to an even number of rectangles. On the right the surface $S^{BB}(\eta)$ is obtained by desingularizing $S^{\times}(\eta)$ on the portion of the fiber where four rectangles meet. Note that the topology of the complex changes in this process. However its interior is still transverse to the vector field generating the geodesic flow (green).

Proposition 3.3 *For Σ a hyperbolic surface, γ a geodesic multicurve, and η an Eulerian coorientation of γ, the associated surface $S^{BB}(\eta)$ is embedded in $T^1\Sigma$, it is bounded by $-\overleftrightarrow{\gamma}$, and its interior is transverse to the orbits of the geodesic flow $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$.*

Proof The surface $S^{BB}(\eta)$ is obtained by desingularizing $S^{\times}(\eta)$, so it is embedded. Its boundary coincide with the boundary of $S^{\times}(\eta)$, so it is (with orientation) $-\overleftrightarrow{\gamma}$. Finally, the desingularization preserves the transversality to $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$. Since $S^{\times}(\eta)$ is positively transverse to $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ away from its boundary, so is $S^{BB}(\eta)$. □

3.3 Birkhoff sections and Schwartzman–Fried–Sullivan theory

Our goal here is to present a criterion for the existence of a Birkhoff section in a given homology class. Such a criterion exists when the Birkhoff section has no boundary (in this case we call it a global cross section), and it goes back to Schwartzman. It can be adapted to Birkhoff sections using a blow-up construction.

Definition 3.4 Let M be a compact 3-manifold and let $(\varphi^t_X)_{t\in\mathbb{R}}$ be a flow on M generated by a smooth nonvanishing vector field X. (Note that X must be tangent to the boundary ∂M, which must therefore be toric.) A *global cross section* for $(M, (\varphi^t_X)_{t\in\mathbb{R}})$ is a compact orientable surface with boundary S such that:

- S is embedded in M with $S \cap \partial M = \partial S$.
- S is positively transverse to X.
- Every orbit of X intersects S. Note that the time to reach S is a continuous (and hence, bounded) function on M.

When such a global cross section exists, there is a well defined, bijective first-return map f on S and the first-return time τ is bounded from above and below by compactness. In this case M fibers over the circle with fiber S, so that M equipped with the vector field X is homeomorphic to $S\times[0,1]/(p,1)\sim(f(p),0)$ equipped with $\tau(p)\frac{\partial}{\partial z}$, where $f(p)=\varphi^{\tau(p)}(p)$ is the first-return map and $\frac{\partial}{\partial z}$ denotes the vector field tangent to the last coordinate. The dynamics of the flow $(\varphi_X^t)_{t\in\mathbb{R}}$ are then, up to the time-reparametrization function τ, the dynamics of the map f.

The following remark is folklore; see for example the discussion at the beginning of [42, Section 3]. It suggests that questions of existence of global cross sections are of algebraic nature.

Proposition 3.5 *For $(M,(\varphi^t)_{t\in\mathbb{R}})$ a flow, and S_1 and S_2 two global cross sections, there is an isotopy along orbits of $(\varphi^t)_{t\in\mathbb{R}}$ that sends S_1 on S_2 if and only if S_1 and S_2 represent the same class in $H_2(M,\partial M;\mathbb{Z})$.*

Proof The direct implication $\Longrightarrow$ is obvious. For the converse, let $\widehat{M}$ be the infinite cyclic cover of M associated to the class $[S_1]=[S_2]\in H_2(M,\partial M;\mathbb{Z})$ $(=H^1(M;\mathbb{Z})$ by Lefschetz duality). By construction, the surface S_1 lifts to $\mathbb{Z}$ distinct parallel copies $(S_1^{(n)})_{n\in\mathbb{Z}}$. The flow $(\varphi^t)_{t\in\mathbb{R}}$ lifts to a flow $(\hat{\varphi}^t)_{t\in\mathbb{R}}$ in $\widehat{M}$. Since S_1 intersects all orbits of $(\varphi^t)_{t\in\mathbb{R}}$, every orbit of $(\hat{\varphi}^t)_{t\in\mathbb{R}}$ intersects each of the surfaces $(S_1^{(n)})_{n\in\mathbb{Z}}$ one after the other.

Now, S_2 also lifts to $\mathbb{Z}$ parallel copies in $\widehat{M}$ with the same property. In particular every orbit of $(\hat{\varphi}^t)_{t\in\mathbb{R}}$ intersects exactly once each of the surfaces $S_1^{(0)}$ and $S_2^{(0)}$. Hence for $p\in S_1^{(0)}$, we can define t_p to be the unique time so that $\hat{\varphi}^{t_p}(p)\in S_2^{(0)}$. The isotopy $(f_s\colon p\mapsto\hat{\varphi}^{st_p}(p))_{s\in[0,1]}$ hence connects $S_1^{(0)}$ to $S_2^{(0)}$ along orbits of $(\hat{\varphi}^t)_{t\in\mathbb{R}}$. Projecting back to M yields the result. □

Note that if we are given a global cross section S, it intersects all orbits positively. So, taking homology classes, we see that the class $[S]\in H_2(M,\partial M;\mathbb{Z})$ intersects positively all homology classes of periodic orbits of the flow. One may wonder whether the above remark can be turned into a sufficient condition: when does a given homology class σ in $H_2(M,\partial M;\mathbb{Z})$ contain a global section?

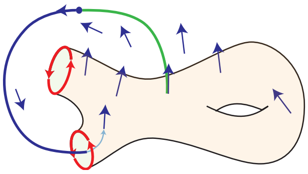
The answer has been given by Sol Schwartzman [38] and Francis Fuller [23], and rephrased by Dennis Sullivan [40]. The quicker way to express it requires to consider invariant measures as currents and to consider their homology classes: given an X-invariant probability measure μ, the associated 1-current c_μ is the linear functional on the space $\Omega^1(M)$ of 1-forms defined by $c_\mu(\lambda)=\int_M\lambda(X(p))\,d\mu(p)$. Since μ is invariant, c_μ is closed as a current, and hence it induces a homology class $[c_\mu]$ in $H_1(M;\mathbb{R})$. The latter is called the *Schwartzman asymptotic cycle* associated to μ. The set of all asymptotic cycles is denoted by $\mathcal{S}\mathrm{chw}_X$. It is a convex subset of $H_1(M;\mathbb{R})$ which contains the classes of the periodic orbits (consider the Dirac linear invariant measures carried by periodic orbits). The following criterion is due to Schwartzman in the case M has no boundary, and to Fried when ∂M is nonempty [38; 19]. Here $\langle\cdot,\cdot\rangle_{(M,\partial M)}$ denotes the intersection pairing $H_2(M,\partial M;\mathbb{R})\times H_1(M;\mathbb{R})\to\mathbb{R}$. Note that $H_2(M,\partial M;\mathbb{Z})\subseteq H_2(M,\partial M;\mathbb{Z})\otimes\mathbb{R}=H_2(M,\partial M;\mathbb{R})$ by the universal coefficient theorem for homology.

Theorem 3.6 (Schwartzman, Fuller, and Fried) *Let M be a 3-manifold with toric boundary equipped with a nonvanishing vector field X tangent to ∂M. A class σ in $H_2(M, \partial M; \mathbb{Z})$ contains a global section for (M, X) if and only if for every asymptotic cycle $c_\mu \in \mathcal{S}\mathrm{chw}_X$ one has $\langle \sigma, c_\mu \rangle_{(M,\partial M)} > 0$.*

Now we turn to Birkhoff sections. Recall from the introduction:

Definition 3.7 For M a compact, orientable 3-manifold with no boundary, X a nonvanishing vector field on M whose flow is denoted by $(\varphi_X^t)_{t\in\mathbb{R}}$, an *embedded Birkhoff section* for $(M, (\varphi_X^t)_{t\in\mathbb{R}})$ is a compact orientable surface S embedded in M such that

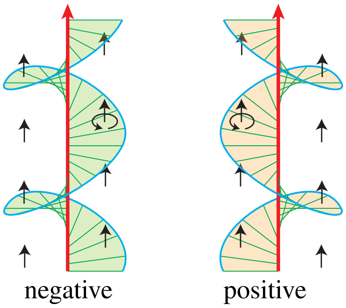
- the interior of S is positively transverse to X,
- its boundary ∂S is tangent to X,
- we have $\varphi_X^{[0,T]}(S) = M$ for some $T > 0$.



The second condition implies that the boundary of S is the union of finitely many periodic orbits of X. Note that one sometimes allows the boundary of S to be immersed instead of embedded, as in [9]. In such case we say that S is an *immersed Birkhoff section*.

The first and second conditions in the definition of a Birkhoff section may look hard to realize at the same time, but actually it is not the case: in a flow box oriented so that the vector field is vertical, the general picture of an embedded Birkhoff section near its boundary is that of one helicoidal staircase. Since the interior of a Birkhoff section S is transverse to X, it is cooriented by X.

Since M is oriented, this induces an orientation on S, and in turn an orientation of ∂S. On the other hand, ∂S is a collection of periodic orbits of X, so it is oriented by X. For every component β of ∂S, we can then define the multiplicity of β as the algebraic number of times one sees β_i in ∂S. Since we restrict our attention to embedded Birkhoff sections, this multiplicity is always ±1. We call a Birkhoff section *positive* (resp. *negative*) if every boundary component has multiplicity $+1$ (resp. -1).



The connection with global cross sections comes from the following remark: starting from a nonsingular flow X on a compact 3-manifold M with no boundary, and given a finite collection β of periodic orbits of X, one can consider the *normal blow-up* of M along β, denoted by M_β. It is obtained from M by removing the 1-submanifold β and replacing it by its unit normal bundle $\nu^1_X(\beta)$. In this construction, each component of β is replaced by a torus. If X is of class C^1, it extends to $\nu^1_X(\beta)$ via its differential, so that M_β is equipped with a continuous vector field X_β.

Now if S is a global cross section for $(M_\beta, (\varphi^t_{X_\beta})_{t\in\mathbb{R}})$, one can change it by an isotopy in an arbitrarily small neighborhood of ∂M_β, so that every boundary component of ∂S is either

- a meridian circle of a boundary torus, that is, the normal bundle to a point $p \in \beta$, or
- a longitude of a boundary torus, that is, its projection in M is an immersion.

After such an isotopy, by blowing down the components of ∂S into orbits of X, we obtain an immersed Birkhoff section for $(M, (\varphi^t_X)_{t\in\mathbb{R}})$ whose boundary is in β. So global cross sections for $(M_\beta, (\varphi^t_{X_\beta})_{t\in\mathbb{R}})$ up to isotopy induce Birkhoff sections whose boundary is in β up to isotopy fixing the boundary.

Conversely, starting from a Birkhoff section S, one can blow up its boundary and obtain a global cross section on the blown-up 3-manifold.

Therefore, provided one can understand the Schwartzman asymptotic cycles after blowing up a periodic orbit, one can adapt the Schwartzman–Fried criterion to the existence of Birkhoff sections. This was done by Fried and even precised by Hryniewicz [19, Theorem N], [29], as we now explain. In our context of a vector field X on a 3-manifold M with a specified finite set β of periodic orbits, every X-invariant measure can be split into two parts: one that is supported on $M \setminus \beta$ and then descends to a X_β-invariant measure on M_β, and one part that corresponds to a combination of Dirac linear X-invariant measures on the components of β. This second part has to be replaced on M_β by an X_β-invariant measure on $\nu^1_X(\beta)$. Since a flow on a 2-torus is in general not uniquely ergodic, the unit normal bundle $\nu^1_X(\beta)$ admits several X_β-invariant measures. However, a given class σ in $H_2(M, \beta; \mathbb{Z})$ induces a class, also denoted by σ, in $H_2(M_\beta, \partial M_\beta; \mathbb{Z})$. All asymptotic cycles associated to all X_β-invariant measures concentrated on $\nu^1_X(\beta)$ have the same pairing with σ, which corresponds to the rotation number of $X_\beta|_{\nu^1_X(\beta)}$ with respect to the slope induced by $\partial\sigma$. We call this pairing the *self-linking* of β along X associated to the framing given by σ, and denote it by $\langle \partial\sigma, \beta^X \rangle_{\nu^1(\beta)}$.

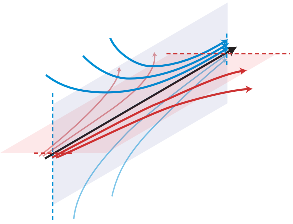
Theorem 3.8 (Schwartzman, Fuller, Fried, and Hryniewicz) *Given are a compact* 3*-manifold M with no boundary, a nonvanishing vector field X on M, and a finite collection β of periodic orbits of X. Then a class σ in $H_2(M, \beta; \mathbb{Z})$ contains an embedded Birkhoff section for $(M, (\varphi^t_X)_{t\in\mathbb{R}})$ if and only if*

- *for every X-invariant measure μ whose support does not intersect β, the corresponding asymptotic cycle $c_\mu \in \mathcal{S}\mathrm{chw}_X$ satisfies $\langle \sigma, c_\mu \rangle_{(M,\beta)} > 0$,*
- *for every component β_i of β, the boundary of $\partial\sigma$ travels plus or minus once along β_i, and one has $\langle \partial\sigma, \beta_i^X \rangle_{\nu^1(\beta_i)} > 0$.*

3.4 Anosov flows

Geodesic flows on unit tangent bundles to hyperbolic surfaces are archetypes of transitive Anosov flows [32; 4]. As such, their asymptotic cycles are easier to understand than those of general flows, as we now explain.

Recall that a flow $(\varphi_X^t)_{t\in\mathbb{R}}$ generated by a vector field X on a 3-manifold is of *Anosov type* if there are two transverse φ_X-invariant 2-foliations $\mathcal{F}^s, \mathcal{F}^u$ on M that intersect along $\mathbb{R}.X$, where X is the generator of the flow, such that $\mathcal{F}^s$ is transversally exponentially contracted by φ_X^t when $t \to +\infty$ and $\mathcal{F}^u$ is transversally exponentially contracted[4] by φ_X^t when $t \to -\infty$.



The leaves of $\mathcal{F}^s$ and $\mathcal{F}^u$ are called *stable* and *unstable manifolds*, respectively.

Recall that two flows are *orbitally equivalent* if there is a homeomorphism sending the oriented orbits of the first flow onto the oriented orbits of the second one. The geodesic flow on a hyperbolic surface is of Anosov type [4]. In particular it is structurally stable, meaning that a small enough perturbation of the generating vector field yields an orbitally equivalent flow. Together with the connectedness of the space of hyperbolic metrics, this implies that the geodesic flows associated to two different hyperbolic metrics are orbitally equivalent [25]. This means that, as long as only the topological properties of orbits are involved, the geodesic flows of all possible hyperbolic metrics on a given surface are equivalent.

Blowing-up some periodic orbits of an Anosov flow does not yield an Anosov flow. However it preserves the pseudo-Anosov character, so we rather work in this context.

Consider the unit disc $\mathbb{D}^2$ in $\mathbb{C}$. For any integer $k \geq 3$ consider the singular 1-foliation $\mathcal{F}_k^1$ on $\mathbb{D}^2$ given by $d(\Re(z^{k/2})) = 0$, and denote by $\mathcal{F}_k^2$ the singular 2-foliation $\mathcal{F}_k^1 \times (0,1)$ on $\mathbb{D}^2 \times (0,1)$. The leaf $\{0\} \times (0,1)$ is singular. Also consider the half-unit disc $\mathbb{U}^2 = \mathbb{D}^2 \cap \{\Im(z) > 0\}$. Consider the singular 1-foliation $\mathcal{F}_\partial^1$ on $\mathbb{U}^2$ given by $d(\Re(s)\cdot\Im(z)) = 0$, and denote by $\mathcal{F}_\partial^2$ the singular 2-foliation $\mathcal{F}_\partial^2 \times (0,1)$ on $\mathbb{U}^2 \times (0,1)$. The leaf $\{0\} \times (0,1)$ is also singular.

Given a compact 3-manifold M with toric boundary, a *foliation with circle-prongs* of M is a 2-foliation with singularities $\mathcal{F}$ of M locally modeled on a standard 2-foliation or on some $\mathcal{F}_k^2$ in the interior of M, and on a standard 2-foliation tangent to ∂M or on $\mathcal{F}_\partial^2$ along ∂M; see Figure 9.

---

[4] Actually this definition corresponds to *topologically Anosov* flows, which is enough for us, as the results we use hold for topologically Anosov flows. Note that it was proven by Shannon that transitive topological Anosov flows are topologically equivalent to smooth Anosov flows [39], so that the topological results on transitive smooth Anosov flows can be used for topologically Anosov flows.

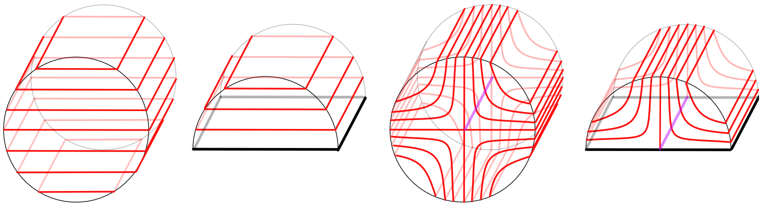


Figure 9: The local picture of a standard 2-foliation in the interior of a 3-manifold (left) and on the boundary (center left). The local picture of the foliation with circle-prong $\mathcal{F}^2_4$ (center right) and the local picture of $\mathcal{F}^2_\partial$ (right).

A flow $(\varphi^t)_{t\in\mathbb{R}}$ on M is of *pseudo-Anosov type* if there are two $(\varphi^t)_{t\in\mathbb{R}}$-invariant foliations with circle-prongs $\mathcal{F}^s, \mathcal{F}^u$ on M that are transverse to each other and intersect along $\mathbb{R}.X$ (except along the singular curves which are common, and parallel to X), where X is the generator of the flow, such that $\mathcal{F}^s$ is transversally exponentially contracted by φ^t when $t \to +\infty$ and $\mathcal{F}^u$ is transversally exponentially contracted[5] by φ^t when $t \to -\infty$. Note that the pseudo-Anosov flows we consider in the sequel are obtained by blowing up periodic orbits of Anosov flows. Hence the circle-prongs of the blown-up foliations are only of type $\mathcal{F}^2_\partial$; the types $\mathcal{F}^2_k$ with $k \geq 3$ do not appear in our context.

Recall that a flow is *transitive* if it has a dense orbit. Geodesic flows on hyperbolic surfaces are transitive. Brunella showed that transitive pseudo-Anosov flows admit finite Markov partitions [7, Theorem 2.1]. Earlier Fried showed that the cone generated by the asymptotic cycles of a flow admitting a finite Markov partition is easy to describe [19, Theorem H]:

Theorem 3.9 (Fried) *Given a compact 3-manifold M with toric boundary and a nonvanishing vector field X on M tangent to ∂M generating a flow $(\varphi^t_X)_{t\in\mathbb{R}}$ admitting a finite Markov partition, there is a finite collection $\{\beta_1, \dots, \beta_n\}$ of periodic orbits of $(\varphi^t_X)_{t\in\mathbb{R}}$ such that $\mathbb{R}_+.\mathcal{S}\mathrm{chw}_X = \mathrm{Conv}(\{\mathbb{R}_+[\beta_i]\}_{i=1,\dots,n})$.*

Combining the above statement with the existence criterion of Theorem 3.8, in the case of geodesic flows we obtain the following result.

Corollary 3.10 *Given a hyperbolic surface Σ and $\vec{\beta}$ a signed collection of periodic orbits of $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ on $T^1\Sigma$, a class σ in $H_2(T^1\Sigma, \vec{\beta}; \mathbb{Z})$ such that $\partial\sigma = \vec{\beta}$ contains a Birkhoff section for $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ if, and only if,*

- *for every periodic orbit $\vec{\alpha}$ of $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$ not in $\vec{\beta}$, one has $\langle \sigma, [\vec{\alpha}]\rangle_{(T^1\Sigma,\vec{\beta})} > 0$,*
- *for every component $\vec{\beta}_i$ of $\vec{\beta}$, one has $\langle \partial\sigma, \vec{\beta}_i^{X_{\mathrm{geod}}}\rangle_{\nu^1(\vec{\beta}_i)} > 0$.*

Actually, Theorem 3.9 states that the infinite set of all periodic orbits in the first item could be replaced by a finite one, but determining this finite set for every signed collection $\vec{\beta}$ does not look trivial to us.

---

[5] Pseudo-Anosov flows correspond to the *expansive flows* of Brunella [7]. Thanks to results of Inaba–Matsumoto and Paternain, both notions coincide, as explained in Brunella's thesis.

3.5 Classes of surfaces with given boundary

We come back to the setting of Theorem A: Σ is a negatively curved surface, γ is a finite collection of periodic geodesics and $\overleftrightarrow{\gamma}$ denotes the symmetric lift of γ. In order to apply Schwartzman, Fuller, Fried, and Hryniewicz's criterion in the form of Corollary 3.10 for finding Birkhoff cross sections bounded by $-\overleftrightarrow{\gamma}$, we need to work in the complement $T^1\Sigma \setminus \overleftrightarrow{\gamma}$ and in particular to determine the space $H_2(T^1\Sigma, \overleftrightarrow{\gamma}; \mathbb{Z})$. In this section we explain that the homology classes of 2-chains bounded by $-\overleftrightarrow{\gamma}$ form an affine space and we give a canonical origin to this space.

Lemma 3.11 *The homology classes of those* 2*-chains whose boundary is* $-\overleftrightarrow{\gamma}$ *form an affine space directed by* $H_1(\Sigma; \mathbb{Z})$.

Proof First we consider the sequence $0 \to H_2(T^1\Sigma; \mathbb{Z}) \xrightarrow{i} H_2(T^1\Sigma, \overleftrightarrow{\gamma}; \mathbb{Z}) \xrightarrow{\partial} H_1(\overleftrightarrow{\gamma}; \mathbb{Z})$, where i is the inclusion map and ∂ is the boundary map. We claim that it is exact.[6] Indeed this is a part of the long exact sequence associated to the pair $(T^1\Sigma, \overleftrightarrow{\gamma})$; see [28, Theorem 2.16], plus the fact that $H_2(\overleftrightarrow{\gamma}; \mathbb{Z})$ is zero.

Now the homology classes of those 2-chains whose boundary is $-\overleftrightarrow{\gamma}$ correspond to the preimages under ∂ of the point $(-1, -1, \dots, -1) \in H_1(\overleftrightarrow{\gamma}; \mathbb{Z}) \simeq \mathbb{Z}^{2|\gamma|}$. Indeed, given two 2-chains with the same boundary, their difference induces a class in $H_2(T^1\Sigma; \mathbb{Z})$. Using the fact that $T^1\Sigma$ is a circle bundle with nonzero Euler class, we get $H_2(T^1\Sigma; \mathbb{Z}) \simeq H_1(\Sigma; \mathbb{Z})$: a nontrivial class in $H_2(T^1\Sigma; \mathbb{Z})$ can be represented by the set of the fibers over a cycle in $H_1(\Sigma; \mathbb{Z})$. □

From Lemma 3.11 we deduce that if we are given an explicit 2-chain S_0 bounded by $-\overleftrightarrow{\gamma}$, the classes of the other 2-chains bounded by $-\overleftrightarrow{\gamma}$ differ from $[S_0]$ by a class in $H_1(\Sigma; \mathbb{Z})$. In our context, there is a natural choice of such an origin S_0, for which the computation of the intersection numbers with asymptotic cycles of the geodesic flow will be easy.

We denote by $S^{\times}_{\pm}$ the rational chain in $C_2(T^1\Sigma, \overleftrightarrow{\gamma}; \mathbb{Q})$ that is half the sum of all elementary rectangles $R^{e,\eta}$ (see Figure 10) and by $\sigma_{\pm}$ its homology class in $H_2(T^1\Sigma, \overleftrightarrow{\gamma}; \mathbb{Q})$,

$$S^{\times}_{\pm} := \frac{1}{2} \sum_{e \in \gamma, \eta_e = \pm} R^{e,\eta_e}, \qquad \sigma_{\pm} := [S^{\times}_{\pm}].$$

In other words, we consider the set of all tangent vectors based at points of γ. Remember that every elementary rectangle is cooriented by the geodesic flow, and hence oriented. Therefore, $S^{\times}_{\pm}$ is also oriented. Its boundary is then exactly $-\overleftrightarrow{\gamma}$ (thanks to the $\frac{1}{2}$ factor). The 2-chain $S^{\times}_{\pm}$ is not a surface since the fibers of the double points of γ are singular. As it is rational the class $\sigma_{\pm}$ might not be realized by a surface, but $2\sigma_{\pm}$ is always an integral class.[7]

---

[6] An erroneous version of this statement is in [19, Lemma 6], where it is claimed that the boundary map is surjective and admits a section. It is not true in general, unless $T^1\Sigma$ is a homology sphere.

[7] Actually, $\sigma_{\pm}$ is realized by a surface if and only if $[\gamma]_2$, the class of γ in $H_1(\Sigma; \mathbb{Z}/2\mathbb{Z})$, is 0. In this case, the homology class of Birkhoff's coorientation η_B (Example 2.8) is 0, and $S^{BB}(\eta_B)$ lies in the class $\sigma_{\pm}$. Also the class $\sigma_{\pm}$ is equal to $\frac{1}{2}[S^{BB}(\eta) + S^{BB}(-\eta)]$ for every Eulerian η. Hence it is always realized as the mean of two surfaces without any assumption on $[\gamma]_2$.

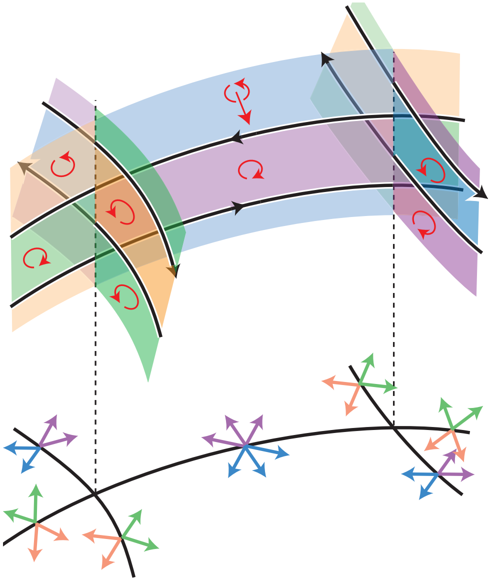


Figure 10: The 2-chain $S^{\times}_{\pm}$ is half of the sum of all rectangles R^{e,η_e}. It is cooriented by the geodesic flow, and hence oriented (in red). Its boundary, taking orientations into account, is $-\overleftrightarrow{\gamma}$.

The class $[S^{\times}_{\pm}]$ yields a canonical origin to the affine space of those 2-chains bounded by $-\overleftrightarrow{\gamma}$, in the sense that it connects the intersection numbers in $T^1\Sigma_{\overleftrightarrow{\gamma}}$ to intersection numbers of the base surface Σ.

Lemma 3.12 *For $\underrightarrow{\alpha}$ a collection of oriented periodic geodesics on Σ, none of which is a component of γ, denote by $\vec{\alpha}$ its lift in $T^1\Sigma$. Then the algebraic intersection $\langle \sigma_\pm, \vec{\alpha}\rangle_{(T^1\Sigma, \overleftrightarrow{\gamma})}$ is equal to $+\frac{1}{2}\operatorname{Len}_\gamma(\underrightarrow{\alpha})$.*

This lemma appears in a different form in [17] where it is used to prove that the linking number of two symmetric collections $\overleftrightarrow{\gamma}_1, \overleftrightarrow{\gamma}_2$ in $T^1\Sigma$ is equal to $-\operatorname{Len}_{\gamma_1}(\gamma_2)$.

Proof Since $S^{\times}_{\pm}$ is positively transverse to the geodesic flow, all intersection points of $\vec{\alpha}$ with $S^{\times}_{\pm}$ contribute positively to the algebraic intersection. Since every rectangle has coefficient $\frac{1}{2}$ in $S^{\times}_{\pm}$, the contribution of every intersection point is $+\frac{1}{2}$. Finally $\vec{\alpha}$ intersects $S^{\times}_{\pm}$ exactly in the fiber of the intersection points of $\underrightarrow{\alpha}$ and γ. □

The connection with intersection norms is now straightforward:

Corollary 3.13 *For $\underrightarrow{\alpha}$ a collection of oriented periodic geodesics on Σ, none of which is a component of γ, the intersection $\langle \sigma_\pm, \vec{\alpha}\rangle_{(T^1\Sigma, \overleftrightarrow{\gamma})}$ is at least equal to $\frac{1}{2}\boldsymbol{x}_\gamma([\underrightarrow{\alpha}])$, with equality if and only if $\underrightarrow{\alpha}$ is an $\boldsymbol{x}_\gamma$-realizing collection of geodesics.*

3.6 Proofs of Proposition D and Theorem A

Let us recall the context: Σ is a hyperbolic surface and γ a finite collection of periodic orbits on Σ. We denote by $\overleftrightarrow{\gamma}$ the symmetric lift of γ in $T^1\Sigma$ and by $T^1\Sigma_{\overleftrightarrow{\gamma}}$ the 3-manifold obtained from $T^1\Sigma$ by blowing up the link $\vec{\gamma}$. It has toric boundary, and it is equipped with the extension, also denoted by $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$, of the geodesic flow. The latter is of pseudo-Anosov type (see Section 3.4).

Denote by π_* the canonical projection from $H_2(T^1\Sigma;\mathbb{R})$ to $H_1(\Sigma;\mathbb{R})$. The next statement is the key property connecting Birkhoff sections and intersection norms.

Lemma 3.14 *If γ is a filling geodesic multicurve on Σ, a class $\sigma \in H_2(T^1\Sigma, \vec{\gamma};\mathbb{Z})$ intersects positively (resp. nonnegatively) every class $[\vec{\alpha}] \in H_1(T^1\Sigma_{\vec{\gamma}};\mathbb{Z})$ for $\underrightarrow{\alpha}$ an oriented periodic geodesic on Σ if and only if the class $\pi_*(\sigma-\sigma_\pm) \in H_1(\Sigma;\mathbb{Z})$ lies in the interior (resp. the closure) of $\frac{1}{2}B^*_{\boldsymbol{x}_\gamma}$.*

Proof For every oriented geodesic $\underrightarrow{\alpha}$ on Σ, by Lemma 3.12, we have

$$\begin{aligned}\langle\sigma,\vec{\alpha}\rangle_{(T^1\Sigma,\vec{\gamma})} &= \langle\sigma-\sigma_\pm,\vec{\alpha}\rangle_{(T^1\Sigma,\vec{\gamma})} + \langle\sigma_\pm,\vec{\alpha}\rangle_{(T^1\Sigma,\vec{\gamma})} = \langle\sigma-\sigma_\pm,\vec{\alpha}\rangle_{(T^1\Sigma,\vec{\gamma})} + \tfrac{1}{2}\operatorname{Len}_\gamma(\underrightarrow{\alpha})\\ &= \langle\pi_*(\sigma-\sigma_\pm),\underrightarrow{\alpha}\rangle_\Sigma + \tfrac{1}{2}\operatorname{Len}_\gamma(\underrightarrow{\alpha}).\end{aligned}$$

Hence $\langle\sigma,\vec{\alpha}\rangle_{(T^1\Sigma,\vec{\gamma})}$ is positive if and only if $-\langle\pi_*(\sigma-\sigma_\pm),\underrightarrow{\alpha}\rangle_\Sigma$ is smaller than $\frac{1}{2}\operatorname{Len}_\gamma(\underrightarrow{\alpha})$.

Now the term $-\langle\pi(\sigma-\sigma_\pm),\underrightarrow{\alpha}\rangle_\Sigma$ depends only on the class $[\underrightarrow{\alpha}] \in H_1(\Sigma;\mathbb{Z})$, while the other term $+\frac{1}{2}\operatorname{Len}_\gamma(\underrightarrow{\alpha})$ is larger that $\frac{1}{2}\boldsymbol{x}_\gamma([\underrightarrow{\alpha}])$, with equality if $\underrightarrow{\alpha}$ is $\boldsymbol{x}_\gamma$-realizing (Corollary 3.13).

We then treat separately the cases $[\underrightarrow{\alpha}] \neq 0$ and $[\underrightarrow{\alpha}] = 0$ in $H_1(\Sigma;\mathbb{Z})$.

Since there is an $\boldsymbol{x}_\gamma$-realizing geodesic in every nonzero homology class, $-\langle\pi_*(\sigma-\sigma_\pm),\underrightarrow{\alpha}\rangle_\Sigma < +\frac{1}{2}\operatorname{Len}_\gamma(\underrightarrow{\alpha})$ for all nonnullhomologous geodesics $\underrightarrow{\alpha}$ if and only if $-\langle\pi_*(\sigma-\sigma_\pm),a\rangle_\Sigma < \frac{1}{2}\boldsymbol{x}_\gamma(a)$ for every nonzero homology class. In the same way, $-\langle\pi_*(\sigma-\sigma_\pm),\underrightarrow{\alpha}\rangle_\Sigma \leq +\frac{1}{2}\operatorname{Len}_\gamma(\underrightarrow{\alpha})$ for all nonnullhomologous geodesics $\underrightarrow{\alpha}$ if and only if $-\langle\pi_*(\sigma-\sigma_\pm),a\rangle_\Sigma \leq \frac{1}{2}\boldsymbol{x}_\gamma(a)$ for every nonzero homology class.

If $\underrightarrow{\alpha}$ is nullhomologous, $\frac{1}{2}\operatorname{Len}_\gamma(\underrightarrow{\alpha}) > 0$ since the multicurve γ is filling, and $-\langle\pi_*(\sigma-\sigma_\pm),\underrightarrow{\alpha}\rangle_\Sigma = 0$.

Summarizing the two previous paragraphs, we find that the class σ intersects positively (resp. nonnegatively) the class of every periodic orbit of the geodesic flow (in the complement of $\vec{\gamma}$) if and only if for every class $a \in H_1(\Sigma;\mathbb{Z})$ we have the inequality $-\langle\pi_*(\sigma-\sigma_\pm),a\rangle_\Sigma < \frac{1}{2}\boldsymbol{x}_\gamma(a)$ (resp. $-\langle\pi_*(\sigma-\sigma_\pm),a\rangle_\Sigma \leq \frac{1}{2}\boldsymbol{x}_\gamma(a)$), which means exactly that the point $-\pi_*(\sigma-\sigma_\pm)$ belongs to the interior (resp. the closure) of $\frac{1}{2}B^*_{\boldsymbol{x}_\gamma}$. Since the latter is symmetric about the origin, this amounts to $\pi_*(\sigma-\sigma_\pm)$ belonging to the interior (resp. the closure) of $\frac{1}{2}B^*_{\boldsymbol{x}_\gamma}$. □

We can now assemble all blocks and prove our main results.

Proof of Proposition D For η an Eulerian coorientation, we consider the Birkhoff–Brunella surface $S^{\boldsymbol{BB}}(\eta)$ given by Definition 3.2. By Proposition 3.3 its interior is transverse to the orbits of the geodesic flow in $T^1\Sigma$ while its boundary consists (with orientation) of $-\vec{\gamma}$. One checks that every elementary rectangle $R^{e,\eta}$ contributes to -1 to the Euler characteristics, and hence $\chi(S^{\boldsymbol{BB}}(\eta))$ is $-|E(\gamma)|$. Since γ is seen as a graph of degree 4, one has $|E(\gamma)| = 2|V(\gamma)|$, so that $\chi(S^{\boldsymbol{BB}}(\eta)) = -2|V(\gamma)|$.

If two Eulerian coorientations η_1 and η_2 are cohomologous, the class $[S^{BB}(\eta_1)-S^{BB}(\eta_2)] \in H_2(T^1\Sigma;\mathbb{Z})$ projects by π onto $[\eta_1-\eta_2]=0$. Since π_* is actually an isomorphism we have $[S^{BB}(\eta_1)-S^{BB}(\eta_2)]=0$, which in turn implies $[S^{BB}(\eta_1)]=[S^{BB}(\eta_2)]$ in $H_2(T^1\Sigma,\vec{\gamma};\mathbb{Z})$.

Finally, if $S^{BB}(\eta_1)$ and $S^{BB}(\eta_2)$ are both Birkhoff sections of $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$ and are homologous, one can blow-up their boundary components (which are the same orbits), and Proposition 3.5 claims that the flow actually realizes an isotopy between the blown-up surfaces. By blowing down, we obtain the desired isotopy away from the boundary. □

Proof of Theorem A Given a hyperbolic surface Σ and a geodesic multicurve γ that fills Σ, Definition 3.2 yields a map that associates to every Eulerian coorientation η of γ a surface $S^{BB}(\eta)$ bounded by $-\vec{\gamma}$ and whose interior is transverse to $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$. Moreover, Proposition D states that if two Eulerian coorientations η_1,η_2 are cohomologous and the surfaces $S^{BB}(\eta_1), S^{BB}(\eta_2)$ are Birkhoff sections for $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$, then they are actually isotopic along the flow. Therefore the map S^{BB} projects to an injective map $[S^{BB}]$ that takes a cohomology class of Eulerian coorientations to an isotopy class of surfaces transverse to $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$.

Lemma 3.11 claims that the homology classes of (rational) 2-chains bounded by $-\vec{\gamma}$ form an affine space directed by $H_1(\Sigma;\mathbb{Q})$. The class $\sigma_\pm$ defined in Section 3.5 gives a canonical origin to this space. It is a half-integral class, and its double $2\sigma_\pm$ is congruent to $[\gamma]_2$ mod 2. Therefore the set $2H_1(\Sigma;\mathbb{Z})$ of the doubles of all integral classes corresponds to the sublattice of $H_1(\Sigma;\mathbb{Z})$ of those points congruent to $[\gamma]_2$ mod 2.

Theorem C states that all classes $[\eta]$ for η Eulerian belong to the closure of $B^*_{x_\gamma}$, and every integral point in $B^*_{x_\gamma}$ that is congruent to $[\gamma]_2$ mod 2 is realized by the class of an Eulerian coorientation. This means that the domain of $[S^{BB}]$ is exactly the integral classes in $B^*_{x_\gamma}$ that are congruent to $[\gamma]_2$ mod 2.

What remains to prove is that the restriction of $[S^{BB}]$ to the interior of $B^*_{x_\gamma}$ has its image in the realm of isotopy classes of Birkhoff sections, and that it is surjective.

By Schwartzman, Fuller, Fried, and Hryniewicz's criterion in the form of Corollary 3.10, a class $\sigma \in H_2(T^1\Sigma,\vec{\gamma};\mathbb{Z})$ whose boundary is $[-\vec{\gamma}]$ contains a Birkhoff cross section if and only if it pairs negatively with all classes $\vec{\alpha}$ of periodic orbits of $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$, plus it links negatively with all boundary components (the >0 in Corollary 3.10 are all replaced by <0 because of the signs of all boundary components).

By Lemma 3.14 the first condition is equivalent to the difference $\pi_*(\sigma-\sigma_\pm)$ lying inside $\frac{1}{2}B^*_{x_\gamma}$, or equivalently to $2\pi_*(\sigma-\sigma_\pm)$ lying inside $B^*_{x_\gamma}$.

Concerning the second condition in Corollary 3.10, one has to check that, if η is an Eulerian coorientation such that $[\eta]$ lies inside $B^*_{x_\gamma}$, for every component $\vec{\gamma}_i$ of $\vec{\gamma}$, one has $\langle \partial[S^{BB}(\eta)], \vec{\gamma}_i^{X_{\text{geod}}}\rangle_{\nu^1(\vec{\gamma}_i)} > 0$. As explained just before Theorem 3.8, $\vec{\gamma}_i^{X_{\text{geod}}}$ denotes any X_{geod}-invariant measure in the boundary component of the blow-up of $\vec{\gamma}_i$. One such invariant measure is carried by the trace of the stable manifold of $\vec{\gamma}_i$, so that one only has to prove that $\partial S^{BB}(\eta)\cap\nu^1(\vec{\gamma}_i)$ intersects the trace of the stable manifold of $\vec{\gamma}_i$.

Let us work by contrapositive and assume that $\partial S^{BB}(\eta) \cap \nu^1(\vec{\gamma}_i)$ does not intersect the stable manifold of $\vec{\gamma}_i$. Consider all double points that are met when traveling along the oriented curve $\underset{\rightarrow}{\gamma}_i$ on Σ. If at least one of them is alternating for η then one sees in Figure 7 that $S^{BB}(\eta)$ rotates from top to bottom (or bottom to top), so that it intersects the stable manifold of $\vec{\gamma}_i$ in a neighborhood of the considered double point. Therefore $\underset{\rightarrow}{\gamma}_i$ has only vertices that are nonalternating for η. Moreover, looking at Figure 8, one sees that at every vertex, the coorientation of the transverse component must be opposite to that of $\underset{\rightarrow}{\gamma}_i$. Therefore the pairing $\eta(\underset{\rightarrow}{\gamma}_i)$ is 0, yielding $[\eta]([\underset{\rightarrow}{\gamma}_i]) = 0$, and so $[\eta]$ does not lie in the interior of $B^*_{\boldsymbol{x}_\gamma}$.

The two previous paragraphs imply that the restriction of $[S^{BB}]$ to the interior of $B^*_{\boldsymbol{x}_\gamma}$ has its image in the realm of isotopy classes of Birkhoff sections, and that it is surjective, thus concluding the proof. □

One may wonder what happens in Theorem A when γ is not filling.[8] In this case, there exists at least one geodesic α not intersecting γ. The two oriented lifts of α yield two periodic orbits $\vec{\alpha}$ and $\overleftarrow{\alpha}$ of $(\varphi^t_{\mathrm{geod}})_{t\in\mathbb{R}}$. These two lifts are anti-isotopic in the complement of $\overleftrightarrow{\gamma}$: the isotopy obtained by rotating the tangent vectors by an angle from 0 to π transports $\vec{\alpha}$ to $-\overleftarrow{\alpha}$. This implies that a surface cannot be positively transverse to $\vec{\alpha}$ and $\overleftarrow{\alpha}$ simultaneously. Therefore $-\overleftrightarrow{\gamma}$ bounds no Birkhoff section. However, the dual unit ball $B^*_{\boldsymbol{x}_\gamma}$ may or may not contain integral points in its interior, depending on γ. So there is no simple extension of Theorem A when γ is not filling, except by saying that $-\overleftrightarrow{\gamma}$ cannot bound a Birkhoff section.

4 Extension to orientable 2-orbifolds

We explain here how the results extend to 2-dimensional orbifolds. Actually Propositions B and D, and Theorem C extend directly. The only point that is not straightforward is Theorem A, which requires an additional argument.

Definition 4.1 [41, Chapter 13] A *Riemannian orientable 2-dimensional orbifold* $\mathbb{O}$ is given by an orientable topological surface $\Sigma_{\mathbb{O}}$ together with an atlas $(U_\alpha, \varphi_\alpha)_{\alpha\in A}$ of charts of the form

$$\varphi_\alpha \colon U_\alpha \to D_\alpha/(\mathbb{Z}/k_\alpha\mathbb{Z}),$$

with D_α a 2-dimensional Riemannian disc on which $\mathbb{Z}/k_\alpha\mathbb{Z}$ acts by rotations, and such that the chart transition maps $\varphi_\alpha \circ \varphi_\beta^{-1}$ are isometries.

Actually the orbifolds to which our theorems extend are the hyperbolic ones. Such a 2-orbifold is always *good* in the sense of Thurston, namely it is a quotient of a hyperbolic surface by a finite automorphism group.

For our purpose we define the *first homology group* $H_1(\mathbb{O};\mathbb{R})$ to be simply $H_1(\Sigma_{\mathbb{O}};\mathbb{R})$. Then the definition of intersection norms extends directly and Proposition B and Theorem C hold.

---

[8] In a previous version of this article, it was claimed that Theorem A also holds in this case. This is false, as was noted by Marty.

We now turn to Proposition D and Theorem A. First we have to define unit tangent bundles to orbifolds and geodesic flows. If D is a Riemannian disc on which $\mathbb{Z}/k\mathbb{Z}$ acts by rotation (with a fixed point), then $\mathbb{Z}/k\mathbb{Z}$ also acts on the unit tangent bundle T^1D. The action on T^1D is free, since the vectors tangent to the fixed point are rotated. Hence the quotient $T^1D/(\mathbb{Z}/k\mathbb{Z})$ is a 3-manifold (actually it is a solid torus).

Definition 4.2 Given a Riemannian orientable 2-orbifold $\mathbb{O} = (\Sigma_{\mathbb{O}}, (U_\alpha, \varphi_\alpha)_{\alpha\in A})$, its *unit tangent bundle* is the 3-manifold $T^1\mathbb{O}$ defined by the atlas $(\hat{U}_\alpha, \hat{\varphi}_\alpha)_{\alpha\in A}$, where $\hat{U}_\alpha = T^1U_\alpha$ and $\hat{\varphi}_\alpha(x,v) = (\varphi_\alpha(x), d(\varphi_\alpha)_x(v))$. It is equipped with a canonical projection $\pi\colon T^1\mathbb{O}\to\mathbb{O}$. If $\mathbb{O}$ is of the form Σ/Γ for some hyperbolic surface Σ, then $T^1\mathbb{O}$ is simply the quotient $(T^1\Sigma)/\Gamma$. The *geodesic flow* on $T^1\mathbb{O}$ is defined as in the nonsingular case by $\varphi^t_{\mathrm{geod}}(\gamma(0),\dot\gamma(0)) = (\gamma(t),\dot\gamma(t))$, where γ is any geodesic with speed 1.

With these definitions, the constructions of Section 3.2 (the BB-surface $S^{BB}(\eta)$ associated to an Eulerian coorientation) can be transposed and Lemmas 3.1, 3.11, and 3.12 remain true.

Now, for $\mathbb{O}$ a hyperbolic 2-orbifold, the unit tangent bundle $T^1\mathbb{O}$ is a 3-manifold, and $H_2(T^1\mathbb{O};\mathbb{R}) \simeq H_1(\mathbb{O};\mathbb{R})$. Indeed closed curves in $\Sigma_{\mathbb{O}}$ lift by π^{-1} to closed surfaces in $T^1\mathbb{O}$. The fact that the unit tangent to a conic disc $D/(\mathbb{Z}/k\mathbb{Z})$ is a torus whose core is the singular fiber implies that cohomologous curves lift to cohomologous surfaces, so that π^{-1} induces a well defined map $\pi_*^{-1}\colon H_1(\mathbb{O};\mathbb{R})\to H_2(T^1\mathbb{O};\mathbb{R})$. The orbifold Euler characteristics of $\mathbb{O}$ is negative by hyperbolicity, so that the Euler number of $T^1\mathbb{O}$ (as a Seifert fibered space) is also negative, and hence the map π_*^{-1} is an isomorphism.

Now Corollary 3.13 holds, but Lemma 3.14 needs to be adapted. Firstly remark that if $\Sigma_{\mathbb{O}}$ is a homology sphere, $\boldsymbol{x}_\gamma$ is the zero-function, so there is no possible interesting version of Lemma 3.14 in this case. Secondly, if $\Sigma_{\mathbb{O}}$ is not a homology sphere, Lemma 3.14 holds, but one argument needs to be developed, namely:

Lemma 4.3 *For $\mathbb{O}$ a Riemannian orientable 2-orbifold and γ a geodesic on $\mathbb{O}$, for every nonzero homology class a in $H_2(\mathbb{O};\mathbb{R})$, there is an $\boldsymbol{x}_\gamma$-realizing geodesic in a.*

Proof Let β be an $\boldsymbol{x}_\gamma$-realizing curve such that $[\beta]=a$. As in the case of a standard surface we want to strengthen β to make it geodesic without changing the geometric intersection with γ. Far from the conic points, one can perform isotopies that shorten β with respect to the hyperbolic metric. Since γ is geodesic, these isotopies cannot increase the number of intersection (that is, no Reidemeister II move is involved).

Around a conic point, one can work in a local conic chart. This amounts to work on a standard disc where everything in invariant under a rotation. Then one can also perform length-decreasing isotopies in an equivariant way, and this does not increase the number of intersection points with γ. □

Proposition D holds with no modification in the proof, and Theorem A has to be changed into Theorem E in order to treat the case of an orbifold whose underlying surface is a sphere.

Proof of Theorem E Suppose that $\Sigma_{\mathbb{O}}$ is a sphere. Then $T^1\Sigma_{\mathbb{O}}$ is a rational homology sphere (in this case, $H_1(T^1\Sigma_{\mathbb{O}};\mathbb{Z})$ is finite, but not reduced to the trivial group, unless $\Sigma_{\mathbb{O}}$ is a sphere with three conic points of respective orders $2, 3$, and 7). If γ is filling, then the class $\sigma_\pm$ intersects every asymptotic cycle, so it contains a Birkhoff section. Since $H_2(T^1\Sigma_{\mathbb{O}};\mathbb{Z})$ is trivial, all Birkhoff sections are homologous, and hence isotopic relatively to their boundary.

If γ is not filling, then there exists a geodesic α not intersecting γ on $\Sigma_{\mathbb{O}}$. Both its oriented lifts do not intersect $S^\times_\pm$, and hence there is an asymptotic cycle whose algebraic intersection with $\sigma_\pm$ is zero. Hence the class $\sigma_\pm$ contains no Birkhoff section. Since it is the unique class with boundary $-\vec{\gamma}$, there is no Birkhoff section bounded by $-\vec{\gamma}$ at all.

Finally if $\Sigma_{\mathbb{O}}$ is not a sphere and γ is filling, the norm $\boldsymbol{x}_\gamma$ is nondegenerate, and the proof of Theorem A translates directly. □

5 Questions

On intersection norms If Σ is a flat torus, then the minimal intersection is always realized by geodesics, which are unique in their homology class. Hence if the collection γ is the union of k geodesics $\gamma_1, \dots, \gamma_k$, then $i_\gamma(\alpha) = \sum_{i=1}^k i_{\gamma_i}(\alpha)$. This implies that the dual ball B^*_γ coincides with the Minkowski sum $B^*_{\gamma_1} + \cdots + B^*_{\gamma_k}$. Since the segment $[-1,1]\times\{0\} \subset \mathbb{R}^2$ is the dual unit ball $B^*_{\boldsymbol{x}_\gamma}$ for γ the vertical circle on the torus, every segment containing 0 in the middle is the dual unit ball of some closed circle on the torus. Therefore every convex polygon in $\mathbb{R}^2$ whose vertices are integral and congruent mod 2 is of the form $B^*_{\boldsymbol{x}_\gamma}$ for some γ. This was already remarked by Thurston [42] and by Schrijver [37]. In higher dimension the situation is probably more intricate.

Question 5.1 Which polyhedra of $\mathbb{R}^{2g}$ with integer vertices can be realized as the dual unit ball $B^*_{\boldsymbol{x}_\gamma}$ for some γ in Σ_g?

A partial answer is given by Abdoul Karim Sane [35], who proves that some polyhedra in $\mathbb{R}^4$ cannot be dual unit ball of any intersection norm on a genus 2-surface.

Also, if Σ is a torus and γ is a union of geodesics, then the above remarks imply that the number of self-intersection points of γ is exactly $\frac{1}{4}$ of the area of $B^*_{\boldsymbol{x}_\gamma}$ (check in Figure 1). Is there an analogous statement in higher genus?

Question 5.2 Which information concerning γ can be read on $B^*_{\boldsymbol{x}_\gamma}$? Is the number of self-intersection points of γ a certain function defined on $B^*_{\boldsymbol{x}_\gamma}$?

This information is interesting since this number is exactly the opposite of the Euler characteristic of every Birkhoff cross section bounded by $\vec{\gamma}$. Note that the number of self-intersection points is homogenous of degree 2, so we should look for degree 2 functions on polyhedra in $\mathbb{R}^{2g}$: does it correspond to some symplectic capacity?

Motivated by our application we only defined the intersection norm for a collection of immersed curves, but one can directly extend it for an arbitrary embedded graph. One can wonder which properties extend to this case and which information on the embedded graphs are encoded in this norm. For example when the graph is Eulerian (ie all vertices have even degree) the connection with Eulerian coorientations remains.

On Birkhoff cross sections Our constructions and our classification result deal only with Birkhoff cross sections bounded by a *symmetric* collection of periodic orbits of the geodesic flow, that is, invariant under the involution $(p, v) \mapsto (p, -v)$. However the only restriction *a priori* for being the boundary of a Birkhoff cross section is to be a boundary, that is, to be nullhomologous. Our results here say nothing about the classification, or even the existence, of Birkhoff cross sections with arbitrary nullhomologous boundary. In this case, the theory of Schwartzman, Fuller, Thurston, and Fried, and the remarks of Sections 3.3 and 3.5 still apply, so that these sections still correspond to the point inside a certain polytope in $H^1(\Sigma; \mathbb{R})$. However we have no analog for the coorientations and the explicit constructions derived from them.

Question 5.3 Is there a natural generalization of the polytope $B^*_{\boldsymbol{x}_\gamma}$ to nonsymmetric finite collections $\vec{\gamma}$ of closed orbits of the geodesic flow $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$, so that integer points in this polytope classify surfaces bounded by $\vec{\gamma}$ and transverse to $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$?

In the case of the flat torus, this question is answered in [13, Theorem 3.12] where a polygon $P_{\vec{\gamma}}$ classifying transverse surfaces bounded by $\vec{\gamma}$ is defined for *every* nullhomologous collection $\vec{\gamma}$.

What would probably unlock the situation in the higher genus case would be to have, for every nullhomologous collection $\vec{\gamma}$, *one* explicit surface bounded by $\vec{\gamma}$ (not necessarily transverse), that is, an analog of $\sigma_\pm$ when $\vec{\gamma}$ is not symmetric. Such an explicit point allows us to compute its intersection with every other periodic orbit $\vec{\alpha}$ of $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$. These intersection numbers are all we need in order to describe explicitly the asymptotic directions of $(\varphi^t_{\text{geod}})_{t\in\mathbb{R}}$ in $T^1\Sigma \setminus \vec{\gamma}$. Generalizing the constructions of [12] is a possibility here.

More generally, one can wonder whether there exists a generalization to all flows of the intersection norm $\boldsymbol{x}_\gamma$ in the following sense:

Question 5.4 For every 3-dimensional flow X, is there an object that describes all isotopy classes of Birkhoff cross sections?

A starting point would be to try with an Anosov flow that is not the geodesic flow, and see whether Gauss linking forms could play this role [24].

Appendix Thurston's theorem on integral seminorms

Our goal in this section is to state and prove Thurston's theorem [42, Theorem 2] affirming that every integral seminorm F defined on a lattice $L \simeq \mathbb{Z}^n$ is the pointwise maximum of a finite set Φ of linear functionals (ie homomorphisms $L \to \mathbb{Z}$). In addition, we strengthen the conclusion of the theorem in the case that F is equivalent modulo an integer $m \geq 1$ to a given homomorphism $\mu\colon L \to \mathbb{Z}_m$, affirming that in this case we can let Φ contain only homomorphisms that are equivalent to μ modulo m as well.

Note that other proofs of Thurston's theorem have been given [21, Theorem 5; 34]. Here, we state the theorem in a way that involves only integer numbers (rather than reals, although the version for real numbers follows as a corollary). The proof we give is similar to Thurston's original argument, but is written in greater detail and, like the statement of the theorem, relies only on the lattice and its dual, rather than extending the seminorm to a real vector space. In fact, the proof yields an effective method for obtaining, for each integral seminorm F and each vector v, a functional $\varphi \leq F$ that coincides with F at v.

To facilitate the exposition we introduce the concept of a *narrow set* with respect to an integral seminorm F, which is any finite subset $X \subseteq V$ such that F is linear on the semigroup spanned by X.

A.1 Definitions and statement of Thurston's theorem

Recall that a *lattice* L is a finitely generated free abelian group. Its *dual lattice* L^* is the group of homomorphisms $L \to \mathbb{Z}$. Note that $L \simeq L^* \simeq \mathbb{Z}^n$ for some $n \in \mathbb{N}$, called the *rank* of L. The elements of L and L^* will be called *vectors* and *functionals*, respectively. A *basis* of a lattice L is an n-tuple $X = (x_i)_{i<n} \subseteq L$ such that every element of L can be expressed by a unique integral combination of the elements x_i.

An *integral seminorm* on a lattice L is a function $F\colon L \to \mathbb{Z}$ with the following two properties:

- **Positive homogeneity** $F(\lambda v) = \lambda F(v)$ for all $v \in L$ and all scalars $\lambda \in \mathbb{N}$.
- **Subadditivity** $F(v + w) \leq F(v) + F(w)$ for all $v, w \in L$.

(Note that we allow $F(-v) \neq F(v)$, and even $F(v) < 0$.)

Note first that every finite nonempty set of functionals $\Phi \subseteq L^*$ determines a integral seminorm M_Φ on L given by

$$M_\Phi(v) = \max_{\varphi \in \Phi} \varphi(v). \tag{3}$$

Thurston's theorem asserts that in fact every integral seminorm is of this form.

Theorem A.1 (Thurston's theorem on integral seminorms) *Every integral seminorm F on a lattice L is of the form*

$$F(v) = \max_{\varphi \in B_F^*} \varphi(v), \tag{4}$$

where $B_F^* \subseteq L^*$ is the ***dual unit ball*** *of F, that is, the set of all functionals $\varphi \in L^*$ satisfying $\varphi(v) \leq F(v)$ for all $v \in L$.*

Remark A.2 The dual unit ball of any integral seminorm F is finite, since the coefficients of a functional $\varphi \in B_F^*$ with respect to basis $E = (e_i)_{0 \leq i < n}$ are bounded by $\varphi_i = \varphi(e_i) \leq F(e_i)$ and $-\varphi_i = \varphi(-e_i) \leq F(-e_i)$.

Remark A.3 For any finite set Φ of functionals on a lattice L we have

$$M_{\Phi} = M_{\mathrm{ext}(\Phi)}, \tag{5}$$

where $\mathrm{ext}(\Phi)$ denotes the set of *extremal points* of the set Φ, that is, those points $\varphi \in \Phi$ that cannot be obtained as a (rational) convex combination of the elements of $\Phi \setminus \{\varphi\}$. Equation (5) holds since every point of Φ is a convex combination of the extremal points of Φ. In particular, (4) in Thurston's theorem is equivalent to

$$F(v) = \max_{\varphi \in \mathrm{ext}(B_F^*)} \varphi(v). \tag{6}$$

Remark A.4 The more commonly formulated version of Thurston's theorem, involving real numbers, is as follows. A *seminorm* on a real vector space V is a subadditive, positively homogeneous function $F\colon V \to \mathbb{R}$, where positive homogeneity means that $F(\lambda v) = \lambda F(v)$ for all vectors $v \in \mathbb{R}^n$ and scalars $\lambda \in \mathbb{R}_{\geq 0}$. Its *dual unit ball* B_F^* is the set of (real-valued) functionals $\varphi \in V^*$ that satisfy $\varphi \leq F$ pointwise. In this setting, Thurston's theorem asserts that any real seminorm F on a vector space $V \simeq \mathbb{R}^n$ taking integer values on some rank-n lattice $L \subseteq V$ is of the form

$$F(v) = \max_{\varphi \in \Phi} \varphi(v),$$

where Φ is the set of linear functionals $\varphi \in B_F^*$ that take integer values on L. This version of Thurston's theorem follows readily from Theorem A.1.

A.2 Proof of Thurston's theorem

The proof is based on a method for verifying that a functional φ on a lattice L is in the dual unit ball of an integral seminorm F after evaluating both functions at finitely many vectors. To describe this method, we introduce the concept of *narrow sets*.

To define this concept, we first recall some additional standard terminology. Let L be a rank-n lattice. A *sublattice* of L is a subgroup of L (and is itself a lattice of rank $\leq n$ by the Smith normal form theorem), and a *semigroup* in L is any subset $S \subseteq L$ that is closed with respect to finite sums (including the empty sum). Any set $X \subseteq L$ spans a sublattice L_X and a semigroup S_X consisting, respectively, of all integral or positive (ie nonnegative) integral combinations of elements of X.

Now we can define narrow sets. Note that, in essence, what we are trying to show in Thurston's theorem is that every integral seminorm is a piecewise-linear function.

Definition A.5 A subset X of a lattice L is *narrow* with respect to a seminorm F defined on L, or F-narrow, if on the semigroup S_X spanned by X the function F is linear, that is, it coincides with some functional $\varphi \in L_X^*$.

Note that there is at most one functional $\varphi \in L_X^*$ that coincides with F on X, and in fact, exactly one if X is linearly independent. To determine whether F coincides with φ on the whole semigroup S_X we have the following criterion.

Proposition A.6 (interior ray test) *Consider a seminorm F on a lattice L, a vector tuple $X = (x_i)_{i \in k} \subseteq L$, and a functional $\varphi \in L_X^*$ that is greater than or equal to F at the vectors x_i (and thus, at all vectors of the semigroup S_X). Take an integer combination $c^+ = \sum_i \alpha_i x_i$ with strictly positive coefficients $\alpha_i > 0$, so that*

$$F(c^+) = F\Big(\sum_i \alpha_i x_i\Big) \le \sum_i \alpha_i F(x_i) = \varphi(c^+). \tag{7}$$

Then the following are equivalent:

(a) $F(c^+) \ge \varphi(c^+)$,

(b) $F \ge \varphi$ *on the lattice* L_X,

(c) $F = \varphi$ *on the semigroup* S_X (*and hence* X *is* F*-narrow*).

The name of this result stems from the fact that the ray spanned by c^+ lies in the interior of the cone of positive combinations of the vectors x_i (in the rational vector space $L_X \otimes_{\mathbb{Z}} \mathbb{Q}$). Proposition A.6 ensures that the function F coincides with the linear function φ on the whole cone if it coincides along this single ray.

Proof We need just show that (a) implies (b), since the other forward implications are evident. Suppose, then, that (a) holds, and take any vector $v \in L_X$. We have to show that $F(v) \ge \varphi(v)$, and we do so as follows.

Recall the picture of the interior ray described above. Since the ray spanned by c^+ is in the interior of the cone spanned by X, it follows that there exists a vector $c \in S_X$ such that the ray spanned by c^+ lies between those spanned by c and by v. More precisely, the sum $v + c$ is a positive multiple of c^+.

Claim A.7 *There exists a vector $c \in S_X$ and a number $\lambda \in \mathbb{N}$ such that $v + c = \lambda c^+$.*

Proof of claim Recall that $c^+ = \sum_i \alpha_i x_i$ for some strictly positive integers $\alpha_i > 0$, and since $v \in L_X$, we can also write $v = \sum_i \beta_i x_i$ using integers β_i. Take a number $\lambda \in \mathbb{N}$ such that $\lambda \alpha_i \ge \beta_i$ for all i. Then we have $\lambda c^+ = v + c$ where $c = \sum_i (\lambda \alpha_i - \beta_i) x_i$ is a vector of S_X since $\lambda \alpha_i - \beta_i \ge 0$ for all i. □

Since F is subadditive and φ is additive, from the equation $\lambda c^+ = c + v$ we infer that if the inequality $F \le \varphi$ holds at c (which it does since $c \in S_X$) and also holds strictly at v (let us assume this, for a contradiction), then it also holds strictly at the vector λc^+. However, we know that $F(\lambda c^+) \ge \varphi(\lambda c^+)$ by the hypothesis (a). Therefore the inequality $F \le \varphi$ cannot hold strictly at v, which means that $F(v) \ge \varphi(v)$, as we had to show. □

Now we are ready to prove Thurston's theorem. Let F be an integral seminorm on a lattice L, and take a vector $v \in L$. We have to show that there exists a functional $\varphi \in B_F^*$ such that $\varphi(v) = F(v)$. And for this, we may assume that v is a *primitive element* of L (that is, that it cannot be written as a multiple $v = \lambda w$ of an element $w \in L$ by an integer $\lambda > 1$), since every element of a lattice is a positive multiple of some primitive element (again, by the Smith normal form theorem).

To prove that $F(v) = \varphi(v)$ for some $\varphi \in B_F^*$, it suffices to show that the vector v is contained in the semigroup S_X generated by some narrow basis X of L, because this means that F coincides with some functional $\varphi \in L^*$ on S_X, and this functional is in the dual unit ball B_F^* by Proposition A.6. Therefore, to finish the proof of Thurston's theorem, it is enough to establish the following result.

Proposition A.8 *If F is an integral seminorm on a lattice L then every primitive vector $v \in L$ is the first element of some F-narrow basis.*

The proof is constructive: if the integral seminorm F is given as an oracle (or "black box") that outputs the value $F(v)$ for any given input vector $v \in L$, we will show how to obtain, after invoking this oracle finitely many times, an F-narrow basis X containing v and the corresponding functional $\varphi \in L^*$ that coincides with F on S_x, and therefore satisfies $\varphi(v) = F(v)$, and is in the dual unit ball B_F^*.

Proof of Theorem A.1 Let $X = (x_i)_{0 \le i < n}$ be a basis of the lattice L such that $x_0 = v$. (Every primitive integral vector is part of a basis, which can be obtained by putting in Smith normal form the one-column matrix of coordinates of v with respect to an initial arbitrary basis of L.)

In general the basis X is not narrow, but we can modify it to make it narrow as follows. We proceed by induction on the dimension. Suppose that for some $k < n$, the k-tuple $X_k = (x_i)_{0 \le i < k}$ is known to be narrow. We may test whether X_{k+1} is narrow by evaluating F on the vector $x_k' := x_k + w$, where $w = \sum_{0 \le i < k} x_i$. Note that

$$F(x_k') \le F(x_k) + F(w). \tag{8}$$

Claim A.9 (increment test) *X_{k+1} is narrow if and only if equality holds in* (8).

Proof of claim Let φ be the unique functional on the lattice $L_{X_{k+1}}$ that coincides with F on X_{k+1}. The vector x_k' is the sum of all the vectors of X_{k+1}, thus, by Proposition A.6, X_{k+1} is narrow if and only if the inequality $F(x_k') \le \varphi(x_k')$ is an equality. However, this inequality is equivalent to (8) since $\varphi(x_k') = \varphi(x_k) + \varphi(w) = F(x_k) + F(w)$. (Here we used the equation $\varphi(w) = F(w)$, which holds because w is a combination of the tuple X_k, that is assumed narrow.) □

If the increment test is not passed, we replace the vector x_k in X by the vector x_k', obtaining a new basis X', and we redo the test. (Note that this replacement is an elementary operation on X, therefore X' is another basis of L.) Since we may need to repeat this replacement many times, we define $x_k^{(t)} = x_k + tw$ for $t \in \mathbb{N}$, and we denote by $X^{(t)}$ the basis obtained from X by replacing x_k with $x_k^{(t)}$.

Claim A.10 *For a large enough $t \in \mathbb{N}$, the tuple $X_{k+1}^{(t)}$ is narrow.*

Proof of claim By the increment test described in Claim A.9 above, it suffices to show that the inequality

$$F(x_k^{(t+1)}) \leq F(x_k^{(t)}) + F(w)$$

is an equality for large enough t. To prove this we consider the function

$$f(t) = F(x_k^{(t)}) = F(x_k + tw).$$

Its discrete derivative $f'(t) := f(t+1) - f(t)$ is integer-valued, increasing (since f is convex), and bounded above by $F(w)$. Therefore f' eventually stabilizes at a constant value. In fact, it stabilizes at the value $F(w)$. We see this by comparing f with the known function $g(t) = F(tw) = tF(w)$, whose difference with f is bounded by the inequality

$$g(t) = F(x_k + tw - x_k) \leq F(x_k + tw) + F(-x_k) = f(t) + F(-x_k). \qquad \square$$

By this process we find a narrow basis X containing v as its first element. By Proposition A.6, it follows that the unique functional $\varphi \in L^*$ that coincides with F on X is in the dual unit ball B_F^* and satisfies $\varphi(v) = F(v)$, as we had to show. $\square$

A.3 Thurston's theorem for seminorms of a given class modulo *m*

For an integer $m \geq 1$, denote by $\mathbb{Z}_m$ the group of integers modulo m, and let $\pi_m \colon \mathbb{Z} \to \mathbb{Z}_m$ be the quotient map. An integer-valued function F on a lattice L is *congruent* to a group homomorphism $\mu \colon L \to \mathbb{Z}_m$ if the function $F_{\operatorname{mod} m} := \pi_m \circ F$ is equal to μ.

Our goal now is to prove the following extension of Thurston's theorem.

Theorem A.11 *Every integral seminorm F on a lattice L that is congruent modulo a certain integer $m \geq 1$ to a given homomorphism $\mu \colon L \to \mathbb{Z}_m$ is of the form*

$$F(v) = \max_{\substack{\varphi \in B_F^* \\ \varphi_{\operatorname{mod} m} = \mu}} \varphi(v).$$

To prove this result we use the following lemma.

Lemma A.12 *Let F be an integral seminorm on a lattice L, and let φ be an extremal functional of the dual unit ball B_F^*. Then there exists a basis X of L such that F coincides with φ on the semigroup S_X.*

Proof Let $(\psi_i)_i$ be the functionals of the dual unit ball B_F^* excluding φ.

We claim first that there exists some primitive vector $v \in L$ such that $\psi_i(v) < \varphi(v)$ for all i. Indeed, extremality of the functional φ implies that it cannot be written as a rational convex combination of the functionals ψ_i. By the Farkas lemma, it follows that there is a vector $v \in L$ (which can be taken primitive) such that $\varphi(v) > \psi_i(v)$ for all i.

Fixed the vector $v \in L$, we apply Proposition A.8, which ensures that the lattice L admits an F-narrow basis X containing the vector v. Since X is F-narrow, the function F coincides with some functional $\widetilde{\varphi} \in L^*$ on the semigroup S_X (and in particular, at the vector v). Moreover, this functional $\widetilde{\varphi}$ is in the dual unit ball B_F^* by Proposition A.6. We conclude that $\widetilde{\varphi} = \varphi$ because φ is strictly greater than all the other functionals $\psi_i \in B_F^*$ at the vector v. □

To finish, let us prove Theorem A.11.

Proof of Theorem A.11 By Remark A.3, it suffices to show that every extremal point of the dual unit ball B_F^* is congruent to μ modulo m. Let φ be an extremal functional of B_F^*. By Lemma A.12, there exists a basis X of L such that $F = \varphi$ on S_X. Suppose, for a contradiction, that $\varphi_{\operatorname{mod} m} \neq \mu$. This means that there is a vector $v \in L$ such that $\varphi_{\operatorname{mod} m}(v) \neq \mu(v)$. Thus for each vector v' of the set $v + mL$ we have

$$\varphi(v') \equiv \varphi(v) \not\equiv \mu(v) \equiv \mu(v') \mod m$$

since both μ and $\varphi_{\operatorname{mod} m}$ vanish on the lattice mL. Take a vector $v' \in (v + mL)$ whose coordinates with respect to the basis X are positive, so that $v' \in S_X$, and hence we have

$$F(v') = \varphi(v') \not\equiv \mu(v') \mod m,$$

contradicting the hypothesis $F_{\operatorname{mod} m} = \mu$. □

References

[1] **N A'Campo**, *Le groupe de monodromie du déploiement des singularités isolées de courbes planes, I*, Math. Ann. 213 (1975) 1–32 MR

[2] **N A'Campo**, *Generic immersions of curves, knots, monodromy and Gordian number*, Inst. Hautes Études Sci. Publ. Math. 88 (1998) 151–169 MR

[3] **I Agol**, **N M Dunfield**, *Certifying the Thurston norm via* $\mathrm{SL}(2,\mathbb{C})$*-twisted homology*, from "What's next?—the mathematical legacy of William P Thurston" (D P Thurston, W P Thurston, editors), Ann. of Math. Stud. 205, Princeton Univ. Press (2020) 1–20 MR

[4] **D V Anosov**, *Geodesic flows on closed Riemannian manifolds of negative curvature*, Trudy Mat. Inst. Steklov. 90 (1967) 209 MR In Russian; translated in Proc. Steklov. Inst. Math. 90 (1969) 1–235

[5] **G D Birkhoff**, *Proof of Poincaré's geometric theorem*, Trans. Amer. Math. Soc. 14 (1913) 14–22 MR

[6] **G D Birkhoff**, *Dynamical systems with two degrees of freedom*, Proc. Natl. Acad. Sci. USA 3 (1917) 314–316

[7] **M Brunella**, *Expansive flows on three-manifolds*, PhD thesis, SISSA, Trieste (1992) Available at `https://iris.sissa.it/bitstream/20.500.11767/3929/1/1963_2672_Tesi_Brunella.pdf`

[8] **M Brunella**, *On the discrete Godbillon–Vey invariant and Dehn surgery on geodesic flows*, Ann. Fac. Sci. Toulouse Math. 3 (1994) 335–344 MR

[9] **V Colin**, **P Dehornoy**, **U Hryniewicz**, **A Rechtman**, *Generic properties of 3-dimensional Reeb flows: Birkhoff sections and entropy*, Comment. Math. Helv. 99 (2024) 557–611 MR

[10] **G Contreras**, **M Mazzucchelli**, *Existence of Birkhoff sections for Kupka–Smale Reeb flows of closed contact 3-manifolds*, Geom. Funct. Anal. 32 (2022) 951–979 MR

[11] **M Cossarini**, *Discrete surfaces with length and area and minimal fillings of the circle*, PhD thesis, Instituto Nacional de Matemática Pura e Aplicada (2018) Available at `https://w3.impa.br/~mbel/tese/MarcosCossarini-Tese-2018.pdf`

[12] **P Dehornoy**, *Genus-one Birkhoff sections for geodesic flows*, Ergodic Theory Dynam. Systems 35 (2015) 1795–1813 MR

[13] **P Dehornoy**, *Geodesic flow, left-handedness and templates*, Algebr. Geom. Topol. 15 (2015) 1525–1597 MR

[14] **P Dehornoy**, *Which geodesic flows are left-handed?*, Groups Geom. Dyn. 11 (2017) 1347–1376 MR

[15] **P Dehornoy**, **A Rechtman**, *Vector fields and genus in dimension 3*, Int. Math. Res. Not. 2022 (2022) 3262–3277 MR

[16] **P Dehornoy**, **M Shannon**, *Almost equivalence of algebraic Anosov flows*, preprint (2019) arXiv 1910.08457

[17] **W Duke**, **O Imamoḡlu**, **A Tóth**, *Modular cocycles and linking numbers*, Duke Math. J. 166 (2017) 1179–1210 MR

[18] **J Franks**, *Generalizations of the Poincaré–Birkhoff theorem*, Ann. of Math. 128 (1988) 139–151 MR

[19] **D Fried**, *The geometry of cross sections to flows*, Topology 21 (1982) 353–371 MR

[20] **D Fried**, *Transitive Anosov flows and pseudo-Anosov maps*, Topology 22 (1983) 299–303 MR

[21] **D Fried**, *Fibrations of S^1 with pseudo-Anosov monodromy*, from "Thurston's work on surfaces" (A Fathi, F Laudenbach, P V, editors), Princeton Univ. Press (2009) 215–224

[22] **S Friedl**, **S Vidussi**, *The Thurston norm and twisted Alexander polynomials*, J. Reine Angew. Math. 707 (2015) 87–102 MR

[23] **F B Fuller**, *On the surface of section and periodic trajectories*, Amer. J. Math. 87 (1965) 473–480 MR

[24] **E Ghys**, *Right-handed vector fields & the Lorenz attractor*, Jpn. J. Math. 4 (2009) 47–61 MR

[25] **M Gromov**, *Three remarks on geodesic dynamics and fundamental group*, Enseign. Math. 46 (2000) 391–402 MR

[26] **S M Gusein-Zade**, *Dynkin diagrams of the singularities of functions of two variables*, Funkcional. Anal. i Priložen. 8 (1974) 23–30 MR In Russian; translated in Funct. Anal. Appl. 8 (1974) 295–300

[27] **S M Gusein-Zade**, *Monodromy groups of isolated singularities of hypersurfaces*, Uspehi Mat. Nauk 32 (1977) 23–65 MR In Russian; translated in Russian Math. Surveys 32 (1977) 23–69

[28] **A Hatcher**, *Algebraic topology*, Cambridge Univ. Press (2002) MR

[29] **U L Hryniewicz**, *A note on Schwartzman–Fried–Sullivan theory, with an application*, J. Fixed Point Theory Appl. 22 (2020) art. id. 25 MR

[30] **U Hryniewicz**, **P A S Salomão**, *On the existence of disk-like global sections for Reeb flows on the tight 3-sphere*, Duke Math. J. 160 (2011) 415–465 MR

[31] **M Ishikawa**, *Tangent circle bundles admit positive open book decompositions along arbitrary links*, Topology 43 (2004) 215–232 MR

[32] **H Jacques**, *Les surfaces à courbures opposées et leurs lignes géodésiques*, J. Math. Pures Appl. 4 (1898) 27–74

[33] **T Marty**, *Anosov flows and Birkhoff sections*, PhD thesis, Université Grenoble Alpes (2021) Available at `https://theses.hal.science/tel-03510071`

[34] **M de la Salle**, *On norms taking integer values on the integer lattice*, C. R. Math. Acad. Sci. Paris 354 (2016) 611–613 MR

[35] **A K Sane**, *Intersection norms and one-faced collections*, C. R. Math. Acad. Sci. Paris 358 (2020) 941–956 MR

[36] **A Schrijver**, *Circuits in graphs embedded on the torus*, Discrete Math. 106/107 (1992) 415–433 MR

[37] **A Schrijver**, *Graphs on the torus and geometry of numbers*, J. Combin. Theory Ser. B 58 (1993) 147–158 MR

[38] **S Schwartzman**, *Asymptotic cycles*, Ann. of Math. 66 (1957) 270–284 MR

[39] **M Shannon**, *Dehn surgeries and smooth structures on* 3*-dimensional transitive Anosov flows*, PhD thesis, Université de Bourgogne, Dijon (2020) Available at `https://www.theses.fr/2020UBFCK035`

[40] **D Sullivan**, *Cycles for the dynamical study of foliated manifolds and complex manifolds*, Invent. Math. 36 (1976) 225–255 MR

[41] **W T Thurston**, *The geometry and topology of three-manifolds* (1979)

[42] **W P Thurston**, *A norm for the homology of* 3*-manifolds*, Mem. Amer. Math. Soc. 339, Amer. Math. Soc., Providence, RI (1986) 99–130 MR

[43] **V Turaev**, *A norm for the cohomology of* 2*-complexes*, Algebr. Geom. Topol. 2 (2002) 137–155 MR

Laboratoire d'analyse et de mathématiques appliquées, Université Paris-Est Créteil
Créteil, France

Institut de Mathématiques de Marseille, Aix-Marseille Université
Marseille, France

`marcos.cossarini@u-pec.fr, pierre.dehornoy@univ-amu.fr`

`https://www.i2m.univ-amu.fr/perso/pierre.dehornoy`

Received: 21 September 2020 Revised: 19 September 2024

Algebraic & Geometric Topology **25**:8 (2025) 4547–4583
DOI: 10.2140/agt.2025.25.4547
Published: 20 November 2025

Thin knots and the cabling conjecture

ROBERT DEYESO III

The cabling conjecture of González-Acuña and Short states that only cable knots admit Dehn surgery to a manifold containing an essential sphere. We approach this conjecture for thin knots using Heegaard Floer homology, primarily via immersed curves techniques inspired by Hanselman's work on the cosmetic surgery conjecture. We show that almost all thin knots satisfy the cabling conjecture, with a possible exception coming from a (conjecturally nonexistent) collection of thin, hyperbolic, L-space knots. This result serves as a reproof that the cabling conjecture is satisfied by alternating knots.

57K18

1 Introduction

For a knot K in S^3, let $S^3_r(K)$ denote r-sloped Dehn surgery along K. If $S^3_r(K)$ is a reducible manifold, meaning it contains an essential 2-sphere, we will call r a reducing slope. The primary example of a reducible surgery to keep in mind is when K is the (p,q)-cable of some knot K' and r is given by the cabling annulus. In this case, we have $S^3_{pq}(K) \cong L(p,q) \# S^3_{q/p}(K')$. The cabling conjecture asserts that this is the only example of a reducible surgery.

Conjecture 1.1 (cabling conjecture, Gonzalez-Acuña–Short [8]) *If K is a knot in S^3 which has a reducible surgery, then K is a cabled knot and the reducing slope is given by the cabling annulus.*

The cabling conjecture is satisfied by many classes of knots. Torus knots, as cables of the unknot, were shown to satisfy the conjecture in [25], and satellite knots [39] and alternating knots [24] satisfy the conjecture as well. Additionally, genus-1 knots [3], strongly invertible knots [6], symmetric knots [18], and knots with low bridge number [12] satisfy the conjecture (for a survey of known results and techniques see [2].) Since the conjecture is satisfied by torus and satellite knots, it remains to consider hyperbolic knots. Our aim is narrower than this however, as we will look at hyperbolic knots that are considered "thin" due to the simpler structure of their knot Floer complexes.

We will present knot Floer homology in more detail in Section 2, but for now recall that $\widehat{\mathrm{HFK}}(K)$ with coefficients in $\mathbb{F}_2$ is a bigraded vector space with Alexander and Maslov gradings, respectively denoted by A and M. A knot K is *Floer homologically thin* if the generators of $\widehat{\mathrm{HFK}}(K)$ all have the same

---

$\delta = A - M$ grading. This family contains alternating knots [28], and the more generalized quasialternating knots [34]. We say K is an *L-space knot* if it admits a surgery to a (Heegaard Floer) L-space, which is a manifold with the simplest Heegaard Floer homology. Using Heegaard Floer homology via immersed curves techniques, we show that:

Theorem 1.2 *If a thin, hyperbolic knot K in S^3 admits a reducible surgery, then K is an L-space knot and the reducing slope must be $r = 2g(K) - 1$ after mirroring K if necessary.*

While stated for thin, hyperbolic knots, this theorem holds more generally for noncabled knots. This is because we use the Matignon–Sayari genus bound, stated below, allowing us to consider only $r \le 2g(K) - 1$. The case where $r > 2g(K) - 1$ can be handled using the techniques in this paper to conclude that $K = T(2, n)$, but perhaps more immediate is the result of Dey that cables of nontrivial knots are not thin [5]. Since the only alternating, L-space knots are the $T(2, n)$ torus knots [33], Theorem 1.2 provides an immersed curves reproof that alternating knots satisfy the cabling conjecture.

Corollary 1.3 *Alternating knots satisfy the cabling conjecture.*

It is conjectured that the only thin, L-space knots are the torus knots $T(2, n)$. Provided this is true, there would not exist thin, hyperbolic, L-space knots and so Theorem 1.2 would show that all thin knots satisfy the cabling conjecture. Regardless, Bodish and the author have since generalized the absolute grading computations in Section 5.1 to circumvent this condition to show that:

Theorem 1.4 [1] *Thin knots satisfy the cabling conjecture.*

Part of the proof strategy for Theorem 1.2 involves obstructing an $\mathbb{R}P^3$ connected summand, and so we get the following corollary with identical proof to that of [20, Corollary 1.5].

Corollary 1.5 *If K is a thin, hyperbolic knot, then $S^3 \setminus \nu K$ does not contain properly embedded punctured projective planes.*

When K is a nontrivial knot in S^3 with reducible surgery $S^3_r(K)$, the surgery decomposes as a connected sum and the reducing slope satisfies $r \neq 0$ due to [7]. We saw from the cabled knot example that the reducing slope is an integer and one of the connected summands is a lens space. The former and latter conditions occur for all reducible surgeries due to [9; 10], respectively. A reducible surgery can admit at most three connected summands due to the combined efforts of [21; 38; 40], in which case two summands are lens spaces and the remaining summand is an integer homology sphere. Since $S^3_r(K)$ must have a nontrivial lens space summand, the integral reducing slope r satisfies $r \neq -1, 0, 1$. In [23], Matignon and Sayari provide the following genus bound if K is noncabled:

$$1 < |r| \le 2g(K) - 1.$$

Heegaard Floer homology satisfies a Künneth formula for connected sums, and has proved very useful in general for studying Dehn surgery. If surgery along K produces a connected sum of precisely two lens spaces, then K must be a cabled knot due to [11]. Further, [11] together with [4] shows that a hyperbolic knot in S^3 cannot admit both a lens space surgery and a reducible surgery. Hom, Lidman, and Zufelt showed that a hyperbolic, L-space knot can admit at most one reducing slope, and the slope must be $2g(K)-1$ after mirroring the knot to make the slope positive [20]. They also established a periodicity structure to the Heegaard Floer homology of a reducible surgery, which is invaluable to the proof strategy of Theorem 1.2. We will involve these constraints via bordered invariants in the form of immersed curves.

Lipshitz, Ozsváth, and Thurston introduced bordered Heegaard Floer invariants for manifolds with boundary in [22]. With $M_1 \cup_h M_2$ denoting a gluing of two such manifolds, they prove a pairing theorem involving the two bordered invariants that recovers the Heegaard Floer homology of the glued-together manifold (see Section 2.2 for more details). In the torus boundary case, Hanselman, Rasmussen, and Watson reinterpreted these bordered invariants as collections of immersed curves in the punctured torus, and proved an analogous pairing theorem. That is, they show that the hat flavor of Heegaard Floer homology of $M_1 \cup_h M_2$ is the Lagrangian intersection Floer homology of the immersed curves invariants for M_1 and M_2. In [13], Hanselman used this package to obtain obstructions for cosmetic surgeries along knots in S^3, and our approach in this paper is largely inspired by this work.

Organization We only consider surgeries with positive slopes, and mirror knots to achieve this whenever necessary. All manifolds are assumed to be compact, connected, oriented 3-manifolds, and the coefficients in Floer homology are taken to belong to $\mathbb{F} = \mathbb{F}_2$. We will denote closed manifolds by X or Y, and manifolds with (typically torus) boundary by M. Figures containing immersed curves invariants will have the curves for $S^3 \setminus \nu K$ in red and the curves for the filling solid torus in blue or purple.

Section 2 summarizes the relevant background from knot Floer homology and Heegaard Floer homology. It also contains an overview of immersed curves invariants, their general properties and form for thin knots, as well as their associated pairing theorem and how to compute Maslov grading differences.

Section 3 expands on the relative Maslov grading for immersed curves invariants of complements of thin knots. Along the way we set up formulas for components of the grading difference formula in terms of $\tau(K)$.

Section 4 uses these relations to generate obstructions to periodicity for various cases of r in relation to $\tau(K)$ and $g(K)$. It hosts a sizable collection of lemmas for the cases with $|\tau(K)| < g(K)$, for which referencing Figure 12 is highly advised.

Section 5 resolves the remaining cases where $|\tau(K)| = g(K)$, including some that use absolute grading information. Once again, Figure 14 may be useful for following the arguments. Afterward, all lemmas are collected to handle the proof of the theorem.

Acknowledgements I am grateful to Tye Lidman for his unending encouragement, patience, and insight as an advisor. I would also like to thank Steven Sivek for pointing out an oversight in the regions used to compute the H's and V's of knots with $\tau(K) < 0$. I was partially supported by NSF grant DMS-1709702 while at NC State and NSF RTG grant DMS-2038103 while at Iowa.

2 Background material

We will assume the reader is familiar with the $\widehat{\mathrm{HF}}$ and HF^+ constructions of Heegaard Floer homology for 3-manifolds [32], and knot Floer homology $\widehat{\mathrm{HFK}}$ for knots in S^3 (with associated full knot Floer complex CFK^∞) [30; 37].

2.1 $\widehat{\mathrm{HF}}$ for reducible surgeries

Let us identify $\mathrm{Spin}^c(S^3_r(K))$ with $\mathbb{Z}/r\mathbb{Z}$ as in [35, Subsection 2.4], and denote the correspondence using $[s] \in \mathrm{Spin}^c(S^3_r(K))$ for $[s] \in \mathbb{Z}/r\mathbb{Z}$. We will also choose equivalence classes for elements of $\mathbb{Z}/r\mathbb{Z}$ as centered about 0, so that, for example, $\mathbb{Z}/r\mathbb{Z} = \{-\frac{r-1}{2}, \dots, 0, \dots, \frac{r-1}{2}\}$ if r is odd. As an abuse of notation, we will commonly use s for the representative of $[s]$ that falls within this range.

The following lemma is a simplified version of a more general Floer homology periodicity result for HF^+ of a general reducible 3-manifold from [20]. Basically, we should expect to see repeated behavior among the spin^c summands of $\widehat{\mathrm{HF}}(S^3_r(K))$ if the surgery is reducible.

Lemma 2.1 (periodicity) *Suppose $S^3_r(K) \cong X \# Y$, where X is an L-space and $|H^2(Y)| = k < \infty$. Then for any $[s] \in \mathrm{Spin}^c(S^3_r(K))$ and $\alpha \in H^2(S^3_r(K)) \cong \mathbb{Z}/r\mathbb{Z}$, we have*

$$\widehat{\mathrm{HF}}(S^3_r(K), [s + k\alpha]) \cong \widehat{\mathrm{HF}}(S^3_r(K), [s])$$

as relatively graded $\mathbb{F}$ vector spaces.

Proof Let $[s] \in \mathrm{Spin}^c(S^3_r(K))$ restrict to $[s_i] \in \mathrm{Spin}^c(X)$ and $[s_j] \in \mathrm{Spin}^c(Y)$. We see that $\widehat{\mathrm{HF}}(X, [s_i]) \cong \mathbb{F}$ since X is an L-space, and so the Künneth formula for $\widehat{\mathrm{HF}}$ [31, Theorem 1.5] implies

$$\begin{aligned}\widehat{\mathrm{HF}}(S^3_r(K), [s]) &\cong H_*(\widehat{\mathrm{CF}}(X, [s_i]) \otimes_{\mathbb{F}} \widehat{\mathrm{CF}}(Y, [s_j])) \\ &\cong \widehat{\mathrm{HF}}(Y, [s_j]).\end{aligned}$$

For any $\alpha \in \mathbb{Z}/r\mathbb{Z}$, we have that $[s + k\alpha]$ restricts to $[s_j]$ in $\mathrm{Spin}^c(Y)$. Then because $\widehat{\mathrm{HF}}(S^3_r(K), [s])$ is independent of $[s_i]$, we obtain

$$\widehat{\mathrm{HF}}(S^3_r(K), [s + k\alpha]) \cong \widehat{\mathrm{HF}}(Y, [s_j]) \cong \widehat{\mathrm{HF}}(S^3_r(K), [s])$$

as relatively graded $\mathbb{F}$ vector spaces. □

We also need to gather some integral invariants of K involved with the mapping cone formula that relates $\mathrm{CFK}^\infty(K)$ to $\mathrm{HF}^+(S^3_r(K))$ [35]. For $s \in \mathbb{Z}$, recall the subcomplexes and quotient complexes of the $\mathbb{Z} \oplus \mathbb{Z}$-filtered full knot Floer complex $\mathrm{CFK}^\infty(K)$

$$\mathcal{A}^+_s = C\{\max\{i, j - s\} \geq 0\} \quad \text{and} \quad \mathcal{B}^+_s = C\{i \geq 0\}.$$

Notice $\mathcal{B}_s^+ \cong \mathrm{CF}^+(S^3)$ by definition. There are also chain maps $\mathfrak{v}_s^+ : \mathcal{A}_s^+ \to \mathcal{B}_s^+$ and $\mathfrak{h}_s^+ : \mathcal{A}_s^+ \to \mathcal{B}_{s+r}^+$ between these subcomplexes. Take homology to obtain $A_s^+ = H_*(\mathcal{A}_s^+)$ and $B_s^+ = H_*(\mathcal{B}_s^+) \cong \mathrm{HF}^+(S^3)$, and induced maps v_s^+ and h_s^+. Let $\mathcal{T}^+$ denote $\mathrm{HF}^+(S^3)$, and notice that $U^N(A_s^+) \cong \mathcal{T}^+$ for sufficiently large N. By restricting both v_s^+ and h_s^+ to this submodule, we obtain $\overline{v}_s^+$ and $\overline{h}_s^+$. The integral invariants of K that we desire are due to [26], and are defined by

$$V_s = \mathrm{rank}(\ker \overline{v}_s^+) \quad \text{and} \quad H_s = \mathrm{rank}(\ker \overline{h}_s^+).$$

These terms have simple behavior when K is alternating because of the "staircase" part of $\mathrm{CFK}^\infty(K)$ due to [28]. This holds more generally for thin knots due to [36], but we will have an alternative geometric way of computing these terms later in Section 2.2. By [20, Lemma 2.3], the maps v_s^+ and h_{-s}^+ agree on homology after identifying $A_s^+ \cong A_{-s}^+$ (essentially reversing the roles of i and j above) so that $V_s = H_{-s}$. These integer invariants are by definition nonnegative, and also satisfy the following lemma.

Lemma 2.2 [26, Lemma 2.4] *The V_s form a nonincreasing sequence and the H_s form a nondecreasing sequence, so that*

$$V_s \geq V_{s+1} \quad \text{and} \quad H_s \leq H_{s+1} \qquad \text{for all } s \in \mathbb{Z}.$$

For a rational homology sphere Y, we can write

$$\mathrm{HF}^+(Y, \mathfrak{s}) \cong \mathcal{T}^+ \oplus \mathrm{HF}_{\mathrm{red}}(Y, \mathfrak{s}),$$

where $\mathcal{T}^+ \cong \mathbb{F}[U, U^{-1}]/\mathbb{F}[U]$ denotes the "tower" submodule. The *d-invariants* $d(Y, \mathfrak{s})$, sometimes called the *Heegaard Floer correction terms*, record the smallest absolutely graded element of $\mathcal{T}^+ \subseteq \mathrm{HF}(Y, \mathfrak{s})$ [27]. These invariants satisfy a few symmetries, such as spin^c conjugation symmetry $d(Y, \mathfrak{s}) = d(Y, \overline{\mathfrak{s}})$ and orientation reversal $d(-Y, \mathfrak{s}) = -d(Y, \mathfrak{s})$, as well as additivity for connected sums. It is normalized so that $d(S^3, \mathfrak{s}_0) = 0$, and is recursively determined for lens spaces in [27, Proposition 4.8]. In [26], the d-invariants of rational surgeries are shown to be determined by the invariants V_s and H_s together with the d-invariants of a lens space that depends on homological data. We state a special case of the more general result for our purposes.

Proposition 2.3 [26, Proposition 1.6] *Suppose r is integral and positive, and fix $0 \leq s < r - 1$. Then*

$$d(S_r^3(K), [s]) = d(L(r, 1), [s]) - 2\max\{V_s, V_{r-s}\}.$$

Among many of its applications, this result enables the following lemma.

Lemma 2.4 [20, Lemma 2.5] *For all $s \in \mathbb{Z}$, the integers V_s and H_s are related by*

$$H_s - V_s = s.$$

We will involve the d-invariants later in Section 5.1 when necessary.

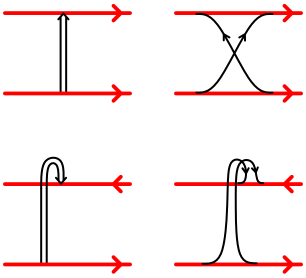


Figure 1: Edges of a grading arrow either follow or oppose the orientations of the attached curve components.

2.2 $\widehat{\mathrm{HF}}$ via immersed curves

Bordered Heegaard Floer homology, introduced by Lipshitz, Ozsváth, and Thurston, provides a relative version of the hat flavor of Heegaard Floer homology for a compact manifold M with boundary. As our only manifolds with boundary in this paper have torus boundary, some of the subtleties of the general bordered theory will be glossed over — please refer to [22] for further details. A *bordered* manifold (M, ϕ) is a compact manifold M with boundary and an orientation-preserving diffeomorphism $\phi\colon \mathbb{T}^2 \to \partial M$. They associate an algebra $\mathcal{A}$ to $\mathbb{T}^2$, and define two bordered invariants related to (M, ϕ): a type-D structure $\widehat{\mathrm{CFD}}(M, \phi)$ that is a left differential module over $\mathcal{A}$, and a type-A structure $\widehat{\mathrm{CFA}}(M, \phi)$ that is a right $\mathcal{A}_\infty$ module.

These two bordered invariants may be "paired" together via the box tensor product, a computable model for the $\mathcal{A}_\infty$ tensor product, providing a cut-and-paste style of recovering $\widehat{\mathrm{HF}}$ for a 3-manifold Y by decomposing Y along a surface. Dubbed the pairing theorem, we will invoke it on bordered invariants in immersed curves form (see Theorem 2.8) due to Hanselman, Rasmussen, and Watson [15; 16]. For a bordered manifold (M, ϕ) with torus boundary, we will specify ϕ by choosing a parameterization (α, β) of ∂M, and also fix a basepoint $z \in \partial M$. They recast the type-D structure $\widehat{\mathrm{CFD}}(M, \alpha, \beta)$ as $\widehat{\mathrm{HF}}(M)$ — a collection of immersed curves in $T_M = \partial M \setminus z$, possibly decorated with local systems, defined up to regular homotopy of the curves. When $M = S^3 \setminus \nu K$, we will often take ϕ described by the Seifert-framed meridian-longitude basis $\{\mu, \lambda\}$.

Remark We caution the reader regarding the similarity of the notation $\widehat{\mathrm{HF}}(Y)$ and $\widehat{\mathrm{HF}}(M)$, with Y a closed manifold and M a compact manifold with boundary. The former invariant is a graded vector space over $\mathbb{F}$, whereas the latter is a (possibly decorated) immersed curve in $\partial M \setminus \{z\}$.

The manifolds in this paper all happen to be *loop type* (see [17]), which means that their associated immersed curves invariant has trivial local systems. If the invariant has multiple curve components, then they are connected by pairs of edges which we denote with a grading arrow as in [15, Definition 28]. These are presented in Figure 1, and while domains involving grading arrows do not contribute to the differential, they are considered when determining Maslov grading differences. When $M = S^3 \setminus \nu K$,

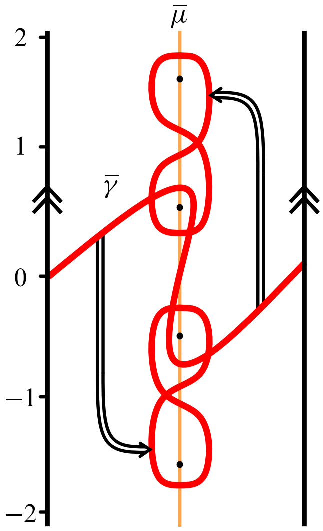


Figure 2: An example of $\widehat{\mathrm{HF}}(M)$ for a hypothetical thin knot K in $\overline{T}_M$. Integral heights are indicated, showing that the curves capture $g(K) = 2$, $\tau(K) = 1$, and $\epsilon(K) = 1$.

we can lift $\widehat{\mathrm{HF}}(M)$ to the infinite cylindrical cover $\overline{T}_M$, where each lifted marked point resides within a neighborhood of the lift of the meridian $\overline{\mu}$. The lifts of the marked points will be taken to lie at purely half-integral heights, and we will isotope the lifted curve components so that their horizontal tangencies occur at integral heights. Precisely one of the curves wraps around the cylinder, and we will use $\overline{\gamma}$ to denote this component. While $\overline{\gamma}$ is generally immersed, we will see that $\overline{\gamma}$ is embedded for thin knot complements. Figure 2 shows a centered lift of the invariant for the complement of a hypothetical example of a thin knot K.

Recall that $\widehat{\mathrm{HFK}}(K)$ detects $g(K)$ due to [29]. Looking in $\overline{T}_M$, genus detection manifests itself in $\widehat{\mathrm{HF}}(M)$ by ensuring that some curve component crosses at height $g(K)$. The immersed curves also satisfy a very powerful constraint related to a conjugation symmetry. For invariants of knot complements of S^3, this means that the curves are invariant under rotation by π.

Theorem 2.5 [15, Theorem 7] *The invariant $\widehat{\mathrm{HF}}(M)$ is symmetric under the elliptic involution of ∂M. Here, the involution is chosen so that z is a fixed point.*

With a horizontally or vertically simplified basis for $\mathrm{CFK}^-(K)$ (see [19, Section 3] for specifics regarding these bases that all knots admit), the procedure of [15, Proposition 47], which is the immersed curves version of [22, Theorem 11.31], allows one to construct $\widehat{\mathrm{HF}}(M)$ from $\mathrm{CFK}^-(K)$. In the special case when $\mathrm{CFK}^-(K)$ is simultaneously horizontally and vertically simplified, $\mathrm{HFK}^-(K)$ is generated by pairing (see Theorem 2.8 below) $\widehat{\mathrm{HF}}(M)$ with $\overline{\mu}$ in $\overline{T}$, and the differentials are recovered using bigons containing modified lifts of the marked point. This is not much of a constraint for us, since thin knots always admit a simultaneously horizontally and vertically simplified basis due to [36, Lemma 7]. In this lemma, Petkova shows that when the vertical and horizontal arrows in $\mathrm{CFK}^-(K)$ have length one, then $\mathrm{CFK}^-(K)$ consists of acyclic box complexes C and a staircase complex C_l. The following is a restatement of these conditions in immersed curves form.

Lemma 2.6 [36, Lemma 7] *If K is thin, then the lifted curve invariant associated to $S^3 \setminus \nu K$ satisfies:*

- *The essential component $\bar{\gamma}$ winds between adjacent basepoints, the height of which is determined by $\tau(K)$, before ultimately wrapping around the cylinder (corresponding to the staircase complex C_l).*
- *Every other component is a simple figure-eight, enclosing vertically adjacent basepoints (corresponding to the acyclic box complexes C).*

This lifted curve invariant also encodes numerical and concordance invariants of K. For example, the Seifert genus is given by the height of the tallest curve component by genus detection above. After following $\bar{\gamma}$ around the cylinder, the height of the first intersection that $\bar{\gamma}$ makes with $\bar{\mu}$ is precisely the Ozsváth–Szabó invariant $\tau(K)$. This is because this intersection corresponds to the distinguished generator of vertical homology whose Alexander grading is $\tau(K)$. Hom's ϵ invariant may also be determined by observing what $\bar{\gamma}$ does next. The essential curve either turns downwards, upwards, or continues straight corresponding to $\epsilon(K)$ being 1, -1, and 0, respectively. These two invariants determine the slope $\bar{\gamma}$ outside of a thin vertical strip surrounding the lifts of the marked point, given by $2\tau(K)-\epsilon(K)$.

Definition 2.7 Let e_n denote the number of simple figure-eight components at height n of $\widehat{\mathrm{HF}}(M)$, viewed in $\overline{T}_M$.

We have $e_{-n}=e_n$ due to Theorem 2.5, and Figure 2 provides an example with $e_0=0$ and $e_{-1}=e_1=1$. Equipped with their properties, we now turn to the main reason for involving bordered invariants in the form of immersed curves. The following is the immersed curves reformulation of the bordered pairing theorem.

Theorem 2.8 [15, Theorem 2] *Consider the gluing $M_1 \cup_h M_2$, where the M_i are compact, oriented 3-manifolds with torus boundary and $h\colon \partial M_2 \to \partial M_1$ is an orientation reversing homeomorphism for which $h(z_2)=z_1$. Then*

$$\widehat{\mathrm{HF}}(M_1 \cup_h M_2) \cong \mathrm{HF}\big(\widehat{\mathrm{HF}}(M_1), h(\widehat{\mathrm{HF}}(M_2))\big),$$

where intersection Floer homology is computed in T_{M_1} and the isomorphism is one of relatively graded vector spaces that respects the Spin^c *decomposition.*

More precisely, $\mathrm{HF}\big(\widehat{\mathrm{HF}}(M_1), h(\widehat{\mathrm{HF}}(M_2))\big)$ decomposes over spin^c structures and carries a relative Maslov grading on each spin^c summand. Theorem 2.8 places these in correspondence with the spin^c decomposition on $\widehat{\mathrm{HF}}(M_1 \cup_h M_2)$, and also ensures the relative Maslov gradings agree. This is best seen when viewing

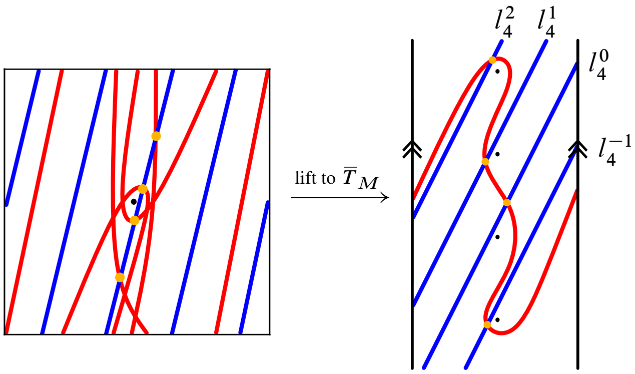


Figure 3: The pairing of $\widehat{\mathrm{HF}}(S^3 \setminus \nu T(2,5))$ and $h(\widehat{\mathrm{HF}}(D^2 \times S^1))$, whose intersection Floer homology is $\widehat{\mathrm{HF}}\big(S^3_4(T(2,5))\big)$.

Dehn surgery as such a gluing, continuing to use M for $S^3 \setminus \nu K$. We have $S^3_r(K) = M \cup_{h_r} (D^2 \times S^1)$ with h_r the slope-r gluing map. Then Theorem 2.8 provides

$$\widehat{\mathrm{HF}}(S^3_r(K)) \cong \mathrm{HF}\big(\widehat{\mathrm{HF}}(M), h_r(\widehat{\mathrm{HF}}(D^2 \times S^1))\big).$$

The spin^c decomposition is recovered by using r vertically adjacent lifts of $h_r(\widehat{\mathrm{HF}}(D^2 \times S^1)$, which is the precise number required to lift every intersection from T_M to $\overline{T}_M$ without duplicates. This is motivated by the example in Figure 3, showing the pairing of curves that recovers $\widehat{\mathrm{HF}}\big(S^3_4(T(2,5))\big)$. The invariant for the solid torus simply consists of a horizontal essential curve, and so $h_4(\widehat{\mathrm{HF}}(D^2 \times S^1)$ is a slope-4 curve in the punctured torus. We have four lifts of $h_4(\widehat{\mathrm{HF}}(D^2 \times S^1)$, each generating intersections in correspondence with the four spin^c summands of $\widehat{\mathrm{HF}}(S^3_4(K))$. These lifts are selected at heights in correspondence with the selected representatives of $\mathbb{Z}/r\mathbb{Z}$ from Section 2. These are -1, 0, 1, and 2 for the example in Figure 3, and motivate the following definition when lifting further to the tiled-plane cover $\widetilde{T}$.

Definition 2.9 Let $l^s_r = h_r(\widehat{\mathrm{HF}}(D^2 \times S^1)$ denote the slope-r line in $\widetilde{T}$ that crosses lifts $\tilde{\mu}$ at heights congruent to $s \pmod r$. These are selected so that each l^s_r crosses at height s in the same column of $\widetilde{T}$, with s taken to be the representative of $[s]$ that falls within the $\mathbb{Z}/r\mathbb{Z}$ range.

In this way, Theorem 2.8 implies

$$\widehat{\mathrm{HF}}(S^3_r(K), [s]) \cong \mathrm{HF}(\widehat{\mathrm{HF}}(S^3 \setminus \nu K), l^s_r).$$

As in the discussion following [13, Theorem 14], the Lagrangian intersection Floer homology has dimension equal to the minimal geometric intersection number of the immersed curves. In particular, using length-minimizing, or "pulled-tight", representatives for curve invariants by regular homotopy that avoid basepoints forces the differential to be identically zero, and so we may determine $\dim\big(\widehat{\mathrm{HF}}(S^3_r(K), [s])\big)$ by counting intersections between $\widehat{\mathrm{HF}}(M)$ and l^s_r. In general this count is modified by any immersed annuli cobounded by the paired curves, but this is only possible if $r = 0$ since $S^3 \setminus \nu K$ is Seifert-framed. As 0-surgery cannot yield a reducible manifold, no immersed annuli appear.

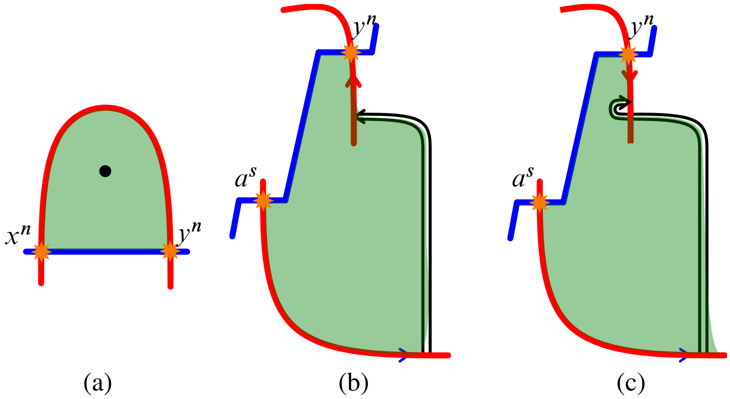


Figure 4: Bigons between intersections of $\widehat{\mathrm{HF}}(M)$ (in red) and l_r^s (in blue) that are used to determine the relative Maslov grading. Example (a) does not involve a grading arrow, while (b) (without a cusp) and (c) (with a cusp) do.

To incorporate the relative Maslov grading, we may compute grading differences between generators belonging to the same spin^c structure using a formula from [13]. Suppose x and y are two intersections belonging to the same $[s] \in \mathrm{Spin}^c(S_r^3(K))$, arising from intersections between $\widehat{\mathrm{HF}}(M)$ and l_r^s. Further, let P be the bigon from y to x whose boundary consists of a (not necessarily smooth) path from y to x in $\widehat{\mathrm{HF}}(M)$, concatenated with a path from x to y in l_r^s. Defined this way, the boundary of P is a closed path that is smooth apart from right corners at x and y, and possibly one or more cusps (possible when traversing grading arrows between components of $\widehat{\mathrm{HF}}(S^3 \setminus \nu K)$. The following formula follows from the conversion of bordered invariants into immersed curves, keeping track of grading contributions from relevant Reeb chords [15, Section 2.2].

Proposition 2.10 *Suppose x, y, and P are defined as above. Let $\mathrm{Rot}(P)$ denote $\frac{1}{2\pi}$ times the total counterclockwise rotation along the smooth sections of P. Alternatively this is $\frac{1}{2\pi}\big(2\pi - a\frac{\pi}{2} - c\pi\big)$, where a denotes the number of corners and c the number of cusps traversed. Let $\mathrm{Wind}(P)$ denote the net winding number of P around enclosed basepoints, and finally let $\mathrm{Wght}(P)$ be the sum of weights (counted with sign) of all grading arrows traversed by P. Then*

$$M(x) - M(y) = 2\,\mathrm{Wind}(P) + 2\,\mathrm{Wght}(P) - 2\,\mathrm{Rot}(P).$$

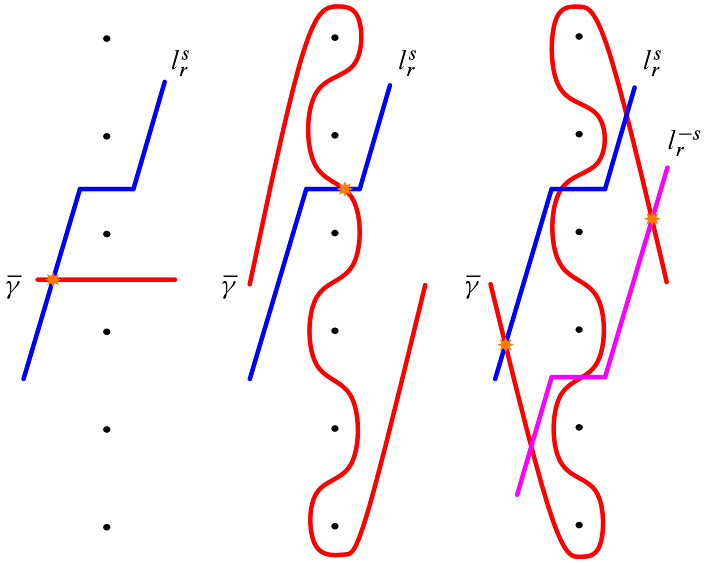
If l_r^s intersects a simple figure-eight component at height n, it generates a *right intersection* y^n and a *left intersection* x^n. Figure 4 shows off the three types of bigons that will typically appear. The first type has P connecting a right and left intersection of the same simple figure-eight. The bigon encloses a single basepoint with positive winding number, total counterclockwise rotation along smooth sections as π, and no contribution from traversed grading arrows. These traits imply $M(x^n) - M(y^n) = 1$. The second and third types are the more interesting ones, and have the same winding number of enclosed basepoints, but the rotation and grading arrow contributions to $M(y^n) - M(a^s)$ initially appear to be different. We will see later that for these bigons, the $2\,\mathrm{Wght}(P) - 2\,\mathrm{Rot}(P)$ component of the grading difference is the same.

3 Thin knots and Maslov grading differences

Throughout this section, let K be a thin knot and let M denote $S^3 \setminus \nu K$. To enable swift grading comparisons later on, let us designate a reference intersection associated to $[s] \in \mathrm{Spin}^c(S^3_r(K))$. We will define a *vertical intersection* to be an intersection between l^s_r and a vertical segment of $\overline{\gamma}$ within the neighborhood of $\overline{\mu}$, provided they exist. If s satisfies $0 \leq |s| < |\tau(K)|$, then such an intersection occurs and we will denote it using a^s. Alternatively, if $|s| \geq \tau(K) \geq 0$ then any intersection between l^s_r and $\overline{\gamma}$ is outside any neighborhood of the lifts of the marked point in $\overline{T}_M$. In this case l^s_r intersects $\overline{\gamma}$ once if $\tau(K) \geq 0$, and so a^s will denote this lone intersection. When $\tau(K) < 0$ and $s \geq 0$, we let a^s denote the intersection between l^r_s and $\overline{\gamma}$ to the left of $\overline{\mu}$. Analogously when $\tau(K) < 0$ and $s < 0$, we will have a^s be the intersection between l^r_s and $\overline{\gamma}$ to the right of $\overline{\mu}$. It is likely helpful to reference Figure 5 for these different possibilities. While cumbersome, this scheme allows us to label the intersection that often corresponds via the pairing theorem to a generator with the least Maslov grading.

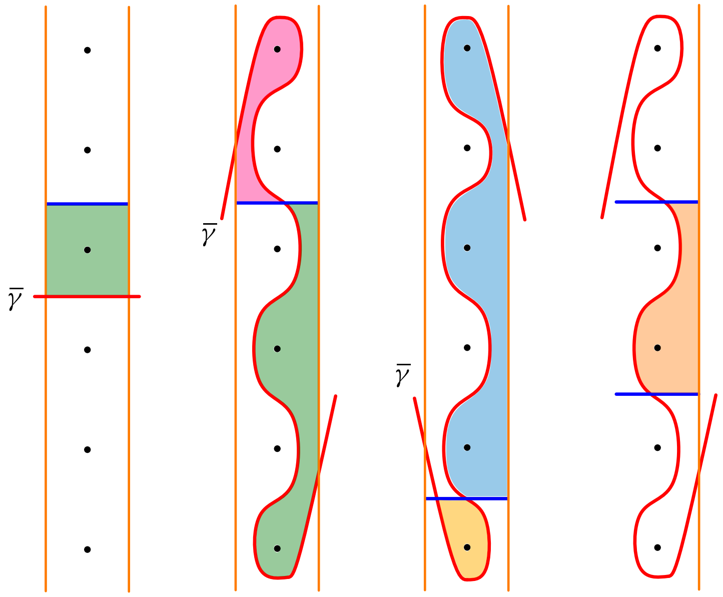
It will also be particularly useful to know the winding number of enclosed lifts of the marked point of specific regions. Consider the neighborhood of $\overline{\mu}$ in $\overline{T}_M$ that contains the lifts of the marked points, which is also wide enough to enclose the vertical segments of $\overline{\gamma}$. Intersect $\overline{\gamma}$ with a horizontal line l^s slightly longer than this neighborhood at height s, so that these segments together bound regions enclosing basepoints.

When $\tau(K) > 0$, we will define H_s to be the number of enclosed lifts of the marked point in the region bounded above by l^s, on the side(s) by the neighborhood of $\overline{\mu}$, and below by $\overline{\gamma}$. If the region is empty, then



(a) $\tau(K) = 0$ (b) $\tau(K) > 0$ (c) $\tau(K) < 0$

Figure 5: The possibilities for the reference intersection a^s (denoted by stars). (c) has two curves representing $s \geq 0$ (intersection with the blue curve) and $s < 0$ (intersection with the purple curve). The case when $\tau(K) > 0$ and $|s| \geq \tau(K)$ is similar to (a).



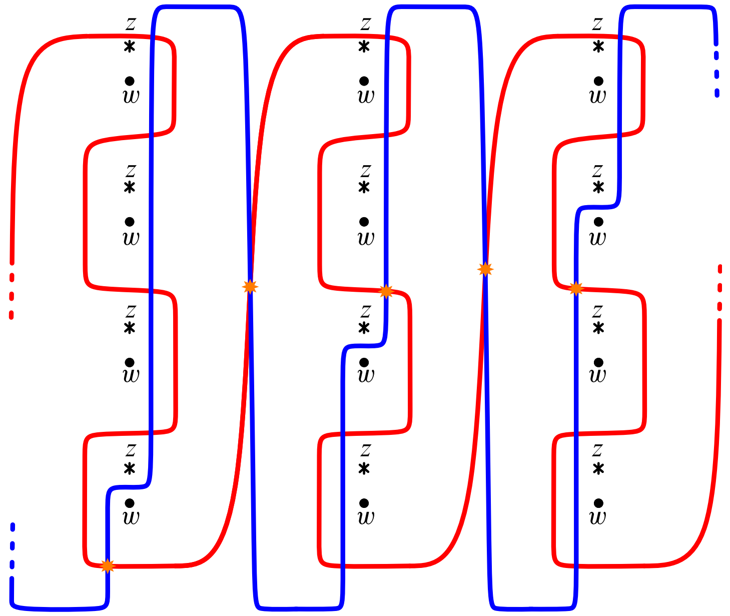
(a) $\tau(K)=0$ (b) $\tau(K)>0$ (c) $\tau(K)<0$ (d)

Figure 6: The regions in the discussion above whose winding numbers determine the H's and V's of a knot. Green regions correspond to H's and pink regions correspond to V's when $\tau(K)>0$. Otherwise when $\tau(K)<0$, the difference in winding numbers between the cyan and yellow regions are used. With L_s counting the winding numbers for the yellow region and U_s counting the winding numbers for the cyan region, we have $H_s=L_s-U_s$ and $V_s=0$ for $s\geq 0$. The region in (d) exhibits $H_s-V_s=s$.

$H_s=0$. Analogously, there is often a region where l^s bounds from below and $\overline{\gamma}$ bounds from above, and so we will denote the number of enclosed lifts of the marked point of such a region by V_s. These regions are depicted in parts a and b of Figure 6, where green regions correspond to H's and pink regions correspond to V's. Due to Theorem 2.5, we have both $H_{-s}=V_s$ and $H_s-V_s=H_s-H_{-s}=\frac{1}{2}(s-(-s))=s$.

Remark This relationship between the H's and V's is no coincidence. In [14], Hanselman establishes the HF^+ immersed curves theory for knot complements of S^3, recovering the $+$-flavored mapping cone diagram. With simple enough curve invariants ($\overline{\gamma}$ makes no self-intersections — see [14, Corollary 12.6]), the tower summands τ^+ of the A_s's and B_s's correspond to specific intersections between $\overline{\gamma}$ and l_r^s. Additionally, the V's and H's then correspond to the number of lifts of the marked point in bigons between these specific intersections. In our case, slight pointed-homotopies of curves yield equivalent intersections that provide the regions above (see Figure 7).

When $\tau(K)<0$, multiple regions are needed to compute H's and V's since the base of the tower in A_s^+ no longer corresponds to an intersection x with $A(x)=s$. The intersection corresponding to the base of the tower is similar to the reference intersection defined before Figure 5. When $s\geq 0$, the base of the tower in A_s^+ corresponds to a generator x with $A(x)=-\tau(K)$. The bigon between it and the



$$\begin{tikzcd}
A^+_{-1} \arrow[dr, "h^+_{-1}"] & & A^+_0 \arrow[dl, "v^+_0"'] \arrow[dr, "h^+_0"] & & A^+_1 \arrow[dl, "v^+_1"'] \\
& B^+ & & B^+ &
\end{tikzcd}$$

Figure 7: Modified curves $\widehat{\mathrm{HF}}(S^3 \setminus (T(2,5)))$ (in red) and l_1 (in blue) in $\widetilde{T}$ to recover the complexes and maps between them associated to $\mathrm{CFK}^\infty(T(2,5))$ in the mapping cone calculating $\widehat{\mathrm{HF}}(S^3_1(T(2,5)))$. Intersections corresponding to surviving generators in homology are represented with orange asterisks.

nonvertical intersection corresponding to the tower B^+ in the codomain of v_s^+ contains no marked points. On the other side, we traverse two bigons (split when the filling curve crosses $\bar{\mu}$ at height s) to reach the nonvertical intersection corresponding to the tower B^+ in the codomain of h_s^+. This agrees with the yellow and cyan bigons in Figure 6, and in short $V_s = 0$ and $H_s = s$ when $s \geq 0$. Alternatively when $s < 0$, the base of the tower in A_s^+ corresponds to a generator x with $A(x) = \tau(K)$, and we likewise have $V_s = -s$ and $H_s = 0$. In proofs to come, we may use U_s and L_s to denote the number of enclosed marked points in the upper (cyan) bigon or the lower (yellow) bigon, where $H_s = L_s - U_s$ and $V_s = 0$ for $s \geq 0$. Also, it is clear that L_s is an increasing function of s and U_s is a decreasing function of s when their respective bigons are defined.

From the discussion in the previous section, we know that the form of $\widehat{\mathrm{HF}}(M)$ is very restricted. Our goal is to leverage this to constrain gradings on $\widehat{\mathrm{HF}}(S^3_r(K),[s]) \cong \mathrm{HF}(\widehat{\mathrm{HF}}(M), l_r^s)$ to obstruct reducible surgeries. We use multisets, which are sets with repetition allowed, to collect these relative Maslov gradings. As mentioned after Definition 2.9, we will think of intersections $y \in \widehat{\mathrm{HF}}(M) \pitchfork l_r^s$ and generators y of $\mathrm{HF}(\widehat{\mathrm{HF}}(M), l_r^s)$ interchangeably.

Definition 3.1 Let $[s] \in \mathrm{Spin}^c(S^3_r(K))$ be arbitrary with reference intersection a^s. For any generator y of $\mathrm{HF}(\widehat{\mathrm{HF}}(M), l^s_r)$, let $M_{\mathrm{rel}}(y)$ denote the grading difference $M(y) - M(a^s)$. We define the desired multiset by

$$\mathrm{MR}^{[s]} := \{M_{\mathrm{rel}}(y) \mid y \in \widehat{\mathrm{HF}}(M) \pitchfork l^s_r\}.$$

Further, let $\mathrm{Width}(\mathrm{MR}^{[s]})$ denote the difference between the largest and smallest elements of this multiset.

Remark As defined, $\mathrm{MR}^{[s]}$ is a collection of integral Maslov gradings differences, and so in general is not an invariant of the pair $(S^3_r(K), [s])$. However, $\mathrm{Width}(\mathrm{MR}^{[s]})$ is an invariant of the pair $(S^3_r(K), [s])$ since $\mathrm{MR}^{[s]}$ can be made to agree with the multiset containing absolute Maslov gradings by uniformly translating all elements by some element of $\mathbb{Q}$. Likewise, the cardinality of $\mathrm{MR}^{[s]}$ and multiplicities of its elements (after uniformly translating so that 0 is the smallest element) are also invariants of the pair $(S^3_r(K), [s])$.

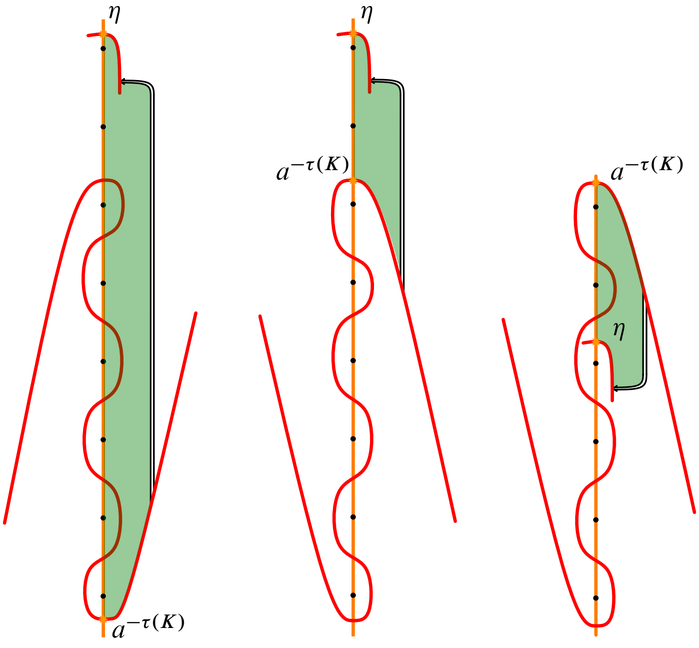
Next, we establish lemmas that enable us to swiftly compute grading differences. For a bigon P between intersections of $\widehat{\mathrm{HF}}(M)$ and l^s_r, we will determine the grading difference contribution of $2\,\mathrm{Wght}(P) - 2\,\mathrm{Rot}(P)$. This is done by considering an analogous, regularly homotopic bigon P_K between intersections of $\widehat{\mathrm{HF}}(M)$ and $\bar{\mu}$ that correspond to generators of $\widehat{\mathrm{HFK}}$ under pairing. We show that the quantities $2\,\mathrm{Wght}(P) - 2\,\mathrm{Rot}(P)$ and $2\,\mathrm{Wght}(P_K) - 2\,\mathrm{Rot}(P_K)$ agree, and computing the latter in terms of the knot Floer invariant $\tau(K)$.

Lemma 3.2 *Let y^n be a right intersection belonging to a simple figure-eight at height n of $\widehat{\mathrm{HF}}(M)$, let a be an intersection from a different component of $\widehat{\mathrm{HF}}(M)$ and l^s_r, and suppose P is a bigon between them. If K is thin, then* $2\,\mathrm{Wght}(P) - 2\,\mathrm{Rot}(P) = -1 - \tau(K) - |n|$.

Proof In the infinite cylinder $\overline{T}_M$, we can represent $\bar{\mu}$, the lift of the meridian of T_M, as the vertical line that pierces each lift of the marked point in $\overline{T}_M$. Let $a^{-\tau(K)}$ be the last intersection that $\bar{\gamma}$ makes with $\bar{\mu}$ before wrapping around $\overline{T}_M$. Because $\widehat{\mathrm{HF}}(M)$ is invariant under the action by the hyperelliptic involution, the weights of the grading arrows connecting $\bar{\gamma}$ to the simple figure-eights at heights n and $-n$ are equivalent. From this we can assume that n is nonnegative, and use $|n|$ in future formulas otherwise.

Lift $\widehat{\mathrm{HF}}(M)$ to $\overline{T}$ for convenience, and intersect it with $\bar{\mu}$. If we place z and w basepoints to the left and right, respectively, of every lift of the marked point, then $\widehat{\mathrm{HFK}}(K) \cong \mathrm{HF}(\widehat{\mathrm{HF}}(M), \bar{\mu})$ due to [15, Theorem 51]. This pairing is depicted in Figure 8. The formula in Proposition 2.10 still holds with the adjustment that Wind is modified to count the net winding number of enclosed w basepoints, denoted by Wind_w.

Since $\widehat{\mathrm{HF}}(M)$ has a simple figure-eight component at height n, there must be a generator η of $\widehat{\mathrm{HFK}}(K)$ with $A(\eta) = n + 1$. Let P_K be the bigon from $a^{-\tau(K)}$ to η that traverses the grading arrow connecting the relevant components of $\widehat{\mathrm{HF}}(M)$, visible in Figure 8 with $\tau(K) \geq 0$ and $\tau(K) < 0$, respectively. To determine $\mathrm{Wght}(P_K)$ directly would require care for the orientations of the grading arrow. However



(a) (b) (c)

Figure 8: The bigon P_K between $a^{-\tau(K)}$ and η, formed from path components in $\widehat{\mathrm{HF}}(M)$ and $\overline{\mu}$. (a) shows this for $\tau(K) \geq 0$ and (b) shows this for $A(\eta) > -\tau(K) > 0$. However for (c) with $A(\eta) \geq -\tau(K) > 0$, the bigon P_K runs from η to $a^{-\tau(K)}$.

since we are after a different term, we can abuse notation by having every grading arrow connect to the right side of a simple figure-eight, regardless of its orientation. Essentially, any change that $\mathrm{Wght}(P_K)$ experiences between the two ways of attaching the grading arrow is inverted and absorbed by $\mathrm{Rot}(P)$, so that $2\,\mathrm{Wght}(P_K) - 2\,\mathrm{Rot}(P_K)$ remains unchanged.

If $\tau(K) \geq 0$ so that $A(a^{-\tau(K)}) < n$, we have

$$M(\eta) - M(a^{-\tau(K)}) = 2\,\mathrm{Wind}_w(P_K) + 2\,\mathrm{Wght}(P_K) - 2\,\mathrm{Rot}(P_K).$$

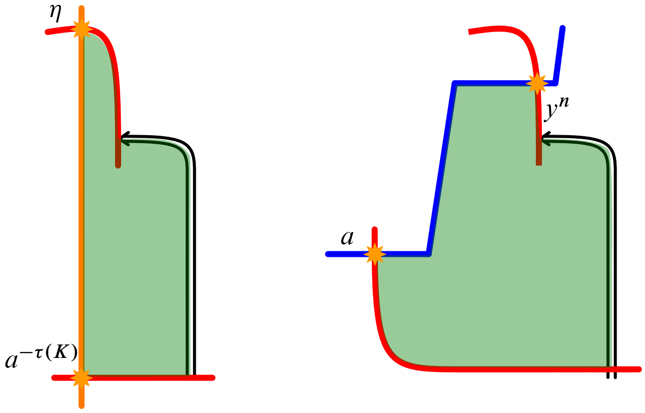
However since K is thin, it follows that

$$M(\eta) - M(a^{-\tau(K)}) = A(\eta) - A(a^{-\tau(K)}) = A(\eta) + \tau(K).$$

Then $2\,\mathrm{Wght}(P_K) - 2\,\mathrm{Rot}(P_K) = A(\eta) - 2\,\mathrm{Wind}(P_K) + \tau(K)$. Since $\mathrm{Wind}(P_K) = A(\eta) + \tau(K)$, we have $2\,\mathrm{Wght}(P_K) - 2\,\mathrm{Rot}(P_K) = -A(\eta) - \tau(K) = -1 - \tau(K) - n$.

If $\tau(K) < 0$, the above computation follows through for $A(\eta) > -\tau(K)$, but the case for $A(\eta) \leq -\tau(K)$ differs slightly. In this situation P_K is a bigon from η to $a^{-\tau(K)}$ that also traverses the grading arrow in reverse, visible in Figure 8. Traveling the grading arrow in reverse means that we have

$$M(a^{-\tau(K)}) - M(\eta) = 2\,\mathrm{Wind}_w(P_K) - 2\,\mathrm{Wght}(P_K) - 2\,\mathrm{Rot}(P_K),$$



(a) (b)

Figure 9: Tilting bigons to show they have equivalent net clockwise rotation along their boundaries. (a) The bigon P_K from $a^{-\tau(K)}$ to η. (b) The bigon P from a to y^n.

and so

$$\begin{aligned}-2\,\mathrm{Wght}(P_K)-2\,\mathrm{Rot}(P_K)&=M(a^{-\tau(K)})-M(a^{\eta})-2\,\mathrm{Wind}_w(P)\\&=-\tau(K)-(n+1)-2(-\tau(K)-(n+1))\\&=1+\tau(K)+n.\end{aligned}$$

Due to the shape of P_K, the bigon has a cusp near the grading arrow regardless of how it connects these components, and so $\mathrm{Rot}(P_K)=0$. Then we have

$$2\,\mathrm{Wght}(P_K)-2\,\mathrm{Rot}(P_K)=2\,\mathrm{Wght}(P_K)+2\,\mathrm{Rot}(P_K)=-1-\tau(K)-n,$$

as claimed.

With the formula established for P_K, we will now show that it is satisfied for a bigon between generators of $\mathrm{HF}(\widehat{\mathrm{HF}}(M),l_r^s)$ with similar attributes. Let y^n be a right intersection from the simple figure-eight at height n and let a be an intersection from a vertical segment of $\bar{\gamma}$ and l_r^s. With P denoting the bigon from a to y^n, we see that P must traverse the same grading arrow that P_K traversed, and so $\mathrm{Wght}(P)=\mathrm{Wght}(P_K)$. Additionally, it is straightforward to see that $\mathrm{Rot}(P)=\mathrm{Rot}(P_K)$ after tilting the bigons as well, with visual given in Figure 9. □

The following proposition considers left and right intersections of a simple figure-eight whose height n is less than $|\tau(K)|$. There is then a nearby vertical intersection a^n, and we will see that these three intersections have little difference in grading.

Proposition 3.3 *Let K be thin and have M denote $S^3\setminus\nu K$. Further, let x^n and y^n be left and right intersections belonging to a simple figure-eight of $\widehat{\mathrm{HF}}(M)$ with height $0\le n<|\tau(K)|$, and let a^n be the nearby vertical generator. Then $-1\le M(y^n)-M(a^n)\le 0$ and $0\le M(x^n)-M(a^n)\le 1$.*

Proof If P is the bigon between a^n and y^n, we have $2\,\mathrm{Wght}(P) - 2\,\mathrm{Rot}(P) = -1 - \tau(K) - |n|$ due to Lemma 3.2. Due to the hyperelliptic involution invariance of $\widehat{\mathrm{HF}}(M)$, we can take $0 \leq n < |\tau(K)|$. We have $\mathrm{Wind}(P)$ is H_n if $\tau(K) \geq 0$ or U_n if $\tau(K) < 0$, the values of which depend on the parity of n and $\tau(K)$ when K is thin. The simple structure of $\overline{\gamma}$ for a thin knot together with a counting argument for $\tau(K) > 0$ yields

$$H_n = \begin{cases} \frac{1}{2}(n+\tau(K)) & \text{if parity}(n) = \text{parity}(\tau(K)), \\ \frac{1}{2}(n+\tau(K)+1) & \text{if parity}(n) \neq \text{parity}(\tau(K)). \end{cases}$$

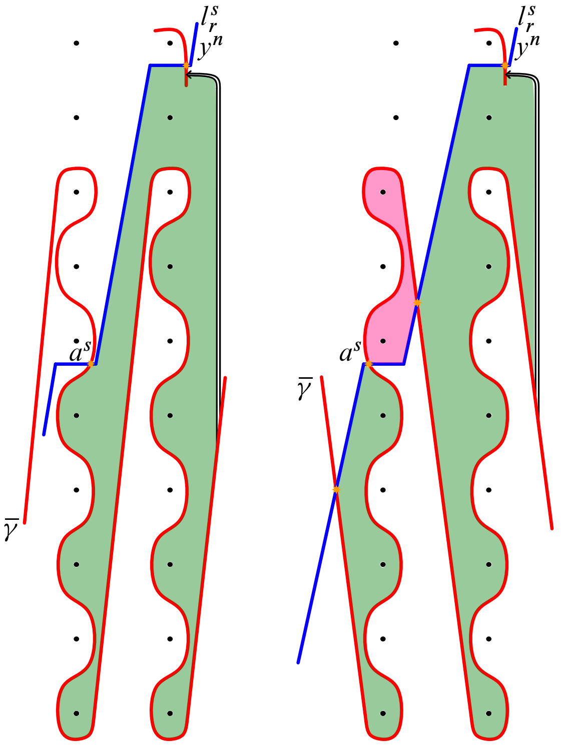
Then for $\tau(K) > 0$ we have $M(y^n) - M(a^n) = 2H_n - 1 - \tau(K) - n$ implies $M(y^n) - M(a^n)$ is either -1 or 0. Since $M(x^n) - M(y^n) = 1$, we see that $M(x^n) - M(a^n)$ is either 0 or 1, handling the $\tau(K) > 0$ case.

When $\tau(K) < 0$, the bigon P runs from y^n to a^n, encloses U_n lifts of the marked points, traverses the grading arrow in reverse, and has $\mathrm{Rot}(P) = 0$. Figure 6 shows that U_n with $\tau(K) < 0$ is the same as $V_n = H_{-n}$ with $\tau(K) \geq 0$, except using $-\tau(K)$ or $-\tau(K) - 1$ in the formula above. Using Lemma 3.2 and the $-\tau(K)$ modified formula for H_{-n}, we have $M(a^n) - M(y^n) = 2H_{-n} + 1 + \tau(K) + n$. This is either 1 or 0, and so $M(y^n) - M(a^n)$ is either -1 or 0 and analogously $M(x^n) - M(a^n)$ is either 0 or 1. □

Since $M(x^n) - M(y^n) = 1$, these possibilities happen in pairs. A simple figure-eight at height $n < |\tau(K)|$ contributes either $\{M_{\mathrm{rel}}(a^n), M_{\mathrm{rel}}(a^n) - 1, M_{\mathrm{rel}}(a^n)\} \subseteq \mathrm{MR}^{[s]}$ or $\{M_{\mathrm{rel}}(a^n), M_{\mathrm{rel}}(a^n), M_{\mathrm{rel}}(a^n) + 1\} \subseteq \mathrm{MR}^{[s]}$. An example of this to keep in mind is when looking at large surgery on the figure-eight knot 4_1. In this situation we have $\{0, -1, 0\} = \mathrm{MR}^{[0]}$, and the right intersection contributing -1 to $\mathrm{MR}^{[0]}$ actually has the smallest relative Maslov grading. Proposition 3.3 then allows us to determine which intersection associated to $[s] \in \mathrm{Spin}^c(S^3_r(K))$ has the smallest relative Maslov grading depending on $\text{parity}(\tau(K))$:

- If $\tau(K) \geq 0$, $\text{parity}(s) = \text{parity}(\tau(K))$, and there is a right intersection y^s, then $M_{\mathrm{rel}}(y^s) = -1$ is the smallest relative grading of $\mathrm{MR}^{[s]}$.
- If $\tau(K) \geq 0$, $\text{parity}(s) = \text{parity}(\tau(K))$, and there is no simple figure-eight at height s, then $M_{\mathrm{rel}}(a^s) = 0$ is the smallest relative grading of $\mathrm{MR}^{[s]}$.
- If $\tau(K) \geq 0$ and $\text{parity}(s) \neq \text{parity}(\tau(K))$, then $M_{\mathrm{rel}}(a^s) = 0$ is the smallest relative grading of $\mathrm{MR}^{[s]}$.
- If $\tau(K) < 0$, then $M_{\mathrm{rel}}(a^s) = 0$ is the smallest relative grading of $\mathrm{MR}^{[s]}$.

The last component of the grading difference formula in Proposition 2.10 to determine is $\mathrm{Wind}(P)$. Lift both $\widehat{\mathrm{HF}}(M)$ and each l^s_r to the tiled plane $\widetilde{T}$, and let the 0th column be the neighborhood of the lift $\widetilde{\mu}$ for which each l^s_r intersects $\widetilde{\mu}$ at height $[s]$. For $[s] \in \mathbb{Z}/r\mathbb{Z}$ define $w_s = \frac{n-[s]}{r}$, with n the largest natural number satisfying $0 \leq n \leq g(K) - 1$ and $n \equiv s \pmod{r}$. This number represents the number of columns of marked points in $\widetilde{T}$ between a^s and a potential furthest right intersection y^n. Further, because the slopes we consider satisfy $r \leq 2g(K) - 1$, we have $w_s \geq 0$. While it is certainly possible that a simple figure-eight component may not exist at this height, it is still sufficient for the following strategy to suppose otherwise.



(a) (b)

Figure 10: Example bigons P between a^s and y^n, showing the contributions from each column to $\mathrm{Wind}(P)$ for (a) $\tau(K) \geq 0$ and (b) $\tau(K) < 0$ with $s \geq 0$.

Proposition 3.4 *For a given $[s] \in \mathbb{Z}/r\mathbb{Z}$, let a^s be the chosen reference intersection and y^n be a right intersection of a furthest possible figure-eight component. If $\tau(K) \geq 0$, then*

$$\mathrm{Wind}(P) = H_s + \sum_{i=1}^{w_s} (s + ir).$$

If $\tau(K) < 0$, then

$$\mathrm{Wind}(P) = \begin{cases} \sum\limits_{i=0}^{w_s} (s + ir) & \text{if } [s] \geq 0, \\ \sum\limits_{i=1}^{w_s} (s + ir) & \text{if } [s] < 0, \end{cases}$$

where all sums are taken to be zero if empty.

When $\tau(K) \geq 0$, the contribution to $\mathrm{Wind}(P)$ from the 0^{th} column of $\widetilde{T}$ is H_s. The contribution from the i^{th} column is $H_{s+ir} - V_{s+ir} = s + ir$, and is shown in Figure 10. When $\tau(K) < 0$, we have the two different reference intersections a^s depending on s influencing whether there is a contribution from the 0^{th} column. Regardless, in every column the contribution to $\mathrm{Wind}(P)$ is $H_{s+ir} - V_{s+ir} = H_{s+ir} = s + ir$. Since these terms are always nonnegative, it follows that the smallest relative grading belongs to an intersection in the 0^{th} column.

4 Initial cases with $|\tau(K)| < g(K)$

Our objective is to build a collection of lemmas required to prove the main theorem. These vary depending on r in relation to $g(K)$, and on $\tau(K)$ and its parity. The primary technique involves comparing the various values of $\mathrm{Width}(\mathrm{MR}^{[s]})$ to obstruct periodicity (see Lemma 2.1), typically done by showing that $\mathrm{Width}(\mathrm{MR}^{[s']})$ is maximal if $[s']$ is the spin^c structure associated to the line that crosses height $g(K)-1$. At other times the widths will agree up to translation, but the multiplicity of specific elements of the grading multisets will not.

Recall that Theorem 2.8 identifies $\widehat{\mathrm{HF}}(S^3_r(K),[s]) \cong \mathrm{HF}(\widehat{\mathrm{HF}}(M), l^s_r)$. In order to halve the amount of comparisons to make, we leverage the fact that $\widehat{\mathrm{HF}}(S^3_r(K),[s]) \cong \widehat{\mathrm{HF}}(S^3_r(K),[-s])$ [31]. In immersed curves form, Theorem 2.5 implies that intersections between $\widehat{\mathrm{HF}}(M)$ and l^s_r in negative columns of $\widetilde{T}$ are in correspondence with intersections of $\widehat{\mathrm{HF}}(M)$ and l^{-s}_r that belong to positive columns of $\widetilde{T}$ (see Figure 11). Also, intersections associated to the self-conjugate spin^c structure(s) [0] (and possibly $[r/2]$) are symmetric in this way by default.

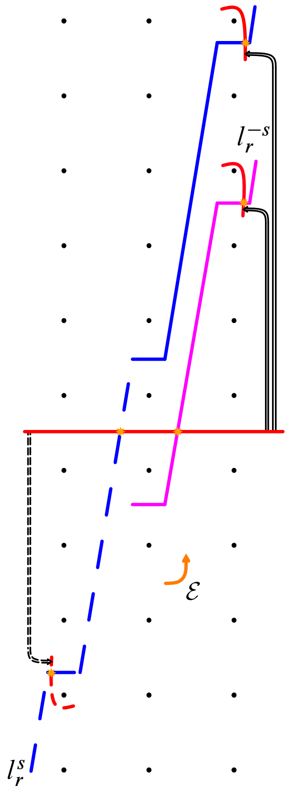


Figure 11: The elliptic involution, denoted by $\mathcal{E}$, on $\partial M \setminus \{z\}$ affects the lift of $\widehat{\mathrm{HF}}(M)$ to the tiled plane by placing intersections of $\widehat{\mathrm{HF}}(M)$ and l^s_r in negative columns (dashed blue line) in correspondence with intersections of $\widehat{\mathrm{HF}}(M)$ and l^{-s}_r in positive columns (solid purple line).

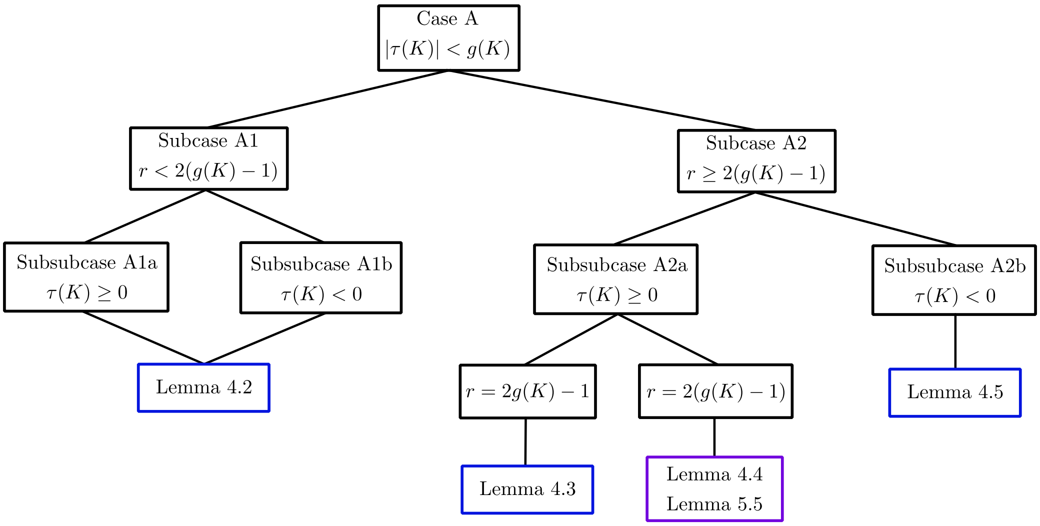


Figure 12: The case flowchart for the arguments in this section. Blue boxes indicate that the contained lemmas only appeal to relative grading information, while purple boxes use some of the absolute grading material from Section 5.

Recall that the smallest element of $\mathrm{MR}^{[s]}$ is the relative grading of an intersection belonging to the 0^{th} column of $\widetilde{T}$, which is either the reference intersection a^s or a nearby right/left intersection. This means that we can capture $\mathrm{Width}(\mathrm{MR}^{[s]})$ by considering nonnegative intersections associated to both $[s]$ and $[-s]$. Note that since $\mathrm{parity}([s]) = \mathrm{parity}([-s])$, the need to translate a multiset by 1 is consistent if it arises.

Definition 4.1 The multiset $\mathrm{MR}_+^{[s]}$ consists of the relative gradings of intersections between $\widehat{\mathrm{HF}}(M)$ and l_r^s that belong to nonnegative columns of $\widetilde{T}$. We define $\mathrm{MR}_-^{[s]}$ analogously, and notice that $\mathrm{Width}(\mathrm{MR}^{[s]}) = \max\{\mathrm{Width}(\mathrm{MR}_+^{[s]}), \mathrm{Width}(\mathrm{MR}_-^{[s]})\}$.

Due to how genus detection is expressed by $\widehat{\mathrm{HF}}(M)$, either $\bar{\gamma}$ achieves height $g(K)$ (equivalent to $|\tau(K)| = g(K)$), or only a simple figure-eight at height $g(K)-1$ achieves this desired height (equivalent to $|\tau(K)| < g(K)$). We will divide the problem among these two cases, starting with the latter. The ensuing case analysis is admittedly complicated, but hopefully Figure 12 makes it more palatable.

Case A $(|\tau(K)| < g(K))$ Since $|\tau(K)| < g(K)$, there exists a simple figure-eight component at height $g(K)-1$. Let $[s']$ be the spin^c structure for which $l_r^{s'}$ intersects this simple figure-eight, which means $w_{s'} = (g(K)-1-s')/r$. Our potential reducing slopes of $1 < r \leq 2g(K)-1$ divide this case into two subcases. When $r \geq 2(g(K)-1)$, we equivalently have $[s'] = g(K)-1$ and $w_{s'} = 0$. Otherwise $r < 2(g(K)-1)$, or equivalently $w_{s'} > 0$, which is the easier starting point.

Subcase A1 $(r < 2(g(K)-1))$ In this situation, we will show that $\mathrm{Width}(\mathrm{MR}^{[s']})$ is maximal.

Lemma 4.2 *Suppose K is thin, $|\tau(K)| < g(K)$, and $1 < r < 2(g(K)-1)$. Then there exists an $[s'] \in \mathrm{Spin}^c(S_r^3(K))$ for which every $[s] \neq [\pm s']$ satisfies $\mathrm{MR}^{[s]} \not\cong \mathrm{MR}^{[s']}$ up to translation.*

Proof For some $[s] \neq [\pm s']$, the largest possible relative grading that $\mathrm{MR}_{+}^{[s]}$ can achieve is associated to an intersection of some hypothetical simple figure-eight at largest height. Looking at the terms in the grading difference formula for a bigon from a^s to such a generator, we see that the $2\,\mathrm{Wind}(P)$ term satisfies $2\,\mathrm{Wind}(P) \geq 2n$ while the other term is $-1-\tau(K)-n$. For this reason, we will suppose that $\widehat{\mathrm{HF}}(M)$ has a simple figure-eight at height n, taken to be the largest integer satisfying both $n < g(K)-1$ and $n \equiv s \pmod r$. Let P' be the bigon between $a^{s'}$ and y^{g-1}, and P the bigon between a^s and y^n. Because the choice of $a^{s'}$ depends on $\tau(K)$, we will handle the $\tau(K) \geq 0$ subcase first before handling the $\tau(K) < 0$ subcase.

Subsubcase A1a ($\tau(K) \geq 0$) Due to Lemma 3.2 and Propositions 3.3–3.4, $\mathrm{Width}(\mathrm{MR}_{+}^{[s]})$ is nearly determined by $M_{\mathrm{rel}}(y^n)$. We have $M_{\mathrm{rel}}(y^n) \leq \mathrm{Width}(\mathrm{MR}_{+}^{[s]}) \leq M_{\mathrm{rel}}(y^n)+1$, with either equality depending on whether a^s is the smallest relatively graded intersection. To compare widths, we compute

$$M_{\mathrm{rel}}(y^{g-1}) = 2\Big(H_{s'} + \sum_{i=1}^{w_{s'}} (s'+ir)\Big) - 1 - (s'+w_{s'}r),$$

and likewise

$$M_{\mathrm{rel}}(y^{n}) = 2\Big(H_{s} + \sum_{i=1}^{w_{s}} (s+ir)\Big) - 1 - (s+w_{s}r).$$

Their difference is then

$$\begin{aligned} M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^n) &= 2\Big(H_{s'} + \sum_{i=1}^{w_{s'}} (s'+ir) - \Big(H_s + \sum_{i=1}^{w_s} (s+ir)\Big)\Big) - (s'+w_{s'}r-(s+w_sr)) \\ &= 2\Big((H_{s'}-H_s) + \sum_{i=1}^{w_{s'}} (s'+ir) - \sum_{i=1}^{w_s} (s+ir)\Big) - (s'-s) - r(w_{s'}-w_s). \end{aligned}$$

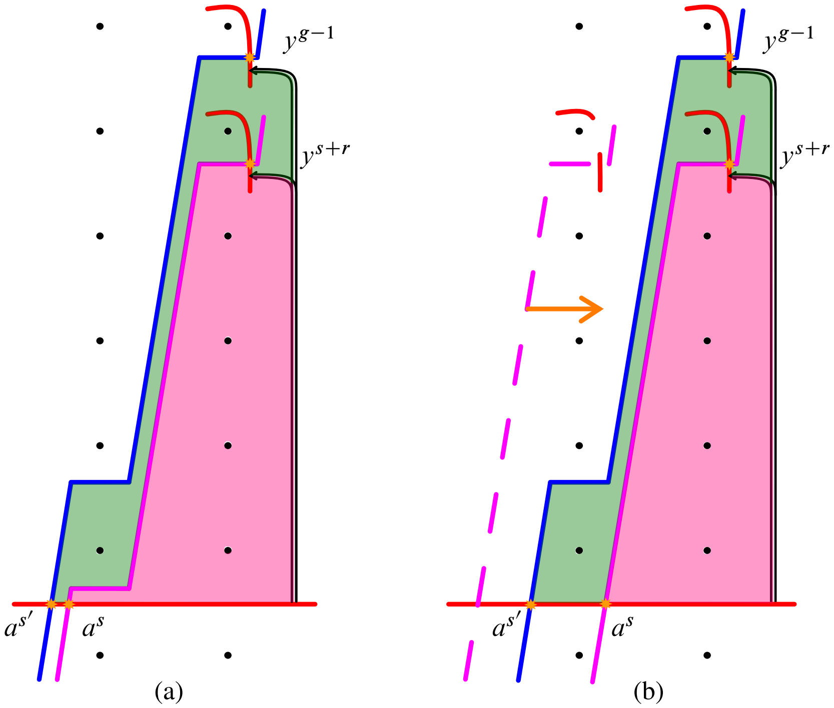
If $s < s'$ so that $w_s = w_{s'}$, then

$$\begin{aligned} M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^n) &= 2((H_{s'}-H_s) + w_{s'}(s'-s)) - (s'-s) \\ &= 2(H_{s'}-H_s) + (2w_{s'}-1)(s'-s) \\ &\geq 1, \end{aligned}$$

since $w_{s'} > 0$ and $s' > s$ implies that $H_{s'} \geq H_s$.

If $s > s'$ so that $w_s = w_{s'}-1$, then shifting P one column to the right in $\widetilde{T}$ (see Figure 13) provides

$$\begin{aligned} M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^n) &= 2\Big((H_{s'}-H_s) + \sum_{i=1}^{w_{s'}} (s'+ir) - \sum_{i=1}^{w_s} (s+ir)\Big) - (s'-s) - r(w_{s'}-w_s) \\ &= 2\Big((H_{s'}-H_s) + \sum_{i=1}^{w_{s'}} (s'+ir) - \sum_{i=2}^{w_{s'}} (s+(i-1)r)\Big) - (s'+r-s) \\ &= 2\Big((H_{s'}-H_s) + (s'+r) + \sum_{i=2}^{w_s'} (s'+ir) - \sum_{i=2}^{w_{s'}} (s+(i-1)r)\Big) - (s'+r-s) \\ &= 2\big((H_{s'}+s-H_s) + (s'+r-s) + (w_{s'}-1)(s'+r-s)\big) - (s'+r-s) \\ &= 2(H_{s'}+(s-H_s)) + (2w_{s'}-1)(s'+r-s) \\ &= 2(H_{s'}-V_s) + (2w_{s'}-1)(s'+r-s) \\ &= 2(H_{s'}-H_{-s}) + (2w_{s'}-1)(s'+r-s), \end{aligned}$$



(a) (b)

Figure 13: Example bigons P' (split-shaded green and pink) and P (shaded pink) when $w_{s'} = 1$. (a) has $s < s'$, while (b) has $s > s'$ together with the single column shift to the right.

where the second line is by column shift. Notice that $s' + s - 1 \leq 2(H_{s'} - H_{-s}) \leq s' + s$ depending on the parities of s and s' together with $s > s'$. Then we have

$$\begin{aligned} M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^{n}) &= 2(H_{s'} - H_{-s}) + (2w_{s'} - 1)(s' + r - s) \\ &\geq s' + s - 1 + (2w_{s'} - 1)(s' + r - s) \\ &\geq s' + s - 1 + s' + r - s \\ &= 2s' - 1 + r \\ &> 1, \end{aligned}$$

since $w_{s'} > 0$ and $s' < \frac{r-1}{2}$ if there exists an $s > s'$.

In both situations, we see that $M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^{n}) \geq 1$. If this difference is greater than one, then

$$\mathrm{Width}(\mathrm{MR}^{[s']}) \geq \mathrm{Width}(\mathrm{MR}_{+}^{[s']}) \geq M_{\mathrm{rel}}(y^{g-1}) > M_{\mathrm{rel}}(y^{n}) + 1 \geq \mathrm{Width}(\mathrm{MR}_{+}^{[s]}).$$

This handles the possibility where we need to translate $\mathrm{MR}_{+}^{[s]}$ by 1, so suppose $M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^{n}) = 1$. This is possible only if $H_s = H_{s'}$, $w_{s'} = 1$, and $s = s' - 1$, which altogether imply that $s = \tau(K)$. However, the widths only match if $\mathrm{Width}(\mathrm{MR}_{+}^{[s]}) = M_{\mathrm{rel}}(y^{n}) + 1$. This condition is equivalent to having $\mathrm{parity}(s) \neq \mathrm{parity}(\tau(K))$, which is a contradiction. Therefore $\mathrm{Width}(\mathrm{MR}^{[s']}) > \mathrm{Width}(\mathrm{MR}_{\pm}^{[s]})$, which completes the $\tau(K) \geq 0$ subcase.

Subsubcase A1b $(\tau(K) < 0)$ Recall that the reference intersection a^s has no nearby left/right intersections belonging to a simple figure-eight. This means that a^s has the smallest relative grading of $\mathrm{MR}^{[s]}$, and so $\mathrm{Width}(\mathrm{MR}_+^{[s]}) = M_{\mathrm{rel}}(y^n) + 1$. From Proposition 3.4 we see

$$\mathrm{Wind}(P) = \begin{cases} s + \sum_{i=1}^{w_s} (s + ir) & \text{if } s \geq 0, \\ \sum_{i=1}^{w_s} (s + ir) & \text{if } s < 0. \end{cases}$$

If $0 \leq s < s'$, then proceeding as before we have

$$\begin{aligned} M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^n) &= 2\Bigl(s' + \sum_{i=1}^{w_{s'}} (s' + ir) - \Bigl(s + \sum_{i=1}^{w_s} (s + ir)\Bigr)\Bigr) - (s' - s) - r(w_{s'} - w_s) \\ &= 2(s' - s + w_{s'}(s' - s)) - (s' - s) \\ &= (2w_{s'} + 1)(s' - s) \\ &\geq 3. \end{aligned}$$

If $s < s' \leq 0$, then

$$M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^n) = (2w_{s'} - 1)(s' - s) \geq 1.$$

If $s > s'$, then as before we have $w_s = w_{s'} - 1$. If $s > s' \geq 0$, then

$$\begin{aligned} M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^n) &= 2\Bigl((s' - s) + \sum_{i=1}^{w_{s'}} (s' + ir) - \sum_{i=1}^{w_s} (s + ir)\Bigr) - (s' - s) - r(w_{s'} - w_s) \\ &= 2\Bigl((s' - s) + \sum_{i=1}^{w_{s'}} (s' + ir) - \sum_{i=2}^{w_{s'}} (s + (i-1)r)\Bigr) - (s' + r - s) \\ &= 2(s' + (s' + r - s) + (w_{s'} - 1)(s' + r - s)) - (s' + r - s) \\ &= 2s' + (2w_{s'} - 1)(s' + r - s) \\ &\geq 1, \end{aligned}$$

where the second line is by column shift. In the event that $0 \geq s > s'$, we get

$$\begin{aligned} M_{\mathrm{rel}}(y^{g-1}) - M_{\mathrm{rel}}(y^n) &= 2\Bigl(\sum_{i=1}^{w_{s'}} (s' + ir) - \sum_{i=1}^{w_s} (s + ir)\Bigr) - (s' - s) - r(w_{s'} - w_s) \\ &= 2\Bigl(\sum_{i=1}^{w_{s'}} (s' + ir) - \sum_{i=2}^{w_{s'}} (s + (i-1)r)\Bigr) - (s' + r - s) \\ &= 2(s' + r + (w_{s'} - 1)(s' + r - s)) - (s' + r - s) \\ &= 2(s + w_{s'}(s' + r - s)) - (s' + r - s) \\ &= 2s + (2w_{s'} - 1)(s' + r - s) \\ &\geq 2s + s' + r - s \\ &\geq (s' + s) + r \\ &\geq 1, \end{aligned}$$

where the second line is by column shift. In every inequality we have $M_{\mathrm{rel}}(y^{g-1}) > M_{\mathrm{rel}}(y^n)$. Then

$$\mathrm{Width}(\mathrm{MR}^{[s']}) \geq \mathrm{Width}(\mathrm{MR}_+^{[s']}) = M_{\mathrm{rel}}(y^{g-1}) + 1 > M_{\mathrm{rel}}(y^n) + 1 = \mathrm{Width}(\mathrm{MR}_+^{[s]}),$$

for each $[s] \in \mathrm{Spin}^c(S_r^3(K))$. This completes the $\tau(K) < 0$ subsubcase, and the proof of Lemma 4.2. □

Case A2 $(r \geq 2(g(K)-1))$ Recall that in this case we have $w_{s'} = 0$. Let us consider $r = 2g(K)-1$ first. When $\tau(K) \geq 0$, the surgery slope is large enough so that every intersection lies in the 0^{th} column of $\widetilde{T}$. Width alone as an invariant won't be enough, so we will also need to appeal to the multiplicities of the elements of the relative grading multisets. They will be used to show that only spin^c structures with the same parity are unobstructed. When we assume that $S_r^3(K)$ is reducible later on, the fact that r is odd will provide a contradiction with periodicity. When $\tau(K) < 0$, we need far less subtlety.

Lemma 4.3 *Suppose K is thin, $0 \leq \tau(K) < g(K)$, and $r = 2g(K)-1$. Then there exists an $[s'] \in \mathrm{Spin}^c(S_r^3(K))$ for which $[s] \neq [\pm s']$ satisfies $\mathrm{MR}^{[s]} \cong \mathrm{MR}^{[s']}$ up to translation only if* parity([s]) *equals* parity([s']).

Proof The spin^c structure $[s']$ we want to consider has $[s'] = g(K)-1$. Suppose for the sake of contradiction that some $[s] \neq [\pm s']$ satisfies $\mathrm{MR}^{[s]} \cong \mathrm{MR}^{[s']}$ up to translation and parity(s) $\neq$ parity(s'). We know that $r > 1$ forces $g(K) > 1$, and also that each l_r^s intersects $\widehat{\mathrm{HF}}(M)$ exactly once due to this large surgery slope. Because the choice of reference generator $a^{s'}$ depends on $\tau(K)$, let us split into two cases: $\tau(K) \geq 0$ and $\tau(K) < 0$.

Assume $\tau(K) \geq 0$. Because all intersections lie within the 0^{th} column of $\widetilde{T}$, we will instead use the hyperelliptic involution invariance of $\widehat{\mathrm{HF}}(M)$ to only consider $s \geq 0$. If $\widehat{\mathrm{HF}}(M)$ has no simple figure-eight at height s, then $\mathrm{Width}(\mathrm{MR}^{[s]}) = 0$ immediately does not match $\mathrm{Width}(\mathrm{MR}^{[s']}) \geq 1$, so we may as well assume that there is a simple figure-eight at height s. We have

$$M_{\mathrm{rel}}(y^{s'}) = 2H_{s'} - 1 - \tau(K) - s' = s' - 1 - \tau(K)$$

by Lemma 3.2 and Proposition 3.4, since $H_{s'} = s'$ when $\tau(K) \leq g(K) - 1 = s'$. Further,

$$\begin{aligned} M_{\mathrm{rel}}(y^{s'}) - M_{\mathrm{rel}}(y^{s}) &= 2H_{s'} - 1 - \tau(K) - s' - (2H_s - 1 - \tau(K) - s) \\ &= 2(H_{s'} - H_s) - (s' - s). \end{aligned}$$

If $s > \tau(K)$, then $H_s = s$ implies that $M_{\mathrm{rel}}(y^{s'}) - M_{\mathrm{rel}}(y^s) = s' - s \geq 1$. But then

$$\mathrm{Width}(\mathrm{MR}^{[s']}) = M_{\mathrm{rel}}(y^{s'}) + 1 > M_{\mathrm{rel}}(y^s) + 1 \geq \mathrm{Width}(\mathrm{MR}^{[s]}),$$

so $s \leq \tau(K)$ together with $\mathrm{Width}(\mathrm{MR}^{[s]}) = 1$. Notice that $\mathrm{Width}(\mathrm{MR}^{[s']}) = M_{\mathrm{rel}}(y^{s'}) + 1 = s' - \tau(K) > 1$ if $\tau(K) < s' - 1$, and so we are also forced to have either $\tau(K) = s' - 1$ or $\tau(K) = s'$. In both cases we have $\mathrm{Width}(\mathrm{MR}^{[s']}) = 1$. Since using width as an invariant has been exhausted, let us count multiplicities of elements of the $\mathrm{MR}^{[s]}$'s next.

Recall that e_n denotes the number of simple figure-eights at height n of $\widehat{\mathrm{HF}}(M)$. Further, we need $e_s = e_{s'}$ in order to have $|\mathrm{MR}^{[s]}| = |\mathrm{MR}^{[s']}|$. We have assumed that parity($[s]$) $\neq$ parity($[s']$), so one of these two multisets contains -1 and must be translated by 1 to make 0 the smallest element. This translated multiset will then contain 0 with multiplicity $e_{s'}$, while the other multiset will contain 0 with multiplicity $e_{s'}+1$. □

When $r = 2(g(K)-1)$, we will end up having $\mathrm{Width}(\mathrm{MR}^{[s]}) = 1$ for every $[s]$ if $\tau(K)$ is large enough. This means relative grading information alone will not be enough, and so we will return to such cases in Section 5.

Lemma 4.4 *Suppose K is thin, $0 \leq \tau(K) < g(K)-2$, and $r = 2(g(K)-1)$. Then there exists an $[s'] \in \mathrm{Spin}^c(S^3_r(K))$ for which every $[s] \neq [\pm s']$ satisfies $\mathrm{MR}^{[s]} \ncong \mathrm{MR}^{[s']}$ up to translation.*

Proof We again use $s' = g(K)-1$, and notice that when $\tau(K) < g(K)-2$, we have

$$M_{\mathrm{rel}}(y^{s'}) = 2(g(K)-1)-1-\tau(K)-(g(K)-1) = g(K)-2-\tau(K) > 0.$$

This shows that $\mathrm{Width}(\mathrm{MR}^{[s']}) = M_{\mathrm{rel}}(y^{s'})+1 > 1$. Any $[s] \neq [\pm s']$ with $|s| \leq \tau(K)$ has $\mathrm{Width}(\mathrm{MR}^{[s]}) = 1$ due to Proposition 3.3, so suppose $\tau(K) < |s| < s'$. In this case, $\mathrm{Width}(\mathrm{MR}^{[s]}) \leq M_{\mathrm{rel}}(y^s)+1$, but we also have $M_{\mathrm{rel}}(y^{s'}) - M_{\mathrm{rel}}(y^s) = s' - |s| > 0$. Then $\mathrm{Width}(\mathrm{MR}^{[s]}) < \mathrm{Width}(\mathrm{MR}^{[s']})$. □

When $\tau(K) < 0$, the fact that the reference intersection a^s lies outside of the neighborhood of $\tilde{\mu}_0$ is very convenient. This is an example of a *nonvertical intersection*, which is an intersection between l^s_r and $\bar{\gamma}$ that lies outside of a neighborhood of a lift $\tilde{\mu}$.

Lemma 4.5 *Suppose K is thin, $-g(K) < \tau(K) < 0$, and $r \geq 2(g(K)-1)$. Then there exists an $[s'] \in \mathrm{Spin}^c(S^3_r(K))$ for which every $[s] \neq [\pm s']$ satisfies $\mathrm{MR}^{[s]} \ncong \mathrm{MR}^{[s']}$ up to translation.*

Proof Since $w_{s'} = 0$, we again have $[s'] = g(K)-1$. Notice that each l^s_r gives rise to only two nonvertical intersections around the 0th column and intersections at height s when $[s] \neq [\pm s']$. We have s' maximal when $w_{s'} = 0$, so use hyperelliptic involution invariance to assume $0 \leq s < s'$. Recall that $\mathrm{Width}(\mathrm{MR}^{[s]}) = M_{\mathrm{rel}}(y^s)+1$ under the assumptions that $\tau(K) < 0$. The formula for $\mathrm{Wind}(P)$ does not depend on $\tau(K)$, which means

$$\begin{aligned} M_{\mathrm{rel}}(y^{s'}) - M_{\mathrm{rel}}(y^s) &= 2s'-1-\tau(K)-s'-(2s-1-\tau(K)-s) \\ &= s'-s. \end{aligned}$$

Then $\mathrm{Width}(\mathrm{MR}^{[s']}) = M_{\mathrm{rel}}(y^{s'})+1 > M_{\mathrm{rel}}(y^s)+1 = \mathrm{Width}(\mathrm{MR}^{[s]})$, which implies $\mathrm{MR}^{[s]} \ncong \mathrm{MR}^{[s']}$. □

In the following section we address the remaining cases involving $|\tau(K)| = g(K)$, as well as the few unresolved cases of this section. In particular, the cases with $g(K)-2 \leq \tau(K) < g(K)$ and $r = 2(g(K)-1)$ are handled in Lemma 5.5.

5 Remaining cases and absolute gradings

With the case analysis for $|\tau(K)| < g(K)$ out of the way, we turn to the more difficult part. As in Section 4, Figure 14 breaks down the upcoming case analysis.

Case B $(|\tau(K)| = g(K))$ When $|\tau(K)|$ is at its largest, the essential curve $\overline{\gamma}$ suffices to indicate $g(K)$ and we are not guaranteed a simple figure-eight at height $g(K)-1$. For these cases we still choose $[s']$ so that $g-1 \equiv s' \pmod{r}$ and continue to use w_s, except now modifying it to just be the largest multiple of r so that $s + w_s r < g(K)$. The $\tau(K) = -g(K)$ case is easier, so we start there.

Subcase B1 $(\tau(K) = -g(K))$

Lemma 5.1 *Suppose K is thin with $\tau(K) = -g(K)$, and let $1 < r \leq 2g(K)-1$. Then there exists an $[s'] \in \mathrm{Spin}^c(S^3_r(K))$ for which every $[s] \neq [\pm s']$ satisfies $\mathrm{MR}^{[s]} \ncong \mathrm{MR}^{[s']}$ up to translation.*

Proof Recall the labeling scheme from Figure 5, and both U_s and L_s defined just before Definition 3.1. The reference intersection a^s is a nonvertical intersection immediately to the left of the 0th column if $s \geq 0$, and is similarly immediately to the right of the 0th column if $s < 0$. In general we will label these generators x^l_s and x^r_s, respectively. Let us dispense with the $w_{s'} = 0$ case first.

Notice that each $\mathrm{MR}^{[s]}$ contains two elements whose difference is precisely $2|s|$. The two bigons we traverse from x^l_s to x^r_s involve the same regions and winding numbers as those in Figure 6 (Part c), and so $M_{\mathrm{rel}}(x^r_s) - M_{\mathrm{rel}}(x^l_s) = 2L_s - 1 - (2U_s - 1) = 2H_s = 2s$. We also see that $2L_s - 1 \leq \mathrm{Width}(\mathrm{MR}^{[s]}) \leq 2L_s$ if $s \leq 0$ and $2U_s - 1 \leq \mathrm{Width}(\mathrm{MR}^{[s]}) \leq 2U_s$ if $s > 0$, with the right-hand, even equalities achieved if an appropriate generator from a simple figure-eight exists at height s. So if some $[s] \neq [\pm s']$ is to achieve

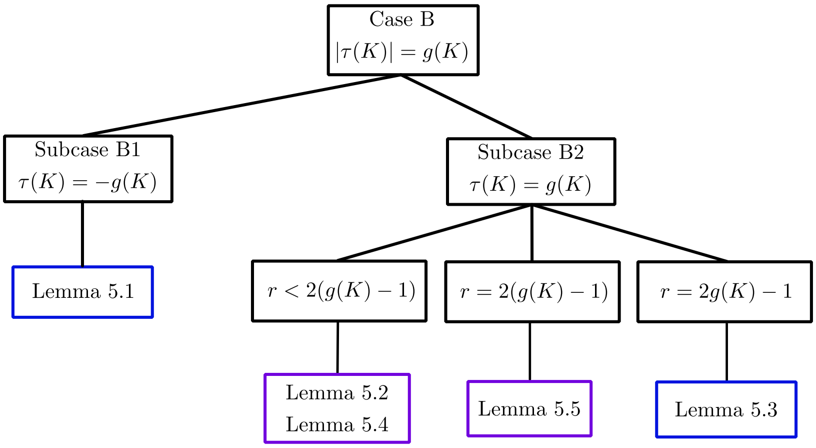


Figure 14: The case flowchart for the arguments in this section. As before, blue boxes indicate that the contained lemmas only appeal to relative grading information, while purple boxes use absolute gradings.

$\mathrm{MR}^{[s]} \cong \mathrm{MR}^{[s']}$ up to translation, we should see that the widths of these multisets agree and that there exist pairs with grading differences $2|s|$ and $2|s'|$. These are only possibly simultaneously true if $U_s = L_{s'}$ and $L_s = U_{s'}$, which forces $s = -s'$ with K thin. Thus, we may assume $w_{s'} > 0$.

If $w_{s'} > 0$, we can appeal to $\mathrm{MR}^{[s']}$ achieving maximal width. Due to the formula for $\mathrm{Wind}(P)$ when $\tau(K) < 0$, the grading difference between consecutive nonvertical intersections between $\overline{\gamma}$ and l_r^s around the i^{th} column is $2(s+ir)$. As before, this happens because $2L_{s+ir} - 1 - (2U_{s+ir} - 1) = 2H_{s+ir} = 2(s+ir)$, which is also positive. Then among nonnegative columns, the vertical intersection in final $(w_s)^{\text{th}}$ column, which we now denote by b^s, has the largest relative grading in $\mathrm{MR}^{[s]}$. This is because the differences around any given column are positive, and the vertical intersection on the right side of the final $(w_s)^{\text{th}}$ column necessarily has a smaller relative grading than b^s. The same reasoning applies to nonpositive columns (one way to see this is to appeal to hyperelliptic involution invariance to note that such a maximally graded vertical intersection belonging to a negative column is in correspondence to one belonging to a positive column associated to the conjugate spin^c structure). Thus, $\mathrm{Width}(\mathrm{MR}^{[s]})$ is either $M_{\mathrm{rel}}(b^s)$ or $M_{\mathrm{rel}}(b^{-s})$ when $w_s > 0$. We will obtain our desired contradiction by comparing $M_{\mathrm{rel}}(b^{s'})$ to every $M_{\mathrm{rel}}(b^s)$ with $s \neq \pm s'$, just as in the lemmas of the previous section.

Chaining the grading differences of vertical intersection pairs from b^s back to a^s, we see that

$$M_{\mathrm{rel}}(b^s) = \begin{cases} 2\Bigg(\displaystyle\sum_{i=0}^{w_s-1} (s+ir) + L_{s+w_s r}\Bigg) - 1 & \text{if } s \geq 0, \\ 2\Bigg(\displaystyle\sum_{i=1}^{w_s-1} (s+ir) + L_{s+w_s r}\Bigg) - 1 & \text{if } s \leq 0, \end{cases}$$

with empty sums taken to be zero as before. Since it can be hectic determining when such a sum is empty, we break into more cases.

When $s < s'$ we have $w_s = w_{s'}$, and it is straightforward to check that

$$M_{\mathrm{rel}}(b^{s'}) - M_{\mathrm{rel}}(b^s) \geq 2(L_{s'+w_{s'}r} - L_{s+w_{s'}r}) > 0.$$

This follows because the various multiples of $s' - s$ are positive if they appear, and because $L_{s'+w_{s'}r} > L_{s+w_{s'}r}$ when $s < s'$.

Let us begin the $s' < s$ cases with $w_{s'} = 1$. For $0 \leq s' < s$ we can once again use a column shift to see

$$M_{\mathrm{rel}}(b^{s'}) - M_{\mathrm{rel}}(b^s) = 2(s' + L_{s'+r} - L_s) > 0,$$

since $s' \geq 0$ and $L_{s'+r} > L_s$. The same inequality holds if $s' \leq 0 < s$, together with dropping the s' term. For $s' < s \leq 0$ with $w_s = 0$, we are forced to have $\mathrm{Width}(\mathrm{MR}^{[s]}) = 2L_s - 1$ if $s \geq 0$ and $\mathrm{Width}(\mathrm{MR}^{[s]}) = 2U_s - 1$ if $s \leq 0$, since $\mathrm{Width}(\mathrm{MR}^{[s']})$ is guaranteed to be odd. For the former we get

$$M_{\mathrm{rel}}(b^{s'}) - M_{\mathrm{rel}}(b^s) = 2(L_{s'+r} - L_s) > 0,$$

since $s < s' + r$. The latter yields

$$M_{\mathrm{rel}}(b^{s'}) - M_{\mathrm{rel}}(b^{s}) = 2(L_{s'+r} - U_s) > 0,$$

since $s > s'$.

Finally we are left with $w_{s'} > 1$ with $s' < s$ (which then implies $w_s = w_{s'} - 1$). If we have $0 \le s' < s$, then the fact that $L_{s'+w_{s'}r}$ is maximal ensures

$$\begin{aligned} M_{\mathrm{rel}}(b^{s'}) - M_{\mathrm{rel}}(b^{s}) &= 2\Biggl(\sum_{i=0}^{w_{s'}-1} (s' + ir) - \sum_{i=0}^{w_s-1} (s + ir)\Biggr) + 2(L_{s'+w_{s'}r} - L_{s+(w_{s'}-1)r}) \\ &= 2\Biggl(s' + \sum_{i=1}^{w_{s'}-1} (s' + ir) - \sum_{i=1}^{w_{s'}-1} (s + (i-1)r)\Biggr) + 2(L_{s'+w_{s'}r} - L_{s+(w_{s'}-1)r}) \\ &= 2s' + 2(w_{s'} - 1)(s' + r - s) + 2(L_{s'+w_{s'}r} - L_{s+(w_{s'}-1)r}) \\ &> 0, \end{aligned}$$

where the second line is by column shift. Analogously, the same inequality holds true if $s' \le 0 < s$ by dropping the $2s'$ term. For $s' < s \le 0$ a single $s' + r - s$ term disappears, but the inequality holds since $s' + r - s > 0$ and $L_{s+w_{s'}r} - L_{s+(w_{s'}-1)r} > 0$.

Then since $M_{\mathrm{rel}}(b^{s'}) > M_{\mathrm{rel}}(b^{s})$ for every configuration of s relative to s' for $w_{s'} > 0$, we have $\mathrm{Width}(\mathrm{MR}^{[s']}) > \mathrm{Width}(\mathrm{MR}^{[s]})$. Together with the argument for $w_{s'} = 0$, this completes the proof. □

Subcase B2 ($\tau(K) = g(K)$) Let us consider $1 < r < 2(g(K) - 1)$ first, delaying the penultimate slope to Lemma 5.5 and the maximal slope to Lemma 5.3. If $r < 2(g(K) - 1)$, then $l_r^{s'}$ intersects $\overline{\gamma}$ more than once for $s' \equiv g(K) - 1 \pmod r$. Our approach involves different arguments depending on whether $l_r^{s'}$ makes nonvertical intersections on both sides of the 0th column. Also, since $\tau(K)$ is positive recall that the reference intersection a^s is once again the vertical intersection belonging to the 0th column.

Lemma 5.2 *Suppose K is thin with $\tau(K) = g(K)$, the surgery slope satisfies $1 < r < 2(g(K) - 1)$, and that there exists a k properly dividing r so that every $[s] \in \mathrm{Spin}^c(S^3_r(K))$ satisfies $\mathrm{MR}^{[s]} \cong \mathrm{MR}^{[s+k]}$ up to translation.*

- *If $r < g(K) - 1$, then $\mathrm{MR}^{[s]} \cong \mathrm{MR}^{[s']}$ up to translation only if $[s] = [-s']$.*
- *If $r \ge g(K) - 1$, then $\tau(K) = g(K) = r = 3$.*

Proof If $r < g(K) - 1$, then the slope of $l_r^{s'}$ is small enough so that intersecting it with $\overline{\gamma}$ produces vertical intersections in at least 3 columns of $\widetilde{T}$. We know $w_{s'} > 0$ since $r < 2(g(K) - 1)$, so suppose $w_{s'} = 1$. We have nonvertical intersections with $\overline{\gamma}$ to the left and right of this column, which we can label $c^{s'}$ and $b^{s'}$, respectively. Then $M_{\mathrm{rel}}(c^{s'}) = 2V_{s'} - 1$ and $M_{\mathrm{rel}}(b^{s'}) = 2H_{s'} - 1$, and so $2H_{s'} - 1 \le \mathrm{Width}(\mathrm{MR}^{[s']}) \le 2H_{s'}$ since $s' > 0$ yields $H_{s'} > V_{s'}$. If some $[s] \ne [\pm s']$ satisfies $\mathrm{MR}^{[s]} \cong \mathrm{MR}^{[s']}$ up to translation then the

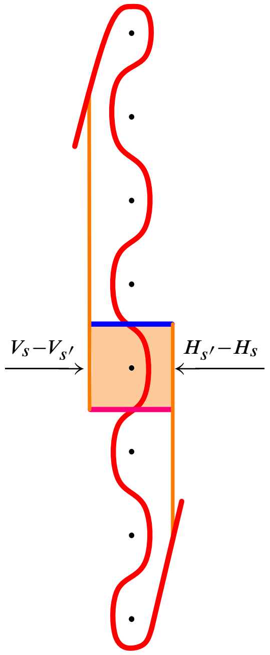


Figure 15: If $H_s = H_{s'}$ and $s = s' - 1$, then $V_s - V_{s'} = 1$ when K is thin.

parities of their widths must agree. Under the same labeling convention for intersections associated to $[s]$, we see that $2V_s - 1 \leq \mathrm{Width}(\mathrm{MR}^{[s]}) \leq 2V_s$ if $s \leq 0$ and $2H_s - 1 \leq \mathrm{Width}(\mathrm{MR}^{[s]}) \leq 2H_s$ if $s \geq 0$.

For $0 < s' < s$, we compute for either parity of width that

$$\mathrm{Width}(\mathrm{MR}^{[s']}) - \mathrm{Width}(\mathrm{MR}^{[s]}) = 2(H_{s'} - V_s).$$

This implies that $V_s = H_{s'}$, which is impossible when $s' > 0$. Similarly, if $s < -s'$ then the analogous statement holds true using H_s.

When $-s' < s < s'$, something interesting occurs. In addition to $l_r^{s'}$, we see that l_r^s successfully makes two nonvertical intersections on both sides of the 0th column. Also since K is thin, it follows that $V_s \leq H_{s'}$ and $H_s \leq H_{s'}$, with equality only possible when $s = -s' + 1$ or $s = s' - 1$, respectively. However, this results in the configuration shown for the latter situation in Figure 15. We see that while the widths of $\mathrm{MR}^{[s]}$ and $\mathrm{MR}^{[s']}$ can agree if $H_s = H_{s'}$, this necessarily results in $V_s \neq V_{s'}$ since K is thin. This is true vice versa as well, and so the multisets cannot both contain the same relative gradings for their respective vertical intersection pairs. Thus we must consider $w_{s'} > 1$, and we will do so following similar computations to those in Lemma 5.1.

When $w_{s'} > 1$ the intersection with largest relative grading in $\mathrm{MR}^{[s]}$ comes from the furthest nonvertical intersection, which we will update and label b^s when $s \geq 0$ or c^s if $s \leq 0$. Since we can use the hyperelliptic involution invariance of $\widehat{\mathrm{HF}}(M)$ to treat such a c^s as b^{-s}, let us only compare $M_{\mathrm{rel}}(b^{s'})$ to the various $M_{\mathrm{rel}}(b^s)$. Chaining grading differences between adjacent nonvertical intersections from b^s back to a^s, we have

$$M_{\mathrm{rel}}(b^s) = 2\Big(H_s + \sum_{i=1}^{w_s-1} (s+ir)\Big) - 1.$$

If $s < s'$ then $w_s = w_{s'}$, and we compute

$$\begin{aligned} M_{\mathrm{rel}}(b^{s'}) - M_{\mathrm{rel}}(b^s) &= 2\Big(H_{s'} - H_s + \sum_{i=1}^{w_s-1} (s'+ir) - \sum_{i=1}^{w_s-1} (s+ir)\Big) \\ &= 2(H_{s'} - H_s + (w_{s'}-1)(s'-s)) \\ &\geq 1. \end{aligned}$$

We have equality only if $H_{s'} = H_s$, $s = s'-1$, and $w_{s'} = 2$. In this case, we have $\mathrm{parity}(s'+2r) = \mathrm{parity}(s')$ regardless of r. However $\mathrm{parity}(s'+2r) \neq \mathrm{parity}(\tau(K))$, and so $\mathrm{parity}(s) = \mathrm{parity}(\tau(K))$. This implies $\mathrm{Width}(\mathrm{MR}^{[s]}) \leq M_{\mathrm{rel}}(b^s)$, which means we cannot have $s < s'$.

If $s > s'$, then with $w_s = w_{s'} - 1$ we obtain

$$\begin{aligned} M_{\mathrm{rel}}(b^{s'}) - M_{\mathrm{rel}}(b^s) &= 2\Big(H_{s'} - H_s + \sum_{i=1}^{w_{s'}-1} (s'+ir) - \sum_{i=1}^{w_s-1} (s+ir)\Big) \\ &= 2\Big(s'+r + H_{s'} - H_s + \sum_{i=2}^{w_{s'}-1} (s+ir) - \sum_{i=2}^{w_{s'}-1} (s+ir)\Big) \\ &= 2(s'+r) - 2(H_s - H_{s'}) + 2(w_{s'}-2)(s'+r-s), \end{aligned}$$

where the second line is by column shift. Now $s - s' \leq 2(H_s - H_{s'}) \leq s - s' + 1$ when K is thin by careful inspection of these regions. This implies

$$\begin{aligned} M_{\mathrm{rel}}(b^{s'}) - M_{\mathrm{rel}}(b^s) &= 2(s'+r) - 2(H_s - H_{s'}) + 2(w_{s'}-2)(s'+r-s) \\ &\geq 2(s'+r) - (s-s'+1) + 2(w_{s'}-2)(s'+r-s) \\ &= 2s' + r - 1 + (2w_{s'}-3)(s'+r-s) \\ &> 1. \end{aligned}$$

Altogether, these grading comparisons are enough to see that $\mathrm{MR}^{[s]} \not\cong \mathrm{MR}^{[s']}$ up to translation when $r < g(K) - 1$. Next we look at the cases with larger surgery slopes.

If $r = g(K) - 1$, then $w_{s'} = 1$ and $s' = 0$. Due to hyperelliptic involution invariance, we can assume $s \neq s'$ satisfies $s < 0$. Notice that $\mathrm{MR}^{[s']}$ contains $2H_{s'} - 1$ with multiplicity at least two since $l_r^{s'}$ generates nonvertical intersections on both sides of the 0th column and $V_{s'} = H_{s'}$. The only way that $\mathrm{MR}^{[s]}$ could contain this grading with multiplicity greater than one is if a simple figure-eight component in the 1st column has an intersection with l_r^s. The nearby vertical intersection a^{s+r} has $M_{\mathrm{rel}}(a^{s+r}) = 2H_{s'} - 2$, and so we would require $\mathrm{parity}(s+r) \neq \mathrm{parity}(\tau(K))$ in order for an intersection with a simple figure-eight to have the desired grading. However $s + r = \tau(K) - 2$, and so $\mathrm{MR}^{[s]}$ cannot contain $2H_{s'} - 1$ more than once. Thus, no $[s] \neq [s']$ satisfies $\mathrm{MR}^{[s]} \cong \mathrm{MR}^{[s']}$ up to translation when $r = g(K) - 1$.

We still have $w_{s'} = 1$ if $r > g(K) - 1$, but now $s' < 0$. The crux of the argument in the previous case relied on $H_{s'} > 1$. This holds more generally when $r < 2g(K) - 3$, except now $\mathrm{MR}^{[s']}$ need only contain $2H_{s'} - 1$ once. Since $\mathrm{Width}(\mathrm{MR}^{[s']}) \geq 3$, the above argument still applies to show that $\mathrm{MR}^{[s]} \cong \mathrm{MR}^{[s']}$ up to translation only if $[s] = [s' \pm 1]$. This forces $k = 1$, which in turn forces $r \leq 3$ so that there cannot exist $[s' + \alpha k]$ with $\mathrm{Width}(\mathrm{MR}^{[s'+\alpha k]}) = 1$. The possibility $r = 2$ is handled exactly as in the $r = g(K) - 1$ argument, and so we must have $r = 3$. This means $s' = -1$, and so $\tau(K) = g(K) = r = 3$.

If $r = 2g(K) - 3$, then only $l_r^{\pm s'}$ generates nonvertical intersections between columns of $\widetilde{T}$. All other l_r^s intersect $\widehat{\mathrm{HF}}(M)$ only in the 0^{th} column, which means that every $\mathrm{Width}(\mathrm{MR}^{[s]}) = 1$. Since $\mathrm{parity}(s') = \mathrm{parity}(\tau(K))$ when $r = 2g(K) - 3$, we must have $e_{s'} = 0$. This is because a simple figure-eight component at this height would contribute an intersection with relative grading -1 to $\mathrm{MR}^{[s']}$, which would yield $\mathrm{Width}(\mathrm{MR}^{[s']}) = 2$ and prevent periodicity. We have $\dim \widehat{\mathrm{HF}}(S_r^3(K), [s']) = 3 + 2e_{s'+r}$, and so some $[s' + k]$ satisfying $\mathrm{MR}^{[s'+k]} \cong \mathrm{MR}^{[s']}$ up to translation forces $1 + 2e_{s'+k} = 3 + 2e_{s'+r}$, or $e_{s'+k} = e_{s'+r} + 1$. If necessary, translate $\mathrm{MR}^{[s'+k]}$ so that 0 is the smallest element. The only way that the multiplicities of 0 and 1 agree is if $\mathrm{MR}^{[s]}$ contains more 0's than 1's, which happens only when $\mathrm{parity}(s' + k) = \mathrm{parity}(\tau(K))$. But $\mathrm{parity}(s') = \mathrm{parity}(\tau(K))$ as well, which is a contradiction since k is odd when r odd. □

We return to the two unhandled cases of $\tau(K) = g(K) = r = 3$ and $r = 2(g(K) - 1)$ shortly in Section 5.1, and for now are left with the case where $r = 2g(K) - 1$. Because $\tau(K) = g(K)$, there is no guaranteed simple figure-eight at height $g(K) - 1$. This small difference is enough of an issue if K is an L-space knot, since each $\mathrm{MR}^{[s]} = \{0\}$ means $\widehat{\mathrm{HF}}$ cannot provide an obstruction. With existing techniques, we can only show the following:

Lemma 5.3 *Suppose K is thin with $\tau(K) = g(K)$, and let $r = 2g(K) - 1$. If there exists a k properly dividing r such that every $[s] \in \mathrm{Spin}^c(S_r^3(K))$ satisfies $\widehat{\mathrm{HF}}(S_r^3(K), [s]) \cong \widehat{\mathrm{HF}}(S_r^3(K), [s + k])$, then K is an L-space knot.*

Proof Suppose for the sake of contradiction that K is not an L-space knot, meaning that

$$\dim \widehat{\mathrm{HF}}(S_r^3(K), [s]) > 1$$

for some $[s] \in \mathrm{Spin}^c(S_r^3(K))$. Each l_r^s intersects $\bar{\gamma}$ precisely once since $r \geq 2g(K) - 1$. In order to have $\dim \widehat{\mathrm{HF}}(S_p^3(K), [s]) > 1$ for some $[s]$, we need for $\widehat{\mathrm{HF}}(M)$ to have a simple figure-eight component at height s. Let t be the height of the lowest simple figure-eight component. We have e_t many simple figure-eights at height t, and so we must also have $e_{t+k} = e_t$ many simple figure-eight components at height $t + k$ to satisfy

$$\dim \widehat{\mathrm{HF}}(S_r^3(K), [t]) = \dim \widehat{\mathrm{HF}}(S_r^3(K), [t + k]).$$

If $\mathrm{parity}([t]) \neq \mathrm{parity}([t + k])$, then one of $\mathrm{MR}^{[t]}$ or $\mathrm{MR}^{[t+k]}$ contains -1 and would need to be translated by 1 to make 0 the smallest element by Proposition 3.3. However, this results in both multisets having unequal multiplicities of 0's and 1's. This would lead to $\mathrm{MR}^{[t]} \ncong \mathrm{MR}^{[t+k]}$ up to translation, and so we must have $\mathrm{parity}([t+k]) = \mathrm{parity}([t])$. However this condition implies that k is even, which contradicts $r = 2g(K) - 1$ being odd. Therefore, K must be an L-space knot. □

5.1 Obstructions from absolute gradings

Until now, we have primarily appealed to information carried by $\widehat{\mathrm{HF}}(S^3_r(K))$ as this is the easier flavor of Heegaard Floer homology computable by immersed curves techniques. However, we now need to involve the absolutely, $\mathbb{Q}$-graded, $+$-flavor of Heegaard Floer homology to handle the remaining cases. When considering $S^3_r(K) = Y \# Z$ with $|H^2(Y)| = k < \infty$, we will use properties of the d-invariants mentioned in Section 2 in order to obtain a relationship between r, k, and the V's associated to K. We initially settle the curious $\tau(K) = g(K) = r = 3$ case, and afterwards assemble the proof of Theorem 1.2.

Lemma 5.4 *Let K be a thin knot with $\tau(K) = g(K) = 3$. Then $S^3_3(K)$ is irreducible.*

Proof If $S^3_3(K)$ is reducible, it must admit an integer homology sphere connected summand Y since $r = 3$ is prime. Using the additivity of the d-invariants, we have

$$d(S^3_3(K), [s]) = d(L(3, \pm 1), [s]) + d(Y).$$

Proposition 2.3 then implies $d(Y) = -2V_0(K) = -2V_1(K)$, which in turn forces $V_0(K) = V_1(K)$. However this is true only for thin knots with even $\tau(K)$, which can be seen using the formula in Proposition 3.3 together with $V_s = H_{-s}$. This forms the desired contradiction. □

Lemma 5.5 *Let K be a thin knot with $\tau(K) \geq g(K) - 2$. Then $S^3_r(K)$ is irreducible when $r = 2(g(K) - 1)$.*

Proof Let $\tau(K) \geq g(K) - 2$, and suppose for the sake of contradiction that $S^3_r(K)$ is reducible for $r = 2(g(K) - 1)$. Then $S^3_r(K)$ admits as connect summands a lens space Y and a summand Z with $|H^2(Z)| = k < \infty$. Since $H_1(S^3_r(K))$ is cyclic and r is even, one of $|H_1(Y)| = \frac{r}{k}$ or k is even. We will show the latter must be true.

Using the immersed curves techniques of the previous section, we see that $\mathrm{Width}(\mathrm{MR}^{[s]}) = 1$ for all $[s]$ when K is thin and $\tau(K) \geq g(K) - 2$. Using $s' \equiv g(K) - 1 \pmod{r}$ again, we are guaranteed to have $\dim \widehat{\mathrm{HF}}(S^3_r(K), [s']) > 1$ since $l^{s'}_r$ either intersects a simple figure-eight at height $g(K) - 1$ when $\tau(K) < g(K)$ or intersects $\overline{\gamma}$ multiple times when $\tau(K) = g(K)$. In order for some $[s' - k]$ to satisfy $\mathrm{MR}^{[s'-k]} \cong \mathrm{MR}^{[s']}$ up to translation, we also require $\mathrm{parity}(s' - k) = \mathrm{parity}(s')$ so that the multiplicities of 0 and 1 agree. This implies k is even.

Let $\pi_Y([s])$ and $\pi_Z([s])$ denote the restrictions of $[s]$ to $\mathrm{Spin}^c(Y)$ and $\mathrm{Spin}^c(Z)$, respectively. Since $\mathrm{Spin}^c(S^3_r(K)) \cong \mathbb{Z}/r\mathbb{Z}$ is $\mathbb{Z}/r\mathbb{Z}$-equivariant [35], we have both

$$\pi_Y\left(\left[s + \frac{r}{k}\right]\right) = \pi_Y([s]) \quad \text{and} \quad \pi_Z([s+k]) = \pi_Z([s]).$$

The two self-conjugate spin^c structures of $S^3_r(K)$ must project onto the lone self-conjugate structure of $L\left(\frac{r}{k},q\right)$, and so

$$\pi_Y([0]) = \pi_Y\left(\left[\frac{r}{2}\right]\right) \in \mathrm{Spin}^c\left(L\left(\frac{r}{k},q\right)\right).$$

Their respective restrictions on Z are distinct, and so let $\pi_Z([0]) = u_e$ and $\pi_Z\left(\left[\frac{r}{k}\right]\right) = u_o$ (subscripts indicate parity of the spin^c structure before restriction). Due to the additivity of d-invariants, we have

$$d(S^3_r(K),[s]) = d\left(L\left(\frac{r}{k},q\right),\pi_Y([s])\right) + d(Z,\pi_Z([s])).$$

Since k is even, we may apply this to the self-conjugate structures to see

$$\begin{aligned} d(S^3_r(K),[0]) - d\left(S^3_r(K),\left[\frac{r}{2}\right]\right) &= \left(d\left(L\left(\frac{r}{k},q\right),[0]\right) + d(Z,u_e)\right) - \left(d\left(L\left(\frac{r}{k},q\right),[0]\right) + d(Z,u_o)\right) \\ &= d(Z,u_e) - d(Z,u_o). \end{aligned}$$

Observe that $\pi_Z\left(\left[\frac{k}{2}\right]\right) = u_0$, and so $\pi_Z\left(\left[\frac{r+k}{2}\right]\right) = u_e$. We likewise have $\pi_Y\left(\left[\frac{k}{2}\right]\right) = \pi_Y\left(\left[\frac{r+k}{2}\right]\right)$, and so

$$d\left(S^3_r(K),\left[\frac{k}{2}\right]\right) - d\left(S^3_r(K),\left[\frac{r+k}{2}\right]\right) = d(Z,u_o) - d(Z,u_e).$$

Using the inductive formula for $d(L(p,q))$ [27, Proposition 4.8], it follows that

$$d(L(r,1),[s]) = \frac{s^2}{r} - s + \frac{r-1}{4}.$$

Summing the prior two equations and using Proposition 2.3 (with $V_{\frac{r-k}{2}} = \max\{V_{\frac{r+k}{2}}, V_{r-\frac{r+k}{2}}\}$) yields

$$\begin{aligned} 2(V_0 - V_{\frac{r}{2}} + V_{\frac{k}{2}} - V_{\frac{r-k}{2}}) &= d(L(r,1),[0]) - d\left(L(r,1),\left[\frac{r}{2}\right]\right) + d\left(L(r,1),\left[\frac{k}{2}\right]\right) - d\left(L(r,1),\left[\frac{r+k}{2}\right]\right) \\ &= -\left(\frac{r^2}{4r} - \frac{r}{2}\right) + \left(\frac{k^2}{4r} - \frac{k}{2}\right) - \left(\frac{(r+k)^2}{4r} - \frac{r+k}{2}\right) \\ &= \frac{r-k}{2}, \end{aligned}$$

Therefore, we have the following relationship between r, k, and the V's associated to K:

$$\frac{r-k}{4} = (V_0 - V_{\frac{r}{2}}) + (V_{\frac{k}{2}} - V_{\frac{r-k}{2}}). \tag{1}$$

Notice that when K is thin and $\tau(K) \geq 0$, we have that

$$V_0 = \begin{cases} \dfrac{\tau(K)+1}{2} & \text{if parity}(\tau(K)) = 1, \\ \dfrac{\tau(K)}{2} & \text{if parity}(\tau(K)) = 0. \end{cases}$$

We will use this to generate contradictions, and break into cases since the values of $V_{\frac{r}{2}}$ and $V_{\frac{r-k}{2}}$ depend on $\tau(K)$. It will also be useful to use the fact that $k \leq \frac{r}{3}$.

Case C1 $(\tau(K) = g(K))$ Here we have $V_{\frac{r}{2}} = 1$ and $V_{\frac{r-k}{2}} > 0$. We see that $V_{\frac{k}{2}} - V_{\frac{r-k}{2}}$ is given by half the distance between $\frac{r-k}{2} - \frac{k}{2}$ since K is thin. Thus,

$$V_{\frac{k}{2}} - V_{\frac{r-k}{2}} = \frac{r}{4} - \frac{k}{2}.$$

This together with (1) above then yields

$$V_0 = \frac{k}{4} + 1.$$

If $\text{parity}(\tau(K)) = 1$, then we have

$$\begin{aligned}\frac{\tau(K)+1}{2} = \frac{k}{4} + 1 \;&\Leftrightarrow\; \frac{\tau(K)-1}{2} = \frac{k}{4}\\ &\Leftrightarrow\; \frac{\tau(K)-1}{2} \le \frac{r}{12}\\ &\Leftrightarrow\; 6(\tau(K)-1) \le 2(g(K)-1)\\ &\Leftrightarrow\; 6(\tau(K)-1) \le 2(\tau(K)-1)\\ &\Rightarrow\; 4\tau(K) \le 4.\end{aligned}$$

However this would imply $g(K) = \tau(K) = 1 \Rightarrow r = 0$, a clear contradiction. If $\text{parity}(\tau(K)) = 0$, then similar reasoning yields $\tau(K) \le 2$, which forces $r = \tau(K) = g(K) = 2$. We return to immersed curves techniques to rule out this case by comparing the multiplicity of elements of $\text{MR}^{[0]}$ and $\text{MR}^{[1]}$. Since $\text{parity}(\tau(K)) = 0$, we must translate $\text{MR}^{[0]}$ by one so that 0 is its smallest element. The multiplicity of 0 in the translated $\text{MR}^{[0]}$ is e_0, and the multiplicity of 0 in $\text{MR}^{[1]}$ is $2e_1 + 2$. However if $S^3_2(K)$ is reducible then Lemma 2.1 forces $e_0 = 2e_1 + 1$ in order for $\dim \widehat{\text{HF}}(S^3_2(K), [1]) = \dim \widehat{\text{HF}}(S^3_2(K), [0])$, generating the desired contradiction.

Case C2 $(\tau(K) = g(K) - 1)$ Once again $V_{\frac{r-k}{2}} > 0$, and in this case we obtain $V_0 = \frac{k}{4}$ since $V_{\frac{r}{2}} = 0$. Together with $r = 2\tau(K)$, the argument of the previous case yields the contradiction $4\tau(K) \le -4$ when $\text{parity}(\tau(K)) = 1$ or $4\tau(K) \le 2$ when $\text{parity}(\tau(K)) = 0$.

Case C3 $(\tau(K) = g(K) - 2)$ We still have $V_{\frac{r}{2}} = 0$, and things are more interesting here since it is possible for $V_{\frac{r-k}{2}} = 0$. This happens only if $k = 2$, in which case (1) becomes

$$\frac{r-2}{4} = V_0 + V_1.$$

Curiously enough $\tau(K) = V_0 + V_1$ for a thin knot, and so this would force $\tau(K) = \frac{r-2}{4} = \frac{2(\tau(K)+1)-2}{4} \Leftrightarrow 4\tau(K) = 2\tau(K) \Rightarrow \tau(K) = 0$. However this forces $r = k$, a contradiction. Then we cannot have $k = 2$, and so $V_{\frac{r-k}{2}} > 0$ and we have $V_0 = \frac{k}{4}$. As with the previous cases, having $r = 2(\tau(K) + 1)$ would yield the contradictions $6(\tau(K)+1) \le 2(\tau(K)+1)$ if $\text{parity}(\tau(K)) = 1$ or $4\tau(K) \le 2$ if $\text{parity}(\tau(K)) = 0$. □

5.2 Proof of Theorem 1.2

Proof Suppose $S^3_r(K)$ is reducible for K thin and hyperbolic. The Matignon–Sayari bound implies that $1 < r \le 2g(K) - 1$ [23, Theorem 1.1], after mirroring the knot if necessary to make the surgery slope positive. Reducibility also gives $S^3_r(K) \cong Y \# Z$ for some lens space Y and some Z with $|H^2(Z)| = k < \infty$. By Lemma 2.1, we have $\widehat{\text{HF}}(S^3_r(K), [s + \alpha k]) \cong \widehat{\text{HF}}(S^3_r(K), [s])$ for arbitrary $[s], \alpha \in \mathbb{Z}/r\mathbb{Z}$. When $r < 2(g(K) - 1)$, Lemmas 4.2, 5.1, and 5.2 apply to show that there exists an $[s'] \in \text{Spin}^c(S^3_r(K))$ such that either $\widehat{\mathrm{HF}}(S^3_r(K),[s'])$ is relatively graded isomorphic only to $\widehat{\mathrm{HF}}(S^3_r(K),[-s'])$, or $\tau(K)=g(K)=r=3$. The latter is prevented by Lemma 5.4, so we proceed with the former. Since $Y \not\cong S^3$, we see that $[s']$ cannot be self-conjugate and also that $|H_1(Y)|=2$. This implies $Y=\mathbb{R}P^3$, as well as $k=|[s']-[-s']|=2\,|[s']|$. However, together this means 4 divides r, which is impossible when $S^3_r(K)$ admits an $\mathbb{R}P^3$ summand with $H_1(S^3_r(K))$ cyclic.

Therefore, we must have $r \geq 2(g(K)-1)$. Lemmas 4.4 and 4.5 cover $-g(K)<\tau(K)<g(K)-2$ and Lemma 5.5 covers $\tau(K)\geq g(K)-2$ for the possibility that $r=2(g(K)-1)$, and so we must have $r=2g(K)-1$. In this situation k is odd since r is odd, which means periodicity will cycle through spin^c structures with different parities. Then Lemmas 4.3, 4.5, and 5.1 apply to fully obstruct reducibility via the above argument if $\tau(K)\neq g(K)$. If $\tau(K)=g(K)$, our techniques have been exhausted and leave just the conclusion of Lemma 5.3, showing that K must be an L-space knot. □

References

[1] **H Bodish**, **R DeYeso III**, *Obstructing reducible surgeries: slice genus and thickness bounds*, preprint (2022) arXiv 2209.01672

[2] **S Boyer**, *Dehn surgery on knots*, from “Handbook of geometric topology”, North-Holland, Amsterdam (2002) 165–218 MR

[3] **S Boyer**, **X Zhang**, *Reducing Dehn filling and toroidal Dehn filling*, Topology Appl. 68 (1996) 285–303 MR

[4] **S Boyer**, **X Zhang**, *On Culler–Shalen seminorms and Dehn filling*, Ann. of Math. 148 (1998) 737–801 MR

[5] **S Dey**, *On Heegaard Floer theoretic properties of knots*, PhD thesis, State University of New York at Buffalo (2021) MR Available at `https://www.proquest.com/docview/2581855057`

[6] **M Eudave Muñoz**, *Band sums of links which yield composite links: the cabling conjecture for strongly invertible knots*, Trans. Amer. Math. Soc. 330 (1992) 463–501 MR

[7] **D Gabai**, *Foliations and the topology of* 3*-manifolds, III*, J. Differential Geom. 26 (1987) 479–536 MR

[8] **F González-Acuña**, **H Short**, *Knot surgery and primeness*, Math. Proc. Cambridge Philos. Soc. 99 (1986) 89–102 MR

[9] **C M Gordon**, **J Luecke**, *Only integral Dehn surgeries can yield reducible manifolds*, Math. Proc. Cambridge Philos. Soc. 102 (1987) 97–101 MR

[10] **C M Gordon**, **J Luecke**, *Knots are determined by their complements*, J. Amer. Math. Soc. 2 (1989) 371–415 MR

[11] **J E Greene**, *L-space surgeries, genus bounds, and the cabling conjecture*, J. Differential Geom. 100 (2015) 491–506 MR

[12] **C M Grove**, *A combinatorial approach to the cabling conjecture*, PhD thesis, The University of Iowa (2016) MR Available at `https://www.proquest.com/docview/1824051835`

[13] **J Hanselman**, *Heegaard Floer homology and cosmetic surgeries in S^3*, J. Eur. Math. Soc. (JEMS) 25 (2023) 1627–1669 MR

[14] **J Hanselman**, *Knot Floer homology as immersed curves*, preprint (2023) arXiv 2305.16271

[15] **J Hanselman**, **J Rasmussen**, **L Watson**, *Heegaard Floer homology for manifolds with torus boundary: properties and examples*, Proc. Lond. Math. Soc. 125 (2022) 879–967 MR

[16] **J Hanselman**, **J Rasmussen**, **L Watson**, *Bordered Floer homology for manifolds with torus boundary via immersed curves*, J. Amer. Math. Soc. 37 (2024) 391–498 MR

[17] **J Hanselman**, **L Watson**, *A calculus for bordered Floer homology*, Geom. Topol. 27 (2023) 823–924 MR

[18] **C Hayashi**, **K Shimokawa**, *Symmetric knots satisfy the cabling conjecture*, Math. Proc. Cambridge Philos. Soc. 123 (1998) 501–529 MR

[19] **J Hom**, *A survey on Heegaard Floer homology and concordance*, J. Knot Theory Ramifications 26 (2017) art. id. 1740015 MR

[20] **J Hom**, **T Lidman**, **N Zufelt**, *Reducible surgeries and Heegaard Floer homology*, Math. Res. Lett. 22 (2015) 763–788 MR

[21] **J Howie**, *A proof of the Scott–Wiegold conjecture on free products of cyclic groups*, J. Pure Appl. Algebra 173 (2002) 167–176 MR

[22] **R Lipshitz**, **P S Ozsvath**, **D P Thurston**, *Bordered Heegaard Floer homology*, Mem. Amer. Math. Soc. 1216, Amer. Math. Soc., Providence, RI (2018) MR

[23] **D Matignon**, **N Sayari**, *Longitudinal slope and Dehn fillings*, Hiroshima Math. J. 33 (2003) 127–136 MR

[24] **W W Menasco**, **M B Thistlethwaite**, *Surfaces with boundary in alternating knot exteriors*, J. Reine Angew. Math. 426 (1992) 47–65 MR

[25] **L Moser**, *Elementary surgery along a torus knot*, Pacific J. Math. 38 (1971) 737–745 MR

[26] **Y Ni**, **Z Wu**, *Cosmetic surgeries on knots in S^3*, J. Reine Angew. Math. 706 (2015) 1–17 MR

[27] **P Ozsváth**, **Z Szabó**, *Absolutely graded Floer homologies and intersection forms for four-manifolds with boundary*, Adv. Math. 173 (2003) 179–261 MR

[28] **P Ozsváth**, **Z Szabó**, *Heegaard Floer homology and alternating knots*, Geom. Topol. 7 (2003) 225–254 MR

[29] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and genus bounds*, Geom. Topol. 8 (2004) 311–334 MR

[30] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and knot invariants*, Adv. Math. 186 (2004) 58–116 MR

[31] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and three-manifold invariants: properties and applications*, Ann. of Math. 159 (2004) 1159–1245 MR

[32] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158 MR

[33] **P Ozsváth**, **Z Szabó**, *On knot Floer homology and lens space surgeries*, Topology 44 (2005) 1281–1300 MR

[34] **P Ozsváth**, **Z Szabó**, *On the Heegaard Floer homology of branched double-covers*, Adv. Math. 194 (2005) 1–33 MR

[35] **P S Ozsváth**, **Z Szabó**, *Knot Floer homology and integer surgeries*, Algebr. Geom. Topol. 8 (2008) 101–153 MR

[36] **I Petkova**, *Cables of thin knots and bordered Heegaard Floer homology*, Quantum Topol. 4 (2013) 377–409 MR

[37] **J A Rasmussen**, *Floer homology and knot complements*, PhD thesis, Harvard University (2003) MR Available at `https://www.proquest.com/docview/305332635`

[38] **N Sayari**, *The reducibility of surgered* 3*-manifolds and homology* 3*-spheres*, Topology Appl. 87 (1998) 73–78 MR

[39] **M Scharlemann**, *Producing reducible* 3*-manifolds by surgery on a knot*, Topology 29 (1990) 481–500 MR

[40] **L G Valdez Sánchez**, *Dehn fillings of* 3*-manifolds and non-persistent tori*, Topology Appl. 98 (1999) 355–370 MR

Department of Mathematics & Statistics, The University of Tennessee at Martin
Martin, TN, United States

`rdeyeso1@utm.edu`

Received: 26 April 2022 Revised: 21 July 2024

Algebraic & Geometric Topology **25**:8 (2025) 4585–4598
DOI: 10.2140/agt.2025.25.4585
Published: 20 November 2025

Linear linkless embeddings: proof of a conjecture by Sachs

LYNN STANFIELD

In 1983, Sachs conjectured that every linklessly embeddable graph has a linear linkless embedding. We prove a stronger statement: every flat embedding of a linkless graph can be linearized.

05C10

1 Introduction

A spatial graph is an embedding of a graph into $\mathbb{R}^3$. Conway and Gordon [2], and Sachs [6], introduced the theory of spatial graphs by showing every embedding of the complete graph K_6 into $\mathbb{R}^3$ contains two cycles which are linked, and every embedding of K_7 into $\mathbb{R}^3$ contains a cycle which is a nontrivial knot. A graph which has a nontrivial link in every embedding is called *intrinsically linked*, and if a graph is not intrinsically linked, it is called *linklessly embeddable*. A significant amount of work followed, including the characterization of linklessly embeddable graphs by Robertson, Seymour, and Thomas [5]: A graph is linklessly embeddable if and only if it does not have a Petersen family minor. These graphs are obtained from the complete graph K_6 by performing a sequence of $Y-\Delta$ and $\Delta-Y$ transforms, as shown in Figure 1.

We are concerned with *linear embeddings*, which are embeddings into $\mathbb{R}^3$ where every edge is a straight line segment. In 1948, Fáry showed the following:

Theorem 1.1 (Fáry's Theorem; see Fáry [3], Stein [8], and Wagner [9]) *All planar graphs have a planar embedding with all edges straight line segments.*

From this, we can easily provide linear linkless embeddings of all apex graphs, which are planar after the removal of some vertex. In 1983, Sachs conjectured every linklessly embeddable graph has a linear linkless embedding [6]. Here we show a stronger statement: every flat embedding can be linearized.

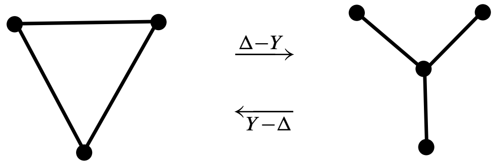


Figure 1: The $Y\Delta$ and ΔY transforms.

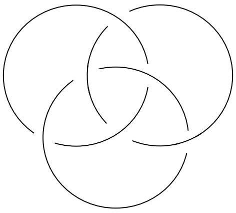


Figure 2: Borromean rings.

Definition 1.2 An embedding ϕ of a graph G is *flat* (*or paneled*) if every cycle $C \in \phi$ bounds a disk with interior disjoint from ϕ.

It is clear from this definition that every flat embedding ϕ must be linkless, as every cycle bounds a disk disjoint from ϕ, but the converse is not true. For example, consider three cycles embedded as the Borromean rings, shown in Figure 2. No two cycles are linked, so this is a linkless embedding. However, it is not flat, since no cycle can bound a disk disjoint from the graph.

However, Robertson, Seymour, and Thomas [5] have shown the following result:

Proposition 1.3 [5] *Every linklessly embeddable graph admits a flat embedding.*

Because of this, we will only work with flat embeddings.

It is very important to clarify how deletions and contractions apply to embeddings. For an embedding ϕ, the edge deletion $\phi\backslash e$ equals $\phi|_{G\backslash e}$. For a vertex deletion, $\phi\backslash v = \phi|_{G\backslash v}$. When we contract an edge e in the graph G, we delete double edges since we want our graph to remain simple.

However, for a topological contraction in the embedding ϕ, we move the ends of the edge towards some point in e and route all edges from its endpoints along the path of the edge e. A contraction of an embedding may contain double edges, and further contractions may create loops. See Figure 3 for an example of a contraction of an embedding which contains double edges. We denote this new embedding by ϕ/e.

In their survey paper, Robertson, Seymour, and Thomas show the following important result:

Proposition 1.4 [5] *An embedding ϕ is flat if and only if, for e a nonloop edge, both embeddings ϕ/e and $\phi\backslash e$ are flat.*

We will call back to this result later. The proposition is discussed in the referenced survey paper.

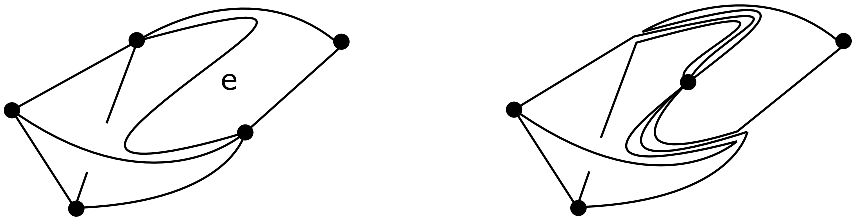


Figure 3: Topological contraction of the edge e.

2 Preliminary lemmas and constructions

The following lemma of Böhme is essential for our further constructions:

Lemma 2.1 [1] *Let ϕ be a flat embedding of a graph G, and $C_1, C_2, \dots, C_k$ be cycles in ϕ. If $C_i \cap C_j$ is connected or empty for all i and j, then there exist disks $\Gamma_1, \dots, \Gamma_k$ such that*

(1) *Γ_i is bounded by C_i,*

(2) *$\Gamma_i \cap \phi = C_i$, and*

(3) *Γ_i and Γ_j have disjoint interiors for all i and j.*

From this point on we will fix G, a linklessly embeddable graph. For an edge $e = xy \in G$, let $M(e) = \{v \in V(G) : v \text{ is adjacent to } x \text{ and } y\}$. We make the following claim:

Lemma 2.2 *Let ϕ be a flat embedding of G. Then for any edge $xy \in G$, there exists an embedded closed ball B_{xy} such that*

(1) *x and y are contained in the interior of B_{xy},*

(2) *all other neighbors of x and y are in the boundary of B_{xy},*

(3) *all edges incident to x or y are contained in B_{xy}, and*

(4) *B_{xy} is otherwise disjoint from $\phi(G)$.*

Proof We can construct this as follows. First, begin with an ϵ-tube around the edge xy, and add a small ball around x and one around y.

Next, we can add ϵ-tubes around all of the edges from x to its neighbors not in $M(xy)$, and the same for y. We do not do this for the vertices in $M(xy)$ because adding both tubes from x and from y to a vertex $m \in M(xy)$ would create genus, and then B_{xy} would not be homeomorphic to a closed ball.

Instead, if $M(xy) = \{m_1, m_2, \dots, m_k\}$, for the cycles $xym_1, xym_2, \dots, xym_k$, the intersection $C_i \cap C_j$ equals $\phi(xy)$ for all $i \neq j$. This is connected, so by Lemma 2.1 there exist disks Γ_i for $i = 1, \dots, k$ bounded by C_i with interiors disjoint from the embedding and each other. Then we can take a small 3D neighborhood around each disk, except m_i is on the boundary of this neighborhood. Union this into the ball B_{xy}. Because their interiors are disjoint, this does not add any genus or interior boundary into B_{xy}, and so the B_{xy} is still homeomorphic to a closed ball.

The ball B_{xy} constructed in this way satisfies properties (1)–(4) by construction. For examples of this construction, see Figures 11, 12, and 14–16 in Section 3. □

We further this construction with the following:

Lemma 2.3 *Given B_{xy} constructed as above, there exists a disk D_{xy} with boundary on the boundary of B_{xy} such that*

(1) *D_{xy} has the vertices of $M(xy)$ on its boundary,*

(2) *D_{xy} intersects the edge xy exactly once, and*

(3) *One component of $B_{xy} - D_{xy}$ contains x and all edges incident to x except xy, and the other contains y and all edges incident to y except xy.*

Proof If $M(xy)$ is empty, take a disk transverse to xy with boundary on the surface of B_{xy}. This disk can be constructed so that condition (3) is satisfied.

If $M(xy)$ is nonempty, we have disks Γ_i for $i = 1, \dots, k$ from before, with a thickening of these disks making up part of B_{xy}. Take a path along these disks from m_i to the midpoint of xy. Thicken this path transverse to Γ_i to meet the surface of B_{xy}. Paste these subdisks together to form a disk D_{xy}. □

We will refer to D_{xy} as the *separating disk* of xy. This disk has a nice property with how it interacts with disks bounded by cycles in the contraction ϕ/xy.

2.1 Intersection of the separating disk with cycle bounded disks

Let ϕ be a flat embedding of a graph G and D_{xy} be constructed as above. Let ϕ/xy be a contraction of this embedding where $v \in D_{xy}$, for v the vertex resulting from the contraction of xy. Note D_{xy} still separates the edges originally incident to x from the ones originally incident to y. We say a cycle *crosses* the disk D_{xy} if the cycle contains the sequence avb, where the edges av and vb are in separate components of $B_{xy} - D_{xy}$.

Proposition 2.4 *In ϕ/xy, for every cycle C not crossing D_{xy}, there exists a disk Γ disjoint from ϕ/xy bounded by C whose interior has no intersection with the interior of D_{xy}.*

Proof The figures for this lemma are accurate up to ambient isotopy, meaning B_{xy} will be shown as a sphere, and D_{xy} as an equatorial disk.

First, consider a cycle C which does not include the vertex v. Let Γ be a disk bounded by C disjoint from the rest of the embedding. The cycle does not intersect the interior of B_{xy}. If the interior of Γ has intersection with the interior of D_{xy}, Γ must pass into the ball B_{xy}, dividing B_{xy} into at least two components. If a region bounded by $\partial B_{xy} \cup \Gamma$ is disjoint from the embedding, push Γ out of B_{xy} through this component without intersecting the graph.

This creates a new disk Γ' which is disjoint from the graph, bounded by C, and disjoint from the interior of B_{xy}. Because the interior of D_{xy} is within the interior of B_{xy}, the disk Γ' has no intersection with the interior of D_{xy}.

If every component bounded by $\partial B_{xy} \cup \Gamma$ contains a portion of the embedding, every component must contain a vertex on the surface of B_{xy}. Then every component bounded by $\partial B_{xy} \cup \Gamma$ contains an edge going to v. Since v sits in one of the components only, Γ cannot be disjoint from the graph, which is a contradiction.

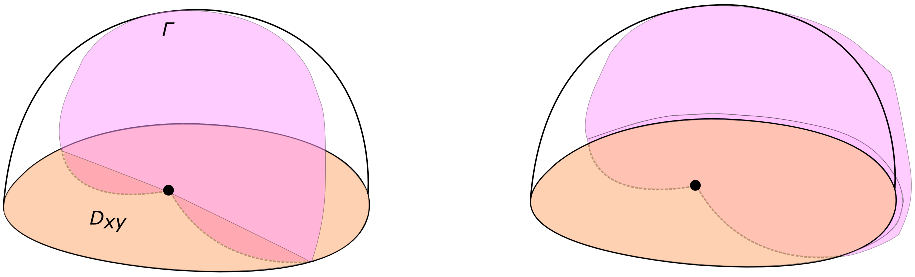


Figure 4: If a region bounded by $\Gamma \cup \partial B_{xy} \cup D_{xy}$ contains no edges, then Γ can be moved to the outside of this sphere without forcing it to intersect the graph. Here the disk is shown moving through the region in the back.

Second, let C be a cycle containing v which does not cross D_{xy}. Assume the interior of any disk Γ bounded by C has intersection with the interior of D_{xy}. We will show this would require the existence of a nontrivial link in ϕ, which is a contradiction. To make the presentation more fluent, we refer to the two components of $B_{xy} - D_{xy}$ as the "top" and "bottom", though they may be differently positioned with respect to a horizontal plane.

The cycle C contains avb, with av and vb both in the same component of $B_{xy} - D_{xy}$. Without loss of generality, assume they are within the bottom of $B_{xy} - D_{xy}$. By assumption, any disk bounded by C and disjoint from $\phi/xy \backslash C$ must intersect D_{xy}. Consider such a disk Γ. The union of surfaces $D_{xy} \cup \Gamma \cup \partial(B_{xy})$ determines disjoint regions within the interior of the top of B_{xy}. Any of these regions which is disjoint from the graph can be used to remove intersections between Γ and D_{xy}, by "pushing" Γ through the region, as in Figure 4. If all interior intersections between Γ and D_{xy} can be removed this way, we get the desired Γ. If not all intersections can be removed, there must be at least two regions in the top of $B_{xy} - D_{xy}$ with boundary included in $D_{xy} \cup \Gamma \cup \partial(B_{xy})$ each containing edges not in C. These edges are on the opposite side of D_{xy} from av and vb, since they are contained in the top. A diagram of the situation is presented in Figure 5. We have a cycle C with edges to v, both on one side of D_{xy}, and two edges to v on the other side of D_{xy} which are separated by the disk bounded by C.

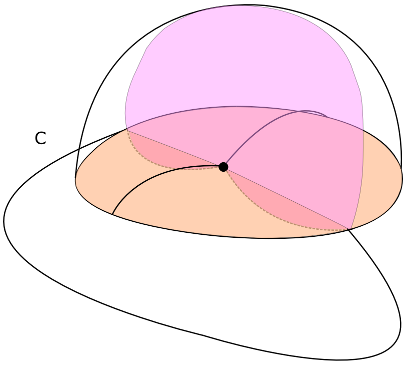


Figure 5: If Γ must have interior intersection with D_{xy}, then there must be at least 2 edges above D_{xy}.

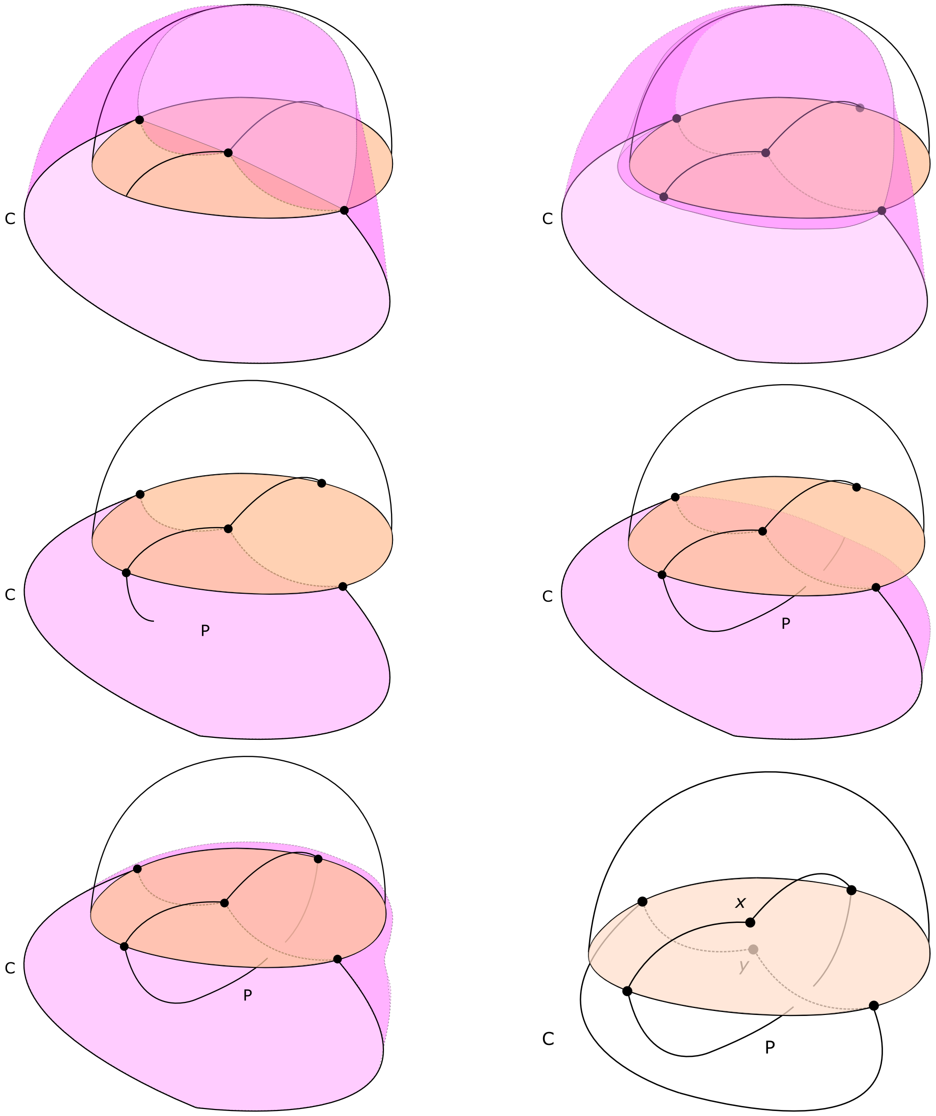


Figure 6: Relative position of Γ, D_{xy}, and $\partial(B_{xy})$. The disk Γ is pink and the disk D_{xy} is orange.

We push the disk out to the surface of the ball. In Figure 6, top row, we show Γ being pushed to the front of B_{xy}. It should be noted we are intentionally forcing our disk to intersect with the graph. By our assumption, a disk Γ bounded by C and disjoint from $\phi/xy\backslash C$ must have interior intersection with D_{xy}. Because D_{xy} is contained in B_{xy}, this means Γ must have interior intersection with B_{xy}. Then by the contrapositive, if we have a disk Γ bounded by C with no interior intersection with B_{xy}, it must have nonempty intersection with $\phi/xy\backslash C$.

Then Γ must intersect an edge going to the boundary of the top of B_{xy}. Push the disk Γ out along every intersection it has with the embedding away from the component of $B_{xy} - D_{xy}$ not containing the edges

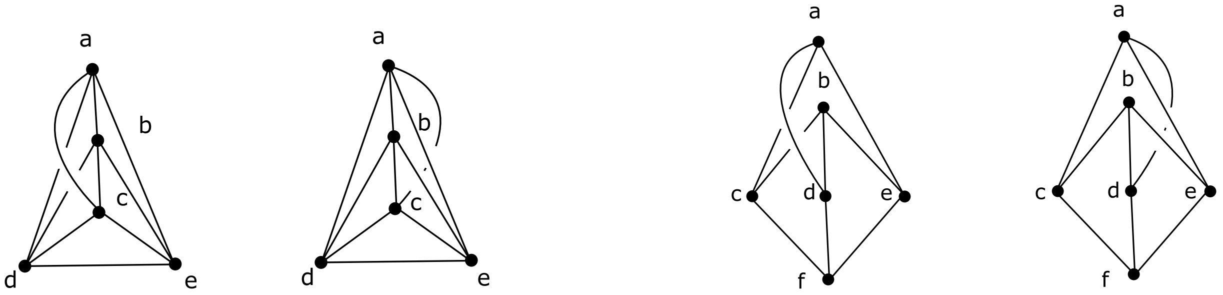


Figure 7: Two flat embeddings of the Kuratowski graphs. Left: K_5. Right: $K_{3,3}$.

in C. Trace these intersections. Eventually, we will no longer have any intersection with the graph. If no path P between vertices on the half of $B_{xy} - D_{xy}$ away from C is traced by intersecting with Γ, the final position for Γ is disjoint from this component of B_{xy}. This would mean Γ is disjoint from the graph, bounded by C, and the interior of Γ is disjoint from the interior of D_{xy}. This is a contradiction of our assumption, so this path P must exist. See Figure 6 for an example of such a path P being traced. But this necessitates the existence of a cycle C' and a cycle containing the path P' through x in ϕ. These must form a nontrivial link in ϕ, shown in Figure 6, bottom right. Then ϕ is not a flat embedding, which is a contradiction. □

We begin with some facts about embeddings of each of K_5 and $K_{3,3}$. These graphs are referred to as the Kuratowski graphs. From Robertson, Seymour, and Thomas's work, we know the following:

Lemma 2.5 [5] *The graphs K_5 and $K_{3,3}$ have exactly two nonambient isotopic flat embeddings.*

The two flat embeddings of K_5 and $K_{3,3}$ are shown in Figure 7. It can be seen that the two differ by the edge not in the maximal planar subgraph.

Lemma 2.6 [5] *If two flat embeddings of a graph G are not ambient isotopic, they must disagree on a K_5 or $K_{3,3}$ subdivision.*

We use the above to show the following simple lemma:

Lemma 2.7 *Let G be a K_5 or $K_{3,3}$ subdivision, and let ϕ_1 and ϕ_2 be the two not ambient isotopic flat embeddings of G. By taking ambient isotopy, ϕ_1 and ϕ_2 may be assumed as in Figure 7. Identify the labeled embeddings together along maximal planar subgraphs, a identified to a and so on, and e_j identified to e_j for the edges in the maximal planar subgraphs. Then the nonagreeing edges, $\phi_1(ax) \cup \phi_2(ax)$, and the cycle disjoint from those vertices form a nontrivial link.*

Proof Assume the embeddings of a K_5 or $K_{3,3}$ subdivision do not form a nontrivial link in the graph obtained by identification. Then the two paths not in the maximal planar graph must bound a union of disks disjoint from the graph. Then one path can be moved to the other through these disks, giving an ambient isotopy between the two distinct embeddings of K_5 or $K_{3,3}$, a contradiction. □

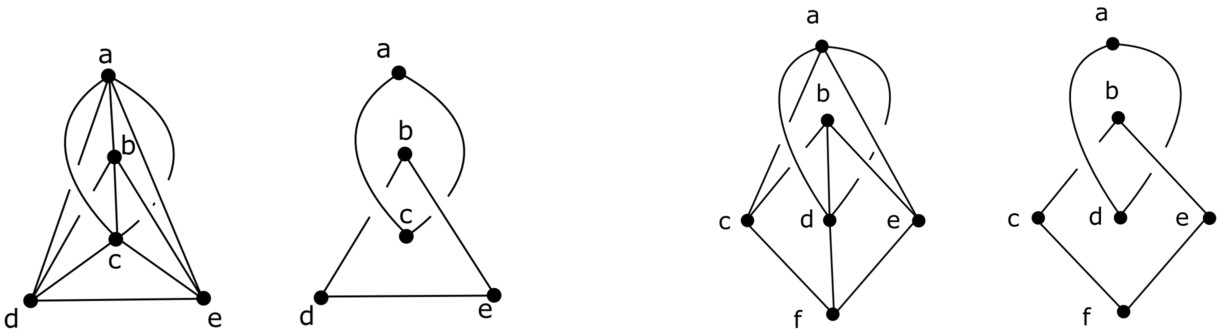


Figure 8: Linked cycles in the identification of the Kuratowski graphs. Left: K_5. Right $K_{3,3}$.

Proposition 2.8 *Let ϕ and ψ be flat embeddings of the simple flat graph G. If there is a B_{xy} such that it is valid in both ϕ and ψ, and the subembeddings $\phi - B_{xy}$ and $\psi - B_{xy}$ are ambient isotopic, then $\phi \cong \psi$.*

Proof By assumption, $\phi - B_{xy} \cong \psi - B_{xy}$. Then we may replace ψ by the embedding after this ambient isotopy, so outside of B_{xy} the embeddings are identically equal.

If ϕ and ψ are not ambient isotopic, they must disagree by some K_5 or $K_{3,3}$ subdivision by Lemma 2.6. Because the embeddings are identical outside of B_{xy}, without loss of generality, the disagreement must involve an edge incident to x. From Lemma 2.7, $\phi(xa) \cup \psi(xa)$ links with the other cycle in our subdivision. However, we know both $\phi(xa)$ and $\psi(xa)$ are contained within B_{xy}. Because B_{xy} is disjoint from the rest of the graph, the linking cycle must contain an edge from y, or the second link component would have to pierce the ball. This is a contradiction, since the only edges inside B_{xy} are those adjacent to x or y. This means the link must be of the form shown in Figure 9.

However, because D_{xy} separates the surface of B_{xy}, if the link as shown existed, then x and the vertex a connected to it are on opposite sides of D_{xy}. The edges from x to a must pass through D_{xy}, which we know does not happen. Then the cycle made with disagreeing edges incident to x from our two embeddings cannot link with a cycle containing an edge incident to y, and so there is no way that $\phi(xa) \cup \psi(xa)$ can link with anything in the graph. Therefore ϕ and ψ cannot differ by any K_5 or $K_{3,3}$ subdivisions, and so $\phi \cong \psi$. □

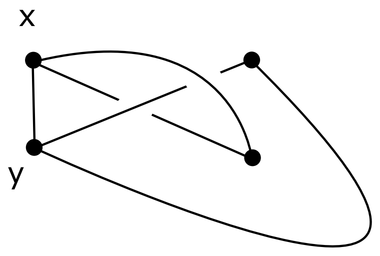


Figure 9: Linking forced by disagreeing embeddings.

2.2 Proof of main result

The following is our main result:

Theorem 2.9 *Every flat embedding of a linklessly embeddable graph G can be linearized.*

Proof We can restrict to connected graphs, since for disconnected graphs the components can be embedded separately. Graphs with up to four vertices are planar. By Fáry's theorem, these graphs have a straight-edge planar embedding. By Lemma 2.6, these graphs have a unique flat embedding. Because the planar embedding is flat, our straight-edge planar embedding gives a linear flat embedding. Assume the statement is true for all graphs with fewer than n vertices.

Let G be a linklessly embeddable graph with n vertices. Let ϕ be a flat embedding of G. Pick an edge $xy \in G$, and construct B_{xy} and D_{xy} in ϕ as described previously. Contract the edge xy so that the new vertex v lies in D_{xy}. This can be done so that the two components of $B_{xy} - D_{xy}$ still completely contain all of the edges they contained before, ie one component containing all edges to x, and the other containing all edges to y. For the vertices which have two edges to v we can assume the edges of the underlying simple graph are within the disk D_{xy}.

Then ϕ/xy is a flat embedding of a graph of order $n-1$, and so the underlying simple graph is linearizable by induction. By induction, there is an ambient isotopy $F : \mathbb{R}^3 \times [0,1] \to \mathbb{R}^3$ such that $F_0(\phi/xy) = \phi/xy$ and $F_1(\phi/xy)$ is a linear embedding of the underlying simple graph. Because ambient isotopy is continuous, we can move the double edges as close to the linear edges in the disk as we wish.

Because $\phi \cong F_1(\phi)$, we replace ϕ with $F_1(\phi)$, which is linear everywhere except inside B_{xy}. In $\mathbb{R}^3$, an ambient isotopy is an orientation-preserving homeomorphism. This means that $F_1(B_{xy})$ still contains all the edges incident to v. Moreover, the edges incident to x and to y are still separated by $F_1(D_{xy})$, since a single component of $B_{xy} - D_{xy}$ unioned with D_{xy} forms a closed ball. For this reason, we still refer to the images of the original ball and disk under the ambient isotopy F_1 by B_{xy} and D_{xy}.

Let ψ be a linear embedding of G such that $\psi(G - \{x, y\}) = (\phi/xy)(G - v)$, and x and y are placed very close to $(\phi/xy)(v)$ on the appropriate side of D_{xy}. This can be linear because x and y being close to v ensures all the edges are close enough to the straight edges in ϕ/xy.

We proceed to show ψ is flat. Recall Proposition 1.4.

By construction, $\psi/xy = \phi/xy$, and is therefore flat. We show now that $\psi\backslash xy$ is flat.

Consider a cycle $C \in \psi\backslash xy$. If C does not contain either x or y, then it corresponds exactly with a cycle in ϕ/xy. By Proposition 2.4, there exists a disk Γ bounded by C, disjoint from $\phi/xy\backslash C$, and disjoint from D_{xy}. Then Γ does not interact with the neighborhood around D_{xy}, which is the only change between the embeddings. Then the same disk Γ from ϕ/xy will also bound C in $\psi\backslash xy$, and is disjoint from the rest of $\psi\backslash xy$.

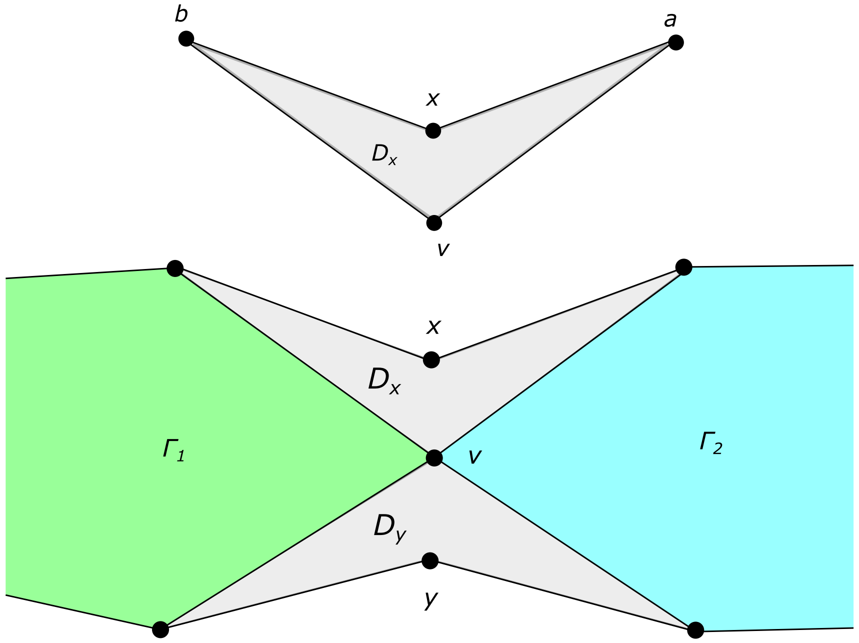


Figure 10: A cycle C in $\phi\backslash xy$ containing both x and y corresponds to two smaller cycles, C_1 and C_2 in ϕ/xy. The top part shows a disk bounded by $xavb$. This is disjoint from $\phi\backslash xy$ by construction. C_1 and C_2 bound disks disjoint from ϕ/xy, Γ_1 and Γ_2. These disks can be extended as shown in the bottom part through disks D_x and D_y to a disk disjoint from $\phi\backslash xy$

If C contains only one of x and y, then C corresponds to some cycle C' in ϕ/xy not crossing D_{xy}, with the x or y replaced by v. Without loss of generality, assume $x \in C$, and ax and xb are part of C. By Proposition 2.4, C' bounds a disk Γ' disjoint from ϕ/xy which has no interior intersection with D_{xy}. Then in $\psi\backslash xy$, Γ' does not intersect any edges incident to y, since they are separated from Γ' by the bounding disk D_{xy}. Because x is so close to the position of v, Γ' is also disjoint from any edges incident to x, since it was disjoint from the edges of v. We can extend Γ' from its edges avb to axb through a disk D bounded by $avbxa$. This is disjoint from the graph by construction, since x is very close to the position of v, and because $xy \notin \psi\backslash xy$. Then after extending Γ' along D to form a disk Γ, C bounds a disk Γ disjoint from the rest of $\psi\backslash xy$.

If C contains both x and y, then C corresponds to two cycles C_1 and C_2 in ϕ/xy meeting at v. By Lemma 2.1, as these cycles have connected intersection, there exist disks Γ_1 and Γ_2 bounded by C_1 and C_2 with interior disjoint from ϕ/xy, such that $\Gamma_1 \cap \Gamma_2 = v$. Then we can mold the edges of $C_1 \cup C_2$ to the edges in C in a similar process to before. These sections are all disjoint from the graph and each other. The union $\Gamma_1 \cup \Gamma_2$ can be extended through D_x and D_y, forming a disk Γ bounded by C with interior disjoint from $\phi\backslash xy$; see Figure 10. Then $\psi\backslash xy$ is flat. By Proposition 1.4, ψ is a flat linear embedding of G.

By Proposition 2.8, ϕ and ψ are exactly the same outside of D_{xy}, and are both flat. Then $\phi \cong \psi$. □

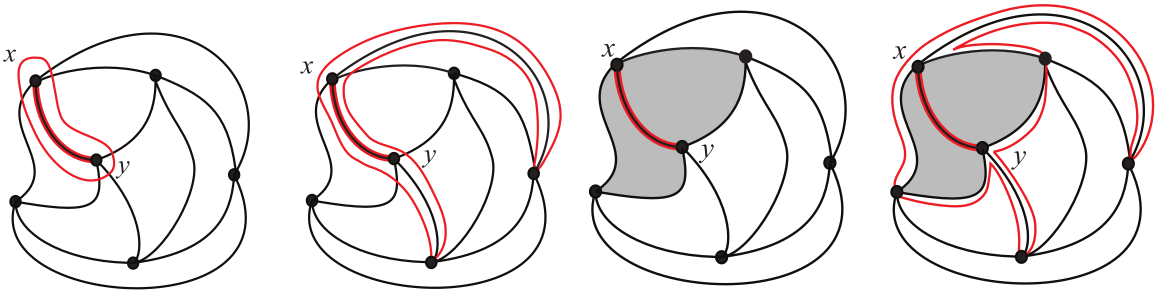


Figure 11: The construction for a planar graph on six vertices, part 1.

3 Examples

3.1 Example 1: planar graph

In Figures 11 and 12, we present an example of our construction. We start with a planar graph on six vertices. Around the highlighted edge xy, we form B_{xy}, whose intersection with the plane is drawn in red. The disks bounded by x, y, and their common neighbors are both contained in the plane, and shaded in gray. Figure 12, center left, shows the contraction to the midpoint of xy, forming the new vertex v. In Figure 12, center right, we have linearized the contraction so that the straight edges from v to the common vertices are the dashed paths in Figure 12, far and center left. Then Figure 12, far right, shows a linear flat embedding of our graph constructed by placing x and y on opposite sides of our separating disk.

3.2 The graph $Q_{13,3}$

Figures 14–16 show another full example of our process for the graph $Q_{13,3}$, shown in Figure 13. This graph has 13 vertices, where every vertex is connected in a sequential cycle. Then, every vertex is connected to the two vertices distance 3 away within the cycle of length 13 — 4 connects to 1 and 7, 2 connects to 12 and 5, etc. Note this graph is triangle-free, so $M(e) = \varnothing$ for all e. In Figure 14, we have a flat embedding.

This embedding is shown to be flat as follows. The graph $Q_{13,3}$ has a maximal planar subgraph with the removal of the edges (1, 11), (11, 12), (9, 12), and (10, 13). By Lemma 2.6, planar graphs have unique

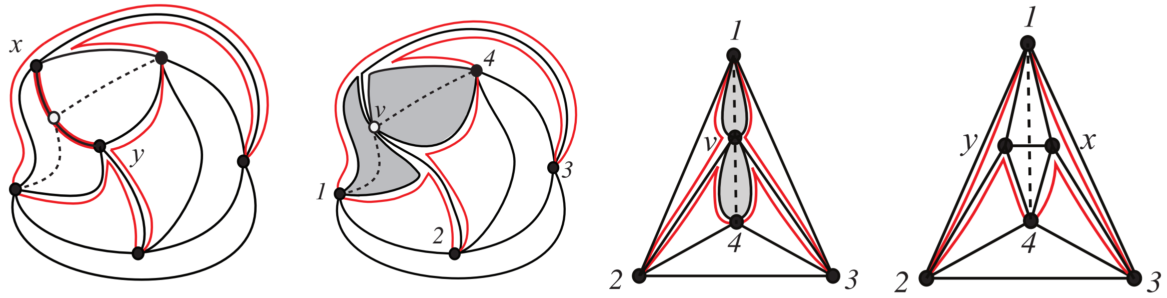


Figure 12: The construction for a planar graph on six vertices, part 2.

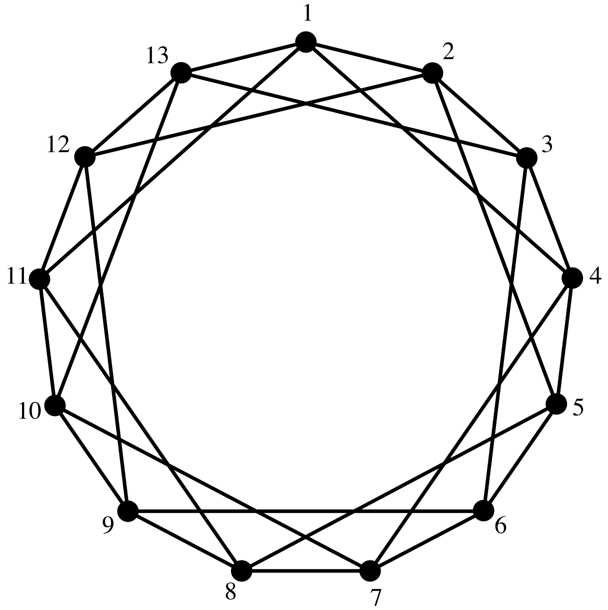


Figure 13: The maximal linklessly embeddable graph $Q_{13,3}$.

flat embeddings, since they do not have any K_5 or $K_{3,3}$ subdivisions. With some simple manipulations, it can be shown that this subgraph in Figure 14 is ambient isotopic to a planar embedding. If we embed one of our removed edges above this embedded plane and another below, we always create a nontrivial link. Then our flat embedding must have all four removed edges embedded above the plane or all four below the plane. In Figure 14, the edges are all embedded above.

We work by contracting the edge $e = (8, 9)$. Figure 15 shows the surface of B_e in green, and D_e in red. Because vertices 8 and 9 have no common neighbors, D_e is just a disk transverse to e placed at its

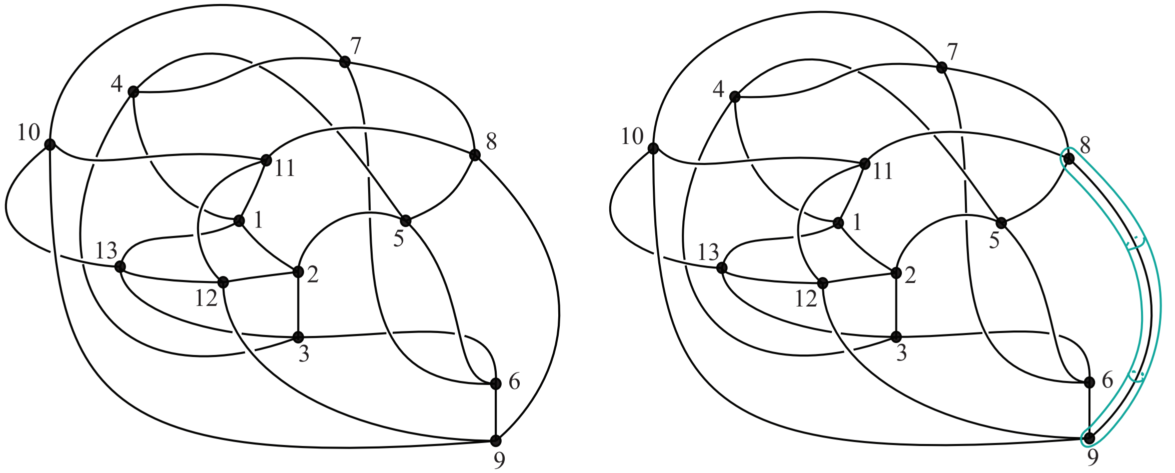


Figure 14: The construction for the graph $Q_{13,3}$, part 1.

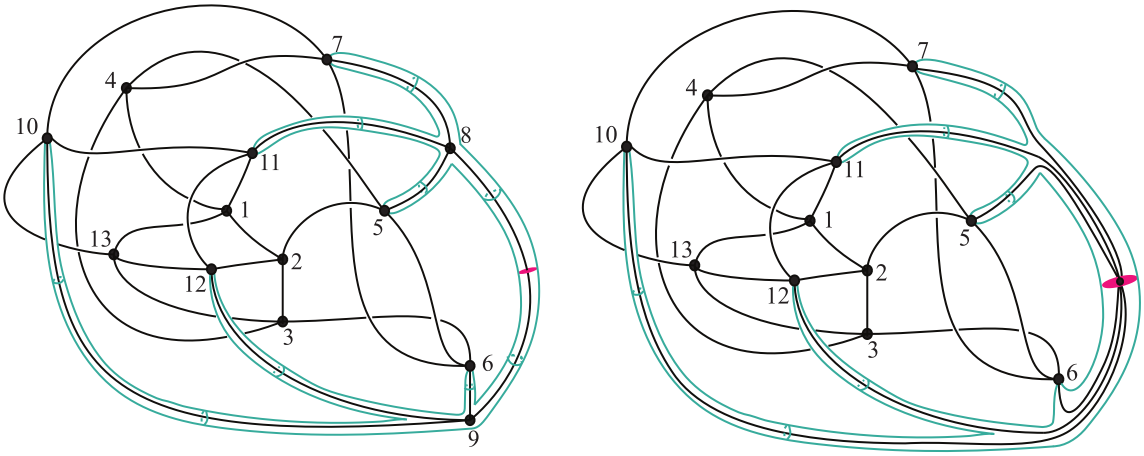


Figure 15: The construction for the graph $Q_{13,3}$, part 2.

midpoint. In Figure 15, we have performed the topological contraction of e to a vertex $v \in D_e$. Again, vertices 8 and 9 have no common neighbors in the graph, so ϕ/e is exactly an embedding of a simple graph. By assumption, there is an ambient isotopy of this embedding which produces a linear embedding. This linear embedding is shown in Figure 16.

We can see here that D_e still separates the neighbors of 9, $\{6, 10, 12\}$, from the neighbors of 8, $\{5, 7, 11\}$. Then in Figure 16, we have placed 8 and 9 on the appropriate sides of D_e, close to the location of v. By our proof, this is a linear flat embedding of $Q_{13,3}$ which is ambient isotopic to the original embedding.

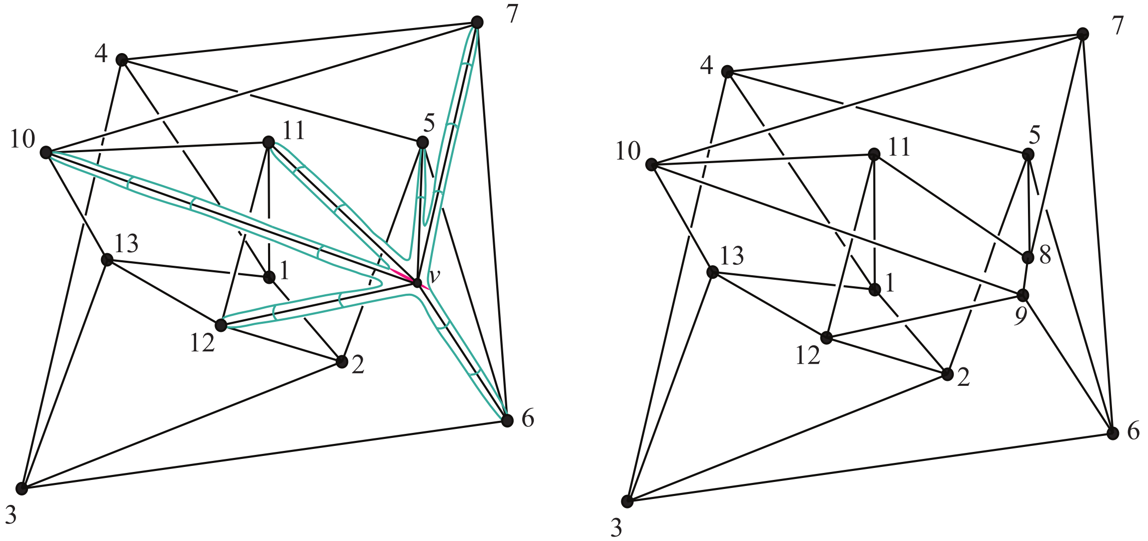


Figure 16: The construction for the graph $Q_{13,3}$, part 3.

4 Conclusion

We proved Sachs' 1983 conjecture on linklessly embeddable graphs, showing they all admit linear linkless embeddings. Further, we showed every flat embedding of a simple graph can be linearized. An obvious next class of embeddings to study is unknotted embeddings of nonintrinsically knotted graphs.

It has been shown by Hughes [4] that not every unknotted embedding of a knotlessly embeddable graph can be realized using straight edges. K_6 is an unknotted graph, and has unknotted embeddings with anywhere from 1 to 10 links. Hughes has shown there are only two nonambient isotopic knotless linear embeddings of K_6, which have 1 and 3 links respectively. Therefore not every knotless embedding can be linear.

Acknowledgements

Thank you to Dr Elena Pavelescu for all the guidance and instruction in completing the thesis [7] this paper is based on and for making the beautiful figures in the examples section, specifically Figures 14–16.

References

[1] **T Böhme**, *On spatial representations of graphs*, from "Contemporary methods in graph theory", Bibliographisches Inst., Mannheim, Germany (1990) 151–167 MR Zbl

[2] **J H Conway**, **C M Gordon**, *Knots and links in spatial graphs*, J. Graph Theory 7 (1983) 445–453 MR Zbl

[3] **I Fáry**, *On straight line representation of planar graphs*, Acta Univ. Szeged. Sect. Sci. Math. 11 (1948) 229–233 MR Zbl

[4] **C Hughes**, *Linked triangle pairs in a straight edge embedding of K_6*, Pi Mu Epsilon J. 12 (2006) 213–218 Zbl

[5] **N Robertson**, **P D Seymour**, **R Thomas**, *A survey of linkless embeddings*, from "Graph structure theory", Contemp. Math. 147, Amer. Math. Soc., Providence, RI (1993) 125–136 MR Zbl

[6] **H Sachs**, *On a spatial analogue of Kuratowski's theorem on planar graphs: an open problem*, from "Graph theory", Lecture Notes in Math. 1018, Springer (1983) 230–241 MR Zbl

[7] **P Stanfield**, *Linear linkless embeddings*: *proof of a conjecture by Sachs*, master's thesis, University of South Alabama (2021) Available at `https://www.proquest.com/docview/2512356289`

[8] **S K Stein**, *Convex maps*, Proc. Amer. Math. Soc. 2 (1951) 464–466 MR Zbl

[9] **K Wagner**, *Bemerkungen zum Vierfarbenproblem*, Jahresber. Dtsch. Math.-Ver. 46 (1936) 26–32 Zbl

Department of Mathematics and Statistics, University of Nevada, Reno
Reno, NV, United States

`stanfield.mathematics@gmail.com`

Received: 1 July 2022 Revised: 24 March 2024

Algebraic & Geometric Topology **25**:8 (2025) 4599–4631
DOI: 10.2140/agt.2025.25.4599
Published: 20 November 2025

Constructing rational homology 3-spheres that bound rational homology 4-balls

LISA LOKTEVA

We present three large families of new examples of plumbed 3-manifolds that bound rational homology 4-balls. These are constructed using two operations, also defined here, that preserve the lack of a lattice embedding obstruction to bounding rational homology balls. Apart from in the cases shown in this paper, it remains open whether these operations are rational homology cobordisms in general.

The new examples include a multitude of families of rational surgeries on torus knots, and we explicitly describe which positive torus knots we now know to have a surgery that bounds a rational homology ball.

While not the focus of this paper, we implicitly confirm the slice-ribbon conjecture for new, more complicated, examples of arborescent knots, including many Montesinos knots.

57K40, 57K41

1 Introduction

In Kirby's Problem 4.5 [14], Casson asks which rational homology 3-spheres bound rational homology 4-balls. While rational homology 3-spheres abound in nature, including the r-surgery $S^3_r(K)$ on a knot K for any $r \in \mathbb{Q} - \{0\}$, very few of them actually bound rational homology balls. In fact, Aceto and Golla showed in [2, Theorem 1.1] that for every knot K and every $q \in \mathbb{Z}_+$, there exist at most finitely many $p \in \mathbb{Z}_+$ such that $S^3_{p/q}(K)$ bounds a rational homology ball. It is hard to answer Casson's question in full generality, but recently a great deal of progress has been made on specific classes of rational homology 3-spheres. For example, in 2007 we learnt the answer for lens spaces [17; 18], in 2020 for positive integral surgeries on positive torus knots [2; 3], and in between we learnt the answer for several other classes on Seifert fibred spaces with three exceptional fibres [15; 16]. We do not yet know the answer for general Seifert fibred spaces with three exceptional fibres. In [19], the author started studying surgeries on algebraic (iterated torus) knots, which are not Seifert fibred but decompose into Seifert fibred spaces when cut along a maximal system of incompressible tori [12].

An important tool to study which 3-manifolds bound rational homology balls is the following well-known corollary of Donaldson's diagonalisation theorem [9, Theorem 1]:

Proposition 1 (corollary of Donaldson's theorem) *Let Y be a rational homology 3-sphere and $Y = \partial X$ for X a negative definite smooth connected oriented 4-manifold. If $Y = \partial W$ for a smooth rational* *homology* 4*-ball* W*, then there exists a lattice embedding*

$$(H_2(X)/\operatorname{Torsion}, Q_X) \hookrightarrow (\mathbb{Z}^{\operatorname{rk} H_2(X)}, -\mathrm{Id}).$$

Here, Q_X is the intersection form on $H_2(X)/\operatorname{Torsion}$. Determining which 3-manifolds in a family $\mathfrak{F}$ bound rational homology 4-balls using lattice embeddings often goes like this:

(i) Find a negative definite filling $X(Y)$ for every $Y \in \mathfrak{F}$.

(ii) Guess the family $\mathfrak{F}' \subset \mathfrak{F}$ of manifolds for which the intersection lattice of the filling (that is, second homology with the intersection form) embeds into the standard lattice of the same rank.

(iii) Show that $\big(H_2(X(Y)), Q_{X(Y)}\big)$ does not embed into $(\mathbb{Z}^{b_2(X(Y))}, -\mathrm{Id})$ for any $Y \in \mathfrak{F} - \mathfrak{F}'$.

(iv) Hopefully prove that Y bounds a rational homology ball for any $Y \in \mathfrak{F}'$.

Every step of this process has the potential to go wrong. For starters, there exist 3-manifolds without any definite fillings [10]. However, lens spaces, surgeries on torus knots and large surgeries on algebraic knots do have definite fillings. In fact, they all bound definite plumbings of disc bundles on spheres. Step (iv) is not guaranteed to work either. For example, $S^3_{-m^2}(K)$ bounds the knot trace $D^4_{-m^2}(K)$ (D^4 with a $(-m^2)$-framed 2-handle glued along K) which has intersection lattice $(\mathbb{Z}, -m^2)$ which embeds into $(\mathbb{Z}, -\mathrm{Id})$, but according to [2, Theorem 1.2], $S^3_{-m^2}(K)$ bounds a rational homology ball for at most two positive integer values of m. However, in [2; 3; 17; 18], the authors managed to find a different filling $X(Y)$ for each Y for which the condition gave a complete obstruction. These $X(Y)$ are plumbings of disc bundles on spheres with a tree-shaped plumbing graph, satisfying the property that the quantity

$$I = \sum_{v \in V} (-w(v) - 3)$$

is negative, where V is the set of vertices of the graph and $w(v)$ is the weight of v.

Steps (ii) and (iii) can sometimes be done at the same time, but often, like in [3] where $\mathfrak{F}$ is the set of positive integral surgeries on positive torus knots, they cannot. It is then important to eliminate embeddable cases early in order to proceed with step (iii). Theorem 1.1 in [3], the classification of positive integral surgeries on positive torus knots bounding rational homology balls, lists five families that are Seifert fibred spaces with three exceptional fibres. They bound a negative definite star-shaped plumbing with three legs. Families (1)–(3) have two complementary legs, that is two legs whose weight sequences are Riemenschneider dual (defined, for the reader's convenience, in Section 2). All such 3-manifolds that bound a rational homology ball have been classified by Lecuona in [16]. Family (5) contains two exceptional graphs which were known to bound rational homology balls both because they arise as boundaries of tubular neighbourhoods of rational cuspidal curves in [7] and because they are surgeries on torus knots $T(p,q)$ where $q \equiv \pm 1 \pmod{p}$, which were studied in [2]. However, family (4) took the authors of [3] a while to find, in the meantime thwarting their attempts at step (iii). Eventually they found family (4) using a computer. This allowed them to finish off their lattice embedding analysis, but family (4) still looked surprising and strange and raised the question of "How could we have predicted its existence?"

1.1 GOCL and IGOCL moves

This work came out of widening the perspective and asking which boundaries of 4-manifolds, described by plumbing trees with negative definite intersection forms and low I, bound rational homology balls. As a preliminary question, we asked ourselves which plumbing trees generate an embeddable intersection lattice. We looked at what the graphs of 3-manifolds we know to bound rational homology balls look like and tried to see if there are any common patterns. In [15, Remark 3.2], Lecuona describes how to get all lens spaces that bound rational homology balls from the linear graphs $(-2,-2,-2)$, $(-3,-2,-3,-3,-3)$, $(-3,-2,-2,-3)$ and $(-2,-2,-3,-4)$ using some modifications. (She restates Lisca's result in [17] in the language of plumbing graphs rather than fractions p/q for $L(p,q)$.) In this paper we define a couple of moves called GOCL and IGOCL moves on embedded plumbing graphs that preserve embeddability and generalise the moves described by Lecuona. From this point of view, Lecuona's list simply turns into a list of IGOCL and GOCL moves that keep the graph linear. The IGOCL move was also used by Jonathan Simone in [24] under the name of expansions. The GOCL move is a generalisation of Lisca's expansions in [17]. The names *GOCL* and *IGOCL* are acronyms for (*inner*) *growth of complementary legs* and these names come from the fact that repeated application of these moves results in pairs of complementary legs (Definition 9) seemingly grown onto the original graph.

We may ask ourselves if these moves preserve the property of the described 3-manifolds bounding rational homology balls. There is unfortunately no obvious rational homology cobordism between two 3-manifolds differing by a GOCL or an IGOCL move. We can however prove that repeated applications of these moves to the embeddable linear graphs $(-3,-2,-3,-3,-3)$, $(-3,-2,-2,-3)$ and $(-2,-2,-3,-4)$ give 3-manifolds bounding rational homology balls. This results in the following theorem:

Theorem A *All* 3*-manifolds described by the plumbing graphs in Figures* 1, 2 *and* 3 *bound rational homology balls.*

There are several methods to prove this theorem. The easiest one, which we can call the *simple method*, uses the following proposition, which follows from the long exact sequence of the pair combined with Poincaré duality and the universal coefficient theorem:

Proposition 2 (simple method) *If a* 4*-manifold* X *consists of one* 0*-handle*, n 1*-handles*, *and* n 2*-handles*, *and if* ∂X *is a rational homology* S^3, *then* X *is a rational homology ball.*

We apply this by showing that we may perform two integral surgeries on the three-manifolds described by these plumbing trees and obtain $(S^1 \times S^2) \# (S^1 \times S^2)$. This method works for every one of the families of Figures 1, 2 and 3.

A more refined method is to show that the above plumbed 3-manifolds are double covers of S^3 branched over a χ-slice link, that is, a link bounding a surface S of Euler characteristic 1 in D^4 [8, Definition 1].

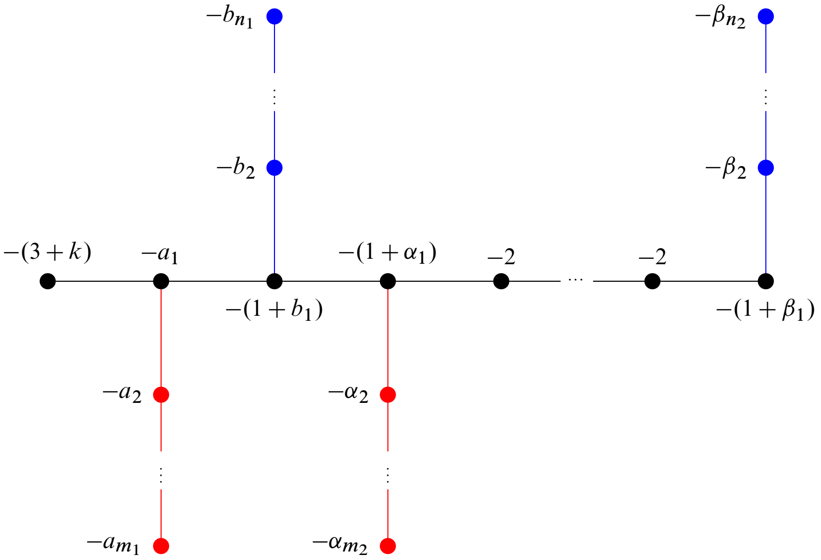


Figure 1: The graphs obtainable by performing IGOCL and GOCL moves on the linear graph $(-3,-2,-3,-3,-3)$. The length of the chain of -2's is $k \geq 0$, $(a_1,\ldots,a_{m_1})$ and $(\alpha_1,\ldots,\alpha_{m_2})$ are complementary sequences, and $(b_1,\ldots,b_{n_1})$ and $(\beta_1,\ldots,\beta_{n_2})$ are complementary sequences.

By [8, Proposition 5.1], the double cover of D^4 branched over a surface of Euler characteristic 1 is a rational homology 4-ball. We use this method, described in Section 3.2 and more specifically Proposition 16, for the graphs in Figures 1 and 2. Given the way we construct S, this method amounts to the simple method, but with the extra step of showing that we can perform the surgeries equivariantly under an involution.

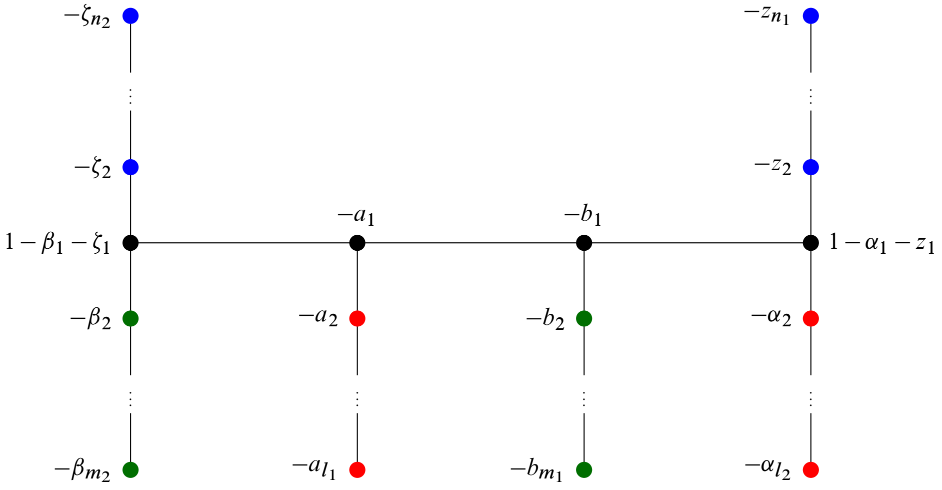


Figure 2: The form of all graphs obtainable from $(-3,-2,-2,-3)$ using GOCL moves. The sequences $(a_1,\ldots,a_{l_1})$ and $(\alpha_1,\ldots,\alpha_{l_2})$ are complementary, as well as the sequences $(b_1,\ldots,b_{m_1})$ and $(\beta_1,\ldots,\beta_{m_2})$, and the sequences $(z_1,\ldots,z_{n_1})$ and $(\zeta_1,\ldots,\zeta_{n_2})$.

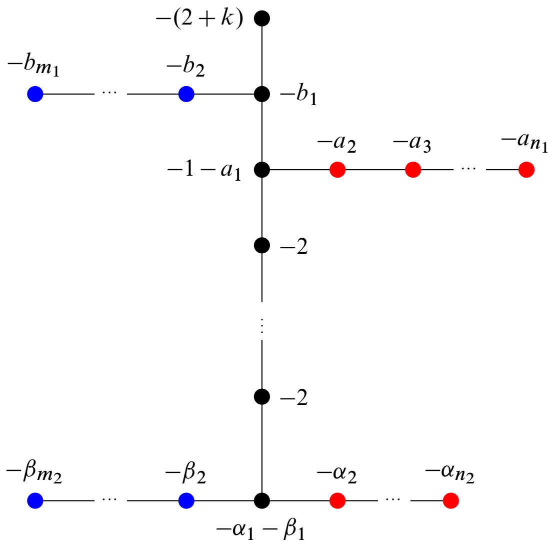


Figure 3: The form of all graphs obtainable from $(-2,-2,-3,-4)$ by GOCL and IGOCL moves. The length of the chain of -2's is $k \geq 0$, $(a_1,\dots,a_{m_1})$ and $(\alpha_1,\dots,\alpha_{m_2})$ are complementary sequences, and $(b_1,\dots,b_{n_1})$ and $(\beta_1,\dots,\beta_{n_2})$ are complementary sequences.

This extra step is quite challenging, and we do not at this moment know if we can perform it on the family of Figure 3. The bonus of using this more difficult method is that we obtain new examples of links that are χ-ribbon in the process.

The families of Figures 1, 2 and 3, together with the one generated by GOCL moves from $(-2,-2,-2)$, include all lens spaces bounding rational homology balls. The $(-2,-2,-2)$ family, however, only includes linear graphs already found by Lisca. On the other hand, the families of Figures 1, 2 and 3 also contain more complicated graphs. In [1], Aceto defines the linear complexity of a plumbing tree to be the minimal number of vertices we need to remove in order to get a linear graph. Our families have linear complexities up to 2. Many papers, eg [1; 2; 3; 15; 24], that use lattice embeddings to obstruct plumbed 3-manifolds from bounding a rational homology ball have used arguments of the form "if my graph Γ is embeddable, then this other linear graph obtained from Γ is embeddable, and we know what those look like", which gets harder to do the further Γ is away from being linear. Thus, we do not yet really have lattice embedding obstructions for families of graphs of complexity greater than 1. The families of Theorem A include many graphs of Seifert fibred spaces. They include family (4) in [3] and predict its existence because family (4) is just the intersection between the set of graphs in Figures 1, 2 and 3 and the negative definite plumbing graphs of positive integral surgeries on positive torus knots.

As mentioned above, there is no obvious rational homology cobordism between the 3-manifolds described by two plumbing graphs differing by a GOCL or an IGOCL move. This is interesting in comparison with the case in the works by Aceto [1] and Lecuona [16]. Lecuona shows that given a plumbing graph Γ, we can modify it to a graph Γ' by subtracting 1 from the weight of a vertex v and attaching two complementary legs $(-a_1, \dots, -a_m)$ and $(-b_1, \dots, -b_n)$ (see Section 2 or [16] for definitions) to v, and the 3-manifolds Y_Γ and $Y_{\Gamma'}$ described by the graphs will be rational homology cobordant, that is, bound a rational homology 4-ball if and only if the other one does. Thus, if she wants to know if a $Y_{\Gamma'}$, for Γ' a graph with two complementary legs coming out of the same vertex, bounds a rational homology ball, she can reduce it to the same question for a simpler graph. However, since we do not know if the GOCL and IGOCL moves are rational homology cobordisms, we cannot play this trick for complementary legs growing out of different vertices.

The paper [4] by Akbulut and Larson is another work showing that applying GOCL moves to embedded plumbing graphs bounding certain rational homology balls gives us new plumbed 3-manifolds that bound rational homology balls. The authors show that the families $\Sigma(2, 4n+1, 12n+5)$ and $\Sigma(3, 3n+1, 12n+5)$ of Brieskorn spheres bound rational homology balls. In fact, these families are obtained by applying GOCL moves to the plumbing graphs of $\Sigma(2, 5, 17)$ and $\Sigma(3, 4, 17)$. Just like us in Section 3, they perform a surgery on their spaces and the result of this surgery is the same for each space, in their case a 0-surgery on the figure eight knot. Their result [4, Lemma 2], saying that any integral homology sphere obtained from a surgery on a 0-surgery on a rationally slice knot bounds a rational homology ball, can be viewed as a generalisation of the simple method and Proposition 2 when $n = 1$. With the same technique as Akbulut and Larson, Şavk [23] constructed two more families $\Sigma(2, 4n+3, 12n+7)$ and $\Sigma(3, 3n+2, 12n+7)$ of Brieskorn spheres that bound rational homology balls. These are obtainable from $\Sigma(2, 7, 19)$ and $\Sigma(3, 5, 19)$, respectively, using IGOCL moves.

1.2 Rational surgeries on torus knots

An interesting generalisation of [3, Theorem 1.1], would be to classify all positive rational surgeries on positive torus knots that bound rational homology balls. Theorem A allows us to construct more examples of such surgeries than is sightly to write down. Instead, we may ask ourselves the following question:

Question 3 *For which $1 < p < q$ with $\mathrm{GCD}(p, q) = 1$ is there an $r \in \mathbb{Q}_+$ such that $S^3_r(T(p, q))$ bounds a rational homology ball?*

Section 4 is dedicated to proving the following theorem:

Theorem B *For the following pairs (p, q) with $1 < p < q$ and $\mathrm{GCD}(p, q) = 1$, there is at least one $r \in \mathbb{Q}_+$ such that $S^3_r(T(p, q))$ bounds a rational ball. Here $k, l \geq 0$.*

(1) $(k+2, (l+1)(k+2)+1)$.

(2) $(k+2, (l+2)(k+2)-1)$.

(3) $(2k+3, (l+1)(2k+3)+2)$.

(4) $(2k+3, (l+2)(2k+3)-2)$.

(5) $(k^2+7k+11, k^3+12k^2+45k+51)$.

(6) $(S^{(k)}_{l+1}, S^{(k)}_{l+2})$ *for* $(S^{(k)}_i)$ *a sequence defined by* $S^{(k)}_0 = 1$, $S^{(k)}_1 = 2$, $S^{(k)}_2 = 2k+7$ *and* $S^{(k)}_{i+2} = (k+4)S^{(k)}_{i+1} - S^{(k)}_i$.

(7) $(T^{(k)}_{l+1}, T^{(k)}_{l+2})$ *for* $(T^{(k)}_i)$ *a sequence defined by* $T^{(k)}_0 = 1$, $T^{(k)}_1 = k+2$, $T^{(k)}_2 = k^2+6k+7$ *and* $T^{(k)}_{i+2} = (k+4)T^{(k)}_{i+1} - T^{(k)}_i$.

(8) (U_{l+1}, U_{l+2}) *for* (U_i) *a sequence defined by* $U_0 = 1$, $U_1 = 3$, $U_2 = 14$ *and* $U_{i+2} = 5U_{i+1} - U_i$.

(9) (R_{l+1}, R_{l+2}) *for* (R_i) *a sequence defined by* $R_0 = 1$, $R_1 = 3$, $R_2 = 17$ *and* $R_{i+2} = 6R_{i+1} - R_i$.

(10) (P_{l+1}, P_{l+2}) *for* (P_i) *a sequence defined by* $P_0 = 1$, $P_1 = 4$, $P_2 = 19$ *and* $P_{i+2} = 5P_{i+1} - P_i$.

(11) (Q_{l+1}, Q_{l+2}) *for* (Q_i) *a sequence defined by* $Q_0 = 1$, $Q_1 = 2$, $Q_2 = 9$ *and* $Q_{i+2} = 5Q_{i+1} - Q_i$.

(12) $(A, (n+1)Q+P)$ *for* P *and* Q *such that* $L(Q,P)$ *bounds a rational homology ball (or equivalently* Q/P *lying in Lisca's set* $\mathcal{R}$ [17]), *and* A *a multiplicative inverse to either* Q *or* $nQ+P$ *modulo* $(n+1)Q+P$ *such that* $0 < A < (n+1)Q+P$.

(13) $\bigl((n+1)Q+P, (l+1)((n+1)Q+P)+A\bigr)$ *for* P *and* Q *such that* $L(Q,P)$ *bounds a rational homology ball (or equivalently* Q/P *lying in Lisca's set* $\mathcal{R}$ [17]), *and* A *a multiplicative inverse to either* Q *or* $nQ+P$ *modulo* $(n+1)Q+P$ *such that* $0 < A < (n+1)Q+P$.

(14) (B, P) *for* P *and* Q *such that* $L(Q,P)$ *bounds a rational homology ball (or equivalently* Q/P *lying in Lisca's set* $\mathcal{R}$ [17]), *and* B *a multiplicative inverse to either* $P\lceil Q/P\rceil - Q$ *or* $Q - P\lfloor Q/P\rfloor$ *modulo* P *such that* $0 < B < P$.

(15) $(P, (l+1)P+B)$ *for* P *and* Q *such that* $L(Q,P)$ *bounds a rational homology ball (or equivalently* Q/P *lying in Lisca's set* $\mathcal{R}$ [17]), *and* B *a multiplicative inverse to either* $P\lceil Q/P\rceil - Q$ *or* $Q - P\lfloor Q/P\rfloor$ *modulo* P *such that* $0 < B < P$.

(16) (P, Q) *such that there is a number* n *with* $(P, Q, n) \in \mathcal{R} \sqcup \mathcal{L}$ *for the sets* $\mathcal{R}$ *and* $\mathcal{L}$ *defined in* [3, Theorem 1.1]. (*Note that here* $r = n \in \{PQ, PQ-1, PQ+1\}$, *so we are looking at an integral surgery.*)

The curious reader can use the methods of Section 4 to obtain the surgery coefficients r too.

In Theorem B, case (16) is shown to bound rational homology balls in [3] and reflects the degenerate cases of surgeries on torus knots that are lens spaces or connected sums of lens spaces, cases (12)–(15) are shown to bound rational homology balls in [16] because their graphs have a pair of complementary legs, while the cases (1)–(11) are shown to bound rational homology balls in this paper, using that there exists an r such that $S^3_r(T(p,q))$ bounds a graph in the families of Figures 1, 2 and 3. The authors of [3] classified all positive integral surgeries on positive torus knots that bound rational homology balls. The classification included 18 families, whereof families (6)–(18) are included in our family (16), family (4) in our family (9), and the others in families (1) and (2).

At the moment of writing we do not know of any other positive torus knots having positive surgeries bounding rational homology balls. The pair (8, 19) is in some metric the smallest example not to appear on the list of Theorem B. Thus we may concretely ask:

Question 4 *Is there an* $r \in \mathbb{Q}_+$ *such that* $S^3_r(T(8,19))$ *bounds a rational homology ball?*

We may also note that some positive torus knots have many surgeries that bound rational homology balls. For example, Theorem A allows us to construct numerous finite and infinite families of surgery coefficients $r \in \mathbb{Q}_+$ such that $S^3_r(T(2,3))$ bounds a rational homology ball. All we need to do is to choose weights in the graphs in Figures 3, 2 and 1 so that we get a star-shaped graph with three legs whereof one is (-2) and another is either $(-2,-2)$ or (-3). For example, $S^3_{(11k+20)^2/(22k^2+79k+71)}(T(2,3))$ bounds a plumbing of the shape in Figure 3 with $(b_1,\dots,b_{m_1}) = (4)$ and $(a_1,\dots,a_{n_1}) = (3)$, and thus bounds a rational homology ball for any $k \geq 0$. There are also surgeries on $T(2,3)$ that bound rational balls, but do not have graphs of the shapes of Figures 1, 2 or 3. For example, we have $S^3_{64/7}(T(2,3)) = -S^3_{64}(T(3,22))$, which bounds a rational homology ball because it is the boundary of the tubular neighbourhood of a rational curve in $\mathbb{C}P^2$ [7], but whose lattice embedding contains a basis vector with coefficient 2, which we do not get by applying GOCL or IGOCL moves to $(-2,-2,-3,-4)$, $(-3,-2,-2,-3)$ and $(-3,-2,-3,-3,-3)$. The lattice embedded plumbing graph of $S^3_{64/7}(T(2,3))$ does however fit into family $\mathcal{C}$ of [26] of symplectically embeddable plumbings. Unfortunately, family $\mathcal{C}$ of [26] contains surgeries on $T(2,3)$ that bound rational homology balls as well as ones that do not. For example, $S^3_{169/25}(T(2,3))$ of [26, Section 2.4, Figure 12] does not bound a rational ball despite bounding a plumbing with an embeddable intersection form. A later paper [6] classified which surgeries on $T(2,3)$ appearing in family $\mathcal{C}$, viewed as surface singularity links, bound a rationally acyclic Milnor fibre. Interestingly, all but two of the embedded graphs in that family are generated by applying IGOCL moves to the graph of $S^3_{64/7}(T(2,3))$. However, we do not know if any other members of family $\mathcal{C}$ bound a rational homology ball which is not a Milnor fibre. Hence, the following is a rich open question worth studying:

Question 5 *For which* $r \in \mathbb{Q}_+$ *does* $S^3_r(T(2,3))$ *bound a rational homology ball?*

1.3 Outline

We start with Section 2 by recalling some results on complementary legs and the basics of the lattice embedding setup. In Section 3 we define the GOCL and IGOCL moves and prove Theorem A. In Section 4 we prove Theorem B for the families (1)–(11), while the other families follow directly from [3; 16].

Acknowledgements

This research was carried out by the author when she was a PhD student at the University of Glasgow, funded by the College of Science and Engineering. This is an updated version of a part of her thesis.

The author would like to thank Ana Lecuona and Marco Golla for the numerous helpful discussions on which this article got built. She would also like to thank her PhD examiners Brendan Owens and Peter Feller for their thorough reviews of an earlier version of this article, which significantly improved the exposition.

2 Complementary legs and lattice embeddings

We list some definitions and easy propositions that are helpful to understanding the paper. We recall the definition of lattice embeddings and apply it to plumbing graphs and complementary legs. In this section, we assume that the reader is familiar with plumbings of disc bundles over spheres and how to convert plumbing diagrams into Kirby diagrams. (If not, see [11, Example 4.6.2].)

Notation 6 Given a forest-shaped plumbing graph Γ with weight function $W\colon V \to \mathbb{Z}$, we may associate to it a 4-manifold X_Γ by describing its Kirby diagram. First, we draw a small unknot at each vertex of Γ. Then, for each edge, we create a Hopf linking between the knots corresponding to the edge ends as in Figure 4. We denote the resulting link by L_Γ. Then X_Γ is the simply connected 4-manifold obtained by attaching 2-handles with framing $W(v)$ to the unknot at each vertex v.

We denote the 3-manifold ∂X_Γ by Y_Γ.

Remark A common abuse of terminology is "the plumbing graph Γ bounds a rational homology ball", which means that Y_Γ bounds a rational homology ball.

Let Γ be a forest-shaped plumbing graph. The second homology of X_Γ is the free abelian group $\mathbb{Z}\langle V_1, \dots, V_k\rangle$ on the vertices and the intersection form is

$$\langle V_i, V_j\rangle_{Q_X} = \begin{cases} \text{weight of } V_i & \text{if } i = j, \\ 1 & \text{if } V_i \text{ is adjacent to } V_j, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 7 Let X be a 4-manifold with boundary. A *lattice embedding*

$$f\colon (H_2(X)/\operatorname{Torsion}, Q_X) \hookrightarrow (\mathbb{Z}^N, -\mathrm{Id})$$

is a linear map f such that $\langle V_i, V_j\rangle_{Q_X} = \langle f(V_i), f(V_j)\rangle_{-\mathrm{Id}}$. We will simply write $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_{-\mathrm{Id}}$. If nothing else is specified, then $N = \operatorname{rk} H_2(X)$, the number of vertices in the graph.

Common abuses of notation include "embedding of the graph", meaning an embedding of the lattice $(H_2(X)/\operatorname{Torsion}, Q_X)$, where $X = X_\Gamma$ for a plumbing graph Γ.

Knowing when a lattice embedding exists is useful because of Proposition 1 in the introduction.

We turn our heads to lattice embeddings of specific plumbing graphs, namely pairs of complementary legs.

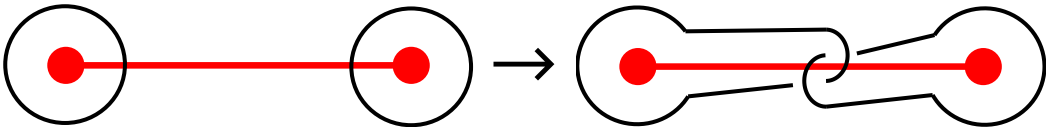


Figure 4: Hopf linking associated to an edge.

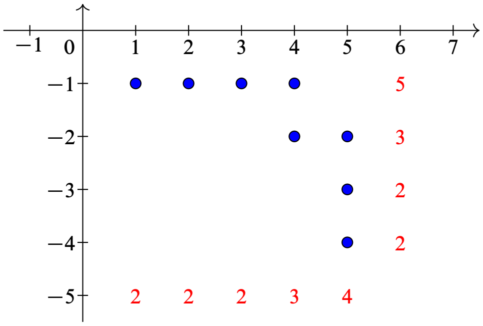


Figure 5: Example of a Riemenschneider diagram representing the complementary fractions $[5,3,2,2]^-$ and $[2,2,2,3,4]^-$.

Definition 8 We define the negative continued fraction $[a_1,\dots,a_n]^-$ as

$$[a_1,\dots,a_n]^- = a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots - \cfrac{1}{a_n}}}.$$

Negative continued fractions often show up in low-dimensional topology because of the slam-dunk Kirby move [11, Figure 5.30], which allows us to substitute a rational surgery on a knot by an integral surgery on a link.

Definition 9 A two-component weighted linear graph $(-\alpha_1,\dots,-\alpha_n), (-\beta_1,\dots,-\beta_k)$ (with α_i, β_j integers greater than or equal to 2) is called a pair of complementary legs if

$$\frac{1}{[\alpha_1,\dots,\alpha_n]^-} + \frac{1}{[\beta_1,\dots,\beta_k]^-} = 1.$$

We call the sequence $(\beta_1,\dots,\beta_k)$ the Riemenschneider dual or complement of the sequence $(\alpha_1,\dots,\alpha_n)$, and we call the fractions $[\alpha_1,\dots,\alpha_n]^-$ and $[\beta_1,\dots,\beta_k]^-$ complementary.

Definition 10 A Riemenschneider diagram is a finite set of points S in $\mathbb{Z}_+ \times \mathbb{Z}_-$ such that $(1,-1) \in S$ and for every point $(a,b) \in S$ but one, exactly one of $(a+1,b)$ or $(a,b-1)$ is in S. If $(n,k) \in S$ is the point with the largest $n-k$, we say that the Riemenschneider diagram represents the fractions $[\alpha_1,\dots,\alpha_n]^-$ and $[\beta_1,\dots,\beta_k]^-$, where α_i is one more than the number of points with $x = i$ and β_j is one more than the number of points with $y = -j$.

Example 11 See Figure 5 for an example of a Riemenschneider diagram.

Proposition 12 (Riemenschneider [22]) *The two fractions represented by a Riemenschneider diagram are complementary.*

Remark Note that given any continued fraction $[\alpha_1,\dots,\alpha_n]^-$ with all $\alpha_i \geq 2$, we may construct a Riemenschneider diagram representing $[\alpha_1,\dots,\alpha_n]^-$ and its Riemenschneider dual.

The following theorem is well known, but we explicitly write out the embedding construction for the reader's convenience.

Proposition 13 *Every pair of complementary legs has a lattice embedding.*

Proof The embedding can be constructed algorithmically from a Riemenschneider diagram. Denote the vertices of the two complementary legs by $(U_1, \dots, U_{m_1})$ and $(V_1, \dots, V_{m_2})$. These vertices generate the second homology of the plumbed 4-manifold described by the graph. We need to send every vertex to an element of $\mathbb{Z}\langle e_1, \dots, e_{m_1+m_2}\rangle$. We will construct this embedding recursively through "partial embeddings", which are maps $f\colon \{U_1, \dots, U_{m_1}, V_1, \dots, V_{m_2}\} \to \mathbb{Z}\langle e_1, \dots, e_{m_1+m_2}\rangle$ such that $\langle f(X), f(X)\rangle \geq$ weight of X.

Order the points in the Riemenschneider diagram so that $P_1 = (1, -1)$, and if $P_i = (a, b)$, then point P_{i+1} is either $(a+1, b)$ or $(a, b-1)$. Now, we recursively build an embedding as follows.

- Start by mapping both U_1 and V_1 to e_1.

- For each nonfinal i, if the current partial embedding is $(u_1, \dots, u_n)$, $(v_1, \dots, v_k)$ (that is, $(U_1, \dots, U_n)$ gets mapped to $(u_1, \dots, u_n)$ and $(V_1, \dots, V_k)$ gets mapped to $(v_1, \dots, v_k)$) and $P_i = (a, b)$ is such that $P_{i+1} = (a+1, b)$, then the new partial embedding will be $(u_1, \dots, u_n + e_{i+1})$, $(v_1, \dots, v_k - e_{i+1}, e_{i+1})$. If $P_{i+1} = (a, b-1)$, then the new partial embedding will be $(u_1, \dots, u_n - e_{i+1}, e_{i+1})$, $(v_1, \dots, v_k + e_{i+1})$ instead.

- If P_i is final and the current partial embedding is $(u_1, \dots, u_n)$, $(v_1, \dots, v_k)$, the new embedding will be $(u_1, \dots, u_n + e_{i+1})$, $(v_1, \dots, v_k - e_{i+1})$. (Or the other way around, as this sign choice is arbitrary.)

It is easy to see that an embedding $(u_1, \dots, u_{m_1})$, $(v_1, \dots, v_{m_2})$ constructed this way will have the properties:

- Each u_i for $i = 1, \dots, n-1$ and v_j for $j = 1, \dots, k$ will be a sum of consecutive basis vectors, all but the last one with coefficient 1, and the last one with coefficient -1. Meanwhile, u_n will be a sum of consecutive basis vectors, all with coefficient 1.

- If the Riemenschneider diagram represents the fractions $[\alpha_1, \dots, \alpha_n]^-$ and $[\beta_1, \dots, \beta_k]^-$, then we have $\langle u_i, u_i\rangle = -\alpha_i$ and $\langle v_j, v_j\rangle = -\beta_j$.

- Since u_i and u_{i+1} have exactly one basis vector in common, one with a positive coefficient and one with a negative one, $\langle u_i, u_{i+1}\rangle = 1$, and similarly $\langle v_i, v_{i+1}\rangle = 1$.

- The other pairs (u_i, u_j) (with $|i-j| > 1$) don't share basis vectors and are thus orthogonal. Similarly, the pairs (v_i, v_j) with $|i-j| > 1$ don't share basis vectors and are thus orthogonal.

- It is easy to show by induction on the construction that $\langle u_i, v_j\rangle = 0$ for all i, j.

These properties show that we are in fact looking at a lattice embedding of the complementary legs. □

Remark In fact, if e_1 is fixed to hit the first vertex of each complementary leg, the rest of the embedding is unique up to renaming of elements and sign of the coefficient [5, Lemma 5.2].

The following facts are useful when dealing with lattice embeddings. We will often use these properties without citing them. The first fact follows from reversing the Riemenschneider diagram, the second from embedding the sequences $(a_m, \dots, a_1)$ and $(b_n, \dots, b_1)$ as in Proposition 13 and mapping the (-1)-weighted vertex to $-e_1$, and the rest from looking at a Riemenschneider diagram.

Proposition 14 *Let $(a_1, \dots, a_m)$ and $(b_1, \dots, b_n)$ be complementary sequences. Then the following hold*:

(1) *The sequences $(a_m, \dots, a_1)$ and $(b_n, \dots, b_1)$ are complementary.*

(2) *The linear graph $(-a_1, \dots, -a_m, -1, -b_n, \dots, -b_1)$ embeds in $(\mathbb{Z}^{m+n}, -\mathrm{Id})$.*

(3) *Either a_m or b_n must equal 2, so assume without loss of generality that $b_n = 2$. Blowing down the -1 in the linear graph $(-a_1, \dots, -a_m, -1, -b_n, \dots, -b_1)$ gives us the linear graph*

$$(-a_1, \dots, -(a_m - 1), -1, -b_{n-1}, \dots, -b_1).$$

This graph is once again a pair of complementary legs linked by a -1, described by the Riemenschneider diagram obtained by removing the last point.

(4) *Repeatedly blowing down the -1 in linear graphs of the form*

$$(-a_1, \dots, -a_m, -1, -b_n, \dots, -b_1)$$

eventually takes us to $(-2, -1, -2)$, or even further to $(-1, -1)$ or (0).

(5) *Similarly, blowing up next to the -1 gives the linear graph*

$$(-a_1, \dots, -(a_m + 1), -1, -2, -b_n, \dots, -b_1)$$

or

$$(-a_1, \dots, -a_m, -2, -1, -(b_n + 1), \dots, -b_1),$$

which are both pairs of complementary legs connected by a -1, described by Riemenschneider diagrams that are expansions of the initial one by one dot.

3 Growing complementary legs on Lisca's graphs

The idea for this work comes from studying the lattice embeddings of linear graphs and other trees that are known to bound rational homology 4-balls. Consider for example Lisca's classification of connected linear graphs that bound rational homology 4-balls [17], in the most convenient form for us described by Lecuona in [15, Remark 3.2]. Every family of embeddable graphs can be obtained from the basic graphs $(-2, -2, -2)$, $(-2, -2, -3, -4)$, $(-3, -2, -2, -3)$ and $(-3, -2, -3, -3, -3)$ by repeated application of two types of moves, one of which is the following: choose a basis vector e hitting exactly two vertices v and w, where w is *final* (Lisca's word for leaf, that is, a vertex of degree 1 [17, page 6]), subtract 1 from

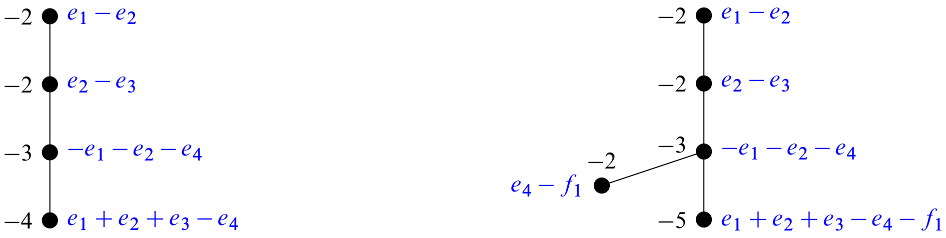


Figure 6: Left: Lisca's $(-2,-2,-3,-4)$ graph with embedding. Right: an expansion of the graph on the left.

the weight of v and attach a new vertex u of weight -2 to w. I will show that we can do away with the assumption that w is final and still get 3-manifolds bounding rational homology 4-balls through repeating this operation.

Example 15 Consider Figure 6 (left), showing an embedding of Lisca's $(-2,-2,-3,-4)$ graph into the standard lattice $(\mathbb{Z}\langle e_1, e_2, e_3, e_4\rangle, -\mathrm{Id})$. Note that e_4 and e_3 hit two vertices each. Choose e_4. We can perform the operation described above by choosing v to be the vertex of weight -4 and w the vertex of weight -3. The result is shown in Figure 6 (right) together with its embedding, which is a kind of "expansion" of the embedding in Figure 6 (left). Our new embedding has two basis vectors hitting exactly two vertices each, namely e_3 and f_1, whereas e_4 now hits three vertices. We may now perform the same operation again on any of these basis vectors, thereby obtaining any graph of the form described in Figure 3, with $k = 0$. We will show that these graphs do not only have lattice embeddings, but also bound rational homology 4-balls.

3.1 GOCL and IGOCL moves

We will now introduce two moves on forest-shaped plumbing graphs with a lattice embedding. Let $\Gamma = (V, E, W)$ be a weighted negative definite graph with lattice embedding $F\colon (V, Q_{X_\Gamma}) \to (\mathbb{Z}^{|V|}, -\mathrm{Id})$. Assume that there is a basis vector e of $\mathbb{Z}^{|V|}$ hitting exactly two vertices A and B in Γ, whose images are v and w, in any order we prefer. Then a *GOCL* (*growth of complementary legs*) *operation* consists of constructing an embedded graph $(\Gamma' = (V', E', W'), F')$ by $V' = V \cup C$, $E' = E \cup \{AC\}$ and $u := F'(C) = -\langle e, v\rangle e - f$, $w' := F'(B) = w - \langle e, v\rangle\langle e, w\rangle f$ and $F'(D) = F(D)$ for all $D \in V - \{B\}$. This move is illustrated in Figure 7. Note that $\langle e, v\rangle\langle e, w\rangle = \langle f, u\rangle\langle f, w'\rangle$. Thus, the GOCL operation substitutes e by f in the set of basis vectors hitting the graph exactly twice and the sign difference between

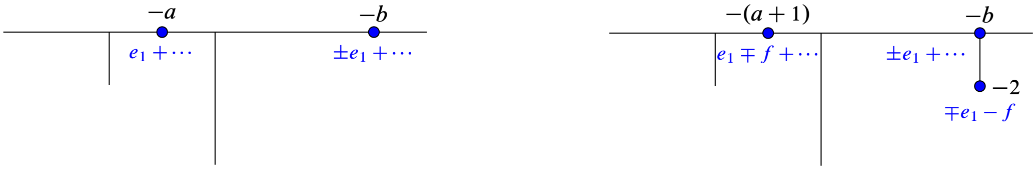


Figure 7: A GOCL move on a graph with a lattice embedding. Left: before GOCL. Right: After GOCL.

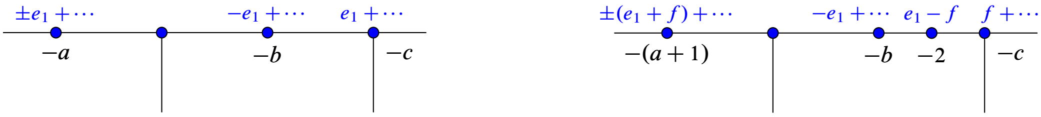


Figure 8: An IGOCL move on a graph with a lattice embedding. Left: before IGOCL. Right: after IGOCL.

the two occurrences of the basis vector is preserved. This operation can therefore be applied repeatedly. If we start with the graph consisting of two vertices of weight -2 and no edges, and the embeddings e_1-e_2 and e_1+e_2, then repeated application of GOCL will simply give us two complementary legs.

The other operation which we will call *IGOCL* (*inner growth of complementary legs*) could be described as growing complementary legs from the inside. Suppose a basis vector e hits exactly three vertices A, B and C in Γ, with their images under the lattice embedding F being u, v and w respectively. Assume also that B and C are adjacent and that $\langle v,e\rangle\langle w,e\rangle=-1$, that is, e hits v and w with opposite signs. Then $\Gamma'=(V',E',W')$ is described by $V'=V\cup\{D\}$, $E'=(E-\{BC\})\cup\{BD,DC\}$, $F'(D)=-\langle v,e\rangle e+\langle v,e\rangle f$, $F'(C)=w-\langle w,e\rangle e+\langle w,e\rangle f$, $F'(A)=u+\langle u,e\rangle f$ and $F'(X)=F(X)$ for all $X\in V-\{A,C\}$. This operation is illustrated in Figure 8. After this operation is performed, we can perform it again on either e or f, but the result is essentially the same. What it does is grow a chain of -2's between two vertices and compensate by subtracting from the weight of a different vertex. If we apply the IGOCL operation on a vector hitting a pair of complementary legs three times, we still get a pair of complementary legs, which explains the name.

3.2 The complicated method to show Theorem A

Now that we have defined the GOCL and IGOCL moves, we want to apply them to Lisca's basic graphs, which are $(-2,-2,-2)$, $(-2,-2,-3,-4)$, $(-3,-2,-2,-3)$, and $(-3,-2,-3,-3,-3)$. We will show for each Lisca graph one by one that the results obtained from repeatedly applying the aforementioned operations always bound rational homology balls. Recall that this can be done for all families using the simple method of Proposition 2. This subsection explains the complicated method to do that, which has the bonus of showing the χ-sliceness of some links including many Montesinos links not treated by Lecuona in [15].

A link in $S^3=\mathbb{R}^3\cup\{\infty\}$ is called *strongly invertible* if it is ambient isotopic to one which both is equivariant with respect to the 180° rotation around the x-axis and fulfils the property that each component intersects the x-axis in exactly two points [20]. Now, recall Notation 6. If Γ is a tree, then the link L_Γ is strongly invertible. Let D be some Kirby diagram of X_Γ such that L_Γ is in the equivariant position. For example, Figure 9 shows a possible diagram D for the plumbing graph in Figure 6 (right).

Let $\pi_D\colon X_\Gamma\to X_\Gamma$ be the involution given by extending this 180° rotation around the x-axis and let $p_D\colon X_\Gamma\to X_\Gamma/\pi_D$ be the quotient map when we identify x with $\pi_D(x)$. By [20, Theorem 3], $X_\Gamma/\pi_D\cong B^4$ and p is a double covering, branched over a surface $S_D\subset B^4$. The surface S_D can be

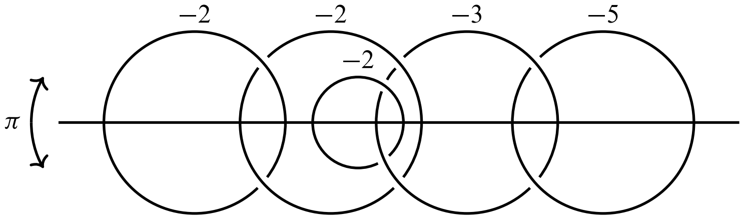


Figure 9: Proof that the attaching link of the graph in Figure 6 (right) is strongly invertible.

drawn by attaching bands to a disc according to the bottom half of the rotation-equivariant drawing, adding as many half-twists as the weight of the corresponding unknot [20]. (See Figure 10.)

By K_D we denote the link $K_D = S_D \cap S^3$. Note that $K_D \neq L_\Gamma$, and also that $K_D = \partial S_D$ and that Figure 10 should be understood with the interior pushed into B^4. Thirdly, note that S_D and K_D depend on the choice of D.

In [8, Definition 1], Donald and Owens define a useful generalisation of sliceness to links. We say that a link in S^3 is χ*-slice* if it bounds a surface of Euler characteristic 1 without closed components in B^4. If the surface has no local maxima, we can call it χ*-ribbon*. The usefulness of the notion of χ-sliceness lies in the following proposition:

Proposition [8, Proposition 2.6] *If L is a nonzero-determinant link which bounds a surface $F \subset B^4$ with no closed components and of Euler characteristic* 1, *then $\Sigma_2(B^4, F)$ is a rational homology ball.*

We now state the proposition at the heart of the complicated method to prove Theorem A:

Proposition 16 (complicated method) *Let Γ be a tree-shaped negative definite plumbing graph, and D an equivariant Kirby diagram of X_Γ as above. If we can equivariantly add n 2-handles with unknotted attaching circles to D and obtain a Kirby diagram D' that describes the* 3*-manifold $\#^n(S^1 \times S^2)$, then K_D is χ-ribbon, and thus $Y_\Gamma \cong \Sigma_2(S^3, K_D)$ bounds a rational homology ball.*

Before we prove this proposition, let us state the following lemma, which is interesting in itself. It follows directly from combining the Smith conjecture (proven by Waldhausen in [27]) and [13, Proposition 5.1].

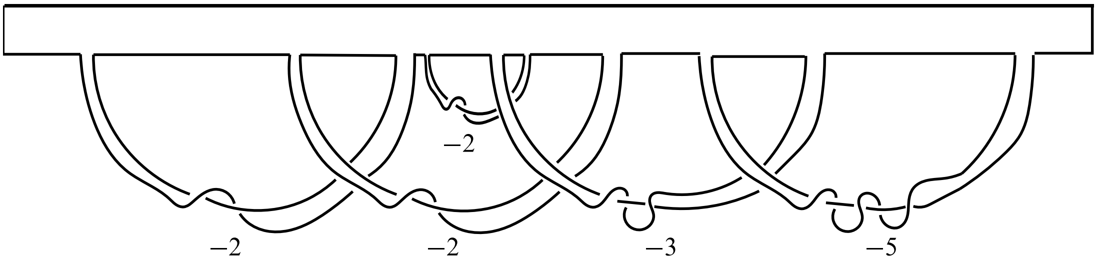


Figure 10: S_D for the diagram D in Figure 9 is a disc with five bands attached. It lives inside B^4 with the boundary lying on S^3, and the interior pushed inside B^4.

Lemma 17 *If L is a link such that $\Sigma_2(S^3, L) \cong \#^n(S^1 \times S^2)$, then L is the unlink of $n+1$ components.*

Proof of Proposition 16 Since $\Sigma_2(S^3, K_{D'}) \cong \#^n(S^1 \times S^2)$, $K_{D'}$ must, by Lemma 17, be the unlink of $n+1$ components. Since $S_{D'}$ is obtained from S_D by attaching bands, the attachment of bands yields a cobordism $(S^3 \times I, B)$ from (S^3, K_D) to $(S^3, K_{D'})$ with only index 1 critical points. Since $K_{D'}$ is the unlink on $n+1$ components, it bounds $n+1$ discs in B^4. Let S be the union of B and these discs. It has no maxima and it retracts onto a graph with $n+1$ vertices and n edges, implying that it has Euler characteristic 1. This shows that K_D is χ-ribbon, and thus, by [8, Proposition 2.6], $Y_\Gamma \cong \partial\Sigma(B^4, S)$, where $\Sigma(B^4, S)$ is a rational homology ball. □

3.3 (−2 , −2 , −2)

This graph has embedding $(e_1 - e_2, e_2 - e_3, -e_1 - e_2)$. The only basis vector hitting twice is e_1 whose both occurrences are in final vertices. Thus applying the GOCL operation keeps the graph linear and all such graphs have been shown by Lisca to be bounding rational homology balls. In fact, these graphs describe the lens spaces $L(p^2, pq \pm 1)$, for $p > q > 0$ with some orientation [17, Lemma 9.2].

3.4 (−3, −2 , −3, −3, −3)

We now show the following proposition using the complicated method described in Section 3.2.

Proposition 18 *Every 3-manifold described by the graph in Figure 1 bounds a rational homology 4-ball.*

In particular, we will apply Proposition 16 to any graph Γ in the family described by Figure 1. We will construct an equivariant (with respect to the 180° rotation around the z-axis) Kirby diagram D' of $S^1 \times S^2$, with all components unknotted and intersecting the z-axis exactly twice, such that removing one of the components gives us a Kirby diagram D of X_Γ.

The proof will be by Kirby calculus, so in Figure 11 we recall the effect of some blow-ups and blow-downs on Kirby diagrams. Recall that if there are k strands of a link component (counted with sign) in a bunch around which we perform a ± 1 blow-up, then the framing of that component increases by $\pm k^2$.

Proof We start with the chain

$$\big(-b_{n_1}, \dots, -b_1, -(2+k), -1, \underbrace{-2, \dots, -2}_{k}, -(1+\beta_1), -\beta_2, \dots, -\beta_{n_2}\big).$$

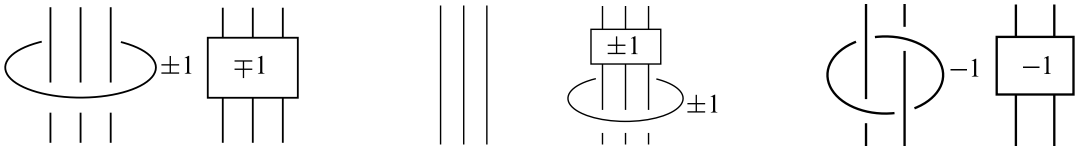


Figure 11: Useful blow-ups and blow-downs. Left: simple blow-down. Middle: simple blow-up. Right: twisted blow-down.

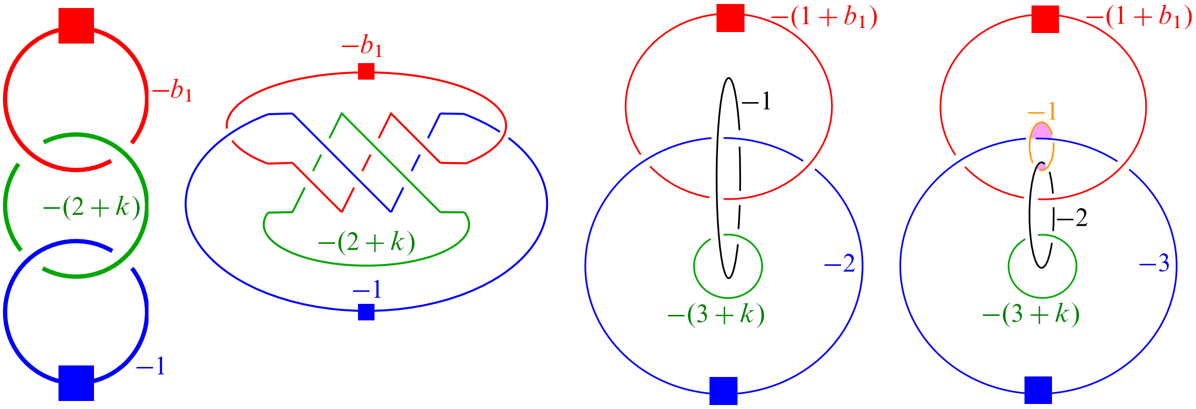


Figure 12: Creating a 3-cycle. First: start. Second: isotope. Third: apply Figure 11 (left). Fourth: result.

It consists of two Riemenschneider dual chains connected by -1, so by Proposition 14(4), it blows down to the (0) chain. Since the graph is a tree, the sign of the crossings doesn't matter yet. We will arrange the crossings around the $-(2+k)$-vertex as in Figure 12 (first image). The chains $(-b_2, \dots, -b_{n_1})$ and $(-2, \dots, -2, -(1+\beta_1), -\beta_2, \dots, -b_{n_2})$, on the other hand, are represented by tiny squares that freely move on their respective components since they are not relevant to the Kirby moves that follow.

We apply the Kirby calculus of Figure 12. The diagram in Figure 12 (third image) can no longer be described by a plumbing graph where every vertex is a link component. In fact, the red, blue, and black components form a triple Hopf link. Let us perform a blow-up at the clasp between the blue and the black components. This is a negative clasp, so we need to perform the twisted blow-up of Figure 11 (right). We obtain Figure 12 (fourth image).

Removing the amber (-1)-knot in Figure 12 (fourth image) yields the Kirby diagram of Figure 1 with $(a_1, \dots, a_{m_1}) = (2)$. To obtain more general tuples $(a_1, \dots, a_{m_1})$, we will have to repeatedly blow up the clasps on the (-1)-weighted component. At the moment, both of these clasps in Figure 12 (fourth image) (in magenta) are negative and can be blown up using Figure 11 (right), an operation that substitutes a clasp by a (-1)-weighted ring with two negative clasps on either side. Thus, repeated blow-ups will give us a figure like Figure 12 (fourth image), but with the amber ring potentially substituted by a longer chain. In any case, the link in this figure is strongly invertible, and removing the (-1)-weighted component yields the Kirby diagram of Figure 1. □

3.5 $(-3, -2, -2, -3)$

In this subsection, we show the following proposition:

Proposition 19 *Every 3-manifold described by the graph in Figure 2 bounds a rational homology 4-ball.*

This is done by applying Proposition 16 to graphs Γ described by Figure 2. In particular, for each member of this family, we will construct a Kirby diagram D' that (1) describes the 3-manifold $(S^1 \times S^2) \# (S^1 \times S^2)$, (2) is equivariant with respect to the 180° rotation around the z-axis, (3) only consists of unknotted components that intersect the z-axis exactly twice, and (4) is such that removing two of the components yields a Kirby diagram of X_Γ.

Proof One Kirby diagram of $(S^1 \times S^2) \# (S^1 \times S^2)$ is a $(0,0)$-framed unlink of two components. By Proposition 14(4), one of the unknots with framing 0 blows up to the chain

$$(-\beta_{m_2}, \dots, -\beta_2, 1-\beta_1, 1-b_1, -b_2, \dots, -b_{m_1}),$$

which can be seen by noting that the above chain is one blow-up away from

$$(-\beta_{m_2}, \dots, -\beta_1, -1, -b_1, \dots, -b_{m_1}).$$

Our first move will be to link the other unknot with framing 0 to the component with framing $1-\beta_1$ using a blow-up, thus obtaining Figure 13 (left). In this figure, the chains $(-b_2, \dots, -b_{m_1})$ and $(-\beta_2, \dots, \beta_{m_2})$ are represented by tiny squares that freely move on their respective components and if there are two on the same one, they could even pass through each other. Figure 13 (middle) is obtained from Figure 13 (left) by a simple isotopy. Note that the purple (-1)-weighted component is linked with the black and the blue ones with negative clasps. We may thus use Figure 11 (right) to blow it up into an arbitrary chain of negative clasps as in Figure 13 (right).

Zoom into the lower part of Figure 13 (right) and note that it looks like Figure 14 (left). It is isotopic to Figure 14 (middle), which clearly blows up to Figure 14 (right). This shows that Figure 13 (right) blows up to Figure 15 (left). Applying an isotopy of the link gives Figure 15 (middle). The green and the black components are now linked positively. The (-1)-blow-up that gets rid of this linking introduces a new component that links to the green and the black components with a negative and positive clasp respectively. Repeated blow-ups thus give us a chain with all clasps negative except the lowest one. We conclude that Figure 15 (right) is a $\mathbb{Z}/2\mathbb{Z}$-equivariant blow-up of Figure 15 (middle) and hence of the $(0,0)$ surgery on

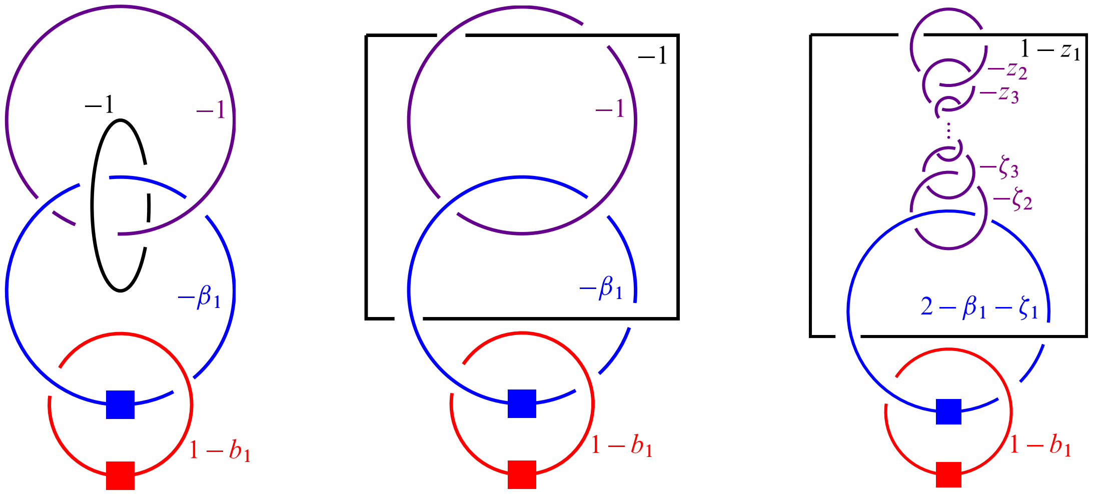


Figure 13: Creating a $\mathbb{Z}_2$-equivariant Kirby diagram of Figure 2 with two extra 2-handles.

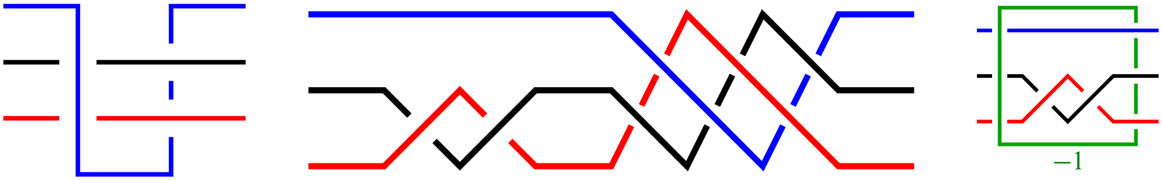


Figure 14: Zooming in on a part of Figure 13 (right) and blowing up.

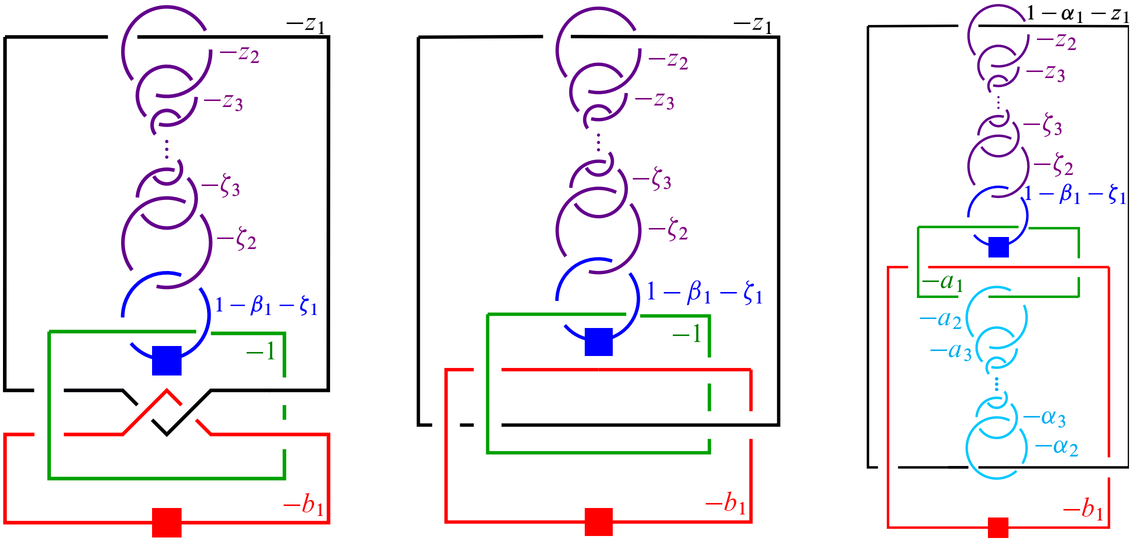


Figure 15: Creating a $\mathbb{Z}/2\mathbb{Z}$-equivariant Kirby diagram of Figure 2 with two extra 2-handles.

the 2-component unlink, but we may also note that removing the two (-1)-weighted components in the "middle" of the purple and turquoise chains gives us a tree-shaped plumbing, namely the one in Figure 2. □

3.6 $(-2,-2,-3,-4)$

We prove the following proposition:

Proposition 20 *Every 3-manifold described by the graph in Figure 3 bounds a rational homology 4-ball.*

This case will be shown in a much simpler way than the one in Sections 3.4 and 3.5. By Proposition 2, it is enough to show that performing two integral surgeries on the graphs in Figure 3 gives us $(S^1\times S^2)\#(S^1\times S^2)$.

Proof Figure 16 (top left) shows the "spine" of Figure 3 as the orange, magenta, green, red, teal, and violet rings. The complementary leg pairs are reduced to dark brown and olive green squares here. The weights of the rings with squares might be affected by the growth of the complementary legs. The blue ring with weight -1 is the first surgery we perform. This first choice of surgery is suggested to us by [5, Figures 17(5) and 17(6)], which provide equivariant surgeries on *linear* expansions of the graph $(-2,-2,-3,-4)$ that yield $(S^1\times S^2)$.

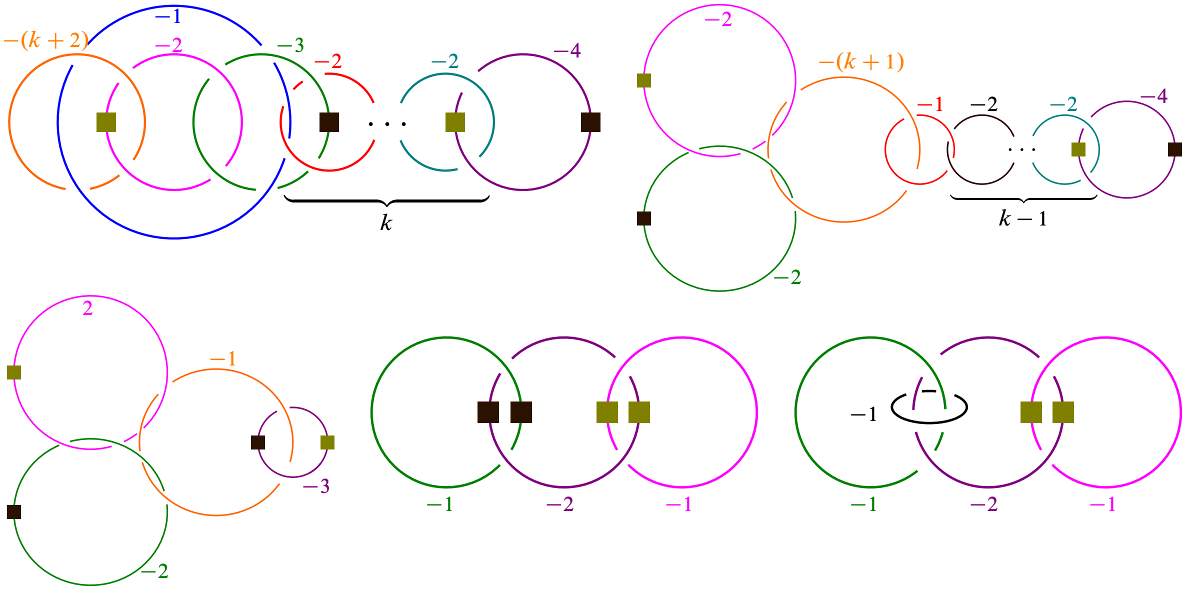


Figure 16: The blue component in the top left and the black one in the bottom right are the two surgeries performed to transform the 3-manifolds in Figure 3 into $(S^1 \times S^2) \# (S^1 \times S^2)$.

Blowing down the (-1)-weighted loop gives us Figure 16 (top right). The red loop has weight -1 and we may blow it down. We continue blowing down the chain of -2's that follows until we reach Figure 16 (bottom left). We blow down the orange loop and obtain Figure 16 (bottom middle).

Everything would be well now if only the little squares were not there, as we would have been able to blow down the figure to one 0-framed unknot. However, because we might have grown two pairs of complementary legs on this figure, Figure 16 (bottom middle) might not actually represent a Kirby diagram with any (-1)-framed loops to blow down.

Instead, we will find another surgery to perform and obtain $(S^1 \times S^2) \# (S^1 \times S^2)$. Here, it's easier to reason backwards. First, note that the 3-manifold in Figure 16 (bottom right) is $(S^1 \times S^2) \# (S^1 \times S^2)$. This can be seen by blowing down the black (-1)-loop, which would separate the green loop (now 0-framed) from the rest. The rest will be a $(-1, -1)$-chain with two complementary legs grown on it, which always blows down to (0). Note that repeated blow-ups of Figure 16 (bottom right) near the black (-1)-framed loop allows us to obtain Figure 16 (bottom middle) with an extra (-1)-framed surgery. This extra (-1)-framed surgery is the one we need to perform in order to obtain $(S^1 \times S^2) \# (S^1 \times S^2)$, completing our proof with the simple method. □

While all the links of Figure 16 appear possible to arrange in an equivariant way, the reader should note that we did not equivariantly add the two extra 2-handles at the same time. Finding an appropriate equivariant diagram D' for using Proposition 16 has proven difficult. It is currently unknown if K_D is χ-slice for any equivariant Kirby diagram D of X_Γ.

4 Proof of Theorem B

In this section, we prove Theorem B by studying the intersection between the graphs in Figures 2, 1 and 3, and plumbing graphs of positive rational surgeries on positive torus knots. First we describe the plumbing graphs of the surgeries on torus knots, and then we go through the intersections with the graphs in Figures 2, 1 and 3 one by one. The change of order compared to Section 3 is because some families obtained from Figure 1 are subfamilies of families obtained from Figure 2.

4.1 Plumbing graphs of rational surgeries on torus knots

In order to find the intersection between the plumbing graphs of rational surgeries on torus knots and the graphs obtained from Lisca's graphs by repeated GOCL and IGOCL moves, we need to know what the plumbing graphs of rational surgeries on torus knots look like. Let $n > 0$ be a *rational* number. We want to find a plumbing graph for $S^3_n(T(p,\alpha))$. We can write $n = [N_1 + p\alpha, N_2, \dots, N_k]^-$ for $N_2, \dots, N_k \geq 2$. The 3-manifold $S_n(T(p,\alpha))$ bounds the 4-manifold in Figure 17, which is positive definite if $n > 0$. Now, we will use the same technique as in [19, Section 3] in order to produce a definite plumbing graph. In the process, we need to measure how far we are from being definite, so the following definition is useful.

Definition 21 The *positive/negative index* of a 4-manifold is the number of positive/negative eigenvalues of its intersection form.

The argument of [19, Section 3] that the blow-ups decrease the surgery coefficient by a constant still holds to show that $S^3_n(T(p,\alpha))$ bounds the 3-manifold described by the graph in Figure 18.

The positive index of this graph is k by the same logic as in [19, Section 3]. To obtain a definite graph, we will need the following generalisation of the algorithm in [19, Figure 2]:

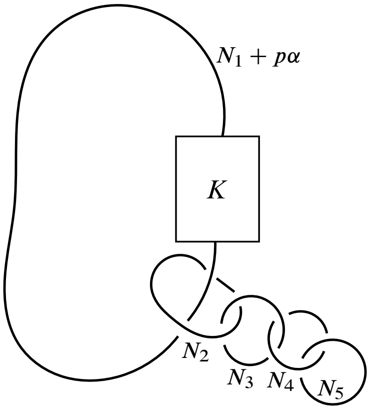


Figure 17: A Kirby diagram with boundary $S^3_n(T(p,\alpha))$ where $n = [N_1 + p\alpha, N_2, \dots, N_k]^-$ and $K = T(p,\alpha)$, here drawn for $k = 5$.

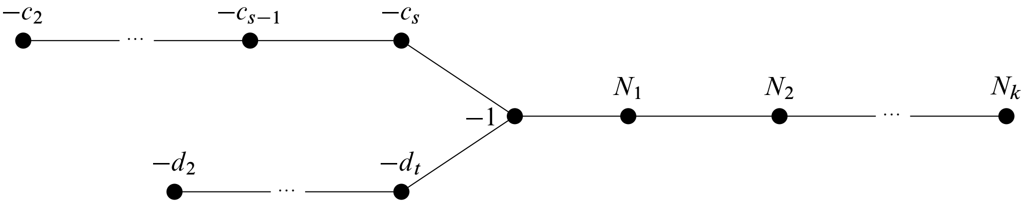


Figure 18: A plumbing graph of $S^3_n(T(p,\alpha))$, where $\alpha > p$. Here $[1, c_2, \dots, c_s]^- = p/\alpha$, $[d_1, \dots, d_t]^- = \alpha/p$ and $[N_1, \dots, N_k]^- = N = n - p\alpha$. In particular, the pair of fractions $([c_2, \dots, c_s]^-, [d_1, \dots, d_t]^-)$ are complementary. Also, we can write $(c_2, c_3, \dots, c_s) = (-2^{[d_1-2]}, a_1 + 1, a_2, \dots, a_r)$ so that $[a_1, \dots, a_r]^-$ and $[d_2, \dots, d_t]^-$ are complementary.

Proposition 22 *Let Γ be a tree-shaped plumbing graph containing a chain (a connected linear subgraph with no nodes, that is, vertices of degree greater than 2) $(-\alpha_1, \dots, -\alpha_k)$, as in Figure 19 (left). Let Γ' be the graph Γ with the chain substituted by the chain $(\beta_1, \dots, \beta_j)$, for complementary fractions $[\alpha_1, \dots, \alpha_k]^-$ and $[\beta_1, \dots, \beta_j]^-$, and the weight of the vertices adjacent to the chain increased by 1. Then $Y_\Gamma = Y_{\Gamma'}$. Moreover, $b^2_+(X_{\Gamma'}) = b^2_+(X_\Gamma) + j$ and $b^2_-(X_{\Gamma'}) = b^2_-(X_\Gamma) - k$.*

Example Before sketching the proof, we will provide an example of the algorithm that we use to change such a chain. Start with the linear graph $(-2, -4, -2)$. Right now, all vertices have negative weights. We want to introduce a positively weighted vertex. Let us perform a 1-blow-up. We obtain $(1, -1, -4, -2)$. Then we blow down the -1 and obtain $(2, -3, -2)$. We perform a 1-blow-up between the 2 and the -3 and obtain $(3, 1, -2, -2)$. Blowing up a 1 again between the last positively weighted vertex and the first negatively weighted one gives us $(3, 2, 1, -1, -2)$. We blow down the -1 to get $(3, 2, 2, -1)$ and again to obtain $(3, 2, 3)$. We note that every time we perform a 1-blow-up, we increase both the positive index and the number of positive vertices by 1, and every time we perform a (-1)-blow-down, we decrease both the negative index and the number of negative vertices by 1. Thus, changing these three negative vertices into three positive ones decreased the negative index by 3 and increased the positive index by 3.

Sketch of the proof Proposition 22 follows from the fact that blow-ups and blow-downs do not change the boundary 3-manifold, together with the following algorithm: (1) performing a 1-blow-up at the right of the rightmost chain element greater than 1, (2) blowing down any (-1)-weighted vertices, and (3) repeating. Following the Riemenschneider diagram, we see that this algorithm gradually substitutes a sequence by its Riemenschneider dual. Blowing up by 1 increases both the positive index and the number of vertices with positive weight by 1, and blowing down a -1 decreases both the number of vertices with

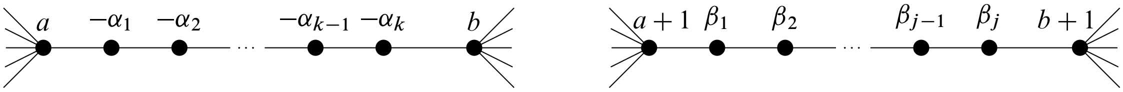


Figure 19: The graphs above bound the same 3-manifold if $[\alpha_1, \dots, \alpha_k]^-$ and $[\beta_1, \dots, \beta_j]^-$ are complementary fractions. Changing a negative chain (on the left) to a positive one (on the right).

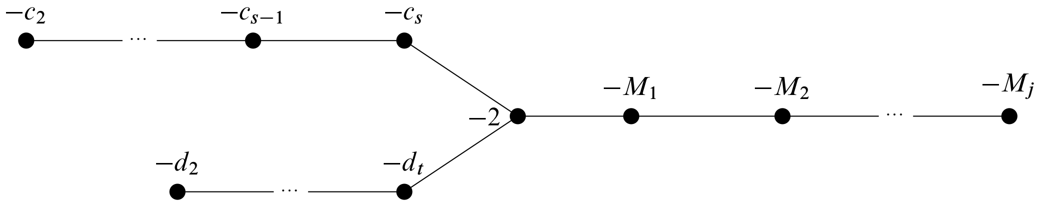


Figure 20: A negative definite plumbing graph of $S^3_n(T(p,\alpha))$, where $N = n-p\alpha > 1$ and $\alpha > p$. Here $[1,c_2,\dots,c_s]^- = p/\alpha$, $[d_1,\dots,d_t]^- = \alpha/p$ and if $N = n-p\alpha = a/b$ with $a,b \in \mathbb{Z}_{>0}$, then $[M_1,\dots,M_j]^- = a/(a-b)$.

negative weight and the negative index by 1. Thus, substituting the k negative-weighted vertices by j positive-weighted ones subtracts k from the negative index and adds j to the positive index. □

If $N > 1$ and thus $N_1 \geq 2$, we can use Proposition 22 to substitute the chain $(N_1,\dots,N_k)$ with its negative Riemenschneider complement $(-M_1,\dots,-M_j)$ and obtain the negative definite graph in Figure 20. If $0 < N < 1$, then the sequence $(N_1,\dots,N_k)$ starts with a 1 possibly followed by some 2's that we can blow down before turning the rest of the chain negative. This will once again give us a negative definite graph, namely the one in Figure 21.

If $N < 0$, that is $N_1 \leq 0$, then turning the positively weighted vertices $(N_2,\dots,N_k)$ negative will not be enough to decrease the positive index to 0. Instead, we will use Proposition 22 to turn the two other legs of our graph positive, and we obtain the graph in Figure 22, which has negative index 1. If $N_1 = 0$, we will perform a 0-absorption [21, Proposition 1.1] and obtain the positive definite graph in Figure 23. If $N_1 \leq 2$, we use Proposition 22 to turn it into a chain of 2's and obtain the graph in Figure 24. If $N_1 = -1$, we simply blow it down and obtain Figure 24, but with the length of the chain of 2's being 0.

In the graphs of Figures 20, 21, 23 and 24, the vertex of degree 3 is called the node. Removing the node splits the graph into three connected components, of which the top left one is called the *torso*, the bottom left one is called the *leg* and the right one is called the *tail*. This vocabulary is chosen to accord with the vocabulary of [19] on iterated torus knots. We also often talk about the torso, leg and tail collectively as legs. This comes from viewing the graphs as general star-shaped graphs rather than graphs of surgeries

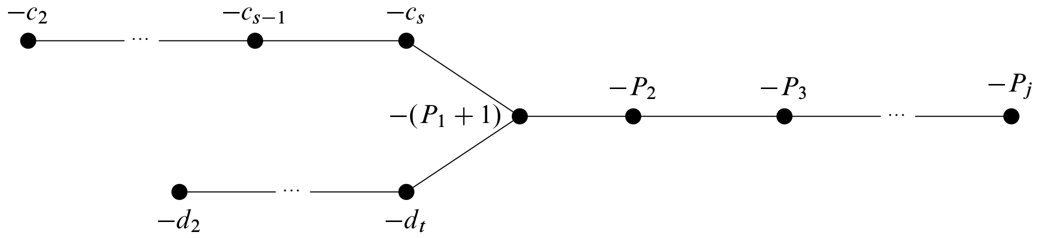


Figure 21: A negative definite plumbing graph of $S^3_n(T(p,\alpha))$, where $\alpha > p$ and $0 < N < 1$. Here $[1,c_2,\dots,c_s]^- = p/\alpha$, $[d_1,\dots,d_t]^- = \alpha/p$ and the fraction $[P_1,\dots,P_j]^-$ is complementary to $1/(1-N) = [N_2,\dots,N_k]^-$. In fact, this means that $N = 1/[P_1,\dots,P_j]^-$.

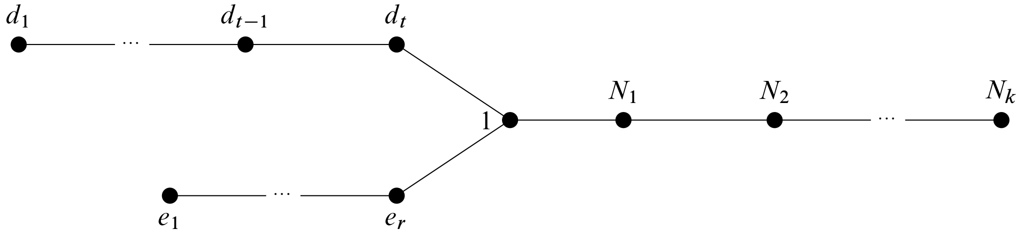


Figure 22: A plumbing graph of $S^3_n(T(p,\alpha))$, where $\alpha > p$ and $N < 0$. Here the negative index is 1, $[d_1,\dots,d_t]^- = \alpha/p$ and $[e_1,\dots,e_r]^-$ is complementary to $[d_2,\dots,d_t]^-$.

on torus knots specifically. (The author recommends looking at a flag of Sicily or the Isle of Man for a more precise metaphor.) This vocabulary is generally used by Lecuona, for instance in [15; 16].

We say that two legs of a star-shaped graph are negatively quasicomplementary if adding one vertex at the end of either leg could make them complementary, and positively quasicomplementary if removing a final vertex from one of the legs could. We say that two legs are quasicomplementary if they are either positively or negatively quasicomplementary. Note that the graphs in Figures 20, 21, 23 and 24 are exactly the star-shaped graphs with three legs whereof two are quasicomplementary. In the rest of this section, we are thus going to look for star-shaped graphs with a pair of quasicomplementary legs among the graphs in Figures 3, 2 and 1. The following very easy to check proposition will come in useful:

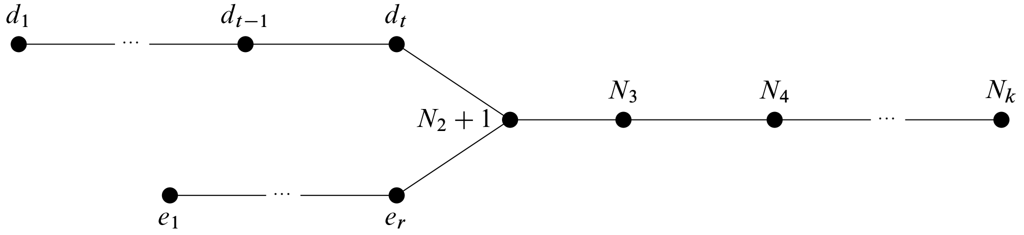


Figure 23: A positive definite plumbing graph of $S^3_n(T(p,\alpha))$, where $\alpha > p$ and $-1 < N < 0$. Here $[d_1,\dots,d_t]^- = \alpha/p$ and $[e_1,\dots,e_r]^-$ is complementary to $[d_2,\dots,d_t]^-$.

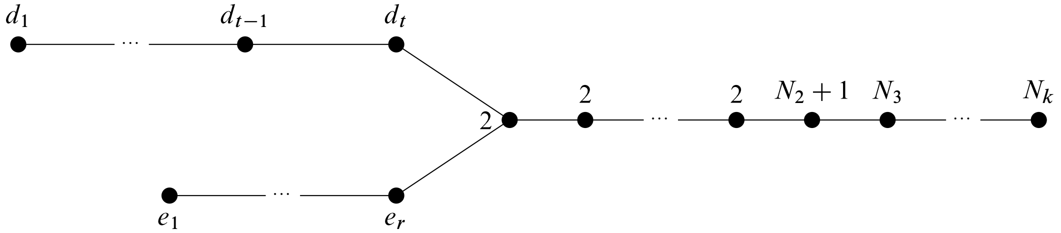


Figure 24: A positive definite plumbing graph of $S^3_n(T(p,\alpha))$, where $\alpha > p$ and $N < -1$. Here $[d_1,\dots,d_t]^- = \alpha/p$ and $[e_1,\dots,e_r]^-$ is complementary to $[d_2,\dots,d_t]^-$, and the tail starts with a chain of -2's of length $-N_1 - 1$.

Proposition 23 *Suppose* $Q/P = [a_1, \dots, a_n]^-$ *and* $(-a_1, -a_2, \dots, -a_n)$ *is either the leg or torso of the plumbing graph of* $S^3_r(T(p,\alpha))$*, a positive rational surgery on a positive torus knot. (Here* $-a_n$ *is the weight of the vertex adjacent to the node.) Then* α/p *is one of the following:*

- Q/P.
- $Q/(Q-P)$.
- $((l+1)Q+P)/Q$ *for some* $l \geq 0$.
- $((l+2)Q-P)/Q$ *for some* $l \geq 0$.

Note that if $\mathrm{GCD}(P,Q) = 1$, then all of these fractions are reduced. However, if $P = Q-1$, then, $\alpha/p = Q/(Q-P) = Q$ is a degenerate case that we ignore.

4.2 $(-3, -2, -2, -3)$

In this subsection, we prove the following:

Proposition 24 *For all torus knots* $T(p,q)$ *in families* (1), (2), (3), (4), (6) *and* (7) *of Theorem B, there exists an* $r \in \mathbb{Q}_+$ *such that* $S^3_r(T(p,q))$ *bounds a rational homology ball.*

This is done by considering the intersections between the graphs in Figures 20, 21, 23 and 24 (rational surgeries on torus knots) and the graphs in Figure 2 (graphs obtainable from $(-3,-2,-2,-3)$ through GOCL moves). Figure 2 is symmetric in the y-axis, so it is enough to try two of the vertices for trivalency, say the one with weight $1-\beta_1-\zeta_1$ and the one with weight $-a_1$.

If we want the vertex with weight $1-\beta_1-\zeta_1$ to be the trivalent vertex in one of Figures 20, 21, 23 and 24, then $l_1 = m_1 = 1$. Hence $\beta_i = \alpha_j = 2$ for all i and j. Also, $l_2 = 1$ or $n_1 = 1$. Suppose that $(-\beta_{m_2}, \dots, -\beta_2)$ is one of the quasicomplementary legs. Proposition 23 would generate that (p,q) belongs to families (1) and (2) in Theorem B. All of these are possible to produce by setting $l_2 = 1$, which frees us up to choose $(\zeta_2, \cdots, \zeta_{n_2})$ completely freely.

We consider what happens if the legs other than $(-\beta_2, \dots, -b_{m_2})$ are quasicomplementary. If $n_1 = 1$, all α, β and ζ become -2, giving us a star-shaped graph with two legs containing nothing but -2's, not allowing us out of the families (1), (2), (3) and (4). We consider the case $l_2 = 1$ instead. We have $a_1 = \alpha_1 = 2$. Let $b_1 = k+2$ (so that the leg $(-\beta_2, \dots, -\beta_{m_2}) = (-2, \dots, -2)$ has length k). We investigate if $(-\zeta_2, \dots, -\zeta_{n_2})$ and $(-2, -(k+2), -z_1-1, -z_2, \dots, -z_{n_1})$ can be quasicomplementary. Consider the diagram in Figure 25. The black dots represent the Riemenschneider diagram of $(z_1, \dots, z_{n_1})$ and $(\zeta_1, \dots, \zeta_{n_2})$. The blue dots are added in such a way that together with the black dots they form the Riemenschneider diagram of $(-2, -(k+2), -z_1-1, -z_2, \dots, -z_{n_1})$. Call it the BB diagram. The Riemenschneider diagram of $(-\zeta_2, \dots, -\zeta_{n_2})$ is to the right of the red line. Call it the RR diagram. Now we wonder if we can choose the black dots and k in such a way that the BB diagram is just the RR diagram plus one row or column at the end. However, we see that it is impossible to create a difference of one between the length of one leg and the complement of the other leg.

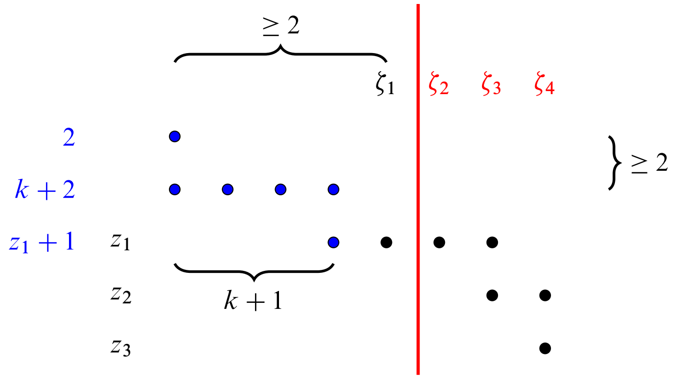


Figure 25: It is impossible to choose k so that the length or height difference between the full Riemenschneider diagram and the diagram to the right of the red line is one.

Consider the vertex labelled $-a_1$ being trivalent instead. This means that $m_1 = 1$. Either $n_1 = 1$ or $l_2 = 1$. First assume that $n_1 = 1$. This means that $\beta_1 = \cdots = \beta_{m_2} = \zeta_1 = \cdots = \zeta_{n_2} = 2$. Either n_2 or m_2 must be 1. No matter the choice, the left leg becomes $(-2, \dots, -2, -3)$ from the outside. If it is included in a pair of quasicomplementary legs, which we can always ensure since we can choose $(\alpha_2, \dots, \alpha_{l_1})$ freely, we use Proposition 23 to get all of families (3) and (4). In the more interesting case (where the leftmost leg is not one of the quasicomplementary ones) $(a_2, \dots, a_{l_1})$ must be quasicomplementary (from the inside) either to $(2, 1+k+\alpha_1, \dots, \alpha_{l_2})$ for some $k \geq 0$ (depicted in Figure 26 (left)) or to $(2+k, 1+\alpha_1, \dots, \alpha_{l_2})$ for some $k \geq 0$ (depicted in Figure 26 (right)), depending on whether n_2 or m_2 is equal to 1.

Let us resolve the first case. In order for $(a_2, \dots, a_{l_1})$ to be quasicomplementary to $(2, 1+k+\alpha_1, \dots, \alpha_{l_2})$ for some $k \geq 0$, the BB diagram should be the same as the RR diagram plus an extra row or column at the end. Since the part to the left of the red line has at least two columns, it must be an extra row. One solution would be $(\alpha_1, \dots, \alpha_{l_2}) = (3)$. If the black diagram has more than one row, we need

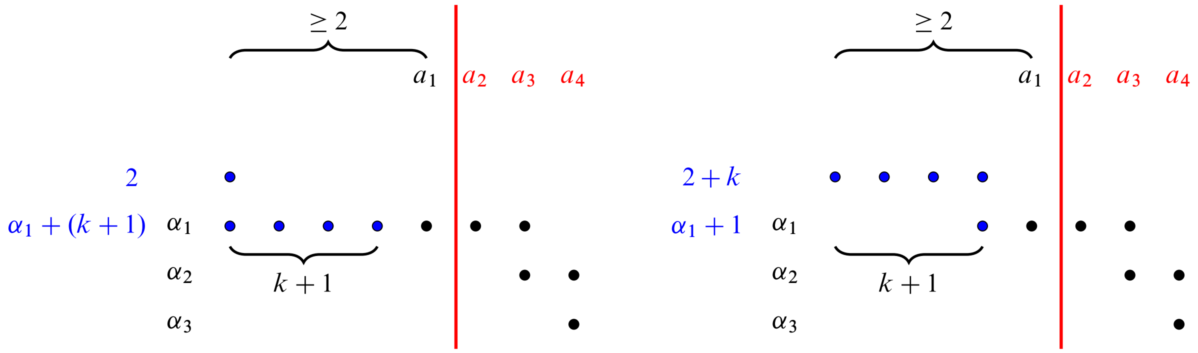


Figure 26: Riemenschneider diagram of the quasicomplementary legs $(-a_2, \dots, -a_{l_1})$ and $(-b_1, 1-\alpha_1-z_1, -\alpha_2, \dots, -\alpha_{l_2})$ in Figure 2 when $m_1 = n_1 = m_2 = 1$ (left) and $m_1 = n_1 = n_2 = 1$ (right).

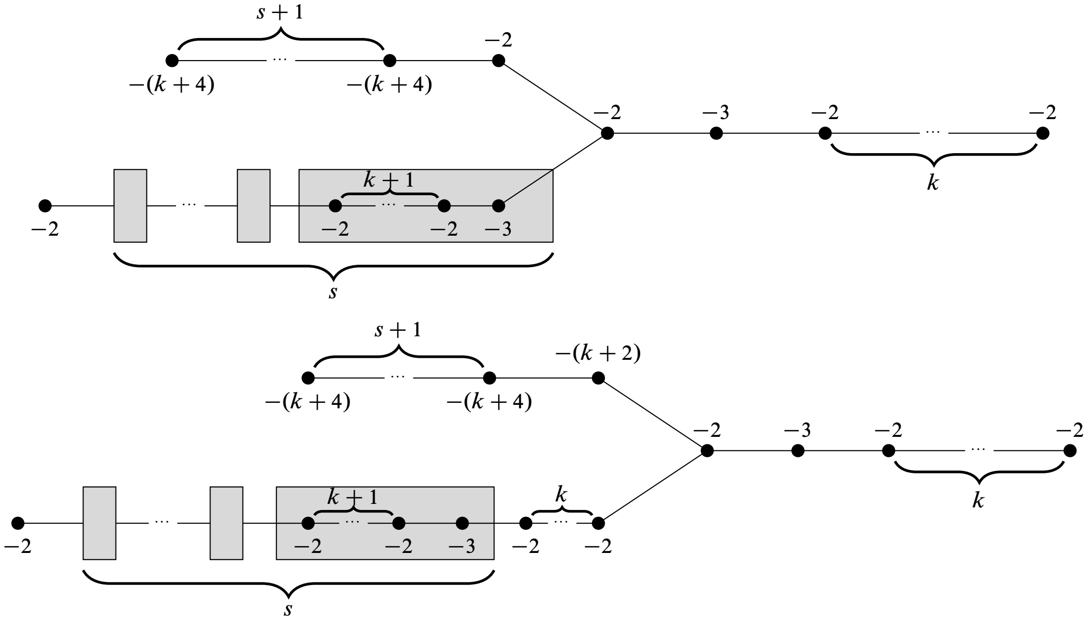


Figure 27: Strange two-parameter families of rational surgeries on positive torus knots that bound rational homology 4-balls.

$\alpha_2 = (k+1)+2+1 = k+4$. We can add as many rows as we want this way. We get that $(\alpha_1, \dots, \alpha_{l_2}) = (3, (k+4)^{[s]})$ and $(a_1, \dots, a_{l_1}) = (2, (3, (2)^{[k+1]})^{[s]}, 2)$, giving us the graph in Figure 27 (top). This graph is of the shape of Figure 24, so it describes $S^3_n(T(p,\alpha))$ for $\alpha/p = [(k+4)^{[s+1]}, 2]^-$ and $N = n - p\alpha = [-1, (2)^{[k+1]}]^- = -(2k+3)/(k+2)$. This corresponds to family (6) in Theorem B. A different formulation of the result is that $S^3_{p\alpha - \frac{2k+3}{k+2}}(T(p,\alpha))$ bounds a rational homology ball for all p and α described by

$$\begin{pmatrix}\alpha\\ p\end{pmatrix} = \begin{pmatrix}k+4 & -1\\ 1 & 0\end{pmatrix}^{s+1} \begin{pmatrix}2\\ 1\end{pmatrix}$$

for some $s, k \geq 0$. If we fix s, then α becomes a degree $s+1$ polynomial in k.

In the second case where $(a_2, \dots, a_{l_1})$ is quasicomplementary to $(2+k, 1+\alpha_1, \dots, \alpha_{l_2})$ for some $k \geq 0$, $(\alpha_1, \dots, \alpha_{l_2}) = (k+3, (k+4)^{[s]})$ for some $s \geq 0$. Then the graph becomes as in Figure 27 (bottom). Now $\alpha/p = [(k+4)^{[s+1]}, k+2]^-$, meaning that $S^3_{p\alpha - \frac{2k+3}{k+2}}(T(p,\alpha))$ bounds a rational homology ball for all p and α described by

$$\begin{pmatrix}\alpha\\ p\end{pmatrix} = \begin{pmatrix}k+4 & -1\\ 1 & 0\end{pmatrix}^{s+1} \begin{pmatrix}k+2\\ 1\end{pmatrix}$$

for some $s, k \geq 0$. This corresponds to family (7) in Theorem B.

If $l_2 = 1$ instead of $n_1 = 1$, then $a_1 = \cdots = a_{l_1} = 2$. We already know that we can choose surgery coefficients when one of the complementary legs consists of only -2's, so we do not need to check that case to formulate Theorem B. In fact we do not need to check further, as any star-shaped graphs with three legs whereof two are quasicomplementary, the third one consisting only of -2's and the node having weight -2 is a positive integral surgery on a positive torus knot, have been classified in [3].

4.3 $(-3, -2, -3, -3, -3)$

In this subsection, we prove the following:

Proposition 25 *For all torus knots $T(p,q)$ in the families* (5) *and* (8) *of Theorem B, there exists an $r \in \mathbb{Q}$ such that $S^3_r(T(p,q))$ bounds a rational homology ball.*

This is done by finding the intersections between the graphs in Figures 20, 21, 23 and 24 (rational surgeries on torus knots) and the graphs in Figure 1 (graphs obtainable from $(-3,-2,-3,-3,-3)$ through GOCL and IGOCL moves).

In Figure 1 there are three possibilities for a trivalent vertex. If we choose the vertex of weight $-a_1$, then $m_2 = 1$ and thus two of the legs are $(-3-k)$ and $(-2,\dots,-2)$. We already know that if one of these is in a quasicomplementary pair, then (p,α) lies in families (1)–(4) in Theorem B, so we get nothing new. Choosing the vertex of weight $-(1+b_1)$ to be trivalent, and noting that we land in families (1)–(4) if the left leg $(-3-k,-2)$ is one of the quasicomplementary ones, does however lead us to find that

$$S^3_{p\alpha-\frac{5}{7}}(T(p,\alpha))$$

bounds a rational homology ball for every

$$\begin{pmatrix}\alpha\\ p\end{pmatrix} = \begin{pmatrix}5 & -1\\ 1 & 0\end{pmatrix}^{s+1}\begin{pmatrix}3\\ 1\end{pmatrix}$$

where $s \geq 0$. This corresponds to family (8) in Theorem B. Choosing the vertex of weight $-(1+\alpha_1)$ to be trivalent gives us $m_1 = n_1 = 1$. If the lower leg $(-2,\dots,-2)$ is included in the pair of quasicomplementary legs, we fall into families (1)–(4) again. We need to investigate when $(-(3+k), -a_1, -(1+b_1))$ can be quasicomplementary to $((-2)^{[b_1-2]}, -3, (-2)^{[k]})$. The Riemenschneider dual of the latter leg is $(-b_1, -(k+2))$, so we need $b_1 = a_1$ and $k+2 = b_1+1$. Note that we also need $a_1 \geq 3$ in order to get a three-legged graph. Let $a = a_1 - 3$. We get $(-(3+k), -a_1, -(1+b_1)) = (-(a+5), -(a+3), -(a+4))$. Our graph is now as in Figure 23. Thus $\alpha/p = [a+5, a+3, a+4]^- = \frac{a^3+12a^2+45a+51}{a^2+7a+11}$. This corresponds to family (5) in Theorem B.

The remaining families follow from [3; 15].

4.4 $(-2, -2, -3, -4)$

In this subsection, we prove the following:

Proposition 26 *For all torus knots $T(p,q)$ in the families* (9), (10) *and* (11) *of Theorem B, there exists an $r \in \mathbb{Q}$ such that $S^3_r(T(p,q))$ bounds a rational homology ball.*

This is done by determining the intersection between the graphs in Figures 20, 21, 23 and 24 and the graphs in Figure 3, that is, between the graphs of surgeries on torus knots and the graphs of Figure 3, which we now know to bound rational homology balls.

To turn Figure 3 into a star-shaped graph, we will need to keep some of the grown complementary legs to length 1. If we let the vertex of weight $-1-a_1$ be trivalent, then $m_1 = 1$ and thus $(\beta_1, \dots, \beta_{m_2})$ consists only of 2's. If $m_2 > 1$ then $n_2 = 1$ and $(a_1, \dots, a_{n_1}) = (2, \dots, 2)$. In order to have trivalency of the $-(1+a_1)$ vertex, $\alpha_1 \geq 3$ is required. It is easy to check that in this case the only legs that can be quasicomplementary are the $(-b_1, -(2+k))$ one and the $(-2, \dots, -2)$ ($\alpha_1 - 2$ times) one. They can either be negatively quasicomplementary, made complementary by adding -3 at the end of the second leg, in which case $\alpha_1 = b_1$ and $k = 0$ have to hold, or they can be positively quasicomplementary, made complementary by removing $-(2+k)$ from the first one, in which case $\alpha_1 - 2 = b_1 - 1$. The first case shows that

$$S^3_{(2b_1^2-2b_1+1)^2/(2b_1^2-b_1+1)}(T(b_1-1, 2b_1-1))$$

bounds a rational homology ball for any $b_1 \geq 3$. The second case shows that

$$S^3_{((k+2)b_1^2-1)^2/((k+2)b_1^2+b_1-1)}(T(b_1, b_1(k+2)-1))$$

bounds a rational homology ball for all integers $b_1 \geq 2$ and $k \geq 0$. Both of these are subfamilies to families (1) and (2) in Theorem B that we will show can in fact be fully realised.

We get more interesting families when we let $m_2 = 1$, because then $(\alpha_1, \dots, \alpha_{n_2})$ can be anything as long as it has something but a 2 somewhere so that $n_1 > 1$. We will get graphs of the form in Figure 28 (left). To make the top and right legs quasicomplementary is easy: we need to choose whether they are to be positively or negatively quasicomplementary and which leg needs an extra vertex or a vertex removed to be complementary, and then we just need to choose $(a_2, \dots, a_{n_1})$ that make it happen. We use Proposition 23 for $Q/P = [2+k, 2]^- = \frac{1}{2}(2k+3)$. This corresponds to the entire families (3) and (4) as well as subfamilies of families (1) and (2) in Theorem B. The top and bottom legs cannot be made quasicomplementary.

The most interesting case to consider is whether the right and the bottom legs can be made quasicomplementary. In Figure 28 (right), the black dots show a Riemenschneider diagram of the complementary sequences $(a_1, \dots, a_{n_1})$ and $(\alpha_1, \dots, \alpha_{n_2})$. Adding the blue dots gives us a Riemenschneider diagram for the sequence

$$(\underbrace{2, \dots, 2}_{k}, \alpha_1 + 2, \alpha_2, \dots, \alpha_{n_2}),$$

with complement $(k+2, 2, a_1, \dots, a_{n_1})$. Considering only the part to the right of the red line gives us a Riemenschneider diagram for $(a_2, \dots, a_{n_1})$ (with the complement $(A_1, \dots, A_{n_3})$). In order for $(2, \dots, 2, \alpha_1 + 2, \alpha_2, \dots, \alpha_{n_2})$ and $(a_2, \dots, a_{n_1})$ to be quasicomplementary, either the picture to the right of the red line and the total picture without the last line, or the total picture and the picture to the right of the red line with an extra column, must be the same. The sequences $(k+2, 2, a_1, \dots, a_{n_1})$

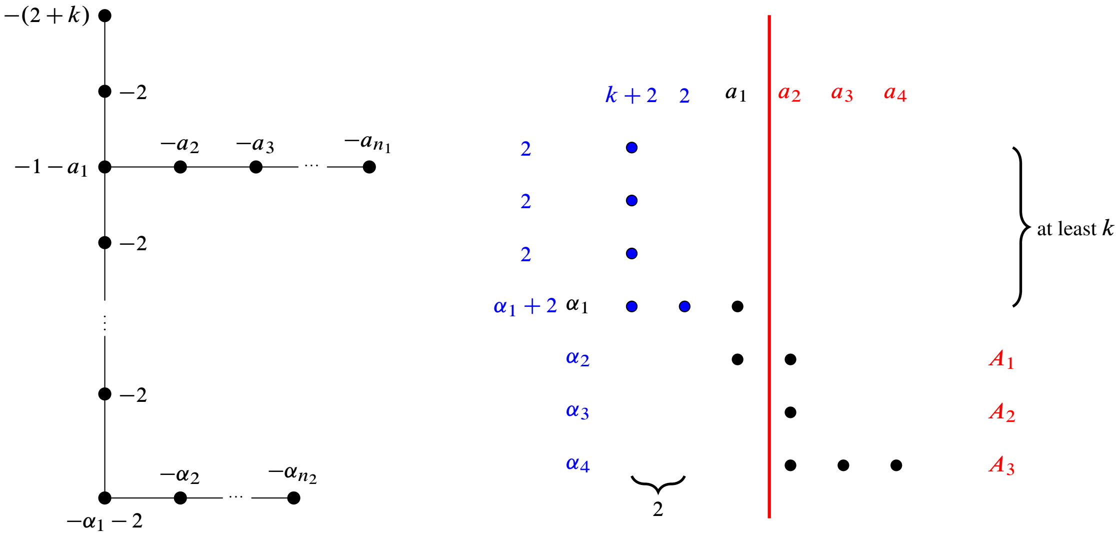


Figure 28: Choosing $m_1 = m_2 = 1$ in Figure 3 gives the star-shaped graph to the left.

and $(a_2, \dots, a_{n_1})$ have length difference 2, removing the second option. The only ways in which $(2, \dots, 2, \alpha_1 + 2, \alpha_2, \dots, \alpha_{n_2})$ and $(A_1, \dots, A_{n_3})$ can have length difference 1 is if any of the following hold:

(1) $k = 0$ and $a_1 = 3$, or

(2) $k = 1$ and $a_1 = 2$.

If $k = 0$ and $a_1 = 3$, then the first row of the total picture has length 3. Thus, in the second total row, to the right of the red line, we need three dots, making a total of four dots. This is a valid solution, namely $(\alpha_1, \dots, \alpha_{n_2}) = (2, 5)$, $(\alpha_1 + 2, \alpha_2, \dots, \alpha_{n_2}) = (4, 5)$, $(A_1, \dots, A_{n_3}) = (4)$ and $(a_1, \dots, a_{n_1}) = (3, 2, 2, 2)$. If we choose to continue and add α_3, that means adding a new row completely to the right of the red line, which must be as long as the second total row, namely four dots. That again gives a valid solution $(\alpha_1, \dots, \alpha_{n_2}) = (2, 5, 5)$, $(\alpha_1 + 2, \alpha_2, \dots, \alpha_{n_2}) = (4, 5, 5)$, $(A_1, \dots, A_{n_3}) = (4, 5)$ and $(a_1, \dots, a_{n_1}) = (3, 2, 2, 3, 2, 2, 2)$. We can continue this process and obtain the solution $(\alpha_1, \dots, \alpha_{n_2}) = (2, (5)^{[l]})$ and

$$(a_1, \dots, a_{n_1}) = ((3, 2, 2)^{[l]}, 2, 2)$$

for all $l \geq 1$. Our legs are positively quasicomplementary, so $\alpha/p = [d_1, \dots, d_t]^- = [(5)^{[l]}, 4]^-$. Since $5 - b/a = (5a - b)/a$, we have that

$$\begin{pmatrix} \alpha \\ p \end{pmatrix} = \begin{pmatrix} 5 & -1 \\ 1 & 0 \end{pmatrix}^l \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

for $l \geq 1$. This corresponds to family (6) in Theorem B. We can compute $N = [0, 3, 2, 2]^- = -\frac{3}{7}$. In other words, if $p_1 = 1$, $p_2 = 4$ and $p_{j+2} = 5p_{j+1} - p_j$ for all $j \geq 0$ [25, Tag A004253], we can say that

$$S^3_{p_j p_{j+1} - \frac{3}{7}}(T(p_j, p_{j+1}))$$

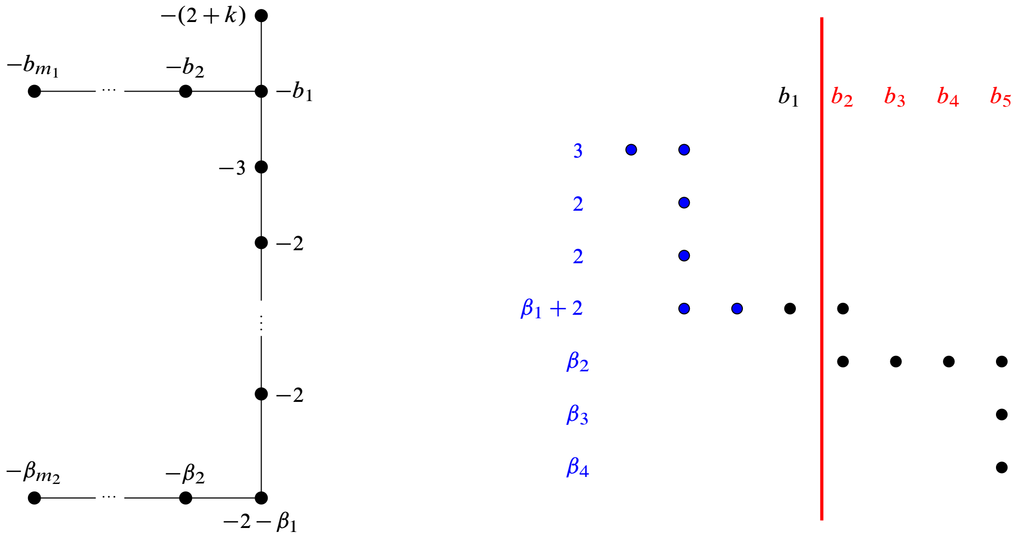


Figure 29: Choosing $n_1 = n_2 = 1$ in Figure 3 gives the star-shaped graph to the left.

bounds a rational homology ball for all $j \geq 1$. In this form it may not be obvious that the numerator of the surgery coefficient is a square, but in fact, $p_j p_{j+1} - \frac{3}{7} = \frac{1}{7} V_{j+1}^2$ for V_j being a sequence defined by $V_1 = 2$, $V_2 = 5$ and $V_{j+2} = 5V_{j+1} - V_j$ for all $j \geq 0$ [25, Tag A003501]. It is a shifted so-called Lucas sequence. The equality can be proven by first proving by induction that $p_{j+2} p_j - p_{j+1}^2 = 3$ for all $j \geq 0$, then noting that $V_{j+1} = p_{j+1} + p_j$ for all $j \geq 0$, and finally combining these equalities.

If $k = 1$ and $a_1 = 2$, the argument goes the same way. The only way for the right and bottom legs to be quasicomplementary is if the Riemenschneider diagram to the right of the red line and the total diagram missing the bottom line coincide. By the same argument as above, it happens if and only if $(\alpha_1, \dots, \alpha_{n_2}) = (3, (5)^{[l]})$ and $(a_1, \dots, a_{n_1}) = (2, (3,2,2)^{[l]}, 2)$ for all $l \geq 0$. In this case $\alpha/p = [(5)^{[l]}, 5, 2]^-$ and $N = [0,2,2,3]^- = -\frac{5}{7}$. This shows that if $Q_1 = 2$, $Q_2 = 9$ and $Q_{j+2} = 5Q_{j+1} - Q_j$ for all $j \geq 1$, then

$$S^3_{Q_j Q_{j+1} - \frac{5}{7}}(T(Q_j, Q_{j+1}))$$

bounds a rational homology ball for all $j \geq 1$. This corresponds to family (7) in Theorem B. Just as before, we can show that

$$Q_j Q_{j+1} - \tfrac{5}{7} = \tfrac{1}{7}(Q_j + Q_{j+1})^2.$$

Returning to Figure 3, we can let the vertex of weight $-b_1$ be the only node. That forces $n_1 = 1$, so $(\alpha_1, \dots, \alpha_{n_2}) = (2, \dots, 2)$. Putting $a_1 = 2$ would give us complete freedom in choosing $(b_2, \dots, b_{m_1})$, so Proposition 23 applied to $Q/P = 2 + k$ gives surgery coefficients n such that $S^3_n(T(k+1, k+2))$, $S^3_n(T(k+2, (l+2)(k+2)-1))$ and $S^3_n(T(k+2, (l+1)(k+2)+1))$ bound rational homology 4-balls. These families correspond to the entire families (1) and (2) in Theorem B. (Note however, that a couple of subfamilies of these will also be realised if we choose $a_1 > 2$ because $(b_2, \dots, b_{m_1}) = (2, \dots, 2)$. These subfamilies have an especially ample supply of choices of surgery coefficients.) If $n_2 > 1$, then

$m_2 = 1$ and $(b_1, \dots, b_{m_1}) = (2, \dots, 2)$. We will have three legs, namely $(-(2+k))$, $((-2)^{[\beta_1 - 2]})$ and $\bigl(-(1+a_1), (-2)^{[k]}, -(2+\beta_1), (-2)^{[a_1-2]}\bigr)$. The first two can be quasicomplementary in two ways, but the generated pairs (p, α) are already known. The first and the third cannot be quasicomplementary. The last two can also not be quasicomplementary if $n_2 > 1$. It is once again more interesting if $n_2 = 1$ and $m_2 > 1$ is allowed. We get the graph in Figure 29 (left). The top and the bottom legs cannot be made quasicomplementary. The left and the bottom legs are the interesting case. Analogously to how we used Figure 28 (right), we can use Figure 29 (right) to show that $k = 0$ and $(\beta_1, \dots, \beta_{m_2}) = (4, (6)^{[l]})$ and $(b_1, \dots, b_{m_1}) = (2, 2, (3, 2, 2, 2)^{[l]}, 2)$. This is in fact family (4) in [3, Theorem 1.1] and family (8) in Theorem B.

Going back to Figure 3, we could also make the vertex of weight $-\alpha_1 - \beta_1$ the only trivalent vertex, but that would require $m_1 = n_1 = 1$, and thus all α_i and all a_j are 2's. The top leg would not be able to be quasicomplementary to a sequence of 2's, and the only way for the left and right legs to be quasicomplementary is if they are the legs (-2) and $(-2, -2)$ in either order. This just gives us two new families of possible surgery coefficients on $T(2,3)$.

References

[1] **P Aceto**, *Rational homology cobordisms of plumbed manifolds*, Algebr. Geom. Topol. 20 (2020) 1073–1126 MR

[2] **P Aceto**, **M Golla**, *Dehn surgeries and rational homology balls*, Algebr. Geom. Topol. 17 (2017) 487–527 MR

[3] **P Aceto**, **M Golla**, **K Larson**, **A G Lecuona**, *Surgeries on torus knots*, *rational balls*, *and cabling*, preprint (2020) arXiv 2008.06760

[4] **S Akbulut**, **K Larson**, *Brieskorn spheres bounding rational balls*, Proc. Amer. Math. Soc. 146 (2018) 1817–1824 MR

[5] **K L Baker**, **D Buck**, **A G Lecuona**, *Some knots in $S^1 \times S^2$ with lens space surgeries*, Comm. Anal. Geom. 24 (2016) 431–470 MR

[6] **M Bhupal**, **A I Stipsicz**, *Weighted homogeneous singularities and rational homology disk smoothings*, Amer. J. Math. 133 (2011) 1259–1297 MR

[7] **J Fernández de Bobadilla**, **I Luengo**, **A Melle Hernández**, **A Némethi**, *Classification of rational unicuspidal projective curves whose singularities have one Puiseux pair*, from "Real and complex singularities" (J-P Brasselet, M A Soares Ruas, editors), Birkhäuser, Basel (2007) 31–45 MR

[8] **A Donald**, **B Owens**, *Concordance groups of links*, Algebr. Geom. Topol. 12 (2012) 2069–2093 MR

[9] **S K Donaldson**, *The orientation of Yang–Mills moduli spaces and* 4*-manifold topology*, J. Differential Geom. 26 (1987) 397–428 MR

[10] **M Golla**, **K Larson**, 3*-manifolds that bound no definite* 4*-manifold*, Math. Res. Lett. 30 (2023) 1063–1080 MR

[11] **R E Gompf**, **A I Stipsicz**, 4*-manifolds and Kirby calculus*, Graduate Studies in Mathematics 20, Amer. Math. Soc., Providence, RI (1999) MR

[12] **C M Gordon**, *Dehn surgery and satellite knots*, Trans. Amer. Math. Soc. 275 (1983) 687–708 MR

[13] **M Hedden**, **Y Ni**, *Manifolds with small Heegaard Floer ranks*, Geom. Topol. 14 (2010) 1479–1501 MR

[14] **R Kirby**, *Problems in low-dimensional topology*, from "Geometric topology" (W H Kazez, editor), AMS/IP Stud. Adv. Math. 2, Amer. Math. Soc., Providence, RI (1997) 35–473 MR

[15] **A G Lecuona**, *On the slice-ribbon conjecture for Montesinos knots*, Trans. Amer. Math. Soc. 364 (2012) 233–285 MR

[16] **A G Lecuona**, *Complementary legs and rational balls*, Michigan Math. J. 68 (2019) 637–649 MR

[17] **P Lisca**, *Lens spaces, rational balls and the ribbon conjecture*, Geom. Topol. 11 (2007) 429–472 MR

[18] **P Lisca**, *Sums of lens spaces bounding rational balls*, Algebr. Geom. Topol. 7 (2007) 2141–2164 MR

[19] **L Lokteva**, *Surgeries on iterated torus knots bounding rational homology* 4*-balls*, Proc. Edinb. Math. Soc. 66 (2023) 557–578 MR

[20] **J M Montesinos**, 4*-manifolds,* 3*-fold covering spaces and ribbons*, Trans. Amer. Math. Soc. 245 (1978) 453–467 MR

[21] **W D Neumann**, *On bilinear forms represented by trees*, Bull. Austral. Math. Soc. 40 (1989) 303–321 MR

[22] **O Riemenschneider**, *Deformationen von Quotientensingularitäten* (*nach zyklischen Gruppen*), Math. Ann. 209 (1974) 211–248 MR

[23] **O Şavk**, *More Brieskorn spheres bounding rational balls*, Topology Appl. 286 (2020) art. id. 107400 MR

[24] **J Simone**, *Classification of torus bundles that bound rational homology circles*, Algebr. Geom. Topol. 23 (2023) 2449–2518 MR

[25] **N J A Sloane**, et al., *The on-line encyclopedia of integer sequences* Available at `http://oeis.org/`

[26] **A I Stipsicz**, **Z Szabó**, **J Wahl**, *Rational blowdowns and smoothings of surface singularities*, J. Topol. 1 (2008) 477–517 MR

[27] **F Waldhausen**, *Über Involutionen der* 3*-Sphäre*, Topology 8 (1969) 81–91 MR

Ångströmlaboratoriet, Uppsala Universitet
Uppsala, Sweden

`lisa.lokteva@math.uu.se`

Received: 8 September 2022 Revised: 3 May 2024

msp

Algebraic & Geometric Topology **25**:8 (2025) 4633–4665
DOI: 10.2140/agt.2025.25.4633
Published: 20 November 2025

Homological stability for the ribbon Higman–Thompson groups

RACHEL SKIPPER
XIAOLEI WU

We generalize the notion of asymptotic mapping class groups and allow them to surject to the Higman–Thompson groups, answering a question of Aramayona and Vlamis in the case of the Higman–Thompson groups. When the underlying surface is a disk, these new asymptotic mapping class groups can be identified with the ribbon and oriented ribbon Higman–Thompson groups. We use this model to prove that the ribbon Higman–Thompson groups satisfy homological stability, providing the first homological stability result for dense subgroups of big mapping class groups. Our result can also be treated as an extension of Szymik and Wahl's work on homological stability for the Higman–Thompson groups to the surface setting.

19D23, 20F36, 20J05, 57M07

Introduction

The family of Thompson's groups and the many groups in the extended Thompson family have long been studied for their many interesting properties. Thompson's group F is the first example of a type F_∞, torsion-free group with infinite cohomological dimension [9], while Thompson's groups T and V provided the first examples of finitely presented simple groups with infinitely many elements. More recently the braided and labeled braided Higman–Thompson groups have garnered attention in part due to their connections with big mapping class groups [8; 11; 4; 28]. In particular, Thumann constructed the ribbon version of Thompson's group V and proved that it is of type F_∞ [30]. We studied the ribbon Higman–Thompson groups $RV_{d,r}$ and their oriented version $RV^+_{d,r}$ in [28]. In fact, we identified them with the so-called labeled braided Higman–Thompson groups and proved that they are all of type F_∞.

The homology of Thompson's groups has also been well-studied. Brown and Geoghegan computed the homology of F in [9]; Ghys and Sergiescu calculated the homology of T in [18]. More recently Szymik and Wahl showed that V is acyclic [29], answering a question due to Brown [7]. One of the key ingredients for their proof was showing that the Higman–Thompson groups $V_{d,1} \hookrightarrow V_{d,2} \hookrightarrow \cdots \hookrightarrow V_{d,r} \hookrightarrow \cdots$ satisfy homological stability for any fixed d. Recall that a family of groups $G_1 \hookrightarrow G_2 \hookrightarrow \cdots \hookrightarrow G_n \hookrightarrow \cdots$ is said to satisfy homological stability if the induced maps $H_i(G_n) \to H_i(G_{n+1})$ are isomorphisms for sufficiently large n. Classical examples of families of groups which satisfy homological stability include symmetric groups [25], general linear groups [24] and mapping class groups of surfaces [19].

Here we extend Szymik and Wahl's work to the class of ribbon Higman–Thompson groups. To accomplish this, we first build a geometric model for the ribbon Higman–Thompson groups using Funar–Kapoudjian's asymptotic mapping class groups [13]. These groups are defined using a rigid structure on a surface minus a Cantor set and they sit naturally inside the ambient big mapping class groups. More recently, Aramayona and Funar [2] generalized the definition to surfaces with nonzero genus. In fact, Aramayona and Funar showed that the half-twist version of their asymptotic mapping class group (see Definition 3.15) is dense in the big mapping class group [2, Theorem 1.3]. Another surprising result of Funar and Neretin says that the half-twist asymptotic mapping class group of a closed surface minus a standard ternary Cantor set is in fact isomorphic to its smooth mapping class group [15, Corollary 2]. Aramayona and Vlamis [3, Question 5.37] asked:

Question *Are there other geometrically defined subgroups of* $\mathrm{Map}(\Sigma_g)$ *which surject to other interesting classes of subgroups of homeomorphism group of the Cantor set, such as the Higman–Thompson groups, Neretin groups, etc?*

We proceed to construct two new classes of asymptotic mapping class groups, one of which answers their question in the case of Higman–Thompson groups while the other family surjects to the symmetric Higman–Thompson groups $V_{d,r}(\mathbb{Z}/2\mathbb{Z})$.

Proposition 3.18 and Theorem 3.20 *Let* Σ *be any compact surface and* $\mathscr{C}$ *be a Cantor set which lies in the interior of a disk in* Σ. *Then the mapping class group* $\mathrm{Map}(\Sigma \setminus \mathscr{C})$ *contains the following two families of dense subgroups: the asymptotic mapping class groups* $\mathscr{B}V_{d,r}(\Sigma)$, *which surject to the Higman–Thompson groups* $V_{d,r}$; *and the half-twist asymptotic mapping class groups* $\mathscr{H}V_{d,r}(\Sigma)$, *which surject to the symmetric Higman–Thompson groups* $V_{d,r}(\mathbb{Z}/2\mathbb{Z})$.

When Σ is the disk, we identify $\mathscr{H}V_{d,r}(\Sigma)$ with the ribbon Higman–Thompson group $RV_{d,r}$ and $\mathscr{B}V_{d,r}(\Sigma)$ with the oriented ribbon Higman–Thompson group $RV^+_{d,r}$ (see Theorem 3.24). Using this geometric model for the ribbon Higman–Thompson groups, we are able to prove the following.

Theorems 4.31 and 4.32 *Suppose* $d \geq 2$. *Then the inclusion maps induce isomorphisms*

$$\iota_{R,d,r}\colon H_i(RV_{d,r}, M) \to H_i(RV_{d,r+1}, M)$$

in homology in all dimensions $i \geq 0$, *for all* $r \geq 1$ *and for all* $H_1(RV_{d,\infty})$*-modules* M. *The same also holds for the oriented ribbon Higman–Thompson groups* $RV^+_{d,r}$.

Remark (1) Here we restrict our main result to the constant coefficient $\mathbb{Z}$ case. Nevertheless, the theorem also holds for some general coefficients by applying [27, Theorem A].

(2) The same method here can also be used to prove that the groups $\mathscr{B}V_{d,r}(\Sigma)$ and $\mathscr{H}V_{d,r}(\Sigma)$ satisfy homological stability. Still, it seems difficult to make it work directly for braided Higman–Thompson groups as we are lacking a good geometric model for them. Ideally, we would realize the braided

> Higman–Thompson groups as some sort of mapping class groups of the disk minus a Cantor set. In fact, since the braided Higman–Thompson groups are subgroups of the oriented ribbon Higman–Thompson groups, we already have a geometric model in some sense. But it is less clear how one can tell when an element of the asymptotic mapping class group lies in the braided Higman–Thompson groups.

To the best of our knowledge, this is the first homological stability result for dense subgroups of big mapping class groups, although density will not play a role in our proof. Our proof uses a recent convenient framework given by Randal-Williams and Wahl [27]. The core of the proof is similar to [29], but with new technical difficulties arising from infinite-type surface topology. In particular, we take advantage of what we call the "mutual link trick", which we abstract from [10] and which we expect to be useful in a number of settings. We hope our result here can be further used to calculate the homology of ribbon Higman–Thompson groups and shed light on the question of whether braided V is acyclic. In fact, our Proposition 4.27 has already been used in [26] to prove that the mapping class groups of the disk minus a Cantor set is acyclic. It is also worth mentioning that the homology of the infinite genus version ribbon Thompson group has been calculated rationally in [14, Theorem 1.2] and integrally in [1, Theorem 1.16]. It in fact has the same homology as the stable homology of mapping class groups.

Outline of paper

In Section 1, we describe the connectivity tools that will be necessary for the remainder of the paper. In Section 2, we introduce the definition of the Higman–Thompson, ribbon Higman–Thompson, and oriented ribbon Higman–Thompson groups using paired forest diagrams to define the elements. In Section 3, we generalize the notion of asymptotic mapping class groups and allow them to surject to the Higman–Thompson groups. And finally, in Section 4, we prove homological stability for the ribbon Higman–Thompson groups and their oriented version.

Notation and conventions

All surfaces in this paper are assumed to be connected and orientable unless otherwise stated. Given a simplicial complex X and a cell $\sigma \in X$, we denote the link of σ in X by $\mathrm{Lk}_X(\sigma)$ and the star of σ by $\mathrm{St}_X(\sigma)$. When the situation is clear, we quite often omit X and simply denote the link by $\mathrm{Lk}(\sigma)$ and the star by $\mathrm{St}(\sigma)$. Recall that X is called n-connected if its homotopy groups are trivial up to dimension n. We also use the convention that (-1)-connected means nonempty and that every space is (-2)-connected. In particular, the empty set is (-2)-connected. Finally, we adopt the convention that elements in groups are multiplied from left to right.

Acknowledgements

The first part of this project was done while Skipper was a visitor in the Unité de mathématiques pures et appliquées at the ENS de Lyon and during a visit to the University of Bonn. She thanks them for their hospitality. She was also supported by the GIF, grant I-198-304.1-2015, *Geometric exponents of random walks and intermediate growth groups* and NSF DMS–2005297 *Group actions on trees and boundaries of trees*.

Part of this work was done when Wu was a member of the Hausdorff Center of Mathematics. At the time, he was supported by Wolfgang Lück's ERC Advanced Grant *KL2MG-interactions* (no. 662400) and the DFG Grant under Germany's Excellence Strategy – GZ 2047/1, Projekt-ID 390685813.

Part of this work was also done when both authors were visiting IMPAN at Warsaw during the Simons Semester *Geometric and analytic group theory*, which was partially supported by the grant 346300 for IMPAN from the Simons Foundation and the matching 2015–2019 Polish MNiSW fund. We would also like to thank Kai-Uwe Bux for inviting us for a research visit at Bielefeld in May 2019 and many stimulating discussions. Special thanks go to Jonas Flechsig for his comments on preliminary versions of the paper. Furthermore, we want to thank Javier Aramayona and Stefan Witzel for discussions, Andrea Bianchi for comments, and Matthew C B Zaremsky for some helpful communications and comments. The authors are also grateful to the referee for many helpful comments which improved the exposition.

1 Connectivity tools

In this section, we review some of the connectivity tools that we need for calculating the connectivity of our spaces. A good reference is [21, Section 2], although not all the tools we use can be found there.

1.1 Complete join

The complete join is useful tool introduced by Hatcher and Wahl [22, Section 3] for proving connectivity results. We review the basics here.

Definition 1.1 A surjective simplicial map $\pi : Y \to X$ is called a *complete join* if it satisfies the following properties:

(1) π is injective on individual simplices.

(2) For each p-simplex $\sigma = \langle v_0, \dots, v_p \rangle$ of X, $\pi^{-1}(\sigma)$ is the join $\pi^{-1}(v_0) * \pi^{-1}(v_1) * \cdots * \pi^{-1}(v_p)$.

Definition 1.2 A simplicial complex X is called weakly Cohen–Macaulay of dimension n if X is $(n-1)$-connected and the link of each p-simplex of X is $(n-p-2)$-connected.

The main result regarding complete join that we will use is the following.

Proposition 1.3 [22, Propostion 3.5] *If Y is a complete join complex over a weakly Cohen–Macaulay complex X of dimension n, then Y is also weakly Cohen–Macaulay of dimension n.*

Remark 1.4 If $\pi\colon Y \to X$ is a complete join, then X is a retract of Y. In fact, we can define a simplicial map $s\colon X \to Y$ such that $\pi \circ s = \mathrm{id}_X$ by sending a vertex $v \in X$ to any vertex in $\pi^{-1}(v)$ and then extending it to simplices. The fact that s can be extended to simplices is granted by the condition that π is a complete join. In particular, we can also conclude that, if Y is n-connected, so is X.

1.2 Bad simplices argument

Let (X, Y) be a pair of simplicial complexes. We want to relate the n-connectedness of Y to the n-connectedness of X via a so-called bad simplices argument; see [21, Section 2.1] for more information. One identifies a set of simplices in $X \setminus Y$ as bad simplices, satisfying the following two conditions:

(i) Any simplex with no bad faces is in Y, where by a "face" of a simplex we mean a subcomplex spanned by any nonempty subset of its vertices, proper or not.

(ii) If two faces of a simplex are both bad, then their join is also bad.

We call simplices with no bad faces good simplices. Bad simplices may have good faces or faces which are neither good nor bad. If σ is a bad simplex, we say a simplex τ in $\mathrm{Lk}(\sigma)$ is good for σ if any bad face of $\tau * \sigma$ is contained in σ. The simplices which are good for σ form a subcomplex of $\mathrm{Lk}(\sigma)$, which we denote by GL_σ and call the good link of σ.

Proposition 1.5 [21, Proposition 2.1] *Let X, Y and GL_σ be as above. Suppose that, for some integer $n \geq 0$, the subcomplex GL_σ of X is $(n-\dim(\sigma)-1)$-connected for all bad simplices σ. Then the pair (X, Y) is n-connected, ie $\pi_i(X, Y) = 0$ for all $i \leq n$.*

We can apply the proposition in the following way.

Theorem 1.6 [21, Corollary 2.2] *Let Y be a subcomplex of a simplicial complex X and suppose the space $X \setminus Y$ has a set of bad simplices satisfying (i) and (ii) above; then:*

(1) *If X is n-connected and GL_σ is $(n-\dim(\sigma))$-connected for all bad simplices σ, then Y is n-connected.*

(2) *If Y is n-connected and GL_σ is $(n-\dim(\sigma)-1)$-connected for all bad simplices σ, then X is n-connected.*

1.3 The mutual link trick

In the proof of [10, Theorem 3.10], there is a beautiful argument for resolving intersections of arcs inspired by Hatcher's flow argument [20]. They attributed the idea to Andrew Putman. Recall Hatcher's flow argument allows one to "flow" a complex to its subcomplex. But in the process, one can only "flow" a vertex to a new one in its link. The mutual link trick will allow one to "flow" a vertex to a new one not in its link provided "the mutual link" is sufficiently connected.

To apply the mutual link trick, we first need a lemma that allows us to homotope a simplicial map to a simplexwise injective one [10, Lemma 3.9]. Recall a simplicial map is called *simplexwise injective* if its restriction to any simplex is injective. See also [16, Section 2.1] for more information.

Lemma 1.7 *Let Y be a compact m-dimensional combinatorial manifold. Let X be a simplicial complex and assume that the link of every p-simplex in X is $(m-p-2)$-connected. Let $\psi\colon Y \to X$ be a simplicial map whose restriction to ∂Y is simplexwise injective. Then, after possibly subdividing the simplicial structure of Y, ψ is homotopic relative ∂Y to a simplexwise injective map.*

Note that, as discussed in [17, Lemma 5.19], there is a mistake in the connectivity bound given in [10] that was corrected in an erratum.

Lemma 1.8 (the mutual link trick) *Let Y be a closed m-dimensional combinatorial manifold and $f\colon Y \to X$ be a simplexwise injective simplicial map. Let $y \in Y$ be a vertex and $f(y) = x$ for some $x \in X$. Suppose x' is another vertex of X satisfying the following conditions:*

(1) $f(\mathrm{Lk}_Y(y)) \le \mathrm{Lk}_X(x')$.

(2) *The mutual link* $\mathrm{Lk}_X(x) \cap \mathrm{Lk}_X(x')$ *is $(m-1)$-connected.*

Then we can define a new simplexwise injective map $g\colon Y \to X$ by sending y to x' and all the other vertices y' to $f(y')$ such that g is homotopic to f.

Proof The conditions that f is simplexwise injective and $f(\mathrm{Lk}_Y(y)) \le \mathrm{Lk}_X(x')$ guarantee that the definition of g can be extended over Y and g is again simplexwise injective.

We need to prove g is homotopic to f. The homotopy will be the identity outside $\mathrm{St}_Y(y)$. Note that, since f is simplexwise injective, $f(\mathrm{Lk}_Y(y)) \le \mathrm{Lk}_X(x)$. Together with condition (1), this gives $f(\mathrm{Lk}_Y(y)) \le \mathrm{Lk}_X(x) \cap \mathrm{Lk}_X(x')$. Since $\mathrm{Lk}_Y(y)$ is an $(m-1)$-sphere and $\mathrm{Lk}_X(x) \cap \mathrm{Lk}_X(x')$ is $(m-1)$-connected, there exists an m-disk B with $\partial B = \mathrm{Lk}_Y(y)$ and a simplicial map $\varphi\colon B \to \mathrm{Lk}_X(x) \cap \mathrm{Lk}_X(x')$ such that φ restricted to ∂B coincides with ψ restricted to $\mathrm{Lk}_Y(y)$. Since the image of B under φ is contained in $\mathrm{St}_X(x)$, which is contractible, we can homotope g, replacing $g|_{\mathrm{St}_Y(y)}$ with φ. Since the image of B under f is also contained in $\mathrm{Lk}_X(x')$, we can similarly homotope f, replacing $f|_{\mathrm{St}_Y(y)}$ with φ. These both yield the same map, so g is homotopic to f. □

2 Higman–Thompson groups and their braided versions

In this section, we first give an introduction to the Higman–Thompson groups and then define their ribbon version.

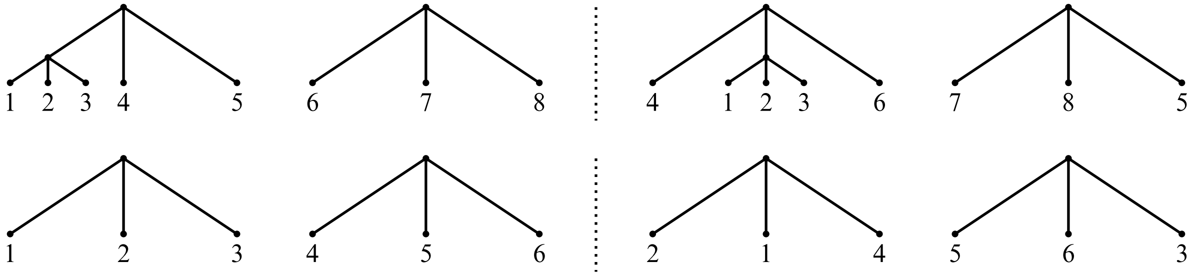


Figure 1: Reduction of the top paired $(3,2)$-forest diagram to the bottom one.

2.1 Higman–Thompson groups

The Higman–Thompson groups were first introduced by Higman as a generalization of the groups [23] given earlier in handwritten, unpublished notes of Richard Thompson. First let us recall the definition of the Higman–Thompson groups. Although there are a number of equivalent definitions of these groups, we will use the notion of paired forest diagrams. First we define a *finite rooted d-ary tree* to be a finite tree such that every vertex has degree $d+1$ except the *leaves*, which have degree 1, and the *root*, which has degree d (or degree 1 if the root is also a leaf). Usually we draw such trees with the root at the top and the nodes descending from it down to the leaves. A vertex v of the tree along with its d adjacent descendants will be called a *caret*. If the leaves of a caret in the tree are leaves of the tree, we will call the caret *elementary*. A collection of r d-ary trees will be called a *(d,r)-forest*. When d is clear from the context, we may just call it an r-forest.

Define a *paired (d,r)-forest diagram* to be a triple (F_-,ρ,F_+) consisting of two (d,r)-forests F_- and F_+ both with l leaves for some l, and a permutation $\rho\in S_l$, the symmetric group on l elements. We label the leaves of F_- with $1,\dots,l$ from left to right, and, for each i, the $\rho(i)^{\text{th}}$ leaf of F_+ is labeled i.

Define a *reduction* of a paired (d,r)-forest diagram to be the following: Suppose there is an elementary caret in F_- with leaves labeled by $i,\dots,i+d-1$ from left to right, and an elementary caret in F_+ with leaves labeled by $i,\dots,i+d-1$ from left to right. Then we can "reduce" the diagram by removing those carets, renumbering the leaves and replacing ρ with the permutation $\rho'\in S_{l-d+1}$ that sends the new leaf of F_- to the new leaf of F_+, and otherwise behaves like ρ. The resulting paired forest diagram (F'_-,ρ',F'_+) is then said to be obtained by *reducing* (F_-,ρ,F_+). See Figure 1 for an idea of reduction of paired $(3,2)$-forest diagrams. The reverse operation to reduction is called *expansion*, so (F_-,ρ,F_+) is an expansion of (F'_-,ρ',F'_+). A paired forest diagram is called *reduced* if there is no reduction possible. Define an equivalence relation on the set of paired (d,r)-forest diagrams by declaring two paired forest diagrams to be equivalent if one can be reached by the other through a finite series of reductions and expansions. Thus an equivalence class of paired forest diagrams consists of all diagrams having a common reduced representative. Such reduced representatives are unique.

There is a binary operation $*$ on the set of equivalence classes of paired (d,r)-forest diagrams. Let $\alpha=(F_-,\rho,F_+)$ and $\beta=(E_-,\xi,E_+)$ be reduced paired forest diagrams. By applying repeated expansions

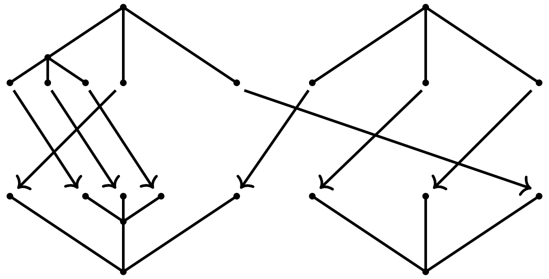


Figure 2: An element of $V_{3,2}$.

to α and β, we can find representatives (F'_-, ρ', F'_+) and (E'_-, ξ', E'_+) of the equivalence classes of α and β, respectively, such that $F'_+ = E'_-$. Then we declare $\alpha * \beta$ to be $(F'_-, \rho'\xi', E'_+)$. This operation is well defined on the equivalence classes and is a group operation.

Definition 2.1 The *Higman–Thompson group $V_{d,r}$* is the group of equivalence classes of paired (d,r)-forest diagrams with the multiplication $*$.

The usual Thompson group V is a special case of Higman–Thompson groups. In fact, $V = V_{2,1}$.

2.2 Ribbon Higman–Thompson groups

For convenience, we will think of the forest F_+ drawn beneath F_- and upside down, ie with the root at the bottom and the leaves at the top. The permutation ρ is then indicated by arrows pointing from the leaves of F_- to the corresponding paired leaves of F_+. See Figure 2 for this visualization of (the unreduced representation of) the element of $V_{3,2}$ from Figure 1.

Now, in the ribbon version of the Higman–Thompson groups, the permutations of leaves are simply replaced by ribbon braids which can twist between the leaves.

Definition 2.2 Let $\mathcal{I} = \coprod_{i=1}^{d} I_i : [0,1] \times \{1, \dots, l\} \to \mathbb{R}^2$ be an embedding, which we refer to as the *marked bands*. A *ribbon braid* is a map $R : ([0,1] \times \{0, 1, \dots, l\}) \times [0,1] \to \mathbb{R}^2$ such that, for any $0 \le t \le 1$, $R_t : [0,1] \times \{1, \dots, l\} \to \mathbb{R}^2$ is an embedding, $R_0 = \mathcal{I}$ and there exists $\sigma \in S_l$ such that $R_1(t)|_{I_i} = I_{\sigma(i)}(t)$ or $R_1(t)|_{I_i} = I_{\sigma(i)}(1-t)$. The usual product of paths defines a group structure on the set of ribbon braids up to homotopy among ribbon braids. This group, denoted by RB_l, does not depend on the choice of the marked bands and it is called the ribbon braid group with l bands. A ribbon braid is *pure* if σ is trivial and we define PRB_l to be the *pure ribbon braid group* with l bands. If we further assume $R_1(t)|_{I_i} = I_{\sigma(i)}(t)$, this subgroup is called the *oriented ribbon braid group* RB_l^+. Similarly, we have the *oriented pure ribbon braid group* PRB_l^+.

Remark 2.3 Note that $\mathrm{RB}_l \cong \mathbb{Z}^l \rtimes B_l$, where the action of the braid group B_l with l strings is induced by the symmetric group action on the coordinates of $\mathbb{Z}^l$. In particular, for the pure ribbon braid group

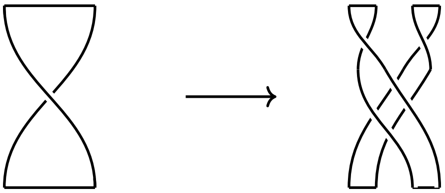


Figure 3: Splitting a ribbon into two ribbons.

PRB_l, $\mathrm{PRB}_l \cong \mathbb{Z}^l \times \mathrm{PB}_l$, where PB_l is the pure braid group with l strings. Under this isomorphism, $\mathrm{RB}_l^+ \cong (2\mathbb{Z})^l \rtimes B_l$ and $\mathrm{PRB}_l^+ \cong (2\mathbb{Z})^l \times \mathrm{PB}_l$.

Definition 2.4 A *ribbon braided paired (d,r)-forest diagram* is a triple $(F_-, \mathfrak{r}, F_+)$ consisting of two (d,r)-forests F_- and F_+ both with l leaves for some l and a ribbon braid $\mathfrak{r} \in \mathrm{RB}_l$ connecting the leaves of F_- to the leaves of F_+.

The expansion and reduction rules for the ribbon braids just come from the natural way of splitting a ribbon band into d components and the inverse operation to this. See Figure 3 for how to split a half twisted band when $d = 2$. Note that not only are the two bands themselves twisted, but the bands are also braided. Everything else will be the same as in the braided case, so we omit the details here. As usual, we define two ribbon braided paired forest diagrams to be equivalent if one is obtained from the other by a sequence of reductions or expansions. The multiplication operation $*$ on the equivalence classes is defined the same way as for $bV_{d,r}$. We direct the reader to [28, Section 2].

Definition 2.5 The *ribbon Higman–Thompson group $RV_{d,r}$* (resp. *oriented ribbon Higman–Thompson group $RV_{d,r}^+$*) is the group of equivalence classes of (resp. oriented) ribbon braided paired (d,r)-forests diagrams with the multiplication $*$.

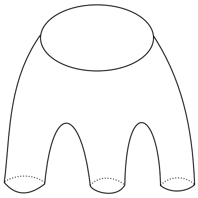
3 Asymptotic mapping class groups related to the ribbon Higman–Thompson groups

The purpose of this section is to generalize the notion of asymptotic mapping class groups and allow them to surject to the Higman–Thompson groups. In particular, we will build a geometric model for the ribbon Higman–Thompson groups which will be crucial for proving homological stability in Section 4. Our construction is largely based on the ideas in [13, Section 2; 2, Section 3].

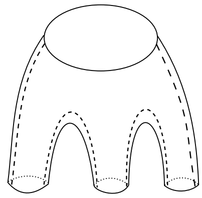
3.1 d-rigid structure

In this subsection, we generalize the notion of a rigid structure to that of a d-rigid structure.

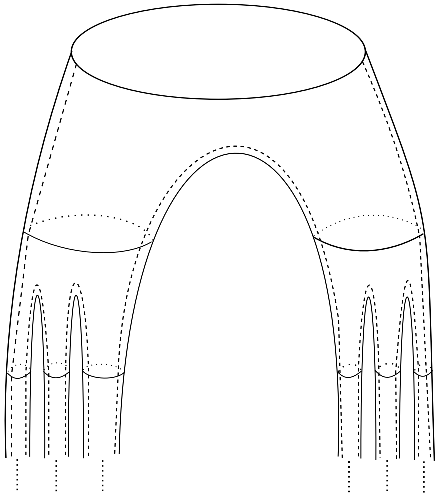
Definition 3.1 A *d-leg pants* is a surface which is homeomorphic to a $(d+1)$-holed sphere.



3-leg pants



3-leg pants with seams



$D^{\infty}_{3,2}$ with seams

Figure 4: 3-leg pants and the surface $D^{\infty}_{3,2}$ with canonical seams.

Recall that the usual pair of pants is a 2-leg pants. We will draw a d-leg pants with one boundary component at the top. In this way, we can conveniently put a counterclockwise total order on the boundary components, making the top component the minimal one. See Figure 4 for an example of a 3-leg pants.

We proceed to build some infinite-type surfaces using some basic building blocks.

Definition 3.2 Let Σ be an compact oriented surface. Call the boundary components of Σ the *based boundary components*. Then $\Sigma^{\infty}_{d,r}$ is the infinite surface, built up as an inductive limit of infinite surfaces $\Sigma_{d,r,m}$ with $m \geq 0$:

(1) $\Sigma_{d,r,0}$ is obtained from Σ by deleting the interior of a disk in Σ. When Σ is a disk D, we declare $D_{d,r,0} = \partial D$.

(2) $\Sigma_{d,r,1}$ is obtained from $\Sigma_{d,r,0}$ attaching a copy of r-leg pants along the newly created boundary of $\Sigma_{d,r,0}$.

(3) For $m \geq 1$, $\Sigma_{d,r,m+1}$ is obtained from $\Sigma_{d,r,m}$ by gluing a pair of d-leg pants to every nonbased boundary circle of $\Sigma_{d,r,m}$ along the top boundary of the pants.

The surface $\Sigma_{d,r,1}$ is called the *base* of $\Sigma^{\infty}_{d,r}$ and the boundary components of $\Sigma^{\infty}_{d,r}$ coming from the base are the *based boundary components*. For each $m \geq 1$, the nonbased boundary components of $\Sigma_{d,r,m}$ naturally embed in $\Sigma^{\infty}_{d,r}$, and we call these the *admissible loops*. We call the admissible loops coming from $\Sigma_{d,r,1}$ the *rooted loops*. The surface $\Sigma^{\infty}_{d,r}$ has a natural induced orientation.

Remark 3.3 The two indices d, r in the definition of $\Sigma^{\infty}_{d,r}$ will be used later to define the Higman–Thompson version of the asymptotic mapping class group (see Definition 3.15), where d is related to the valence of the rooted trees and r is the number of roots in the definition of the Higman–Thompson groups.

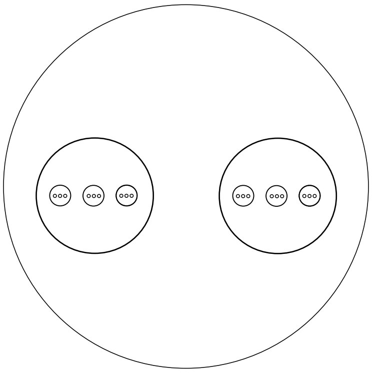


Figure 5: Disk model for the surface $D^\infty_{3,2}$.

Remark 3.4 To define our d-rigid structure, we do not really need $\Sigma_{d,r,0}$. But it will be convenient to have $\Sigma_{d,r,0}$ later, in Definition 3.17, for defining the map from $\Sigma^\infty_{d,r}$ to the tree $\mathcal{T}_{d,r}$.

Definition 3.5 A compact subsurface $A \subset \Sigma^\infty_{d,r}$ is *admissible* if $\Sigma_{d,r,1} \subseteq A$ and all of its nonbased boundaries are admissible. The subsurfaces $\Sigma_{d,r,m}$ are called the *standard admissible subsurfaces* of $\Sigma^\infty_{d,r}$.

Remark 3.6 In the special case where the starting surface is a disk, we will use the notation $\Sigma = D$, $\Sigma_{d,r,m} = D_{d,r,m}$, and $\Sigma^\infty_{d,r} = D^\infty_{d,r}$. See Figure 4 for a picture of the surface $D^\infty_{3,2}$. In this case, we can think of $D^\infty_{d,r}$ as a subsurface of a disk D. More specifically, let $D = \{(x, y) \mid x^2 + y^2 \le 1\}$ and $x_i = (2i - r - 1)/(r + 1)$ for $1 \le i \le r$. We place r disks with center at each $(x_i, 0)$ of radius $r_0 = 1/(4(r + 1))$. Denote these disks by $D_1, \dots, D_r$. The complement of the interior of these r disks in D is homeomorphic to the r-leg pants $D_{d,r,1}$. Now, for each disk D_i, $1 \le i \le r$, we can equally distribute d points in the x-axis inside D_i and place a disk with radius r_0/d^2 centered at each. The complements of the interiors of these d disks in D_i are all d-leg pants. We can continue the process inductively. At the end, the disks converge to a Cantor set which we denote by $\mathcal{C}$. In particular $D^\infty_{d,r}$ is homeomorphic to $D \setminus \mathcal{C}$. We will refer to this as the *puncture model* for $D^\infty_{d,r}$. See Figure 5 for a picture of $D^\infty_{3,2}$ with this model. The advantage of this model is we can view $D^\infty_{d,r}$ and all its admissible subsurfaces directly as subsurfaces of D.

Remark 3.7 Now $\Sigma^\infty_{d,r}$ can be obtained from Σ by attaching a copy of $D^\infty_{d,r}$ to the nonbased boundary component of $\Sigma_{d,r,0}$. In particular, $\Sigma^\infty_{d,r}$ is obtained from Σ by deleting a copy of the Cantor set, and any admissible subsurface of $\Sigma^\infty_{d,r}$ can be viewed directly as a subsurface of Σ using the puncture model. Recall that any two Cantor sets are homeomorphic; hence, by the classification of infinite surfaces [3, Theorem 2.2], $\Sigma^\infty_{d,r}$ is homeomorphic to $\Sigma \setminus \mathcal{C}$, where $\mathcal{C}$ is the standard ternary Cantor set sitting inside some disk in Σ regardless of the choice of d and r.

Definition 3.8 (1) A *suited d-pants decomposition* of the infinite surface $\Sigma^\infty_{d,r}$ is a maximal collection of distinct nontrivial simple closed curves in the interior of $\Sigma^\infty_{d,r} \setminus \Sigma_{d,r,1}$ which are not isotopic to the boundary, pairwise disjoint and pairwise nonisotopic, with the additional property that the complementary regions in $\Sigma_{d,r}^{\infty} \setminus \Sigma_{d,r,1}$ are all d-leg pants.

(2) A *d-rigid structure* on $\Sigma_{d,r}^{\infty}$ consists of two pieces of data:

- a suited d-pants decomposition; and
- a *d-prerigid structure*, ie a countable collection of disjoint line segments embedded into $\Sigma_{d,r}^{\infty} \setminus \Sigma_{d,r,1}$ such that the complement of their union in each component of $\Sigma_{d,r}^{\infty} \setminus \Sigma_{d,r,1}$ has two connected components.

These pieces must be *compatible* in the following sense: firstly, the traces of the d-prerigid structure on each d-leg pants (ie the intersections with pants) are made up of $d+1$ connected components, called *seams*; secondly, each boundary component of the pants intersects with exactly two components of the seams at two distinct points; thirdly, the seams cut each pants into two components. Note that these conditions imply that each component is homeomorphic to a disk. One says then that the suited d-pants decomposition and the d-prerigid structure are *subordinate* to the d-rigid structure.

(3) By construction, $\Sigma_{d,r}^{\infty}$ is naturally equipped with a suited d-pants decomposition, which will be referred to below as the *canonical suited d-pants decomposition*. We also fix a d-prerigid structure on $\Sigma_{d,r}^{\infty}$ (called the *canonical d-prerigid structure*) compatible with the canonical suited d-pants decomposition. See Figure 4. Using the puncture model, the seams of the canonical d-prerigid structure are just the intersections of $[-1,1] \times \{0\}$ with each d-leg pants. The resulting d-rigid structure is called the *canonical d-rigid structure* on $\Sigma_{d,r}^{\infty}$. Very importantly, for each admissible subsurface, the canonical d-rigid structure induces an order on the admissible boundaries. In Figure 4, the induced order on the admissible loops are counterclockwise. Using the puncture model, the admissible loops are ordered from left to right.

(4) The seams cut each component of $\Sigma_{d,r}^{\infty} \setminus \Sigma_{d,r,1}$ into two pieces; we choose the front piece in each component, and these r pieces together form the *visible side* of $\Sigma_{d,r}^{\infty}$.

(5) A suited d-pants decomposition (resp. d-(pre)rigid structure) is *asymptotically trivial* if, outside a compact subsurface of $\Sigma_{d,r}^{\infty}$, it coincides with the canonical suited d-pants decomposition (resp. canonical d-(pre)rigid structure).

Remark 3.9 It is important that the seams cut each d-pants into two components and each component is homeomorphic to a disk, as the mapping class group of a disk is trivial.

Definition 3.10 Let $\Sigma_{d,r}^{\infty}$ and $\overline{\Sigma}_{d,r'}^{\infty}$ be two surfaces with d-rigid structure and let $\varphi\colon \Sigma_{d,r}^{\infty} \to \overline{\Sigma}_{d,r'}^{\infty}$ be a homeomorphism. One says that φ is *asymptotically rigid* if there exists an admissible subsurface $A \subset \Sigma_{d,r}^{\infty}$ such that

(1) $\varphi(A)$ is also admissible in $\overline{\Sigma}_{d,r'}^{\infty}$;

(2) $\varphi|_A$ maps the based boundaries to based boundaries, admissible loops to admissible loops; and

(3) the restriction of $\varphi\colon \Sigma^{\infty}_{d,r} \setminus A \to \Sigma^{\infty}_{d,r'} \setminus \varphi(A)$ is *rigid*, meaning that it respects the traces of the canonical d-rigid structure, mapping the suited d-pants decomposition into the suited d-pants decomposition, the seams into the seams, and the visible side into the visible side.

If we drop the condition that φ should map the visible side into the visible side, φ is called *asymptotically quasirigid*. The surface A is called a *support* for φ.

Remark 3.11 We are not using the word "support" in the usual sense, as the map outside the support defined above might well not be the identity, but the map is uniquely determined up to isotopy by Remark 3.9.

Remark 3.12 In [13, Definition 2.3], they do not actually require that the support contain the base. This will not make a difference, as one can always enlarge the support so that it contains the base.

Remark 3.13 The surface $\Sigma^{\infty}_{d,r+d-1}$ can be identified with the surface $\Sigma^{\infty}_{d,r}$ such that $\Sigma_{d,r+d-1,m} = \Sigma_{d,r,m+1}$ for any $m \geq 1$, and the d-rigid structure of $\Sigma^{\infty}_{d,r}$ coincides with d-rigid on $\Sigma^{\infty}_{d,r+d-1}$ outside $\Sigma_{d,r,2}$. In this way, $\Sigma^{\infty}_{d,r}$ is asymptotically rigid homeomorphic to $\Sigma^{\infty}_{d,r+d-1}$ through the identity map.

Remark 3.14 Let Σ' be a subsurface of $\Sigma^{\infty}_{d,r}$ such that there exist an admissible subsurface A of $\Sigma^{\infty}_{d,r}$ satisfying:

(1) $A \cap \Sigma'$ is a compact surface.

(2) The boundaries of Σ' are disjoint from the admissible boundary components of A.

(3) If an admissible boundary component L of A is contained in Σ', then the punctured disk component of $\Sigma^{\infty}_{d,r}$ cutting along L is also contained in Σ'.

Then Σ' has a naturally induced d-rigid structure. In fact, we can take $A \cap \Sigma'$ to be the base surface and the d-rigid structure can simply be inherited from $\Sigma^{\infty}_{d,r}$. One, of course, can choose different A here, which may give different induced d-rigid structure, but it is unique up to asymptotically rigid homeomorphism.

3.2 Asymptotic mapping class groups surjecting to Higman–Thompson groups

Given a (possibly noncompact) surface Σ, recall the mapping class group of Σ is defined to be the group of isotopy classes of orientation preserving homeomorphisms of Σ that fixes $\partial\Sigma$ pointwise, ie

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

With this, we can now define the asymptotic mapping class group and the half-twist asymptotic mapping class group.

Definition 3.15 The *asymptotic mapping class group* $\mathcal{B}V_{d,r}(\Sigma)$ (resp. *half-twist asymptotic mapping class group* $\mathcal{H}V_{d,r}(\Sigma)$) is the subgroup of $\mathrm{Map}(\Sigma^{\infty}_{d,r})$ consisting of isotopy classes of asymptotically rigid (resp. quasirigid) self-homeomorphisms of $\Sigma^{\infty}_{d,r}$. When Σ is the disk, we sometimes simply denote the group by $\mathcal{B}V_{d,r}$ (resp. $\mathcal{H}V_{d,r}$).

Definition 3.16 Let A be an admissible subsurface of $\Sigma^{\infty}_{d,r}$, and $\mathrm{Map}(A)$ be its mapping class group which fixes the each boundary component pointwise. Each inclusion $A \subseteq A'$ of admissible surfaces induces an injective embedding $j_{A,A'} \colon \mathrm{Map}(A) \to \mathrm{Map}(A')$. The collection forms a direct system, whose direct limit we call the *compactly supported pure mapping class group*, denoted by $\mathrm{PMap}_c(\Sigma^{\infty}_{d,r})$. The group $\mathrm{PMap}_c(\Sigma^{\infty}_{d,r})$ is naturally a subgroup of $\mathcal{B}V_{d,r}(\Sigma)$ and we denote the inclusion map by j.

Definition 3.17 Let $\mathcal{F}_{d,r}$ be the forest with r copies of a rooted d-ary tree and $\mathcal{T}_{d,r}$ be the rooted tree obtained from $\mathcal{F}_{d,r}$ by adding an extra vertex to $\mathcal{F}_{d,r}$ and r extra edges each connecting this vertex to a root of a tree in $\mathcal{F}_{d,r}$. There is a natural projection $q \colon \Sigma^{\infty}_{d,r} \to \mathcal{T}_{d,r}$, such that the pullback of the root is $\Sigma_{d,r,0}$ and the pullbacks of the midpoints of any edges are admissible loops.

Any element in $\mathcal{B}V_{d,r}(\Sigma)$ can be represented by an asymptotically rigid homeomorphism $\varphi \colon \Sigma^{\infty}_{d,r} \to \Sigma^{\infty}_{d,r}$. In particular, we have an admissible subsurface A of $\Sigma^{\infty}_{d,r}$ such that $\varphi|_A \colon (A, \partial_b A) \to (\varphi(A), \varphi(\partial_b A))$ is a homeomorphism. Let F_- be the smallest subforest of $\mathcal{F}_{d,r}$ which contains $q(A) \cap \mathcal{F}_{d,r}$, and F_+ be the smallest subforest of $\mathcal{F}_{d,r}$ which contains $q(\varphi(A)) \cap \mathcal{F}_{d,r}$. Note that F_- and F_+ have the same number of leaves and their leaves are in one-to-one correspondence with the admissible loops of A and $\varphi(A)$. Now let ρ be the map from leaves of F_- to F_+ induced by φ. Together, these define an element $[(F_-, \rho, F_+)] \in V_{d,r}$. We call this map π. One can show π is well defined. Similarly to [13, Proposition 2.4; 2, Propositions 4.2 and 4.6], we now have the following proposition.

Proposition 3.18 *We have the short exact sequences*

$$1 \to \mathrm{PMap}_c(\Sigma^{\infty}_{d,r}) \xrightarrow{j} \mathcal{B}V_{d,r}(\Sigma) \xrightarrow{\pi} V_{d,r} \to 1, \quad 1 \to \mathrm{PMap}_c(\Sigma^{\infty}_{d,r}) \xrightarrow{j} \mathcal{H}V_{d,r}(\Sigma) \xrightarrow{\pi} V_{d,r}(\mathbb{Z}/2\mathbb{Z}) \to 1.$$

Remark 3.19 Here, as in [2], $V_{d,r}(\mathbb{Z}/2\mathbb{Z})$ is the twisted version of the Higman–Thompson group where one allows flipping the subtree below every leaf. See for example [5] for more information.

Proof We will prove the proposition for $\mathcal{B}V_{d,r}(\Sigma)$. The other case is essentially the same. First we show the map π is surjective. Given any element $[(F_-, \rho, F_+)] \in V_{d,r}$, let T_- (resp. T_+) be the tree obtained from F_- (resp. F_+) by adding a single root on the top and r edges connecting to each root of the trees in F_- (resp. F_+). Furthermore, let T'_- (resp. T'_+) be the tree obtained from F_- (resp. F_+) by throwing away the leaves and the open half edge connecting to the leaves. Then let $A_- = q^{-1}(T'_+)$ and $A_+ = q^{-1}(T'_+)$, which are both admissible subsurfaces of $\Sigma^{\infty}_{d,r}$. Now one can produce a homeomorphism $\varphi_0 \colon A_- \to A_+$ which is the identity on the based boundary and maps the admissible loops of A_- to the admissible loops of A_+ following the information from ρ, mapping the visible part to the visible part for each admissible loop. From here, we extend φ_0 to a map $\varphi\colon \Sigma_{d,r}^\infty \to \Sigma_{d,r}^\infty$ such that φ is an asymptotically rigid homeomorphism.

If an element $g \in \mathscr{B}V_{d,r}(\Sigma)$ is mapped to a trivial element $\pi(g) = [(F_-, \rho, F_+)] \in V_{d,r}$, then the two forests F_- and F_+ are the same and the induced map ρ is trivial. This means we can assume the support A for the asymptotically rigid homeomorphism φ_g corresponding to g is the same as $\varphi(A)$, and φ induces identity map on the admissible boundary components. Thus $g \in \mathrm{PMap}_c(\Sigma_{d,r}^\infty)$. Finally, given any element $g \in \mathrm{PMap}_c(\Sigma_{d,r}^\infty)$, it is clear that $\pi \circ j(g) = 1$. □

The mapping class group $\mathrm{Map}(\Sigma_{d,r}^\infty)$ has a natural quotient topology coming from the compact–open topology on $\mathrm{Homeo}^+(\Sigma_{d,r}^\infty, \partial\Sigma_{d,r}^\infty)$. See [3, Sections 2.3 and 4.1] for more information. Aramayona and Funar [2, Theorem 1.3] showed that, when Σ is a closed surface, $\mathscr{H}V_{2,1}(\Sigma)$ is dense in $\mathrm{Map}(\Sigma_{2,1}^\infty)$. We improve their result to the following.

Theorem 3.20 *The groups* $\mathscr{B}V_{d,r}(\Sigma)$ *and* $\mathscr{H}V_{d,r}(\Sigma)$ *are dense in the mapping class group* $\mathrm{Map}(\Sigma_{d,r}^\infty)$.

Proof The proof in [2, Section 6] adapts directly to show that $\mathscr{H}V_{d,r}(\Sigma)$ is dense in $\mathrm{Map}(\Sigma_{d,r}^\infty)$ and so we will not repeat it here. To show $\mathscr{B}V_{d,r}(\Sigma)$ is also dense in $\mathrm{Map}(\Sigma_{d,r}^\infty)$, it suffices to show any element in $\mathscr{H}V_{d,r}(\Sigma)$ can be approximated by a sequence of elements in $\mathscr{B}V_{d,r}(\Sigma)$. Note first that any half-Dehn twists around admissible loops in $\Sigma_{d,r}^\infty$ lies in $\mathscr{H}V_{d,r}(\Sigma)$. In fact, given an admissible loop α, we can choose an admissible subsurface A such that α is an admissible loop of A. Then the half-Dehn twist around α is asymptotic quasirigid with support A; in fact, it is the identity on all the components of $\Sigma_{d,r}^\infty \setminus A$ except at the component containing α, where it rotates 180 degree. Now given an asymptotic quasirigid homeomorphism f of $\Sigma_{d,r}^\infty$ with support A', we can first compose f with half-Dehn twists around those admissible loops of A where f restricted to the component below them switches the front and back. The composition now is an asymptotic rigid homeomorphism. Thus $\mathscr{H}V_{d,r}(\Sigma)$ can be generated by $\mathscr{B}V_{d,r}(\Sigma)$ and half-Dehn twists around the admissible loops in $\Sigma_{d,r}^\infty$; it suffices now to show that any half-Dehn twists around admissible loops in $\Sigma_{d,r}^\infty$ can be approximated by a sequence of elements in $\mathscr{B}V_{d,r}(\Sigma)$. Given an admissible loop L, let h_L be a half-Dehn twist at L. We will construct a sequence of elements $x_i \in \mathscr{B}V_{d,r}(\Sigma)$ such that, for any compact subset K of $\Sigma_{d,r}^\infty$, there exists N such that, for any $j \geq N$, x_j and h_L coincide on K. Recall we have the map $q\colon \Sigma_{d,r}^\infty \to \mathscr{T}_{d,r}$ (see Definition 3.17) that maps the admissible loops to edge middle points in $\mathscr{T}_{d,r}$. Now consider those admissible loops whose images under q lying below $q(L)$ have distance i to $q(L)$. Note that there are d^i such admissible loops. We list them as $L_{i,1}, \dots, L_{i,d^i}$. Let $h_{L_{i,k}}$ be the half-Dehn twists around $L_{i,k}$ and let $x_i = h_L h_{L_{i,1}} \cdots h_{L_{i,d^i}}$; then $x_i \in \mathscr{B}V_{d,r}(\Sigma)$ and the sequence $\{x_i\}$ has the desired property. □

Now recall that, by Remark 3.7, $\Sigma_{d,r}^\infty$ is homeomorphic to $\Sigma \setminus \mathscr{C}$ for any d and r; hence, we have our first result stated in the introduction.

3.3 The asymptotic mapping class group of the disk punctured by the Cantor set

In this last subsection, we want to identify the asymptotic mapping class group $\mathscr{B}V_{d,r}(D)$ with the oriented ribbon Higman–Thompson groups $RV_{d,r}^{+}$ and the half-twist asymptotic mapping class group $\mathscr{H}V_{d,r}(D)$ with the ribbon Higman–Thompson group $RV_{d,r}$. The following lemma appears in [6, Section 2] without a proof, so we provide the details here. Note that what they call the pure ribbon braid group is the oriented pure ribbon braid group in Definition 2.2.

Lemma 3.21 *Let D_k be the $(k+1)$-holed sphere. Then $\mathrm{Map}(D_k)$ can be naturally identified with the pure oriented ribbon braid group PRB_k^{+}.*

Proof Note that D_k can be identified with a disk with k holes. Let ∂_b denote the boundary of the disk. Let $\overline{D}_k$ be a disk with k punctures obtained from D_k by attaching one punctured disk to each hole. The induced map $\mathscr{C}ap\colon \mathrm{Map}(D_k) \to \mathrm{PMap}(\overline{D}_k)$ is the capping homomorphism. Note that $\mathrm{PMap}(\overline{D}_k) \cong \mathrm{PB}_k$. Now, applying [12, Proposition 3.19] and the fact that the Dehn twists around the holes of D_k commute, one sees that the kernel K is a free abelian group of rank k generated by these k Dehn twists. Here the capping homomorphism splits. To prove this, we first embed PB_k into PRB_k^{+} by viewing the pure braid group of k strings as the set of ribbon braids on k bands that have no twists. We can think of D_k as being embedded into $\mathbb{R}^2$ with ∂_b as the unit circle and the k holes in D_k equally distributed inside ∂_b along the x-axis. The intersections of these holes with the x-axis gives k subintervals of the x-axis, denoted by $I_1, \dots, I_k$. We now put the bands representing a pure braid $x \in \mathrm{PB}_k \le \mathrm{PRB}_k^{+}$ in $D \times [0,1]$ which starts and ends at $I_1, \dots, I_k$. Note that the bands here will not twist at all. Now we comb the bands straight from bottom to top. This induces a homeomorphism of $D_k \times \{0\}$ and hence an element in the mapping class group $\mathrm{Map}(D_k)$. One checks that this map is a group homomorphism and injective. Since PB_k acts on K trivially, we have $\mathrm{Map}(D_k) \cong K \times \mathrm{PB}_k \cong \mathbb{Z}^k \times \mathrm{PB}_k \cong \mathrm{PRB}_k^{+}$, where the number of Dehn twists around each boundary component is naturally identified with the number of full twists on each bands. □

To promote Lemma 3.21 so that it works for the ribbon braid group, we need some extra terminology. As in the proof of Lemma 3.21, we identify D_k with the unit disk in $\mathbb{R}^2$ with k small disks removed whose centers are equally distributed on the x-axis. The x-axis cuts the boundary loops of each deleted disk into two components, providing a cell structure on the loops. We will call the part that lies above the x-axis the *visible part*. We define the *rigid mapping class group* $\mathrm{RMap}_{+}(D_k)$ of D_k to be the isotopy classes of homeomorphisms of D_k which fix $\partial_b D_k$ pointwise and map the visible part of the holes to the visible part of the holes. Note elements in $\partial_b D_k$ are allowed to map one boundary hole to another just as in the definition of the asymptotic mapping class group. If we only assume the cell structure on the loops has to be preserved, the resulting group is called *quasirigid mapping class group* D_k and denoted by $\mathrm{RMap}(D_k)$. With these preparations, the following lemma is now clear.

Lemma 3.22 *There is a natural isomorphism between the oriented ribbon braid group RB_k^{+} and $\mathrm{RMap}_{+}(D_k)$ (resp. between the ribbon braid group RB_k and $\mathrm{RMap}(D_k)$).*

Proof As in the proof of Lemma 3.21, we put the element in the (oriented) ribbon braid group between $D \times [0,1]$, then we comb the bands straight from bottom to top, which gives the corresponding element in $\mathrm{RMap}_+(D_k)$ (resp. $\mathrm{RMap}(D_k)$). □

Given two admissible subsurfaces A and A' of $D^\infty_{d,r}$ (possibly with different r) with k admissible boundary components, we want to fix a canonical way to identify a homeomorphism $f\colon A \to A'$ as an element in the ribbon braid group. Note that each boundary loop except the base one inherits a visible side from $D^\infty_{d,r}$. We will use the puncture model for $D^\infty_{d,r}$ going forward.

As above, let D_k be the subsurface of D which is the complement of k disjoint open disks with centers at $a_i = (2i-k-1)/(k+1)$ of radius 2^{-k} for $1 \le i \le k$. Now, given any admissible subsurface A_k of $D^\infty_{d,r}$ with k admissible boundaries, denote the centers from left to right by $c_i \in [0,1]\times\{0\}$ for $1 \le i \le k$ with corresponding radii $r_1, r_2, \dots, r_k$. Now we define an isotopy $\mathcal{N}_{A_k}\colon D\times[0,1] \to D$ such that $\mathcal{N}_{A_k,0} = \mathrm{id}_D$ and $\mathcal{N}_{A_k,1}$ maps A_k to D_k via a homeomorphism. We first shrink the admissible boundary loops of A_k so that they have radius r, where $r = \min\{r_1, \dots, r_k, 2^{-k}\}$. Then we isotope A_k by moving the centers c_i to a_i along $[0,1]\times\{0\}$ in D. And in the last step we enlarge the radii one by one to 2^{-k}. The following lemma is now immediate.

Lemma 3.23 *Let $\phi\colon D^\infty_{d,r} \to D^\infty_{d,r}$ be an asymptotically rigid (resp. quasirigid) homeomorphism which is supported on the admissible subsurface A_k. Write $A'_k = \phi(A_k)$; then:*

1. $\mathfrak{r}_\phi = \mathcal{N}_{A'_k,1} \circ \phi|_{A_k} \circ \mathcal{N}^{-1}_{A_k,1}\colon D_k \to D_k$ *gives an element in the oriented ribbon braid group* RB^+_k *(resp. the ribbon braid group* RB_k*). Conversely, given an element* $\mathfrak{r} \in \mathrm{RB}^+_k$ *(resp.* RB_k*), we have an asymptotic rigid (resp. quasirigid) homeomorphism which is unique up to isotopy, supported on A_k, and maps A_k to A'_k.*
2. *Let A_{k+d} be the admissible subsurface of $D^\infty_{d,r}$ obtained from A_k by adding a d-leg pants and $\phi(A_{k+d}) = A'_{k+d}$. Then the associated oriented ribbon braid (resp. the ribbon braid) of ϕ can be obtained from $\mathfrak{r}_\phi$ by splitting the corresponding band into d bands. Conversely, if we split one band of the ribbon braids into d bands, the isotopy class of the corresponding asymptotic rigid (resp. quasirigid) homeomorphism does not change.*

Note that, for any $d \ge 2$ and $r \ge 1$, we have a natural embedding $\iota_{d,r}\colon D^\infty_{d,r} \to D^\infty_{d,r+1}$ that maps the rooted boundaries of $D^\infty_{d,r}$ to the first r rooted boundaries according to the order. This induces embeddings of groups $i_{\mathcal{H},d,r}\colon \mathcal{H}V_{d,r} \to \mathcal{H}V_{d,r+1}$ and $i_{\mathcal{B},d,r}\colon \mathcal{B}V_{d,r} \to \mathcal{B}V_{d,r+1}$. On the other hand, we also have natural embeddings $i_{R,d,r}\colon RV_{d,r} \to RV_{d,r+1}$ and $i_{R^+,d,r}\colon RV^+_{d,r} \to RV^+_{d,r+1}$ induced by inclusion of roots. We have the following.

Theorem 3.24 *There exist isomorphisms $f_{d,r}\colon \mathcal{H}V_{d,r} \to RV_{d,r}$ such that $f_{d,r+1} \circ i_{\mathcal{H},d,r} = i_{\mathcal{H},d,r+1} \circ f_{d,r}$. Restricting to the subgroups $\mathcal{B}V_{d,r}$, one gets isomorphisms $f_{d,r}\colon \mathcal{B}V_{d,r} \to RV^+_{d,r}$ with the same property.*

Proof Since the two cases are parallel, we will only prove the theorem for $\mathcal{B}V_{d,r}$. We will define two maps $f_{d,r}\colon \mathcal{B}V_{d,r} \to RV^{+}_{d,r}$ and $g_{d,r}\colon RV^{+}_{d,r} \to \mathcal{B}V_{d,r}$ such that $f_{d,r} \circ g_{d,r} = \mathrm{id}$ and $g_{d,r} \circ f_{d,r} = \mathrm{id}$.

Given an element $x \in \mathcal{B}V_{d,r}$, we can define $f_{d,r}$ as follows. Let φ_x be an asymptotically rigid homeomorphism of $D^{\infty}_{d,r}$ representing x with support A_k, where k is the number of admissible loops. By Proposition 3.18, $\pi(x)$ provides an element $[F_-, \sigma, F_+]$ in the Higman–Thompson group $V_{d,r}$, where F_- and F_+ are (d,r)-forests with k leaves. But what we want is a ribbon braid connecting the k leaves. For this we simply apply Lemma 3.23(1) to the map φ_x with support A_k; denote the corresponding element in RB^{+}_{k} by $\mathfrak{r}_{\varphi_x}$. We define $f_{d,r}(x) = [F_-, \mathfrak{r}_{\varphi_x}, F_+]$.

Now, given $y \in RV^{+}_{d,r}$, one can define an element in $\mathcal{B}V_{d,r}$ as follows. Suppose $(F_-, \mathfrak{r}_y, F_+)$ is a representative for y, where F_- and F_+ are (d,r)-forests and $\mathfrak{r}$ is a ribbon braid between the leaves of F_- and F_+. Add a root to F_- (resp. F_+) with an edge connecting to the root of each tree in F_- (resp. F_+) and then throw away the open half edge connecting to the leaves. Denote the resulting tree by T_- (resp. T_+). Now $q^{-1}(T_-)$ and $q^{-1}(T_+)$ give us two admissible subsurfaces A_k and A'_k in $D^{\infty}_{d,r}$, where k is the number of leaves for F_-. And, by Lemma 3.23(1), the ribbon element $\mathfrak{r}_y$ in RB^{+}_{k} gives us an asymptotic rigid homeomorphism ψ_y with support A_k and maps A_k to A'_k.

Now one can check that $f_{d,r} \circ g_{d,r} = \mathrm{id}$ and $g_{d,r} \circ f_{d,r} = \mathrm{id}$. Therefore, the two groups are isomorphic. The fact that the diagram commutes is immediate from the definition. □

4 Homological stability of ribbon Higman–Thompson groups

In this section, we show the homological stability for oriented ribbon Higman–Thompson groups and explain at the end how the same proof applies to the ribbon Higman–Thompson groups.

4.1 Homogeneous categories and homological stability

In this subsection, we review the basics of homogeneous categories and refer the reader to [27] for more details. Note that we adopt their convention of identifying objects of a category with their identity morphisms.

Definition 4.1 [27, Definition 1.3] A monoidal category $(\mathcal{C}, \oplus, 0)$ is called *homogeneous* if 0 is initial in $\mathcal{C}$ and if the following two properties hold:

(H1) $\mathrm{Hom}(A, B)$ is a transitive $\mathrm{Aut}(B)$-set under postcomposition.

(H2) The map $\mathrm{Aut}(A) \to \mathrm{Aut}(A \oplus B)$ taking f to $f \oplus \mathrm{id}_B$ is injective with image

$$\mathrm{Fix}(B) := \{\phi \in \mathrm{Aut}(A \oplus B) \mid \phi \circ (\iota_A \oplus \mathrm{id}_B) = \iota_A \oplus \mathrm{id}_B\},$$

where $\iota_A\colon 0 \to A$ is the unique map.

Definition 4.2 [27, Definition 1.5] Let $(\mathscr{C}, \oplus, 0)$ be a monoidal category with 0 initial. We say that $\mathscr{C}$ is *prebraided* if its underlying groupoid is braided and, for each pair of objects A and B in $\mathscr{C}$, the groupoid braiding $b_{A,B}\colon A \oplus B \to B \oplus A$ satisfies

$$b_{A,B} \circ (A \oplus \iota_B) = \iota_B \oplus A\colon A \to B \oplus A.$$

Definition 4.3 [27, Definition 2.1] Let $(\mathscr{C}, \oplus, 0)$ be a monoidal category with 0 initial and (A, X) a pair of objects in $\mathscr{C}$. Define $W_n(A, X)_\bullet$ to be the semisimplicial set with set of p-simplices

$$W_n(A, X)_p := \mathrm{Hom}_{\mathscr{C}}(X^{\oplus p+1}, A \oplus X^{\oplus n})$$

and with face map

$$d_i\colon \mathrm{Hom}_{\mathscr{C}}(X^{\oplus p+1}, A \oplus X^{\oplus n}) \to \mathrm{Hom}_{\mathscr{C}}(X^{\oplus p}, A \oplus X^{\oplus n})$$

defined by precomposing with $X^{\oplus i} \oplus \iota_X \oplus X^{\oplus p-i}$.

Also say the category $\mathscr{C}$ satisfies (LH3) at a pair of objects (A, X) with *slope* $k \geq 2$ if:

(LH3) For all $n \geq 1$, $W_n(A, X)_\bullet$ is $((n-2)/k)$-connected.

Quite often, we can reduce the semisimplicial complex to a simplicial complex.

Definition 4.4 [27, Definition 2.8] Let A, X be objects of a homogeneous category $(\mathscr{C}, \oplus, 0)$. For $n \geq 1$, let $S_n(A, X)$ denote the simplicial complex whose vertices are the maps $f\colon X \to A \oplus X^{\oplus n}$ and whose p-simplices are $p+1$ sets $\{f_0, \dots, f_p\}$ such that there exists a morphism $f\colon X^{\oplus p+1} \to A \oplus X^{\oplus n}$ with $f \circ i_j = f_j$ for some order on the set, where

$$i_j = \iota_{X^{\oplus j}} \oplus \mathrm{id}_X \oplus \iota_{X^{\oplus p-j}}\colon X = 0 \oplus X \oplus 0 \to X^{\oplus p+1}.$$

Definition 4.5 Let $\mathrm{Aut}(A \oplus X^{\oplus \infty})$ be the colimit of

$$\cdots \xrightarrow{-\oplus X} \mathrm{Aut}(A \oplus X^{\oplus n}) \xrightarrow{-\oplus X} \mathrm{Aut}(A \oplus X^{\oplus n+1}) \xrightarrow{-\oplus X} \mathrm{Aut}(A \oplus X^{\oplus n+2}) \xrightarrow{-\oplus X} \cdots.$$

Then any $\mathrm{Aut}(A \oplus X^{\oplus \infty})$-module M may be considered as an $\mathrm{Aut}(A \oplus X^{\oplus n})$-module for any n, by restriction, which we continue to call M. We say that the module M is abelian if the action of $\mathrm{Aut}(A \oplus X^{\oplus \infty})$ on M factors through the abelianizations of $\mathrm{Aut}(A \oplus X^{\oplus \infty})$, or in other words if the derived subgroup of $\mathrm{Aut}(A \oplus X^{\oplus \infty})$ acts trivially on M.

We are now ready to quote the theorem that we will use.

Theorem 4.6 [27, Theorem 3.4] *Let $(\mathscr{C}, \oplus, 0)$ be a prebraided homogeneous category satisfying* (LH3) *for a pair (A, X) with slope $k \geq 3$. Then, for any abelian $\mathrm{Aut}(A \oplus X^{\oplus \infty})$-module M, the map*

$$H_i(\mathrm{Aut}(A \oplus X^{\oplus n}); M) \to H_i(\mathrm{Aut}(A \oplus X^{\oplus n+1}); M)$$

induced by the natural inclusion map is surjective if $i \leq (n-k+2)/k$, and injective if $i \leq (n-k)/k$.

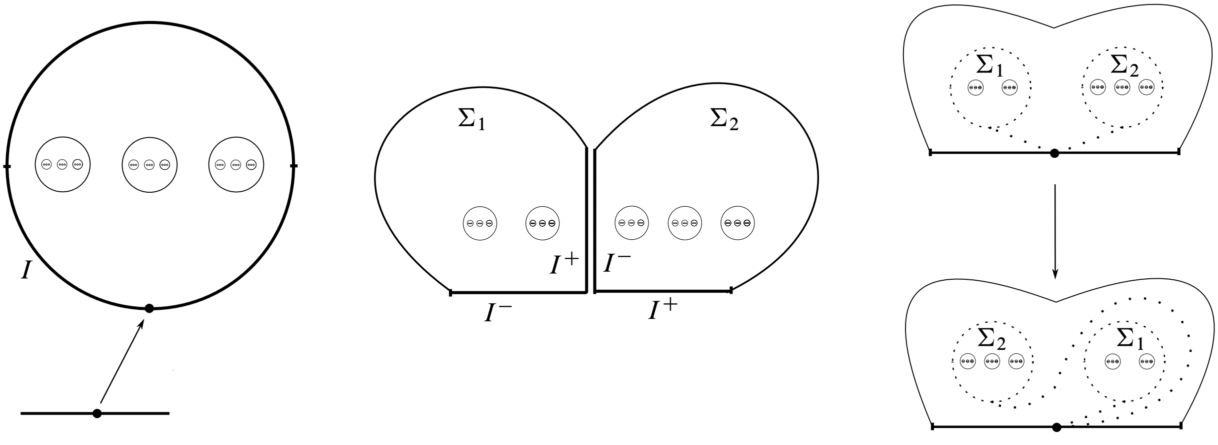


Figure 6: The braided monoidal structure for the category $\mathscr{G}_d$.

4.2 Homogeneous category for the groups $RV^+_{d,r}$

The purpose of this section is to produce a homogeneous category for proving homological stability of the ribbon Higman–Thompson groups $RV^+_{d,r}$. Note that, by Theorem 3.24, this is the same as proving the asymptotic mapping class groups $\mathscr{B}V_{d,r}$ have homological stability. This allows us to define our homogeneous category geometrically. The category is similar to the ones produced in [27, Section 5.6]. Essentially, we replace the annulus or Möbius band by the infinite surface $D^\infty_{d,1}$.

Recall $D^\infty_{d,r}$ is an infinite surface equipped with a canonical asymptotic rigid structure with boundary component denoted by $\partial_b D^\infty_{d,r}$. Let $I = [-1, 1] \subset \partial_b D^\infty_{d,r}$ be an embedded interval as in Figure 6, left. Let $I^- = [-1, 0]$ and $I^+ = [0, 1]$ be subintervals of I. Let $D^\infty_{d,1} \oplus D^\infty_{d,1}$ be the boundary sum of two copies of $D^\infty_{d,1}$ obtained by identifying I^+ of the first copy with I^- of the second copy. Inductively, we could define similarly $\bigoplus_r D^\infty_{d,1}$ for any $r \geq 0$. Here $\bigoplus_0 D^\infty_{d,1}$ is just the standard disk D. Abusing notation, when referring to I^- and I^+ on $\bigoplus_r D^\infty_{d,1}$, we will mean the two copies of I^- and I^+ which remain on the boundary. Thus we have an operation $\oplus$ on the set $\bigoplus_r D^\infty_{d,1}$ for any $r \geq 0$. See Figure 6, center, for a picture of $\bigl(\bigoplus_2 D^\infty_{d,1}\bigr) \oplus \bigl(\bigoplus_3 D^\infty_{d,1}\bigr)$. In fact, $\bigl(\bigl(\bigoplus_r D^\infty_{d,1}\bigr), \oplus\bigr)$ is the free monoid generated by $D^\infty_{d,1}$. Note that $\bigoplus_r D^\infty_{d,1}$ has a naturally induced d-rigid structure and we can identify it with $D^\infty_{d,r}$, which will be of use to us later.

We can now define the category $\mathscr{G}_d$ to be the monoidal category with objects $\bigoplus_r D^\infty_{d,1}$ for $r \geq 0$, $\oplus$ as the operation, and D as the 0 object. So far this is the same as defining the objects as the natural numbers and addition as the operation. When $r = s$, we define the morphisms $\operatorname{Hom}\bigl(\bigoplus_r D^\infty_{d,1}, \bigoplus_s D^\infty_{d,1}\bigr) = \mathscr{B}V_{d,r}$, which is the group of isotopy classes of asymptotically rigid homeomorphisms of $D^\infty_{d,r}$; when $r \neq s$, let $\operatorname{Hom}\bigl(\bigoplus_r D^\infty_{d,1}, \bigoplus_s D^\infty_{d,1}\bigr) = \varnothing$. Note that we did not universally define the morphisms to be the sets of isotopy classes of asymptotically rigid homeomorphisms as we want our category to satisfy cancellation, ie if $A \oplus C = A$ then $C = 0$; see [27, Remark 1.11] for more information. The category $\mathscr{G}_d$ has a natural braiding as in the usual braid group case; see Figure 6, right.

Now, applying [27, Theorem 1.10], we have a prebraided homogeneous category $U\mathscr{G}_d$, which has the same objects as $\mathscr{G}_d$ and morphisms defined as follows: for any $s \le r$, a morphism in $\mathrm{Hom}\big(\bigoplus_s D^\infty_{d,1}, \bigoplus_r D^\infty_{d,1}\big)$ is an equivalence class of pairs $\big(\bigoplus_{r-s} D^\infty_{d,1}, f\big)$, where $f\colon \big(\bigoplus_{r-s} D^\infty_{d,1}\big) \oplus \big(\bigoplus_s D^\infty_{d,1}\big) \to \bigoplus_r D^\infty_{d,1}$ is a morphism in $\mathscr{G}_d$ and $\big(\bigoplus_{r-s} D^\infty_{d,1}, f\big) \sim \big(\bigoplus_{r-s} D^\infty_{d,1}, f'\big)$ if there exists an isomorphism $g\colon \bigoplus_{r-s} D^\infty_{d,1} \to \bigoplus_{r-s} D^\infty_{d,1} \in \mathscr{G}_d$ making the following diagram commute up to isotopy:

$$\begin{tikzcd}
\big(\bigoplus_{r-s} D^\infty_{d,1}\big) \oplus \big(\bigoplus_s D^\infty_{d,1}\big) \arrow[r, "f"] \arrow[d, "g \oplus \mathrm{id}_{\bigoplus_s D^\infty_{d,1}}"'] & \bigoplus_r D^\infty_{d,1} \\
\big(\bigoplus_{r-s} D^\infty_{d,1}\big) \oplus \big(\bigoplus_s D^\infty_{d,1}\big) \arrow[ur, "f'"'] &
\end{tikzcd}$$

We write $\big[\bigoplus_{r-s} D^\infty_{d,1}, f\big]$ for such an equivalence class. Now, by Theorem 4.6, to prove homological stability for the oriented ribbon Higman–Thompson groups, we only need to verify that the category $\mathscr{G}_d$ satisfies (LH3) at the pair $(D, D^\infty_{d,1})$. In fact, by Theorem 3.24, proving the oriented ribbon Higman–Thompson groups satisfy homological stability is the same as proving that the asymptotic mapping class groups $\mathscr{B}V_{d,r}$ satisfy homological stability. Now consider the family of groups

$$\mathrm{Aut}(A \oplus X) \hookrightarrow \mathrm{Aut}(A \oplus X^{\oplus 2}) \hookrightarrow \mathrm{Aut}(A \oplus X^{\oplus 2}) \hookrightarrow \cdots \hookrightarrow \mathrm{Aut}(A \oplus X^{\oplus n}) \hookrightarrow \cdots,$$

where $A = D$ and $X = D^\infty_{d,1}$. By definition, this gives rise to the family of groups $\mathscr{B}V_{d,1} \hookrightarrow \mathscr{B}V_{d,2} \hookrightarrow \cdots \hookrightarrow \mathscr{B}V_{d,n} \hookrightarrow \cdots$. Now we have shown that the category $(\mathscr{G}_d, \oplus, D)$ is a prebraided homogeneous category, so, by Theorem 4.6, it suffices to verify (LH3) at the pair $(D, D^\infty_{d,1})$ to prove our homological stability result. As a matter of fact, we will show that $W_r(D, D^\infty_{d,1})_\bullet$ is $(r-3)$-connected in the next subsection. First, let us further characterize the morphisms in $U\mathscr{G}_d$. Call $0 = I^- \cap I^+$ the basepoint of $\bigoplus_r D^\infty_{d,1}$.

Definition 4.7 Given $s < r$, an injective map $\varphi\colon \big(\bigoplus_s D^\infty_{d,1}, I^+\big) \to \big(\bigoplus_r D^\infty_{d,1}, I^+\big)$ is called an *asymptotically rigid embedding* if it satisfies the following properties:

(1) $\varphi(\partial D^\infty_{d,s}) \cap \partial D^\infty_{d,r} = I^+$.

(2) φ maps $\bigoplus_s D^\infty_{d,1}$ homeomorphically to $\varphi\big(\bigoplus_s D^\infty_{d,1}\big)$ and there exists an admissible surface $A \subset \bigoplus_s D^\infty_{d,1}$ such that $\varphi\colon \bigoplus_s D^\infty_{d,1} \setminus A \to \varphi\big(\bigoplus_s D^\infty_{d,1}\big) \setminus \varphi(A)$ is rigid.

(3) The closure of the complement of $\varphi\big(\bigoplus_s D^\infty_{d,1}\big)$ in $\bigoplus_r D^\infty_{d,1}$ with its induced d-rigid structure is asymptotically rigidly homeomorphic to $\bigoplus_{r-s} D^\infty_{d,1}$.

Lemma 4.8 *For $s < r$, the equivalence classes of pairs $\big[\bigoplus_{r-s} D^\infty_{d,1}, f\big]$ are in one-to-one correspondence with the isotopy classes of asymptotically rigid embeddings of $\big(\bigoplus_s D^\infty_{d,1}, I^+\big)$ into $\big(\bigoplus_r D^\infty_{d,1}, I^+\big)$.*

Remark 4.9 Here isotopies are carried out among asymptotically rigid embeddings.

Proof Given an equivalence class of a pair $\big[\bigoplus_{t-s} D^{\infty}_{d,1}, f\big]$, the restriction map $f|_{\bigoplus_s D^{\infty}_{d,1}}$ is an asymptotically rigid embedding. Any two equivalence classes of pairs will induce the same map $f|_{\bigoplus_s D^{\infty}_{d,1}}$; hence, we have a well-defined map from the set of equivalence pairs to the set of isotopy classes of asymptotically rigid embeddings.

We produce the inverse of the restriction map as follows. If we have an asymptotically rigid embedding $\varphi\colon \big(\bigoplus_s D^{\infty}_{d,1}, I^+\big) \to \big(\bigoplus_r D^{\infty}_{d,1}, I^+\big)$, by part (3) of Definition 4.7, we also have an asymptotically rigid homeomorphism $\phi\colon C \to \bigoplus_{r-s} D^{\infty}_{d,1}$, where C is the closure of the complement of $\varphi\big(\bigoplus_s D^{\infty}_{d,1}\big)$ in $\bigoplus_r D^{\infty}_{d,1}$. Up to isotopy, we can assume $\phi^{-1}|_{I^+}$ coincides with $\varphi|_{I^-}$. Now define a map $\bar{f}\colon \big(\bigoplus_{r-s} D^{\infty}_{d,1}\big) \oplus \big(\bigoplus_s D^{\infty}_{d,1}\big) \to \bigoplus_r D^{\infty}_{d,1}$ by $\bar{f}|_{\bigoplus_{r-s} D^{\infty}_{d,1}} = \phi^{-1}$ and $\bar{f}|_{\bigoplus_s D^{\infty}_{d,1}} = \varphi$. One can check that $\bar{f}$ is an asymptotically rigid homeomorphism. Then $\big(\bigoplus_{r-s} D^{\infty}_{d,1}, \bar{f}\big)$ gives a representative of an equivalence class of pairs. □

4.3 Higher connectivity of the complex $W_r(D, D^{\infty}_{d,1})_\bullet$

We want to prove that the complex $W_r(D, D^{\infty}_{d,1})_\bullet$ is highly connected; see Figure 8 and the paragraph following the proof of Lemma 4.24 for an outline of our general strategy.

Remark 4.10 As explained in the proof of [27, Lemma 5.21], a simplex of $S_r(D, D^{\infty}_{d,1})$ has a canonical ordering on its vertices induced by the local orientation of the surfaces near the parametrized interval in their based boundary. Thus the geometric realization $|W_r(D, D^{\infty}_{d,1})_\bullet|$ is homeomorphic to $S_r(D, D^{\infty}_{d,1})$.

Our first step now is to simplify the complex $S_r(D, D^{\infty}_{d,1})$ further.

Definition 4.11 Given $r \geq 2$, we call a loop $\alpha\colon (I, \partial I) = ([0,1], \{0,1\}) \to \big(\bigoplus_r D^{\infty}_{d,1}, 0\big)$ an *asymptotically rigidly embedded loop* if there exists an asymptotically rigid embedding $\varphi\colon (D^{\infty}_{d,1}, I^+) \to \big(\bigoplus_r D^{\infty}_{d,1}, I^+\big)$ with $\varphi|_{(\partial D^{\infty}_{d,1}, 0)} = \alpha$ up to based isotopy. See Figure 7.

Remark 4.12 When $r = 1$, we just call a loop asymptotically rigidly embedded if it is isotopic to the boundary.

Lemma 4.13 *When $r \geq 2$, a loop $\alpha\colon (I, \partial I) \to \big(\bigoplus_r D^{\infty}_{d,1}, 0\big)$ is isotopic to an asymptotically rigidly embedded loop if and only if there exists an admissible surface $A \subseteq \bigoplus_r D^{\infty}_{d,1}$ such that the admissible loops of A are disjoint from α, the number of admissible loops of A that lie in the disk bounded by α is $1 + a(d-1)$ for some $a \geq 0$, and there exist some admissible loops which do not lie inside the disk bounded by α up to isotopy.*

Proof It is clear that a loop which is isotopic to an asymptotically rigidly embedded loop has the properties given in the lemma.

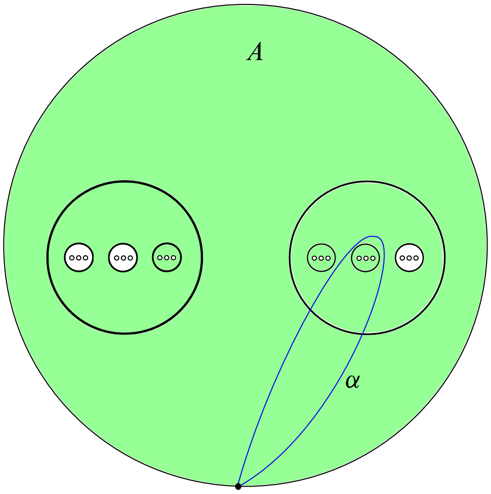


Figure 7: The curve α is an asymptotically rigidly embedded loop, with the green shaded surface A the corresponding admissible subsurface.

For the other direction, we can assume up to isotopy that $\alpha(I) \cap \partial\big(\bigoplus_r D^\infty_{d,1}\big) = I^+$. We know that $D^\infty_{d,r}$ is asymptotically rigidly homeomorphic to $D^\infty_{d,r+d-1}$; thus, the surface bounded by the loop α is asymptotically rigidly homeomorphic to $D^\infty_{d,1}$. Therefore, the number of boundary components of A bounded by the complement disk is $r-1 \bmod d-1$ and thus $D^\infty_{d,r-1}$ is asymptotically rigidly homeomorphic to the complement surface. These two facts together imply α is an asymptotically rigidly embedded loop. □

Now we define the complex $U_r(D, D^\infty_{d,1})$ which is the surface version of the complex U_r given in [29, Section 2.4].

Definition 4.14 For $r \geq 1$, let $U_r(D, D^\infty_{d,1})$ denote the simplicial complex whose vertices are isotopy classes of asymptotically rigidly embedded loops and a set of vertices $\alpha_0, \dots, \alpha_p$ forms a p-simplex if and only if any corresponding asymptotically rigid embeddings $\phi_0, \dots, \phi_p$ form a p-simplex in $S_r(D, D^\infty_{d,1})$.

We denote the canonical map from $S_r(D, D^\infty_{d,1})$ to $U_r(D, D^\infty_{d,1})$ by π. The next lemma follows directly from the definition. In fact, given a set of vertices $\alpha_0, \dots, \alpha_p$, if they form a p-simplex, then any corresponding asymptotically rigid embeddings $\phi_0, \dots, \phi_p$ form a p-simplex in $S_r(D, D^\infty_{d,1})$. But this means that, for any $\psi_i \in \pi^{-1}(\alpha_i)$ for $0 \leq i \leq p$, the collection $\psi_0, \psi_1, \dots, \psi_p$ forms a p-simplex.

Lemma 4.15 *The map π is a complete join.*

By Proposition 1.3, we need only show that $U_r(D, D^\infty_{d,1})$ is highly connected. Similar to [29, Section 2.4], we will produce several other complexes closely related to $U_r(D, D^\infty_{d,1})$. We first have the following complex, which is analogous to the complex U^∞_r in [29, Definition 2.12].

Definition 4.16 Let $U_r^{\infty}(D, D_{d,1}^{\infty})$ be the simplicial complex with vertices given by asymptotically rigidly embedded loops in $\bigoplus_r D_{d,1}^{\infty}$, where $\alpha_0, \alpha_1, \dots, \alpha_p$ form a p-simplex if the punctured disks bounded by them are pairwise disjoint (outside of the basepoint) and there exists at least one admissible loop that does not lie in those disks.

Remark 4.17 (1) The $(r-2)$-skeleton of $U_r^{\infty}(D, D_{d,1}^{\infty})$ is the same as that of $U_r(D, D_{d,1}^{\infty})$. Notice though that $U_r^{\infty}(D, D_{d,1}^{\infty})$ is in fact infinite-dimensional.

(2) Since $\bigoplus_r D_{d,1}^{\infty}$ is asymptotically rigidly homeomorphic to $\bigoplus_{r+d-1} D_{d,1}^{\infty}$, $U_r^{\infty}(D, D_{d,1}^{\infty})$ is isomorphic to $U_{r+d-1}^{\infty}(D, D_{d,1}^{\infty})$ as a simplicial complex.

We also need another complex, which is the surface version of the complex T_r^{∞} of [29, Defintion 2.14]. For convenience, we will orient the admissible loops in $\bigoplus_r D_{d,1}^{\infty}$ so that they bound the punctured disk according to the orientation.

Definition 4.18 An *almost admissible loop* is a loop $\alpha\colon (I, \partial I) \to \bigl(\bigoplus_r D_{d,1}^{\infty}, 0\bigr)$ which is freely isotopic to one of the nonbased admissible loops.

Note that, by Lemma 4.13, an almost admissible loop is an asymptotically rigidly embedded loop.

Definition 4.19 Define the simplicial complex $T_r^{\infty}(D, D_{d,1}^{\infty})$ to be the full subcomplex of $U_r^{\infty}(D, D_{d,1}^{\infty})$ all of whose vertices are almost admissible loops.

Just as discussed in Remark 4.17, $T_r^{\infty}(D, D_{d,1}^{\infty})$ is in fact isomorphic to $T_{r+d-1}^{\infty}(D, D_{d,1}^{\infty})$ as a simplicial complex.

We now want to further characterize the almost admissible loops by building a connection to the usual arc complex. We let A be the quotient $[0,2]/1 \sim 2$. This corresponds to identifying the endpoint 1 of the interval $[0, 1]$ with the basepoint 1 of the circle given by $[1,2]/1 \sim 2$.

Definition 4.20 An injective continuous map $L\colon (A, 0) \to (D_{d,r}^{\infty}, 0)$ is called a *lollipop* on the surface $D_{d,r}^{\infty}$ if $\alpha|_{[1,2]}$ is isotopic to an admissible loop in $D_{d,r}^{\infty}$ and $L|_{[0,1]}$ is an arc connecting the basepoint 0 to the loop $L([1,2])$. The map $L|_{[0,1]}$ is called the *arc part* of the lollipop L and $L|_{[1,2]}$ is called the *loop part*. See Figure 9, where the blue curve is a lollipop.

Lollipops are examples of what Hatcher and Vogtmann [21] refer to as tethered curves.

Lemma 4.21 *The set of isotopy classes of almost admissible loops is in one-to-one correspondence with the set of isotopy classes of lollipops.*

Proof We define a map g from the isotopy classes of lollipops to the isotopy classes of almost admissible loops and show that the map is bijective.

Given a lollipop $L\colon (A,0)\to(D^{\infty}_{d,r},0)$, we can map it to an almost admissible loop $\alpha\colon[0,1]\to(D^{\infty}_{d,r},0)$ as follows. We define $\alpha(0)=0$ and let $\alpha(t)$ run parallel to L outside the region bounded by L. The orientation of α is simply the one that coincides with the loop part of L. Since α can be freely homotoped to the admissible loop $L|_{[1,2]}$, α is almost admissible. Any isotopy of L induces an isotopy of α; hence, the map is well defined.

Now we show g is surjective. Given any almost admissible loop $\alpha\colon[0,1]\to(D^{\infty}_{d,r},0)$, let A be the admissible loop which is freely isotopic to α. Up to isotopy, we can assume that A lies in the interior of the surface bounded by α. Then the surface bounded by α and A must be an annulus. From here one can produce an arc connecting the basepoint 0 to a point in A. Together with A, this provides the lollipop.

Finally, we argue that g is injective. Suppose L_1 and L_2 are two lollipops such that $g(L_1)$ and $g(L_2)$ are isotopic; denote the isotopy by f. By the isotopy extension theorem (see for example [12, Proposition 1.11]) there exists an isotopy $F\colon D^{\infty}_{d,r}\times[0,1]\to D^{\infty}_{d,r}$ such that $F|_{D^{\infty}_{d,r}\times 0}=\mathrm{id}_{D^{\infty}_{d,r}}$ and $F|_{g(L_1)\times[0,1]}=f$. In particular, $F|_{D^{\infty}_{d,r}\times 1}$ maps the almost admissible loop $g(L_1)$ to the almost admissible loop $g(L_2)$. Hence L_1 is isotoped through F to a lollipop which lies in a small neighborhood of L_2 and is bounded by the loop $g(L_2)$. Therefore, one can then isotope L_1 to L_2. □

We now have the following definition of lollipop complex.

Definition 4.22 The *lollipop complex* $L^{\infty}_{r}(D,D^{\infty}_{d,1})$ has vertices as lollipops, and $p+1$ lollipops $L_0,L_1,\dots,L_p$ form a p-simplex if they are pairwise disjoint outside the basepoint 0 and there exists at least one admissible loop which does not lie inside the disks bounded by the L_i.

The following lemma is immediate from Lemma 4.21.

Lemma 4.23 *The complex $L^{\infty}_{r}(D,D^{\infty}_{d,1})$ is isomorphic to $T^{\infty}_{r}(D,D^{\infty}_{d,1})$ as a simplicial complex.*

Lemma 4.24 *Given a p-simplex σ in $L^{\infty}_{r}(D,D^{\infty}_{d,1})$, its link* $\mathrm{Lk}(\sigma)$ *is isomorphic to $L^{\infty}_{r_\sigma}(D,D^{\infty}_{d,1})$ for some $r_\sigma>0$.*

Proof By Lemma 4.23, we can just prove the lemma for $T^{\infty}_{r}(D,D^{\infty}_{d,1})$. Let $\alpha_0,\alpha_1,\dots,\alpha_p$ be the vertices of σ, which are almost admissible loops. Up to isotopy, we can assume they are pairwise disjoint except at the basepoint 0. Now let C be the complement surface of σ, whose based boundary is the concatenation of $\alpha_p,\alpha_{p-1},\dots,\alpha_0$ and ∂D. The surface C has a naturally induced d-rigid structure. In particular, C is asymptotically rigidly homeomorphic to D^{∞}_{d,r_σ} for some $r_\sigma>0$. Thus link $\mathrm{Lk}(\sigma)$ is isomorphic to $T^{\infty}_{r_\sigma}(D,D^{\infty}_{d,1})$. □

$$\begin{tikzcd}
 & & & U_r^\infty(D, D_{d,1}^\infty) & T_r^\infty(D, D_{d,1}^\infty) \arrow[l, "\supseteq"'] \arrow[r, "\cong"] & L_r^\infty(D, D_{d,1}^\infty) \\
W_r(D, D_{d,1}^\infty)_\bullet \arrow[r, "\simeq"] & S_r(D, D_{d,1}^\infty) \arrow[r, "\pi"] & U_r(D, D_{d,1}^\infty) & & & \\
 & & U_r^{(r-2)}(D, D_{d,1}^\infty) \arrow[u, "\subseteq"'] \arrow[r, "\cong"] & (U_r^\infty(D, D_{d,1}^\infty))^{(r-2)} \arrow[uu, "\subseteq"'] & &
\end{tikzcd}$$

Figure 8: A summary of the relationships between the complexes defined so far.

Let us summarize the relationships we have so far between our various complexes, which are illustrated in Figure 8. The leftmost homeomorphism between $W_r(D, D_{d,1}^\infty)_\bullet$ and $S_r(D, D_{d,1}^\infty)$ comes from Remark 4.10. By Lemma 4.15, there is a complete join map π from the complex $S_r(D, D_{d,1}^\infty)$ to the complex of isotopy classes of asymptotically rigidly embedded loops, $U_r(D, D_{d,1}^\infty)$, which implies that both complexes have exactly the same connectivity properties. Thus we can choose to work with $U_r(D, D_{d,1}^\infty)$. Next, Remark 4.17(1) demonstrates that the $(r-2)$-skeleton of $U_r(D, D_{d,1}^\infty)$ is the same as the $(r-2)$-skeleton of the complex of asymptotically rigidly embedded loops in $\bigoplus_r D_{d,1}^\infty$, denoted by $U_r^\infty(D, D_{d,1}^\infty)$. Since our goal is to show $W_r(D, D_{d,1}^\infty)_\bullet$ is weakly Cohen–Macauley of dimension $r-2$ (see Corollary 4.30), this implies we can again shift our focus to $U_r^\infty(D, D_{d,1}^\infty)$. Next, by Definition 4.19, $T_r^\infty(D, D_{d,1}^\infty)$ is a subcomplex of $U_r^\infty(D, D_{d,1}^\infty)$. In the next pages, we will show that this complex is isomorphic to the lollipop complex $L_r^\infty(D, D_{d,1}^\infty)$ as a simplicial complex (Lemma 4.23), and that the lollipop complex (and hence $T_r^\infty(D, D_{d,1}^\infty)$) is contractible with a bad simplices argument (Proposition 4.27). We then use the contractibility of $T_r^\infty(D, D_{d,1}^\infty)$ and a bad simplices argument to prove the complex $U_r^\infty(D, D_{d,1}^\infty)$ is contractible and weakly Cohen–Macaulay of dimension $r-2$ (Proposition 4.28 and Corollary 4.29), implying ultimately that our initial complex is weakly Cohen–Macaulay of dimension $r-2$, as needed.

In Proposition 4.28, we will deduce the connectivity of $U_r^\infty(D, D_{d,1}^\infty)$ using the connectivity of the lollipop complex $L_r^\infty(D, D_{d,1}^\infty)$ by applying a bad simplices argument. Our goal now is to show that $L_r^\infty(D, D_{d,1}^\infty)$ is highly connected. Let us make some definitions first.

Definition 4.25 Given any lollipop $L\colon (A, 0) \to (D_{d,r}^\infty, 0)$, we define the *free height* $\mathfrak{h}_L$ to be the minimal number m such that $L([1, 2])$ is contained in $D_{d,r,m}$ up to free isotopy. We also define the height of an admissible loop to be the minimal number m such that it is contained in $D_{d,r,m}$ (see Definition 3.2).

To analyze the connectivity of $L_r^\infty(D, D_{d,1}^\infty)$, we need the following lemma, which is a direct translation of [29, Lemma 3.8].

Lemma 4.26 *For any $r, p, N \geq 1$, there exists a number $\mathfrak{h}_{r,p,N} \geq 0$, such that, for any p-simplex σ in $L_r^\infty(D, D_{d,1}^\infty)$ and any $\mathfrak{h} \geq \mathfrak{h}_{r,p,N}$, there are at least N lollipops of free height $\mathfrak{h}$ in $L_r^\infty(D, D_{d,1}^\infty)$ that are in $\mathrm{Lk}(\sigma)$.*

Proof Note that, for any vertex L in $L_r^{\infty}(D, D_{d,1}^{\infty})$, $L|_{[1,2]}$ is an admissible loop in $\bigoplus_r D_{d,1}^{\infty}$. Recall the function q defined in Definition 3.17, which maps an admissible loop to an edge midpoint in the tree $\mathscr{T}_{d,r}$. Since each edge has a unique descendent vertex, we can instead map the loop to this vertex which lies in the forest $\mathscr{F}_{d,r}$. Using this connection, we can now choose $\mathfrak{h}_{r,p,N}$ to be the same as in [29, Lemma 3.8]. Then we have at least N admissible loops of height $\mathfrak{h} \geq \mathfrak{h}_{r,p,N}$ which lie in the complement of the surface corresponding to σ in $\bigoplus_r D_{d,1}^{\infty}$. Connecting each of these admissible loops to the basepoint in the complement surface, we get a set of lollipops in $\mathrm{Lk}(\sigma)$. $\square$

We now show that the complex $L_r^{\infty}(D, D_{d,1}^{\infty})$ is in fact contractible. The idea of proof is similar to that of [29, Proposition 3.1] but with significantly more technical difficulty. Intuitively, to define their complex, one only needs information from the loop parts of the lollipops, which are much easier to "make" disjoint in general, but for us, we also have to deal with the arc parts, which could potentially cause more problems.

Proposition 4.27 *The complex $L_r^{\infty}(D, D_{d,1}^{\infty})$ is contractible for any $r \geq 1$.*

Proof The complex $L_r^{\infty}(D, D_{d,1}^{\infty})$ is obviously nonempty. We will show by induction that, for all $k \geq 0$, any map $S^k \to L_r^{\infty}(D, D_{d,1}^{\infty})$ is null-homotopic. Assume $L_r^{\infty}(D, D_{d,1}^{\infty})$ is $(k-1)$-connected.

Let $f\colon S^k \to L_r^{\infty}(D, D_{d,1}^{\infty})$ be a map. As usual, we can assume that the sphere S^k comes with a triangulation such that the map f is simplicial. We first use Lemma 1.7 to homotope f to a map that is simplexwise injective. For that we need that, for every p-simplex σ in $L_r^{\infty}(D, D_{d,1}^{\infty})$, its link $\mathrm{Lk}\,(\sigma)$ is $(k-p-2)$-connected. But, by Lemma 4.24, $\mathrm{Lk}(\sigma)$ can be identified with $L_{r_\sigma}^{\infty}(D, D_{d,1}^{\infty})$ for some $r_\sigma \geq 1$, so it is $(k-1)$-connected and the conditions of Lemma 1.7 are satisfied.

Now, since S^k is a finite simplicial complex, the free height of the vertices of S^k has a maximum value. We first want to homotope f to a new map such that all the vertices have free height at least $\mathfrak{h} = \mathfrak{h}_{r,k,N}$, where $N = v_0 + v_1 + \cdots + v_k + 2$, v_i is the number of i-simplices of S^k, and $\mathfrak{h}_{r,k,N}$ is determined by Lemma 4.26. For that we use a bad simplices argument.

We call a simplex of the sphere S^k bad if all of its vertices are mapped to vertices in $L_r^{\infty}(D, D_{d,1}^{\infty})$ that have free height less than $\mathfrak{h}$. We will modify f by removing the bad simplices inductively, starting with those of the highest dimension. Let σ be a bad simplex of maximal dimension p among all bad simplices. We will modify f and the triangulation of S^k in the star of σ in a way that does not add any new bad simplices. In the process, we will increase the number of vertices by at most 1 in each step, and not at all if σ is a vertex. This implies that, after doing this for all bad simplices, we will have increased the number of vertices of the triangulation of S^k by at most $v_1 + \cdots + v_k$. As S^k originally had v_0 vertices, at the end of the process its new triangulation will have at most $v = v_0 + v_1 + \cdots + v_k$ vertices. There are two cases.

Case 1 $(p = k)$ If the bad simplex σ is of the dimension k of the sphere S^k, then its image $f(\sigma)$ has a complement loop which bounds a surface C asymptotically rigidly homeomorphic to D^{∞}_{d,r_σ} for some $r_\sigma \geq 1$ by Lemma 4.24. Now we can choose a lollipop y in C with free height at least $\mathfrak{h} + 1$. In particular, $f(\sigma) \cup y$ forms a $(k{+}1)$-simplex. We can then add a vertex a in the center of σ, replacing σ by $\partial\sigma * a$ and replacing f by the map $(f|_{\partial\sigma}) * (a \mapsto y)$ on $\partial\sigma * a$. This map is homotopic to f through the simplex $f(\sigma) \cup \{y\}$. We have added a single vertex to the triangulation. Because L has free height $\mathfrak{h} + 1$, we have not added any new bad simplices, and we have removed one bad simplex, namely σ. Moreover, f remains simplexwise injective.

Case 2 $(p < k)$ If the bad simplex σ is a p-simplex for some $p < k$, by maximality of its dimension, the link of σ is mapped to vertices of free height at least $\mathfrak{h}$ in the complement of the subsurface $f(\sigma)$. The simplex σ has $p+1$ vertices, whose images are pairwise disjoint outside the basepoint up to based isotopy. By Lemma 4.26 and our choice of $\mathfrak{h}$, there are at least $N = v + 2$ lollipops $y_1, \dots, y_N$ of free height $\mathfrak{h}$ such that each $f(\sigma) \cup \{y_i\}$ forms a $(p{+}1)$-simplex. As there are fewer vertices in the link than in the whole sphere S^k, and S^k has at most v vertices, by the pigeonhole principle, the loop parts of the vertices in $f(\mathrm{Lk}(\sigma))$ are contained in at most v punctured disks bounded by the corresponding admissible loops with free height $\mathfrak{h}$. As $N = v + 2$, there are at least two of the above vertices y_i and y_j of free height $\mathfrak{h}$ such that any loop parts of vertices in $f(\mathrm{Lk}(\sigma))$ are disjoint from the loop parts of y_i and y_j. We can further assume that the arc parts of y_i and y_j never intersect with any loop part of a vertex in $f(\mathrm{Lk}(\sigma))$. And, up to replacing the loop parts of y_i and y_j with an admissible loop lying inside the disk bounded by the loop parts of y_i and y_j (note that this may increase the free height of y_i and y_j), we can further assume that the arc parts of vertices in $f(\mathrm{Lk}(\sigma))$ are disjoint from the loop parts of y_i and y_j. But, unlike the situation in the proof of [29, Proposition 3.1], a new problem we are facing here is that the arc parts of y_i or y_j might intersect the arc parts of the vertices in $f(\mathrm{Lk}(\sigma))$ even up to isotopy. In particular, given a simplex τ lying in the link of σ, $f(\sigma) \cup f(\tau) \cup y_i$ does not necessarily form a simplex now.

For that we want to apply the mutual link trick (see Lemma 1.8) to remove the intersections of $f(\mathrm{Lk}(\sigma))$ with y_i via a sequence of homotopies. In the process, we will only modify f on $\mathrm{Lk}(\sigma)$ and the new map will still map $\mathrm{Lk}(\sigma)$ to $\mathrm{Lk}_{L_r^{\infty}(D, D^{\infty}_{d,1})}(f(\sigma))$. Recall that f is simplexwise injective. Up to isotopy, we can further choose representatives for vertices in $f(\mathrm{Lk}(\sigma))$ such that the intersection points of vertices in $f(\mathrm{Lk}(\sigma))$ and y_i are isolated. Moreover, we assume the number of intersection points is minimal for each vertex in $f(\mathrm{Lk}(\sigma))$. Now we choose an intersection point x_0 in the arc $y_i([0,1])$ that is closest to $y_i(1)$, and denote the corresponding lollipop by β, which is the image of some vertex $b \in \mathrm{Lk}(\sigma)$. We can choose β' to be a variation of β: β' coincides with β for the most part, except around the intersection point with y_i, we replace it by an arc going around the loop part of y_i. See Figure 9. Now we apply Lemma 1.8, for which we need to check the following two conditions:

(1) $f(\mathrm{Lk}_{S^k}(b)) \leq \mathrm{Lk}_{L_r^{\infty}(D, D^{\infty}_{d,1})}(\beta')$. This follows from our definition of β'. If a vertex v in $f(\mathrm{Lk}_{S^k}(b))$ is disjoint from β, using the fact that the intersection point x_0 is the closest one to $y_i(1)$ and $f(v)$ is disjoint from the loop part of y_i, β' is also disjoint from v.

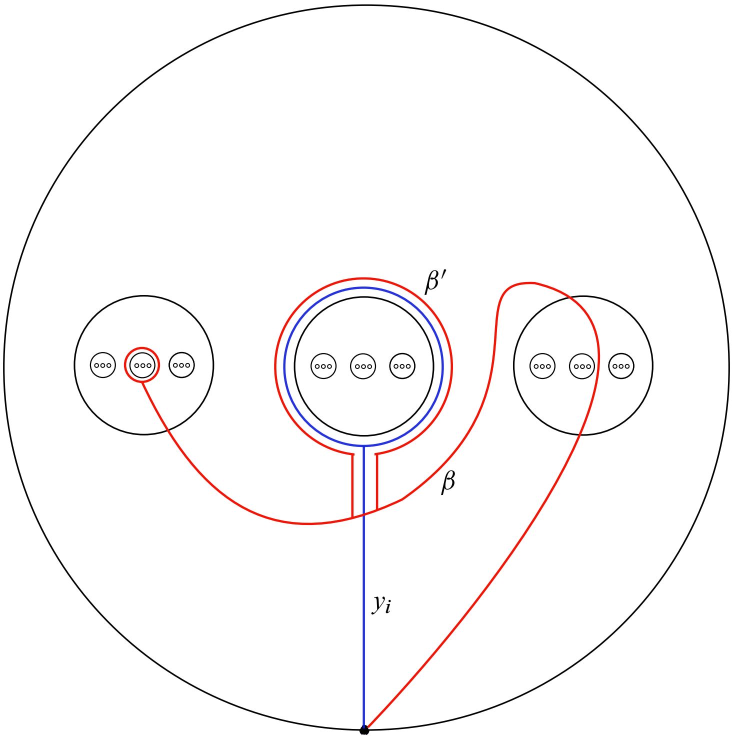


Figure 9: Replacing β by β' to reduce the number of intersection points with y_i.

(2) $\mathrm{Lk}(\beta)\cap\mathrm{Lk}(\beta')$ is $(k-1)$-connected. The lollipops β and β' together will bound a disk which contains the loop part of β and y_i. In any event, the complement of these is a surface asymptotically rigidly homeomorphic to some surface $D^{\infty}_{d,r'}$ for some $r'\ge 1$. By our induction, it is $(k-1)$-connected.

Now Lemma 1.8 says we can homotope f to a new map such that $f(b)=\beta'$ and $f(\mathrm{Lk}(\sigma))$ has fewer intersection points with y_i. Step by step, at the end we have a simplexwise injective map f such that any vertex in $f(\mathrm{Lk}(\sigma))$ only intersects with y_i at the basepoint. In particular, for any $\tau\in\mathrm{Lk}(\sigma)$, $f(\sigma)\cup f(\tau)\cup\{y_i\}$ forms a simplex in $L^{\infty}_{r}(D, D^{\infty}_{d,1})$.

We can then replace f inside the star

$$\mathrm{St}(\sigma)=\mathrm{Lk}(\sigma)*\sigma\simeq S^{k-p-1}*D^{p}$$

by the map $(f|_{\mathrm{Lk}(\sigma)})*(a\mapsto y_i)*(f|_{\partial\sigma})$ on

$$\mathrm{Lk}(\sigma)*a*\partial\sigma\simeq S^{k-p-1}*D^{0}*S^{p-1},$$

which agrees with f on the boundary $\mathrm{Lk}(\sigma)*\partial\sigma$ of the star and is homotopic to f through the map $(f|_{\mathrm{Lk}(\sigma)})*(a\mapsto y_i)*(f|_{\sigma})$ defined on

$$\mathrm{Lk}(\sigma)*a*\sigma\simeq S^{k-p-1}*D^{0}*D^{p}.$$

Now $\mathrm{Lk}(\sigma) * a * \partial(\sigma)$ has exactly one extra vertex a compared to the star of σ, unless σ is just a vertex, in which case $\partial(\sigma)$ is empty and it has the same number of vertices. As y_i has height at least $\mathfrak{h}$, we have not added any new bad simplices. Hence we have reduced the number of bad simplices by one by removing σ.

By induction, we can now assume that there are no bad simplices for f with respect to a triangulation with at most v vertices. With this assumption, we want to cone off f just as we coned off the links in the above argument. We have more than $N = v + 2$ vertices of free height $\mathfrak{h}$ in $L_r^\infty(D, D_{d,1}^\infty)$, and at most v vertices in the sphere. The loop parts of these vertices are admissible loops of height at least $\mathfrak{h}$. By the pigeonhole principle, we know that there are at least two lollipops z_i and z_j of free height $\mathfrak{h}$ such that the punctured disks bounded by their loop parts are disjoint from the punctured disk bounded by any loop part of the lollipops in the vertices of $f(S^k)$. Just as before, we can further assume that the arc parts of z_i and z_j never intersect with any loop part of the vertices in $f(S^k)$, and the arc parts of vertices in $f(S^k)$ are disjoint from the loop part of z_i and z_j. But the same problem appears again, as we want vertices of $f(S^k)$ to be disjoint from the whole lollipop z_i. For that we apply Lemma 1.8 again, and the same proof as before implies that we can homotope f so that its image is disjoint from z_i. In particular, $f(S^k)$ lies in the link of z_i. Hence we can homotope f to a constant map since $\mathrm{St}(z_i)$ is contractible. □

Proposition 4.28 *The complex $U_r^\infty(D, D_{d,1}^\infty)$ is contractible.*

Proof As $T_r^\infty(D, D_{d,1}^\infty)$ is a subcomplex of $U_r^\infty(D, D_{d,1}^\infty)$, we can use a bad simplices argument.

We call a vertex of $U_r^\infty(D, D_{d,1}^\infty)$ bad if it does not lie in $T_r^\infty(D, D_{d,1}^\infty)$ and a simplex bad if all of its vertices are bad. Given a bad p-simplex σ, we need to determine the connectivity of the good link GL_σ (see Section 1.2 for the definition of GL_σ). As in the proof of Lemma 4.24, we have a complement surface C_σ of σ in $D_{d,1}^\infty$. Note that C_σ inherits a d-rigid structure and it is asymptotically rigidly homeomorphic to $\bigoplus_{r_\sigma} D_{d,1}^\infty$ for some $r_\sigma > 0$. In particular, we can now identify GL_σ with $T_{r_\sigma}^\infty(D, D_{d,1}^\infty)$, which is contractible. Thus, by Proposition 1.5, the pair $(U_r^\infty(D, D_{d,1}^\infty), T_r^\infty(D, D_{d,1}^\infty))$ is i-connected for any $i \geq 0$. By Proposition 4.27, $T_r^\infty(D, D_{d,1}^\infty) \cong L_r^\infty(D, D_{d,1}^\infty)$ is contractible, so $U_r^\infty(D, D_{d,1}^\infty))$ is also contractible. □

Corollary 4.29 *The complex $U_r(D, D_{d,1}^\infty)$ is weakly Cohen–Macaulay of dimension $r-2$.*

Proof Note first that a simplicial complex is $(r-3)$-connected if and only if its $(r-2)$-skeleton is. Since $U_r(D, D_{d,1}^\infty)$ has the same $(r-2)$-skeleton as $U_r^\infty(D, D_{d,1}^\infty)$ and $U_r^\infty(D, D_{d,1}^\infty)$ is contractible, so in particular $(r-3)$-connected, indeed $U_r(D, D_{d,1}^\infty)$ is $(r-3)$-connected.

Now let σ be a p-simplex of $U_r(D, D_{d,1}^\infty)$, with vertices $\phi_0, \phi_1, \dots, \phi_p$. We need to check that the link $\mathrm{Lk}_{U_r(D, D_{d,1}^\infty)}(\sigma)$ is $(r-p-4)$-connected. We can assume $p \leq r-3$ as any space is (-2)-connected. Moreover, it suffices to show the $(r-p-3)$-skeleton of $\mathrm{Lk}_{U_r(D, D_{d,1}^\infty)}(\sigma)$ is $(r-p-4)$-connected. Since

$\phi_0, \phi_1, \dots, \phi_p$ form a p-simplex, similar to the proof of Lemma 4.24, the complement surface of σ is asymptotically rigidly homeomorphic to some d-rigid surface D^{∞}_{d,k_σ} for some $k_\sigma > 0$. Then we can identify the $(r-p-3)$-skeleton of $\mathrm{Lk}_{U_r(D,D^{\infty}_{d,1})}(\sigma)$ with the $(r-p-3)$-skeleton of $U^{\infty}_{k_\sigma}(D, D^{\infty}_{d,1})$. Since $U^{\infty}_{k_\sigma}(D, D^{\infty}_{d,1})$ is even contractible, we have the connectivity bound we need. □

Now, by Lemma 4.15 and Proposition 1.3, we have the following.

Corollary 4.30 *The complexes $S_r(D, D^{\infty}_{d,1})$ and $W_r(D, D^{\infty}_{d,1})_\bullet$ are weakly Cohen–Macaulay of dimension $r-2$.*

4.4 Homological stability

We are finally ready to prove the homological stability result.

Theorem 4.31 *Suppose $d \geq 2$. Then the inclusion maps induce isomorphisms*

$$\iota_{R^+,d,r} \colon H_i(RV^+_{d,r}, M) \to H_i(RV^+_{d,r+1}, M)$$

in homology in all dimensions $i \geq 0$ for all $r \geq 1$ and for all $H_1(RV^+_{d,\infty})$-modules M.

Proof From Corollary 4.30, $W_r(D, D^{\infty}_{d,1})$ is $(r-2)$-connected; hence, in particular, the category $\mathcal{G}_d$ satisfies (LH3) at the pair of objects $(D, D^{\infty}_{d,1})$ with slope $k = 3$. By Theorem 4.6, for any abelian RV^+_∞-module M, the map

$$H_i(RV^+_{d,r}; M) \to H_i(RV^+_{d,r+1}; M)$$

induced by the natural inclusion map is an isomorphism if $r \geq 3i + 3$.

But we can improve the stability range as in the proof of [29, Theorem 3.6] by noticing that we have the same canonical isomorphism between $RV^+_{d,r}$ and $RV^+_{d,r+d-1}$. In fact, denoting this isomorphism by $I_{d,r}$, we have the commutative diagram

$$\begin{tikzcd}
RV^+_{d,1+r-1} \arrow[r, "\iota_{R^+,d,r}"] \arrow[d, "I_{d,r}"] & RV^+_{d,1+(r-1)+1} \arrow[d, "I_{d,1+(r-1)+1}"] \\
RV^+_{d,d+r-1} \arrow[r, "\iota_{R^+,d,r+1}"] & RV^+_{d,d+r-1+1}
\end{tikzcd}$$

Given that the vertical maps are isomorphisms and the bottom horizontal maps induce isomorphisms on the i^{th} homology when $d + r - 1 \geq 3i + 3$, the top map must also induce isomorphism on the homology as long as $r \geq 3i + 3$. This has improved the stable range by $d - 1$. Step by step, the map $\iota_{R^+,d,r}$ must induce isomorphisms on homology in all dimensions $i \geq 0$ and for all $r \geq 1$. □

Theorem 4.32 *Suppose $d \geq 2$. Then the inclusion maps induce isomorphisms*

$$\iota_{R,d,r} \colon H_i(RV_{d,r}, M) \to H_i(RV_{d,r+1}, M)$$

in homology in all dimensions $i \geq 0$ for all $r \geq 1$ and for all $H_1(RV_{d,\infty})$-modules M.

Sketch of proof The proof will be exactly the same as that of Theorem 4.31. Note first that, by Theorem 3.24, this is the same as proving the half-twist asymptotic mapping class groups $\mathscr{H}V_{d,r}$ have homological stability. We define the braided monoidal category $\mathscr{G}'_d$ to be the category with objects $\bigoplus_r D^\infty_{d,1}$ for $r \geq 0$, $\oplus$ as the operation, and D as the 0 object. When $r = s$, we define the morphisms $\mathrm{Hom}\bigl(\bigoplus_r D^\infty_{d,1}, \bigoplus_s D^\infty_{d,1}\bigr) = \mathscr{H}V_{d,r}$, which can also be understood as the group of isotopy classes of asymptotically quasirigid homeomorphisms of $\bigoplus_r D^\infty_{d,1}$; when $r \neq s$, let $\mathrm{Hom}\bigl(\bigoplus_r D^\infty_{d,1}, \bigoplus_s D^\infty_{d,1}\bigr) = \varnothing$. We then have a homogeneous category $U\mathscr{G}'_d$, and, to prove homological stability for the sequence of groups $\mathscr{H}V_{d,1} \leq \mathscr{H}V_{d,2} \leq \cdots$, we only need to prove the associated space $W_r(D, D^\infty_{d,1})_\bullet$ — or, in fact, the associated simplicial complex $S_r(D, D^\infty_{d,1})$ — is highly connected. At this point, the complex is slightly different from the oriented case, but still the new complex $S_r(D, D^\infty_{d,1})$ is a complete join over the old complex $U_r(D, D^\infty_{d,1})$. Hence, the connectivity of $S_r(D, D^\infty_{d,1})$ again follows from Corollary 4.29 and Proposition 1.3. □

References

[1] **J Aramayona**, **K-U Bux**, **J Flechsig**, **N Petrosyan**, **X Wu**, *Asymptotic mapping class groups of Cantor manifolds and their finiteness properties*, preprint (2021) arXiv 2110.05318

[2] **J Aramayona**, **L Funar**, *Asymptotic mapping class groups of closed surfaces punctured along Cantor sets*, Mosc. Math. J. 21 (2021) 1–29 MR

[3] **J Aramayona**, **N G Vlamis**, *Big mapping class groups: an overview*, from "In the tradition of Thurston: geometry and topology" (K Ohshika, A Papadopoulos, editors), Springer (2020) 459–496 MR

[4] **J Aroca**, **M Cumplido**, *A new family of infinitely braided Thompson's groups*, J. Algebra 607 (2022) 5–34 MR

[5] **C Bleak**, **C Donoven**, **J Jonušas**, *Some isomorphism results for Thompson-like groups* $V_n(G)$, Israel J. Math. 222 (2017) 1–19 MR

[6] **C-F Bödigheimer**, **U Tillmann**, *Embeddings of braid groups into mapping class groups and their homology*, from "Configuration spaces" (A Bjorner, F Cohen, C De Concini, C Procesi, M Salvetti, editors), CRM Series 14, Ed. Norm., Pisa (2012) 173–191 MR

[7] **K S Brown**, *The geometry of finitely presented infinite simple groups*, from "Algorithms and classification in combinatorial group theory" (G Baumslag, C F Miller, editors), Math. Sci. Res. Inst. Publ. 23, Springer (1992) 121–136 MR

[8] **K S Brown**, *The homology of Richard Thompson's group F*, from "Topological and asymptotic aspects of group theory" (R Grigorchuk, M Mihalik, M Sapir, Z Šunić, editors), Contemp. Math. 394, Amer. Math. Soc., Providence, RI (2006) 47–59 MR

[9] **K S Brown**, **R Geoghegan**, *An infinite-dimensional torsion-free* FP_∞ *group*, Invent. Math. 77 (1984) 367–381 MR

[10] **K-U Bux**, **M G Fluch**, **M Marschler**, **S Witzel**, **M C B Zaremsky**, *The braided Thompson's groups are of type* F_∞, J. Reine Angew. Math. 718 (2016) 59–101 MR Correction in 778 (2021) 219–221

[11] **P Dehornoy**, *The group of parenthesized braids*, Adv. Math. 205 (2006) 354–409 MR

[12] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR

[13] **L Funar**, **C Kapoudjian**, *On a universal mapping class group of genus zero*, Geom. Funct. Anal. 14 (2004) 965–1012 MR

[14] **L Funar**, **C Kapoudjian**, *An infinite genus mapping class group and stable cohomology*, Comm. Math. Phys. 287 (2009) 784–804 MR

[15] **L Funar**, **Y Neretin**, *Diffeomorphism groups of tame Cantor sets and Thompson-like groups*, Compos. Math. 154 (2018) 1066–1110 MR

[16] **S Galatius**, **O Randal-Williams**, *Homological stability for moduli spaces of high dimensional manifolds, I*, J. Amer. Math. Soc. 31 (2018) 215–264 MR

[17] **A Genevois**, **A Lonjou**, **C Urech**, *Asymptotically rigid mapping class groups, I: Finiteness properties of braided Thompson's and Houghton's groups*, Geom. Topol. 26 (2022) 1385–1434 MR

[18] **E Ghys**, **V Sergiescu**, *Sur un groupe remarquable de difféomorphismes du cercle*, Comment. Math. Helv. 62 (1987) 185–239 MR

[19] **J L Harer**, *Stability of the homology of the mapping class groups of orientable surfaces*, Ann. of Math. 121 (1985) 215–249 MR

[20] **A Hatcher**, *On triangulations of surfaces*, Topology Appl. 40 (1991) 189–194 MR

[21] **A Hatcher**, **K Vogtmann**, *Tethers and homology stability for surfaces*, Algebr. Geom. Topol. 17 (2017) 1871–1916 MR

[22] **A Hatcher**, **N Wahl**, *Stabilization for mapping class groups of* 3*-manifolds*, Duke Math. J. 155 (2010) 205–269 MR

[23] **G Higman**, *Finitely presented infinite simple groups*, Notes on Pure Mathematics 8, Australian National University, Canberra (1974) MR

[24] **W van der Kallen**, *Homology stability for linear groups*, Invent. Math. 60 (1980) 269–295 MR

[25] **M Nakaoka**, *Homology of the infinite symmetric group*, Ann. of Math. 73 (1961) 229–257 MR

[26] **M Palmer**, **X Wu**, *On the homology of big mapping class groups*, J. Topol. 17 (2024) art. id. e12358 MR

[27] **O Randal-Williams**, **N Wahl**, *Homological stability for automorphism groups*, Adv. Math. 318 (2017) 534–626 MR

[28] **R Skipper**, **X Wu**, *Finiteness properties for relatives of braided Higman–Thompson groups*, Groups Geom. Dyn. 17 (2023) 1357–1391 MR

[29] **M Szymik**, **N Wahl**, *The homology of the Higman–Thompson groups*, Invent. Math. 216 (2019) 445–518 MR

[30] **W Thumann**, *Operad groups and their finiteness properties*, Adv. Math. 307 (2017) 417–487 MR

Department of Mathematics, University of Utah
Salt Lake City, UT, United States
Shanghai Center for Mathematical Sciences, Fudan University
Shanghai, China

`rachel.skipper@utah.edu`, `xiaoleiwu@fudan.edu.cn`

Received: 31 December 2022 Revised: 23 October 2023

Algebraic & Geometric Topology **25**:8 (2025) 4667–4718
DOI: 10.2140/agt.2025.25.4667
Published: 20 November 2025

A group-theoretic framework for low-dimensional topology
Or: how not to study low-dimensional topology?

SARAH BLACKWELL
ROBION KIRBY
MICHAEL KLUG
VINCENT LONGO
BENJAMIN RUPPIK

A correspondence, by way of Heegaard splittings, between closed, oriented 3-manifolds and pairs of surjections from a surface group to a free group has been studied by Stallings, Jaco and Hempel. This correspondence, by way of trisections, was recently extended by Abrams, Gay and Kirby to the case of smooth, closed, connected, oriented 4-manifolds. We unify these perspectives and generalize this correspondence to the case of links in closed, oriented 3-manifolds and links of knotted surfaces in smooth, closed, connected, oriented 4-manifolds. The algebraic manifestations of these four subfields of low-dimensional topology (3-manifolds, 4-manifolds, knot theory and knotted surface theory) are all strikingly similar, and this correspondence perhaps elucidates some unique character of low-dimensional topology.

57K40; 20F05, 57M05

1 Introduction

All manifolds and submanifolds discussed in this paper are smooth and, with the exception of surfaces in 4-manifolds, oriented. We use decompositions of manifolds in dimensions three and four, possibly together with links, to give a group-theoretic framework for studying these spaces. We begin by reviewing the simplest case of closed 3-dimensional manifolds, where this work has already been carried out by Stallings and Jaco [40; 20].

A *Heegaard decomposition*, or *Heegaard splitting*, of a closed 3-manifold M^3 is a pair of handlebodies H_1 and H_2 embedded inside of M with boundaries a common genus g surface Σ_g such that $M = H_1 \cup_{\Sigma_g} H_2$. Every such 3-manifold admits a Heegaard decomposition (for example by triangulating M and taking a regular neighborhood of the 1-skeleton). By choosing a basepoint on Σ_g, we then obtain the following pushout diagram between fundamental groups, where the maps are induced by inclusion:

$$\begin{tikzcd}
\pi_1(\Sigma_g, *) \arrow[r, two heads] \arrow[d, two heads] & \pi_1(H_1, *) \arrow[d, two heads] \\
\pi_1(H_2, *) \arrow[r, two heads] & \pi_1(M, *)
\end{tikzcd}$$

Note that $\pi_1(H_1, *)$ and $\pi_1(H_2, *)$ are both free groups of rank g, and the maps are surjections. Jaco proved that, given a surjective homomorphism $\phi\colon \pi_1(\Sigma_g, *) \twoheadrightarrow F_g$, there is a unique handlebody $H(\phi)$ with $\partial H(\phi) = \Sigma_g$ such that the map induced on π_1 by inclusion of Σ_g as the boundary agrees with ϕ (see Jaco [19], Lemma 2.3, and also Leininger and Reid [29, Lemma 2.2] for a simpler proof in this case). From this, it follows that a pair of surjective homomorphisms (ϕ_1, ϕ_2) with $\phi_i\colon \pi_1(\Sigma_g, *) \twoheadrightarrow F_g$ determines a 3-manifold $H(\phi_1) \cup_{\Sigma_g} H(\phi_2)$, and that every closed 3-manifold arises in this way. Jaco referred to these pairs of maps as *splitting homomorphisms*.

One concrete application of this is the following group-theoretic recasting of the 3-dimensional Poincaré conjecture. Writing $\pi_1(\Sigma_g, *) = \langle a_1, b_1, \ldots, a_g, b_g \mid [a_1, b_1] \cdots [a_g, b_g] = 1 \rangle$, there is a surjective homomorphism

$$\pi_1(\Sigma_g, *) \twoheadrightarrow \langle x_1, \ldots x_g \rangle \times \langle y_1, \ldots y_g \rangle, \qquad a_i \mapsto (x_i, 1), \quad b_j \mapsto (1, y_j).$$

The Poincaré conjecture is equivalent to the statement that this is the unique surjective homomorphism of these groups modulo precomposing with automorphisms and postcomposing with products of automorphisms (see Hempel [16]). Thus, by Perelman's work [35], this result follows, and we are left in the state where the only known proof of this perhaps innocent-looking group-theoretic result involves a careful analysis of Ricci flow.

In addition to the observation that every closed 3-manifold admits a Heegaard decomposition, there is a corresponding uniqueness theorem called the *Reidemeister–Singer theorem*, which states that any two Heegaard decompositions of a fixed 3-manifold differ by a sequence of simple inverse geometric operations called stabilization and destabilization [36; 39]. Jaco proposed a way of incorporating the Reidemeister–Singer theorem into the construction of 3-manifolds from appropriate pairs (ϕ_1, ϕ_2) to obtain a bijective correspondence [20].

More recently, a 4-dimensional analogue of Heegaard splittings, called *trisections*, together with a corresponding uniqueness theorem has been introduced by Gay and Kirby [8]. A trisection of a closed 4-manifold X^4 is a decomposition $X = X_1 \cup X_2 \cup X_3$ into 4-dimensional 1-handlebodies X_i, which pairwise intersect in genus g handlebodies H_g, and with triple intersection a genus g surface Σ_g. Every smooth, closed, connected, oriented 4-manifold admits a trisection, which is unique up to a stabilization operation [8]. (See Section 4A for a further review of trisections.) The inclusion maps between the various components of a trisection of a 4-manifold induce maps between their fundamental groups, which produces the following commutative diagram, where every face is a pushout and every homomorphism is surjective, and in which the basepoint is chosen to lie on Σ_g:

$$\begin{tikzcd}
& \pi_1(H_g,*) \arrow[rr, two heads] \arrow[dd, two heads] & & \pi_1(X_1,*) \arrow[dd, two heads] \\
\pi_1(\Sigma_g,*) \arrow[ur, two heads] \arrow[rr, two heads, crossing over] \arrow[dd, two heads] & & \pi_1(H_g,*) \arrow[ur, two heads] & \\
& \pi_1(X_3,*) \arrow[rr, two heads] & & \pi_1(X^4,*) \\
\pi_1(H_g,*) \arrow[ur, two heads] \arrow[rr, two heads] & & \pi_1(X_2,*) \arrow[ur, two heads] \arrow[from=uu, two heads, crossing over] &
\end{tikzcd}$$

Abrams, Gay and Kirby [1] noticed that the analogue of being able to recover a 3-manifold from a pair of surjective homomorphisms (ϕ_1, ϕ_2) holds in dimension four via trisections. Namely, given three surjective homomorphisms (ϕ_1, ϕ_2, ϕ_3) with $\phi_i\colon \pi_1(\Sigma_g,*) \twoheadrightarrow F_g$ such that the pairwise pushout of any pair ϕ_i and ϕ_j is a free group F_k, since $\#^k(S^1 \times S^2)$ is the unique closed, orientable 3-manifold with fundamental group F_k (by Perelman's work [35]) we obtain a closed 4-manifold by realizing three handlebodies $H(\phi_i)$, gluing them along their common boundary Σ_g, and filling in their pairwise unions, which are diffeomorphic to $\#^k(S^1 \times S^2)$, with three 4-dimensional 1-handlebodies (uniquely by Laudenbach and Poénaru [28]). They called this triple of maps (which then determine the entire cube pictured above) a *group trisection*, where the object being trisected is the group resulting from pushing out the three maps into a cube (in this case $\pi_1(X^4,*)$).

Additionally, Abrams, Gay and Kirby [1] use the uniqueness theorem for trisections to obtain results analogous to those previously mentioned in dimension three. Namely, they obtain a group-theoretic statement that is equivalent to the smooth 4-dimensional Poincaré conjecture and, by modding out the set of such triples (ϕ_1, ϕ_2, ϕ_3), they obtain a bijection between a group-theoretically defined set and the set of all smooth, closed, connected, oriented 4-manifolds.

Not only can every 3-manifold be split into a union of two handlebodies, but additionally, given a link $L \subset M$, we have a Heegaard splitting $M = H_1 \cup_{\Sigma_g} H_2$ such that the tangles $T_1 = L \cap H_1$ and $T_2 = L \cap H_2$ are trivial (that is, consist of arcs that can all be simultaneously isotoped in H_i into Σ_g). This is called a *bridge splitting* of $L \subset M$. Note that the complement of L in each handlebody is again a handlebody and hence has free fundamental group. In the case of $M = S^3$ with the Heegaard splitting into balls, this is the classical setting of *bridge position* of links (see [38]).

One dimension up, a similar story emerges. A *knotted surface* is a closed (potentially nonorientable or disconnected) surface smoothly embedded in a 4-manifold. Meier and Zupan showed that a knotted surface in a trisected 4-manifold can always be isotoped to be in *bridge position*, meaning that it

$$\{\phi_1,\phi_2\colon \pi_1(\Sigma_g,*) \twoheadrightarrow F_g\}/{\sim} \xrightarrow[\texttt{Heegaard splittings}]{\cong} \{\texttt{3-manifolds}\}/\texttt{diff.}$$

$$\{\phi_1,\phi_2\colon \pi_1(\Sigma_g-\{2b \text{ pts}\},*) \twoheadrightarrow F_{g+b}\}/{\sim} \xrightarrow[\substack{\texttt{bridge}\\ \texttt{splittings}}]{\cong} \{\texttt{links in 3-manifolds}\}/\texttt{diff.}$$

$$\{\phi_1,\phi_2,\phi_3\colon \pi_1(\Sigma_g,*) \twoheadrightarrow F_g\}/{\sim} \xrightarrow[\texttt{trisections}]{\cong} \{\texttt{4-manifolds}\}/\texttt{diff.}$$

$$\{\phi_1,\phi_2,\phi_3\colon \pi_1(\Sigma_g-\{2b \text{ pts}\},*) \twoheadrightarrow F_{g+b}\}/{\sim} \xrightarrow[\substack{\texttt{bridge}\\ \texttt{trisections}}]{\cong} \{\texttt{surfaces in 4-manifolds}\}/\texttt{diff.}$$

Figure 1: A summary of the main results of this paper; the equivalences via bridge splittings and bridge trisections are novel, while the other equivalences are a recasting and unification of previous work. All manifolds are smooth and, with the exception of surfaces in 4-manifolds, oriented, and the diffeomorphisms are orientation-preserving. All maps ϕ_i shown are surjective homomorphisms, and satisfy two algebraic conditions (see Definition 2.1). The homomorphisms ϕ_i in the third and fourth rows need to satisfy the additional condition that they push out pairwise to free groups of an appropriate rank (see Sections 4A and 4B).

intersects the trisected 4-manifold in such a way that the surface inherits its own trisection, called a *bridge trisection* [31; 32]. This is unique up to a stabilization operation [31; 18]. (See Section 4B for a further review of bridge trisections.) Given the existence and uniqueness of such a decomposition in this setting, it is natural to wonder whether knotted surfaces in 4-manifolds can also be given such a group-theoretic framework. Achieving this goal was the initial motivation for this work.

The main results of this paper are bijective correspondences from group-theoretic sets to the set of 3-manifolds together with a link and 4-manifolds together with a knotted surface, and are summarized in Figure 1. Just as the cases of 3-manifolds and 4-manifolds are facilitated by Heegaard splittings and trisections, respectively, our results for links in 3-manifolds and surfaces in 4-manifolds use bridge splittings and bridge trisections, respectively.

In order to get off the ground constructing these spaces from appropriate group homomorphisms, we need to know how to recover a trivial tangle $T(\phi)$ with boundary points $\{p_1,\dots,p_{2b}\}$ in a handlebody $H(\phi)$ with boundary Σ_g from a suitable homomorphism

$$\phi\colon \pi_1(\Sigma_g-\{p_1,\dots,p_{2b}\},*) \twoheadrightarrow F_{g+b},$$

where F_{g+b} is playing the role of the fundamental group of the complement of the trivial tangle. (For the precise algebraic conditions on ϕ needed for this construction, we defer to Definition 2.1.) Section 2 is dedicated to this task. Our first main result (see Theorem 2.10), and the result underlying all of the constructions of spaces from group homomorphisms, is a method for constructing $H(\phi)$ and $T(\phi)$ algorithmically. This method is inspired by the procedure for computing the corresponding homomorphism given the topological data of the surface together with curves indicating the handlebody and the trivial tangle (see Lemma 2.4). When naively trying to construct diagrams for $H(\phi)$ and $T(\phi)$, we run into the possibility of constructing diagrams with too many curves. We fix this using bands to connect curves together, where the combinatorics of how the bands connect curves is guided by a process called Stallings folding [41], whose behavior is guaranteed to serve our purposes by the conditions placed on ϕ (see the proof of Theorem 2.10).

With this construction in hand, the constructions of the maps in Figure 1 are straightforward and surjectivity follows from existence of the various geometric decompositions. In Sections 3 and 4 we discuss in detail the various geometric descriptions, the map in Figure 1, and the various algebraically defined relations that need to be collectively modded out by on the set of homomorphisms in order to obtain a bijection. This latter part involves setting up appropriate relations on the set of homomorphisms that mimic the geometric moves needed in the corresponding uniqueness theorem.

For example, in the case of closed 3-manifolds, by way of Heegaard splittings, all such 3-manifolds can be described by a *Heegaard diagram*, and two Heegaard diagrams result in the same 3-manifold if and only if they are related by a sequence of handleslides, diffeomorphisms of the surface applied to the diagram, and stabilizations (this is a diagrammatic restating of the Reidemeister–Singer theorem; see Theorem 3.1). In this case, we need to mod out our set of homomorphisms by an equivalence relation generated by three relations $\sim_h$, $\sim_m$ and $\sim_s$ (h for handleslide, m for mapping class and s for stabilization) that algebraically mimic the corresponding diagrammatic moves. In Section 3A, we carry out this process for closed 3-manifolds and in Sections 3B, 4A and 4B we do the analogous procedure for links in 3-manifolds, closed 4-manifolds and surfaces in 4-manifolds, respectively. In the "relative" cases of Sections 3B and 4B, there are additional relations needed, corresponding to the different types of stabilizations available in these settings.

It is unclear if this formalism will prove useful in deriving topological results (see our subtitle). However, in Section 5 we give some additional examples and pose some questions regarding potential applications. One curious consequence of our work is that although smoothly knotted surfaces in the 4-sphere cannot be distinguished by fundamental groups (or even their complements), they *can* be distinguished by group trisections (see Corollary 4.9).

Acknowledgments

We were inspired by Stallings' paper *How not to prove the Poincaré conjecture* [40] in more ways than one.

This project started at the 2020 virtual *Summer trisectors workshop*, which was financially supported by the NSF. We are grateful for this wonderful opportunity to exchange ideas, and would like to thank all of the participants for their input. We want to give special credit to Michelle Chu, David Gay, Gabriel Islambouli, Jason Joseph, Chris Leininger, Jeffrey Meier, Puttipong Pongtanapaisan, Arunima Ray and Alexander Zupan. We also would like to sincerely thank the referees for their careful reading of our paper and their insightful comments. Blackwell, Klug and Ruppik would additionally like to thank the Max Planck Institute for Mathematics in Bonn for its hospitality and financial support.

2 Trivial tangles in handlebodies from algebra

Let Σ_g denote the genus g oriented surface with basepoint $*$ and marked points $p_1, \dots, p_{2b}$ as in Figure 2. We abuse notation and let $p_1, \dots, p_{2b}$ also denote the fundamental group elements as pictured. Using the notation $[a, b] = aba^{-1}b^{-1}$ for the commutator of a and b, we have

$$\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) = \langle p_1, \dots, p_{2b}, a_1, b_1, \dots, a_g, b_g \mid p_1 \cdots p_{2b} = [a_1, b_1] \cdots [a_g, b_g] \rangle,$$

where the a_i, b_i are generators of $\pi_1(\Sigma_g, *)$. If $b = 0$, by convention this is the fundamental group of a closed genus g surface,

$$\pi_1(\Sigma_g, *) = \langle a_1, b_1, \dots, a_g, b_g \mid [a_1, b_1] \cdots [a_g, b_g] = 1 \rangle,$$

while for $g = 0$ this is the group of a $2b$-times punctured sphere,

$$\pi_1(S^2 - \{p_1, \dots, p_{2b}\}, *) = \langle p_1, \dots, p_{2b} \mid p_1 \cdots p_{2b} = 1 \rangle.$$

Observe that for $b \geq 1$ this is a free group on $2g + 2b - 1$ generators, although it will be useful for us to instead remember the relation $p_1 \cdots p_{2b} = [a_1, b_1] \cdots [a_g, b_g]$, which we call the *surface relation*.

We take Figure 2 to be our standard model for the genus g oriented surface Σ_g with basepoint $*$ and marked points $p_1, \dots, p_{2b}$. Proponents of the right-hand rule may be disappointed in our model, as each a_i and b_i pair violate this convention, but our choice of surface relation mandates the labels and orientations of the a_i and b_i curves. We choose to write the relation in this form for notational convenience on the algebraic side.

A genus g *handlebody* H is a compact orientable 3-manifold whose boundary is a genus g closed surface, with the property that H can be cut along 2-dimensional disks so that the resulting space is a set of 3-dimensional balls. A b-component *trivial tangle* T in a handlebody H is a collection of b properly embedded arcs in H such that all of the arcs can be simultaneously isotoped into the boundary of H. Given two handlebodies containing trivial tangles (H_1, T_1) and (H_2, T_2) with the property that $\partial H_1 = \partial H_2$ and $\partial T_1 = \partial T_2$, we say that (H_1, T_1) and (H_2, T_2) are *equivalent* if there exists a diffeomorphism $H_1 \to H_2$ mapping T_1 to T_2 that is the identity on ∂H_1. In the special case where T_1 and T_2 are empty, we then say that the handlebodies H_1 and H_2 are equivalent.

We will be concerned with the set of equivalence classes of handlebodies and trivial tangles (H, T) such that $\partial H = \Sigma_g$ and $\partial T = \{p_1, \dots, p_{2b}\}$. Any equivalence class of such a handlebody and trivial tangle can be described by a *diagram* $\mathfrak{D} = (C, S)$ on the surface Σ_g, made up of a collection of g disjoint homologically linearly independent simple closed curves $C = \{C_1, \dots, C_g\}$ (referred to as a *cut system*) together with b pairwise disjoint embedded arcs $S = \{S_1, \dots, S_b\}$ whose endpoints are $\{p_1, \dots, p_{2b}\}$ (referred to as a *shadow diagram*). In [32], this collection of cut system curves together with the shadow arcs is referred to as a “curve-and-arc system”.

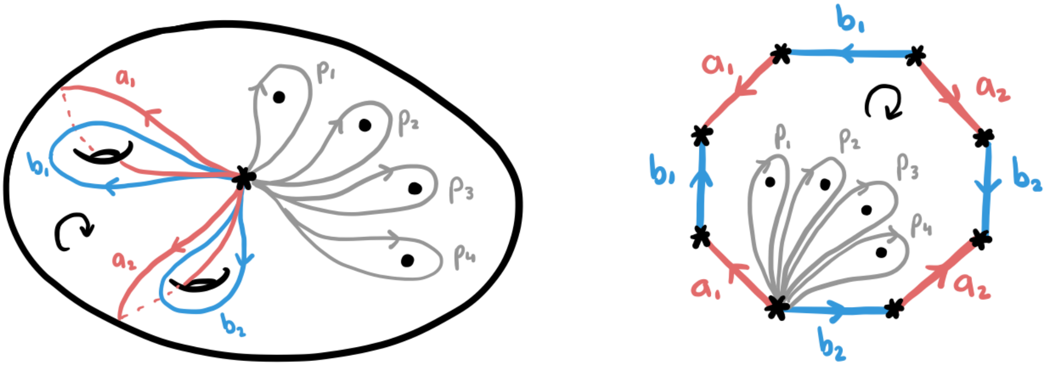


Figure 2: The genus g oriented surface Σ_g with basepoint $*$ and marked points $p_1, \dots, p_{2b}$, represented two ways. Gluing the edges of the polygon on the right gives the surface on the left. The fundamental group $\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$ is generated by a_i, b_i and (abusing notation) p_i.

The handlebody H can be constructed from the cut system by taking $\Sigma_g \times [0,1]$, attaching g 3-dimensional 2-handles to $\Sigma_g \times \{1\}$ along all of the curves $C_1, \dots, C_g$, and attaching a 3-ball to the resulting 2-sphere boundary component. Given a shadow diagram in addition to the cut system, a trivial tangle in the resulting handlebody can be constructed by taking the arcs of the tangle to be the union of $\{p_1, \dots, p_{2b}\} \times \left[0, \frac{1}{2}\right]$ together with the arcs $S_1, \dots, S_b$ in the shadow diagram considered as arcs in $\Sigma_g \times \left\{\frac{1}{2}\right\}$. Conversely, given a handlebody H and a trivial tangle T we can obtain a diagram $\mathfrak{D} = (C, S)$ for (H, T) by choosing a set of g disjoint embedded disks in H that cut H into a 3-ball and letting the cut system C be the boundary of these disks, and taking an isotopy relative to the boundary of T into Σ_g and letting the shadow diagram S denote the end result of this isotopy.

We now give a name to these disks, as well as a few other disks referred to in some of the following proofs. Given a handlebody H and a trivial tangle T, we refer to disjoint properly embedded disks bounded in H by its cut system curves as *cut disks*, and disjoint properly embedded disks that are the endpoint union of a shadow arc and the associated tangle component as *bridge disks*. One choice for these bridge disks is the track $S_i \times \left[0, \frac{1}{2}\right]$. A *bubble disk* in $H - T$ is a properly embedded disk which encloses a bridge disk of the tangle strand; see Figure 3.

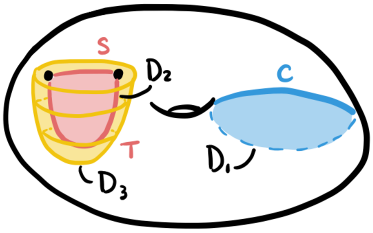


Figure 3: A cut disk D_1 bounded by a cut system curve C, a bridge disk D_2 bounded by a shadow arc S and tangle strand T, and a bubble disk D_3.

2A Bounding homomorphisms

The following definition is motivated by Lemma 2.4, and examples are given in Examples 2.5 and 2.6.

Definition 2.1 (bounding homomorphism) A *bounding homomorphism* is an epimorphism from a (possibly) punctured surface group to a free group

$$\phi\colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow \langle t_1, \dots, t_b, h_1, \dots, h_g\rangle$$

with the following properties:

(1) The image of the subgroup generated by $a_1, b_1, \dots, a_g, b_g$ surjects onto the quotient obtained by setting the $t_i = 1$.

(2) Each p_i maps to a conjugate of one of the t_j, where each of t_j and its inverse t_j^{-1} appears exactly once as the central letter. More precisely, there exists a bijection

$$f\colon \{p_1, \dots, p_{2b}\} \to \{t_1, t_1^{-1}, \dots, t_b, t_b^{-1}\}$$

and there are group elements $g_i \in \langle t_1, \dots, t_b, h_1, \dots, h_g\rangle$ with $\phi(p_i) = g_i f(p_i) g_i^{-1}$.

There are two special cases worth mentioning. If $b = 0$, this is an epimorphism from a closed surface group to a free group. Topologically, this will correspond to having no tangle strands. If $g = 0$, this corresponds to a trivial tangle in the 3-ball. The necessity of properties (1) and (2) will be seen in the proof of Theorem 2.10, but, roughly speaking, property (1) will allow us to distinguish between the handlebody and tangle, and property (2) is a natural condition coming from the proof of Lemma 2.4.

Definition 2.2 (topological realization) A *topological realization* of a bounding homomorphism $\phi\colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow \langle t_1, \dots, t_b, h_1, \dots, h_g\rangle$ is a trivial tangle T in a handlebody H with $\partial H = \Sigma_g$ and $\partial T = \{p_1, \dots, p_{2b}\}$ such that there is an isomorphism $\psi\colon \pi_1(H - T, *) \xrightarrow{\cong} \langle t_1, \dots, t_b, h_1, \dots, h_g\rangle$ that makes the diagram

$$\begin{tikzcd}
 & \pi_1(H - T, *) \arrow[d, "\psi", "\cong"'] \\
\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \arrow[ur, two heads, "\iota_*"] \arrow[dr, two heads, "\phi"'] & \\
 & \langle t_1, \dots, t_b, h_1, \dots, h_g\rangle
\end{tikzcd}$$

commute, where the map $\iota_*\colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow \pi_1(H - T, *)$ is induced by inclusion.

Lemma 2.3 (uniqueness of realization) *Let (H_1, T_1) and (H_2, T_2) be two trivial tangles in handlebodies with $\partial H_1 = \partial H_2 = \Sigma_g$ and $\partial T_1 = \partial T_2 = \{p_1, \dots, p_{2b}\}$ such that there is an isomorphism ρ between the fundamental groups of the tangle complements which makes the diagram*

$$\begin{tikzcd}
 & \pi_1(H_1 - T_1, *) \arrow[d, "\rho", "\cong"'] \\
\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \arrow[ur, two heads, "\iota_{1*}"] \arrow[dr, two heads, "\iota_{2*}"] & \\
 & \pi_1(H_2 - T_2, *)
\end{tikzcd}$$

commute, where the maps $\iota_{i*}\colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow \pi_1(H_i - T_i, *)$ *for* $i = 1, 2$ *are induced by inclusion. Then* (H_1, T_1) *and* (H_2, T_2) *are equivalent.*

Proof We will construct a diffeomorphism $H_1 \to H_2$ mapping T_1 to T_2 that extends the identity on the boundary ∂H_1, by defining it first on 2-cells in the complement of T_1 in H_1, and then extending over 3-balls.

Fix a cut system for the handlebody H_1 and shadow arcs for the tangle T_1, and choose whiskers connecting each curve and arc to the basepoint. Let λ be a based cut system curve which thus bounds a cut disk in H_1. From commutativity of the diagram, we know that λ is homotopically trivial in the tangle complement $H_2 - T_2$, and so by Dehn's lemma it bounds an embedded disk in $H_2 - T_2$. Extend the identity on the boundary to the cut disk bounded by λ in $H_1 - T_1$ by mapping it to the disk obtained by Dehn's lemma in $H_2 - T_2$. In the same manner, extend the map to a complete system of cut disks for the handlebody H_1.

Let η be the based boundary of a closed tubular neighborhood of one of the shadow arcs of T_1. This curve bounds a bubble disk in $H_1 - T_1$; recall Figure 3. Again from commutativity of the diagram, η is null-homotopic in $H_2 - T_2$, and, by another application of Dehn's lemma, bounds a disk. Use these disks to extend the map over all of the bubble disks in $H_1 - T_1$.

To finish the construction, use the Alexander trick to extend the map over the 3-cells in H_1. Observe that each of the bubble disks cuts off a single tangle strand on one of its sides. Combined with the observation that there is a unique trivial 1-strand tangle in the 3-ball, this shows that the diffeomorphism $H_1 \to H_2$ we constructed can be arranged to map T_1 to T_2. □

Now we will set up some notation in preparation for the following lemma. Let $\mathfrak{D} = (C, S)$ be a diagram for a handlebody and trivial tangle (H, T) with $\partial H = \Sigma_g$ and $\partial T = \{p_1, \dots, p_{2b}\}$, together with an ordering of the curves in the cut system $C_1, \dots, C_g$, an ordering of the arcs in the shadow diagram $S_1, \dots, S_b$, and a choice of an orientation for each of the C_i and S_j. Observe that cutting the surface Σ_g along the cut system curves and shadow arcs creates a connected, planar surface, and thus we will be able to choose dual loops h_i and t_j as follows.

For each curve C_i, pick a closed loop h_i based at $*$ that does not intersect the arcs S_k for any k or the curves C_k for $k \neq i$, and that intersects the curve C_i in a single point. Orient h_i so that, at the point of intersection of h_i and C_i, the orientation of h_i followed by the orientation of C_i agrees with the ambient orientation of Σ_g (which we have assumed to be clockwise; see Figure 2).

Similarly, for each arc S_j, choose a loop t_j based at $*$ that intersects S_j in exactly one point and does not intersect any of the other arcs or curves. Orient t_j so that, at the point of intersection of t_j and S_j, the orientation of t_j followed by the orientation of S_j agrees with the ambient orientation of Σ_g. Therefore, we now have $h_i, t_j \in \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$. See Figure 4.

Lemma 2.4 (diagrams to maps) *Using the notation defined above, we have the following*:

(1) *The loops representing* h_i *and* t_j *are well defined in* $\pi_1(H - T, *)$*, independent of choices.*

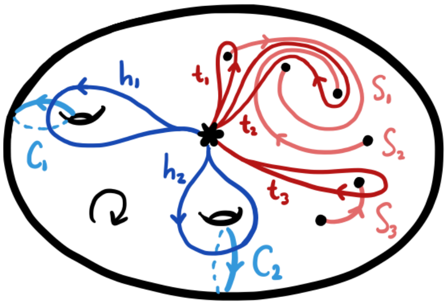


Figure 4: Finding loops h_i dual to the cut system curves C_i, and loops t_j dual to the shadow arcs S_j.

(2) *The map*

$$\psi_{\mathfrak{D}} \colon \pi_1(H-T,*) \to \langle t_1,\dots,t_b,h_1,\dots,h_g\rangle, \qquad h_i \mapsto h_i, \quad t_j \mapsto t_j,$$

is an isomorphism.

(3) *The composition of the map induced by inclusion and* $\psi_{\mathfrak{D}}$,

$$\phi \colon \pi_1(\Sigma_g - \{p_1,\dots,p_{2b}\},*) \xrightarrow{\iota_*} \pi_1(H-T,*) \xrightarrow{\psi_{\mathfrak{D}}} \langle t_1,\dots,t_b,h_1,\dots,h_g\rangle,$$

on an element $\lambda \in \pi_1(\Sigma_g - \{p_1,\dots,p_{2b}\},*)$ *is computed as follows. Represent* λ *by a based, closed, immersed curve on* Σ_g *that is transverse to the cut system curves* C_i *and the shadow arcs* S_i. *We call this curve* λ *again. The image of* λ *under this composition of maps is given by traversing* λ *and building a word in the elements* $t_1,\dots,t_b,h_1,\dots,h_g$ *and their inverses as follows. Start with the empty word. For each intersection between* C_i *and* λ *we concatenate* $h_i^{\pm 1}$ *on the right, and for each intersection between* S_j *and* λ *we concatenate* $t_j^{\pm 1}$ *on the right, where the sign is determined by the sign of the intersection of the oriented curves as in Figure 5.*

(4) *The map* ϕ *in* (3) *is a bounding homomorphism, and, with the choice of isomorphism* $\psi_{\mathfrak{D}}$, *the trivial tangle* (H,T) *is a topological realization of* ϕ.

Proof (1) Observe that the choices of these elements h_i and t_j are not unique when considered as elements in $\pi_1(\Sigma_g,*)$; see Figure 6. However we now prove that they are unique in the group $\pi_1(H-T,*)$. By choosing a collection of disjoint cut disks and bridge disks, and cutting $H-T$ along these disks, we obtain a 3-ball as in Figure 7. From this it follows that the choices of h_i and t_j are unique as elements of $\pi_1(H-T,*)$, because there is a unique homotopy class of curves connecting points in a 3-ball.

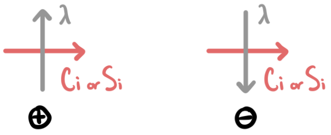


Figure 5: Sign convention for intersection points, where the pink arrow is a cut system curve C_i or shadow arc S_i, and the gray arrow is an element in the fundamental group of the punctured surface, represented by a based, closed immersed curve λ on the surface.

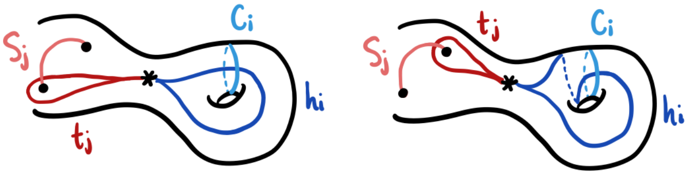


Figure 6: Inequivalent choices of h_i and t_j in the surface minus the points, which become isotopic in the handlebody minus the tangle determined by the C_i and S_j.

(2) Now define the map

$$\langle t_1, \dots, t_b, h_1, \dots, h_g \rangle \to \pi_1(H - T, *)$$

by sending $t_j \mapsto t_j$ and $h_i \mapsto h_i$. We now argue that this map is an isomorphism. First observe that $\pi_1(H - T, *)$ is a free group of rank $g + b$, because $H - T$ deformation retracts onto a spine obtained in the following way. As above, cutting along (a choice of) cut disks and bridge disks results in a 3-ball, and thus the h_i and t_j make up a spine for the tangle complement $H - T$. This also means that the homomorphism above is surjective, and, since free groups are Hopfian, it must be an isomorphism.

(3) Using the spine from the proof of part (2), we apply the cut system–spine duality from [22] to see that the image of an element $\lambda \in \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$ under these maps can be computed by recording intersections of λ with the cut disks and bridge disks, since traversing an edge in the spine corresponds to hitting its dual disk. For curves λ that live on the surface Σ_g, these intersections occur on the boundaries of the disks. See Figure 7.

(4) To check the first condition for a bounding homomorphism, we glue 2-handles to the meridians of the tangle strands to kill the normal closure of the generators t_i, and observe that the h_i make up a spine for the resulting handlebody.

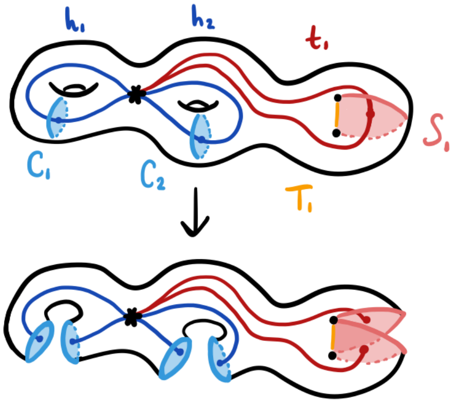


Figure 7: Cutting along cut disks (bounded by cut system curves C_i) and bridge disks (bounded by shadow arcs S_i and tangles T_i) gives a 3-ball. The figure also shows how to build a spine of the tangle complement in the handlebody.

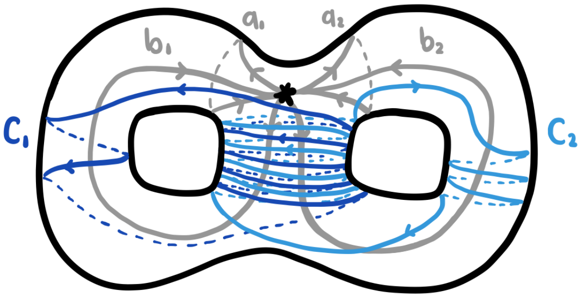


Figure 8: One of the handlebodies in a genus 2 Heegaard splitting of the Poincaré homology sphere, where the other handlebody is the solid genus 2 handlebody filling the interior. The generators a_1, b_1, a_2, b_2 of the fundamental group of the surface are shown in gray, and the two curves C_1, C_2 of the cut system are shown in light and dark blue.

For the second condition, we represent the generators $p_i \in \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$ in the following way. Choose a simple system of rays (or whiskers) r_i connecting $*$ to each point in $\{p_1, \dots, p_{2b}\}$. Then p_i corresponds to a loop that runs out along r_i, goes around the puncture p_i in a negative direction, as dictated by Figure 2, and returns along r_i. On punctured surfaces this is also known as a Hurwitz arc system [25, Section 2.3]. The sequence in which the whisker of the loop around p_i intersects the cut system and tangle shadows will read off a word g_i in the free group. Then running around the puncture reads off one of the generators t_i, followed by returning to $*$ along r_i contributing g_i^{-1}. □

Example 2.5 (handlebody with empty tangle) Here we give an example for the case $b = 0$. In this case the conditions on a bounding homomorphism $\phi\colon \pi_1(\Sigma_g, *) \twoheadrightarrow \langle h_1, \dots, h_g\rangle$ ensure that it is an epimorphism from a surface group to a free group. Consider the genus 2 handlebody shown in Figure 8, which is one of the handlebodies in a genus 2 Heegaard splitting of the Poincaré homology sphere. The bounding homomorphism for this handlebody can be read off by recording the sequence of intersections of the generators a_i and b_i with the curves C_j:

$$\pi_1(\Sigma_2, *) = \langle a_1, b_1, a_2, b_2 \mid [a_1, b_1][a_2, b_2] = 1\rangle \twoheadrightarrow \langle h_1, h_2\rangle$$

is given by

$$a_1 \mapsto h_1^{-1}, \quad b_1 \mapsto (h_1h_2)^5h_1^{-2}, \quad a_2 \mapsto (h_1h_2)^5h_2^3, \quad b_2 \mapsto h_2.$$

Example 2.6 (running example) The following map is a bounding homomorphism which is realized by a trivial 2-bridge tangle in a solid genus 1 handlebody. It appears as the green tangle in the bridge trisection of $\mathbb{RP}^2$ in $\mathbb{CP}^2$ from [32, Figure 2; 23, Figure 3]. See Figure 9. We will use this bounding homomorphism as a running example in the proof of Theorem 2.10:

$$\pi_1(\Sigma_1 - \{p_1, \dots, p_4\}, *) = \langle p_1, p_2, p_3, p_4, a_1, b_1 \mid p_1p_2p_3p_4 = [a_1, b_1]\rangle \twoheadrightarrow \langle t_1, t_2, h_1\rangle$$

is given by

$$p_1 \mapsto t_2h_1t_1h_1^{-1}t_2^{-1}, \quad p_2 \mapsto t_2, \quad p_3 \mapsto h_1t_1^{-1}h_1^{-1}, \quad p_4 \mapsto h_1t_2^{-1}h_1^{-1}, \quad a_1 \mapsto t_2h_1, \quad b_1 \mapsto h_1.$$

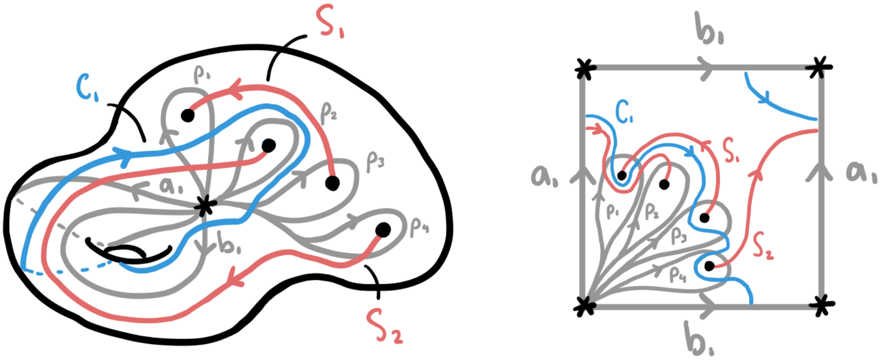


Figure 9: A trivial 2-bridge tangle in a solid genus 1 handlebody, which appears as the green tangle in the bridge trisection of $\mathbb{RP}^2$ in $\mathbb{CP}^2$ from [32, Figure 2; 23, Figure 3].

2B Stallings folding

We now discuss a technique, due to Stallings, called folding [41], which in our context will be used to give a topological realization of any bounding map. There are several applications of folding in the study of finitely generated free groups (eg for the membership problem or determining the index and normality of a subgroup; see [5]). However, for our purposes we only need one application, namely that folding gives a convenient algorithmic method to determine if a set of elements $w_1, \dots, w_k \in F_n$ generate F_n, where F_n is the free group generated by the elements $x_1, \dots, x_n$.

We now describe this algorithm. We begin by forming a directed graph Γ with edges labeled by elements of $\{x_1, \dots, x_n\}$, where Γ is topologically a wedge of k circles and each of the circles is subdivided and labeled according to the words $w_1, \dots, w_k$ as in Figure 10. We can change Γ by a move called a fold to obtain a new such graph.

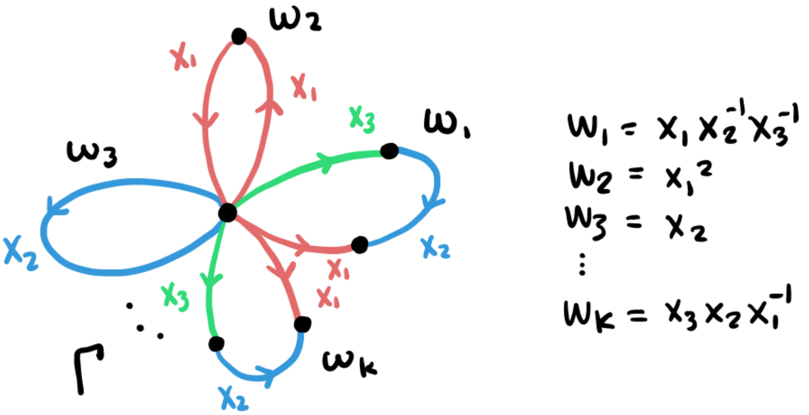


Figure 10: A directed graph Γ with edges labeled by elements of $\{x_1, \dots, x_n\}$, where Γ is topologically a wedge of k circles and each of the circles is subdivided and labeled according to the words $w_1, \dots, w_k$.

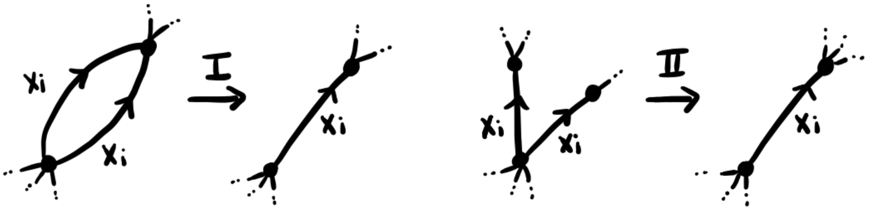


Figure 11: Types of folds. The labels and orientations of the edges being folded must match.

Definition 2.7 (fold) A (*Stallings*) *fold* is a move on a labeled, directed graph which takes two edges and replaces them with one single edge, with the following restrictions. The original edges must

(1) have the same label,

(2) share a vertex, and

(3) be oriented either both in or both out of the shared vertex.

The label and orientation of the new edge are induced by those of the original edges. If the original edges share a second vertex, we call this a *type I fold*, and if not, a *type II fold*. In the case of the type II fold, the unshared vertices are identified together after replacing the original edges with the new edge. See Figure 11.

Lemma 2.8 (Stallings) *The elements $w_1, \dots, w_k \in F_n$ generate F_n if and only if there exists a sequence of folds beginning with the graph Γ given by $w_1, \dots, w_k$ as in Figure 10, and any such sequence of folds terminates in the graph R_n in Figure 12.*

The proof of Lemma 2.8 follows as a special case of [5, Theorem 4.7]. Before we begin the proof of our main technical theorem, we mention one last lemma that will be used to check that a given set of closed curves constitutes a cut system.

Lemma 2.9 *Let $w_1, \dots, w_k$ be elements in a free group F_n with free generating set $x_1, \dots, x_n$ such that $w_1, \dots, w_k$ generate F_n. Let $\exp_{x_i} : F_n \to \mathbb{Z}$ denote the exponent sum homomorphism, that is, the*

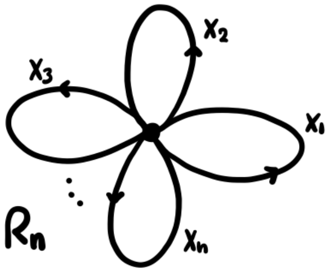


Figure 12: The directed graph R_n is topologically a wedge of n circles, with each circle uniquely labeled by an element x_i.

signed count of occurrences of the letter x_i. *Then the* n *vectors*

$$v_i = (\exp_{x_i}(w_1), \exp_{x_i}(w_2), \dots, \exp_{x_i}(w_k)) \in \mathbb{Z}^k$$

are linearly independent in $\mathbb{Z}^k$.

Proof Consider the $n \times k$ matrix where the rows are given by the vectors v_i. The j^{th} column is made up of all of the exponent sums of the word w_j, and thus computes the image of w_j under the abelianization map $\mathrm{ab}\colon F_n \to \mathbb{Z}^n$, where the images of the generators x_i form the basis of the codomain. Since the words $w_1, \dots, w_k$ generate F_n, the columns of the matrix generate $\mathbb{Z}^n$ and thus its column rank is n. The claim now follows from the equality of row and column rank and the observation that $n \le k$. □

Theorem 2.10 (existence of realization) *For every bounding homomorphism*

$$\phi\colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow \langle t_1, \dots, t_b, h_1, \dots, h_g \rangle,$$

there exists a topological realization (*where the realizing tangle* (H, T) *is unique by Lemma 2.3*) *and*, *further*, *we give a* (*polynomial-time*) *algorithm to construct a diagram for the topological realization.*

Proof The plan is to describe a topological realization of ϕ by an explicit diagram that we will produce in such a way that Lemma 2.4 ensures that it is indeed a topological realization of ϕ. In the first stage we produce a preliminary diagram $\mathfrak{D}$ that, if it were to consist of only a cut system and a shadow diagram, would be a realization of ϕ. This preliminary diagram will not necessarily be unique, and could possibly include “extra” components. After this, in the second stage we will apply Stallings folding from Lemma 2.8 to guide band sums which eliminate any extra components of the preliminary diagram. At this point, we will have obtained a unique diagram, regardless of the choices made while producing the preliminary diagram. Finally, in the third stage we argue why the necessary bands always exist.

First stage (preliminary diagram) For the first stage, it is necessary to initially make the distinction between words in the generators and inverses of generators in a free group, and elements of the free group. For now, when we write $\phi(p_i)$, $\phi(a_j)$, $\phi(b_k)$, we mean the unique freely reduced words representing these elements. To produce $\mathfrak{D}$, look at two (not necessarily freely reduced) words, the first given by concatenating the freely reduced $\phi(p_i)$, namely

$$w_1 = \phi(p_1)\phi(p_2)\cdots\phi(p_{2b}),$$

and the second given by concatenating the freely reduced $\phi(a_i)$, $\phi(b_j)$, namely

$$w_2 = \phi(a_1)\phi(b_1)\cdots\phi(a_g)\phi(b_g).$$

Note that since ϕ is a homomorphism these words are equal as elements of $\langle t_1, \dots, t_b, h_1, \dots, h_g \rangle$.

The first step in drawing the preliminary diagram $\mathfrak{D}$ is to represent Σ_g by a polygon with edges $a_1, b_1, \dots, a_g, b_g$ and punctures as in Figure 13, which will begin our main running example for the proof. (If $g = 0$, view the sphere as the plane with a point at infinity, place all of the punctures in a

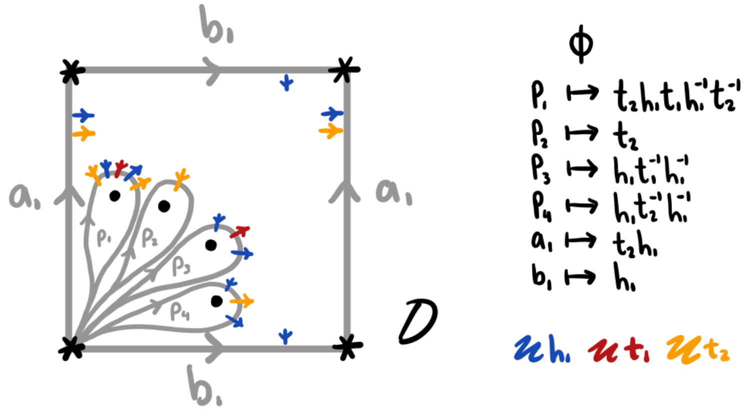


Figure 13: Our main running example for this proof comes from the bounding homomorphism in Example 2.6. Represent $\Sigma_1 = T^2$ by a polygon with punctures. Mark each of the circles representing a_1 and b_1 with "oriented dashes" labeled (using color) by the respective elements in $\phi(a_1)$ and $\phi(b_1)$. Additionally, mark the loops around each punctured point p_i with "oriented dashes" labeled by the respective elements in $\phi(p_1), \dots, \phi(p_4)$. Note that for the oriented dashes we use colors corresponding to h_1, t_1 and t_2 to see this correspondence easily, but keep in mind that throughout this running example we are really recovering the curves C_1, S_1 and S_2, which is why the colors in our end result, Figure 22, differ slightly from those in Example 2.6.

line, and place the basepoint at infinity. See [3, Section 4.2.3] for an example of this.) Mark each of the circles representing $a_1, b_1, \dots, a_g, b_g$ on Σ_g with "oriented dashes" labeled by the respective elements in $\phi(a_1), \phi(b_1), \dots, \phi(a_g), \phi(b_g)$. Additionally, mark the loops around each punctured point p_i on Σ_g with "oriented dashes" labeled by the respective elements in $\phi(p_1), \dots, \phi(p_{2b})$.

The second step is to freely reduce both of the words w_1 and w_2, and, as cancellations occur in the free reductions, draw arcs between the corresponding dashes as in Figure 14. The arcs retain the respective

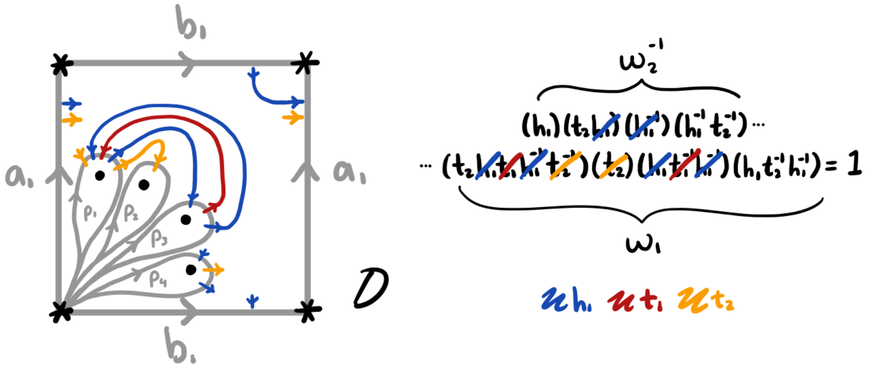

b_1
a_1
a_1
p_1
p_2
p_3
p_4
b_1
D
w_2^{-1}
$(h_1)(t_2h_1)(h_1^{-1})(h_1^{-1}t_2^{-1})\cdots$
$\cdots(t_2h_1t_1h_1^{-1}t_2^{-1})(t_2)(h_1t_1^{-1}h_1^{-1})(h_1t_2^{-1}h_1^{-1}) = 1$
w_1
h_1 t_1 t_2

Figure 14: Running example. Drawing arcs between the dashes as corresponding cancellations occur in the free-reductions of the words w_1 and w_2. There are multiple ways to freely reduce these words; this running example shows one possible choice. In fact, there exists a different choice of cancellations here which would avoid the need to proceed to the second stage.

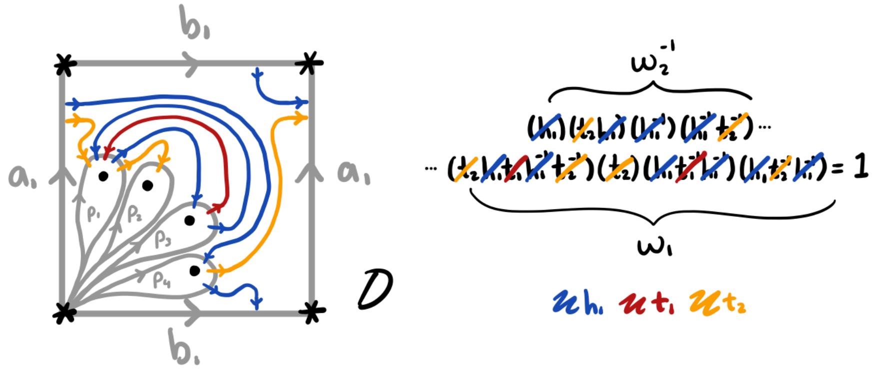


Figure 15: Running example. Continuing to draw arcs between dashes until all cancellations have occurred.

labelings and have orientations induced by the dashes. Let w_1' and w_2' denote the resulting freely reduced words, which are equal as freely reduced words since they are equal as elements of the free group. Therefore $(w_2')^{-1}w_1'$ freely reduces to yield the trivial word.

The final step is to continue to carry out this free reduction down to the trivial word, drawing arcs with each cancellation as in the second step. See Figure 15. This reduction will not necessarily be unique and could produce different diagrams, but this indeterminacy will be fixed in the second stage. The loops around the punctures each intersect an odd number of dashes, which are now each part of an arc. Connect the middle dash to the punctured point, and connect the rest of the dashes in pairs that go around the opposite side of the puncture, as in Figure 16.

We consider the preliminary diagram $\mathscr{D}$ to be the resulting collection of disjoint, oriented arcs and closed curves on Σ_g, each labeled by a generator of $\langle t_1,\dots,t_b,h_1,\dots,h_g\rangle$. In general $\mathscr{D}$ at this stage will consist of too many closed loops and will not give a realization of ϕ. Note that by condition (1) in the definition of a bounding homomorphism, there is at least one closed curve with each label h_j. Similarly, by condition (2), there is exactly one arc with each label t_i. However, in general $\mathscr{D}$ will consist of additional closed curves with labels h_j and t_j.

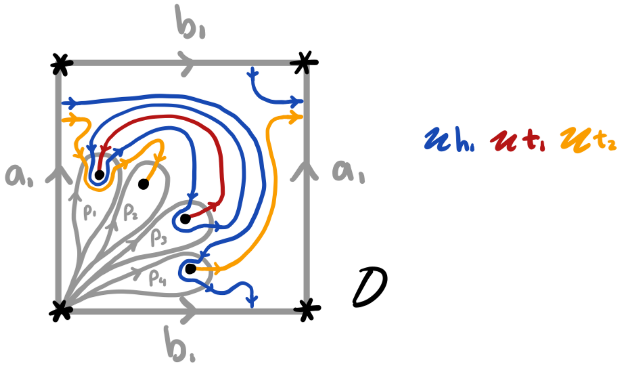


Figure 16: Running example. Connecting the middle dash to the puncture and the other dashes in pairs.

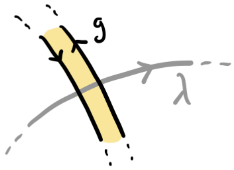


Figure 17: Following the curve a running across a band, we read off the canceling pair of intersections gg^{-1}.

Suppose that at this stage there are no such additional curves in $\mathfrak{D}$; namely, for each h_i there is exactly one closed curve with that label and for each t_j there are no closed curves with that label. We now show that the curves labeled by the elements h_i form a cut system, namely that they are homologically linearly independent.

Let $C_1, \dots, C_g$ denote the closed curves in this case, with corresponding labels $h_1, \dots, h_g$, respectively. The curves $a_1, b_1, \dots, a_g, b_g$ in Figure 2 give a basis for $H_1(\Sigma_g; \mathbb{Z})$. Using that the intersection product on $H_1(\Sigma_g; \mathbb{Z})$ is given by $a_i \cdot b_j = \delta_{ij}$, we find that, by construction, in $H_1(\Sigma_g; \mathbb{Z})$ we have

$$C_i = \exp_{h_i}(\phi(b_1))a_1 + \exp_{h_i}(\phi(a_1))b_1 + \cdots + \exp_{h_i}(\phi(b_g))a_g + \exp_{h_i}(\phi(a_g))b_g,$$

where

$$\exp_{h_i} : \langle t_1, \dots, t_b, h_1, \dots, h_g \rangle \to \mathbb{Z}$$

maps an element to the exponent sum of h_i. By property (1) of being a bounding homomorphism, together with Lemma 2.9, it follows that the elements of C_i are linearly independent in $H_1(\Sigma_g; \mathbb{Z})$ and therefore in this case the curves $C_1, \dots, C_g$ form a cut system.

From this it follows that, in this case, $\mathfrak{D}$ is a diagram for a handlebody and trivial tangle. Furthermore, in this case, we know by Lemma 2.4 that $\mathfrak{D}$ is a diagram for a topological realization of ϕ.

Second stage (Stallings folding) In the second stage, we will modify the preliminary diagram $\mathfrak{D}$ in steps by performing orientation-preserving band sums between components in order to eliminate the extra closed curves, as in Figure 22. These bands may pass through the a_i, b_j and p_k representing the elements in $\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$, but may not pass through the curves and arcs we drew in the previous stage. Assuming for a moment that we successfully banded together all of the curves and arcs, so that there is exactly one curve/arc with each label, the above argument still applies and the resulting diagram will be a diagram for a cut system and trivial tangle (namely, it will consist of a cut system together with a shadow diagram) since each band introduces a pair of canceling intersections as in Figure 17. Furthermore, because the added intersections cancel, applying Lemma 2.4 shows that the resulting diagram is in fact a diagram for a topological realization of ϕ.

The algorithm for finding these bands is as follows. Note that, here and beyond, we use the word *color* to describe the "labels" as mentioned in Definition 2.7, and give the word *label* a new, specific meaning. (In particular, we use *color* colloquially to mean "labeled by a color", as opposed to the graph-theoretic

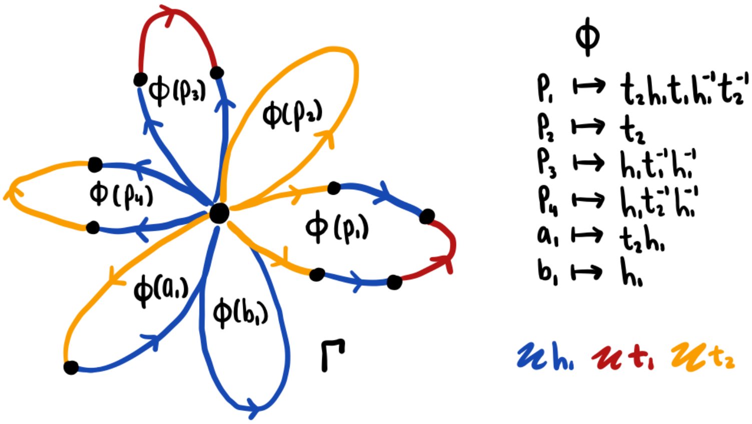


Figure 18: Running example. The graph Γ is topologically a wedge of circles, where each circle is colored by the words $\phi(p_1), \dots, \phi(p_4), \phi(a_1), \phi(b_1)$. Here colors are used to denote this coloring.

notion.) Let Γ be the graph that is topologically a wedge of circles such that each circle is *colored* by the words

$$\phi(p_1), \dots, \phi(p_{2b}), \phi(a_1), \phi(b_1), \dots, \phi(a_g), \phi(b_g)$$

as in Figure 18. Add an additional *label* to each edge of Γ, namely label each edge of Γ with the corresponding closed curve or arc in the preliminary diagram $\mathcal{D}$ as in Figure 19. We will modify the graph Γ by folding and this will dictate how to modify the diagram $\mathcal{D}$ by band sums. At each stage, we will refer to the new graph again by Γ and the new diagram again by $\mathcal{D}$.

Since the elements

$$\phi(p_1), \dots, \phi(p_{2b}), \phi(a_1), \phi(b_1), \dots, \phi(a_g), \phi(b_g)$$

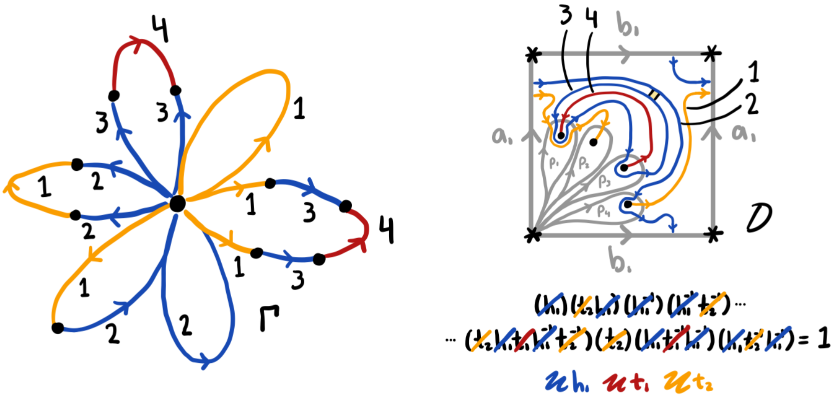

4
3
3
1
1
2
2
1
3
4
1
3
1
2
2
1
2
Γ
3 4
b_1
1
2
a_1
a_1
p_1
p_2
p_3
p_4
D
b_1
$(\cancel{h_1})(\cancel{t_2}\cancel{h_1})(\cancel{h_1^{-1}})(\cancel{h_1^{-1}}\cancel{t_2^{-1}}) \cdots$
$\cdots (\cancel{t_2}\cancel{h_1}\cancel{t_1}\cancel{h_1^{-1}}\cancel{t_2^{-1}})(\cancel{t_2})(\cancel{h_1}\cancel{t_1^{-1}}\cancel{h_1^{-1}})(\cancel{h_1}\cancel{t_2^{-1}}\cancel{h_1^{-1}}) = 1$
h_1 t_1 t_2

Figure 19: Running example. An additional label is added to each edge of Γ which represents the corresponding curve/arc in $\mathcal{D}$. Here numbers are used to denote these labels.

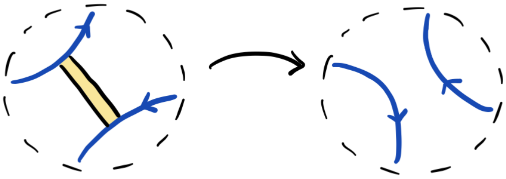


Figure 20: An orientation-preserving band sum between two curves/arcs in a neighborhood within the diagram $\mathscr{D}$. The curves/arcs are the same color but have different labels; that is, they are different curves/arcs in $\mathscr{D}$ that are colored by the same word.

generate the free group $\langle t_1, \ldots, t_b, h_1, \ldots, h_g \rangle$, by Lemma 2.8 there exists a sequence of foldings of the graph Γ to the graph R_n in Figure 12, where $n = b + g$ is the rank of the free group. Choose a sequence of such foldings. Recall that foldings must occur between edges with the same *coloring*, but not necessarily the same *labels* (using the language of the previous paragraph). Whenever two edges of Γ with the same label are folded, the diagram $\mathscr{D}$ is left unchanged and the edge of Γ resulting from the fold is given the same label as the edges it came from. Whenever two edges of Γ with different labels are folded, an orientation-preserving band between the two corresponding curves/arcs is chosen, disjoint from all of the other curves/arcs, and the diagram $\mathscr{D}$ is modified by performing a band sum along this band as in Figure 20. The edge of Γ resulting from the fold is given a new labeling that identifies the

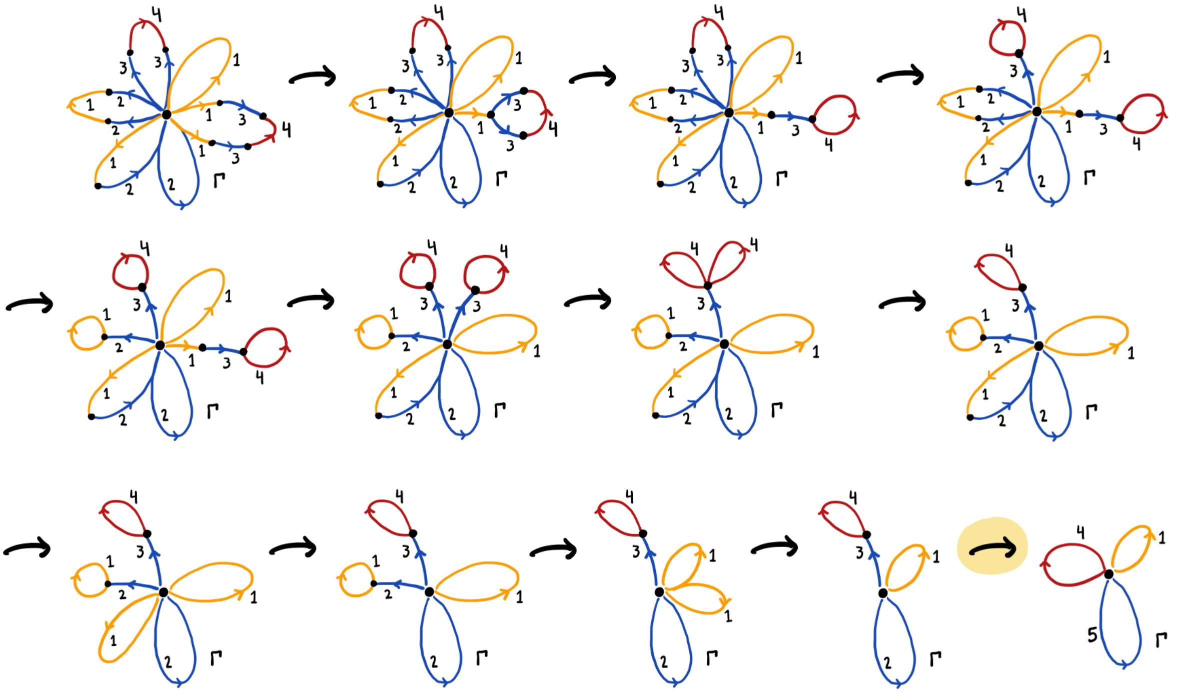


Figure 21: Running example. One possible sequence of Stallings folds. The diagram $\mathscr{D}$ remains unchanged throughout all of the folds except the last fold (highlighted), after which a band sum is performed as in Figure 22.

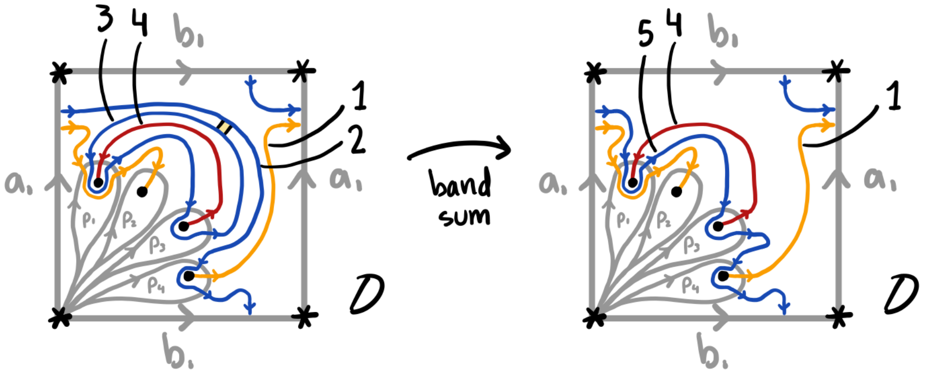


Figure 22: Running example. The curves 2 and 3 on the left are banded together to form the new curve 5 on the right. Now we are finished! In the notation of Example 2.6, the curve 5 is C_1, the arc 4 is S_1, and the arc 1 is S_2.

two banded together curves/arcs. Any other occurrences of the involved labels are modified as well. See Figure 21 for a sequence of folds in our running example, and Figure 22 for the resulting band sum. See Example 2.11 for a different example containing more complicated band sums.

Assuming for a moment that all the required bands do exist, in the final diagram there will be one curve/arc in the diagram for each edge of R_n, since the graph Γ folds down to R_n. It then follows that there will be exactly b arcs and g closed curves in the final diagram $\mathfrak{D}$. The preceding discussion then applies to see that the resulting closed curves form a cut system and the diagram $\mathfrak{D}$ indeed does provide a topological realization of ϕ.

Third stage (existence of bands) We now tackle the problem of the existence of the bands, which will follow from the following claim. Note that in the claim we are not assuming edges have the same color or label, even though our definition of folding requires edges to have the same color. Furthermore, we prove the existence of some "extra" orientation-reversing bands, which we do not need for the second stage of our proof, but we *do* need as part of our inductive argument for the claim. Thus we prove that bands exist in a more general setting, which will imply that the specific bands we want in the previous part of the proof do indeed exist.

Claim *Given two incident edges e_1 and e_2 in the graph Γ at any state of the above procedure, there exists a band between the corresponding curves/arcs that label e_1 and e_2 which is disjoint from all other curves/arcs in the diagram $\mathfrak{D}$. In particular*:

(1) *If the incident edges are oriented so that they both go into their shared vertex, as in Figure 23(I), then there exists an orientation-preserving band from the right of the curve/arc labeling e_1 to the right of the curve/arc labeling e_2.*

(2) *If the incident edges are oriented so that they both go out of their shared vertex, as in Figure 23(II), then there exists an orientation-preserving band from the left of the curve/arc labeling e_1 to the left of the curve/arc labeling e_2.*

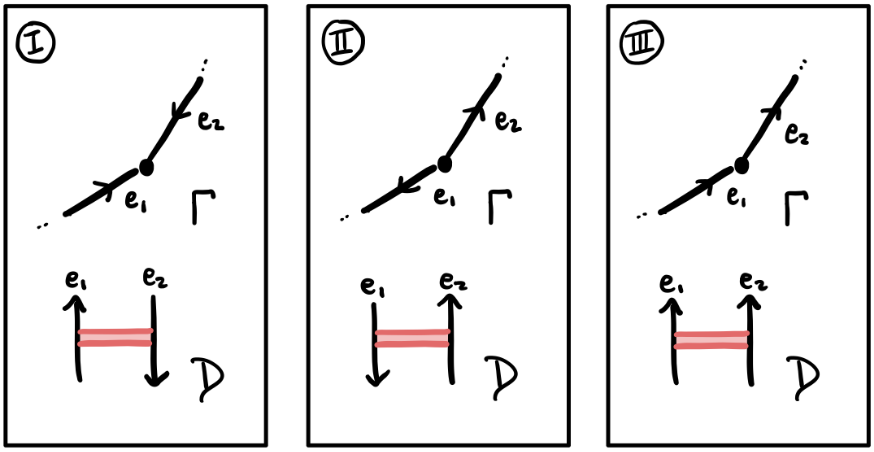


Figure 23: We consider three cases for how the edges e_1 and e_2 are oriented. Here e_1 and e_2 are the names of the edges in the graph Γ, not labels or colors, and we use the same notation to refer to the curves/arcs in $\mathfrak{D}$ that correspond to the edges in Γ.

(3) *If the incident edges are oriented so that one goes into their shared vertex and one goes out*, *as in Figure 23*(*III*), *then there exists an orientation-reversing band from the right of the curve/arc labeling e_1 to the left of the curve/arc labeling e_2.*

Here we are fixing some conventions on how the orientations of incident edges in Γ correspond to orientations in $\mathfrak{D}$, which can be done without loss of generality and in such a way that the cases (I), (II) and (III) are compatible with each other. Also note that in some cases, both an orientation-preserving and an orientation-reversing band might exist between the corresponding curves/arcs. The statement of the claim only contains the existence of those bands which are necessary for the proof.

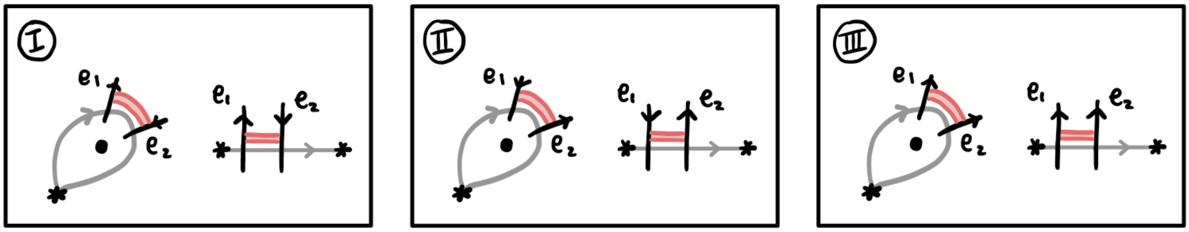


Figure 24: Some orientation checks for the base case of the claim, in which the edges e_1 and e_2 are in the same circle in the graph Γ. We abuse notation and use e_1 and e_2 to also refer to the oriented dashes in $\mathfrak{D}$ that correspond to these edges in Γ. These figures show e_1 and e_2 in the diagram $\mathfrak{D}$, where $*$ is the basepoint and the gray, unlabeled curves/arcs represent the generators a_i, b_i and p_i, in the notation of the previous stages. Here we check whether a disjoint orientation-preserving band (in pink) between e_1 and e_2 can be found in cases (I) and (II), and whether a disjoint orientation-reversing band (in pink) between e_1 and e_2 can be found in case (III).

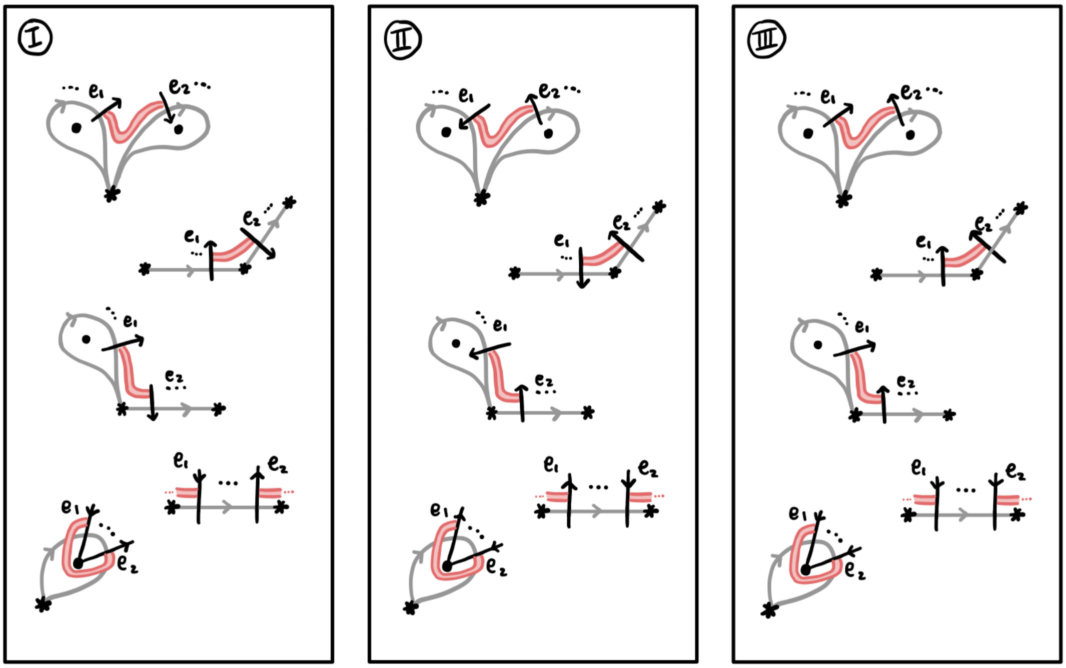


Figure 25: Some orientation checks for the base case of the claim, in which the edges e_1 and e_2 are in different circles in the graph Γ. We abuse notation and use e_1 and e_2 to also refer to the oriented dashes in $\mathfrak{D}$ that correspond to these edges in Γ. These figures show e_1 and e_2 in the diagram $\mathfrak{D}$, where $*$ is the basepoint and the gray, unlabeled curves/arcs represent the generators a_i, b_i and p_i, in the notation of the previous stages. Here we check whether a disjoint orientation-preserving band (in pink) between e_1 and e_2 can be found in cases (I) and (II), and whether a disjoint orientation-reversing band (in pink) between e_1 and e_2 can be found in case (III).

We prove the claim by induction on the number of folds that have been performed in the algorithm, checking throughout that our three cases for orientations hold. We first check that the claim is valid for the preliminary diagram $\mathfrak{D}$ before any folding has been performed. Our initial graph Γ is topologically a wedge of circles, and incident edges can either be in the same circle or in different circles.

If the edges e_1 and e_2 are in the same circle in Γ, then these correspond to oriented dashes in $\mathfrak{D}$ that are right next to each other (except in one case mentioned below). Thus we can draw a band between them which is disjoint from the rest of the diagram. See Figure 24 for some of the orientation checks. If the edges e_1 and e_2 are in different circles in Γ, then they are both connected to the central vertex. They are therefore labeled by the "outermost" curves/arcs in $\mathfrak{D}$ and are connected to the basepoint by an arc which is disjoint from the other curves/arcs. Thickening and joining these arcs then gives a band. See Figure 25 for some of the orientation checks. (The case where the edges are in the same circle, but both connected to the central vertex and not sharing a second vertex is included in Figure 25 rather than Figure 24. Specifically, see the bottom-most two pictures in Figure 25.)

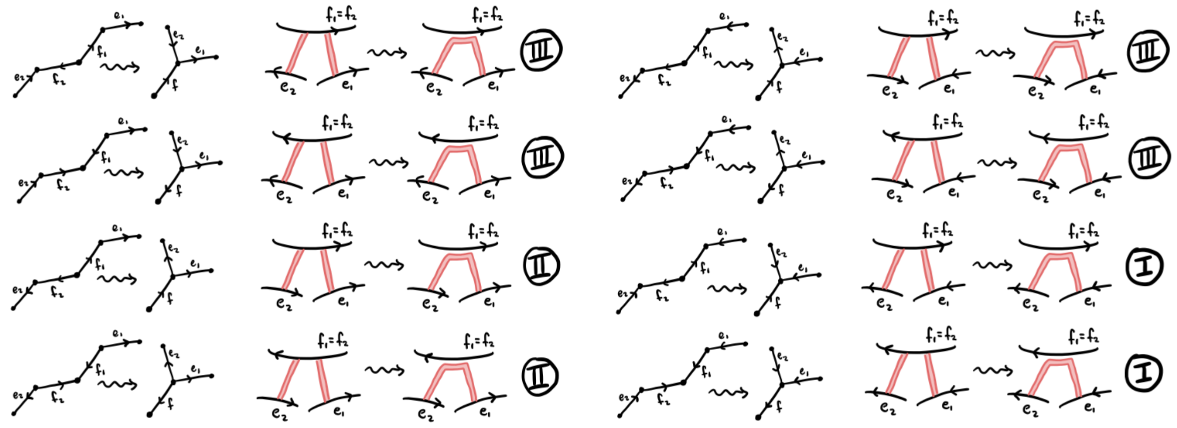


Figure 26: Orientation checks for the inductive step of the claim when the edges f_1 and f_2 have the same label. We abuse notation and use e_1, e_2, f_1, f_2 and f to also refer to the curves/arcs in the diagram that correspond to these edges in Γ. These figures show the possible orientations for e_1, e_2, f_1, f_2 and f relative to each other in Γ (on the left), and how in the corresponding diagrams (on the right), a disjoint orientation-preserving band between e_1 and e_2 can be found inductively when they are oriented as in cases (I) and (II), and a disjoint orientation-reversing band between e_1 and e_2 can be found inductively when they are oriented as in case (III). These pictures suffice for all cases in which e_1 and e_2 are connected to the nonshared vertex of f_1 and f_2, respectively, and are not incident before the fold.

For the inductive step, we verify that the validity of the claim is preserved after folding has occurred. Let $\mathscr{D}$ be the diagram before the fold, $\mathscr{D}'$ be the diagram after the fold, f_1 and f_2 be the edges to be folded, and f be the new folded edge. We need to show that any curves/arcs that label edges that are newly incident after the fold still have a band between them. For ease of explanation, we will slightly abuse notation and use e_1, e_2, f_1, f_2 and f to also refer to the curves/arcs in the diagram that are labeled by these edges.

We first handle the case where the two folded edges f_1 and f_2 have the same label. Note that, in this case, $\mathscr{D} = \mathscr{D}'$. Since f_1 and f_2 have the same label, they correspond to the same curve/arc in $\mathscr{D}$, and existing bands will suffice in all cases except those in which e_1 and e_2 are connected to the nonshared vertex of f_1 and f_2, respectively, and are not incident before the fold. In these cases, observe that a band from e_1 to e_2 can be created by taking the existing band from e_1 to f_1, following along $f_1 = f_2 = f$, and continuing along the existing band from f_2 to e_2. See Figure 26 for the orientation checks.

Finally, assume that the two edges f_1 and f_2 have different labels. In this case, the diagram $\mathscr{D}'$ differs from the diagram $\mathscr{D}$ by an orientation-preserving band sum between f_1 and f_2, which merges these curves/arcs into the same component f. Therefore, as in the previous case, existing bands will suffice in all cases except those in which e_1 and e_2 are connected to the nonshared vertex of f_1 and f_2, respectively, and are not incident before the fold. In these cases, observe that a band from e_1 to e_2 can be created by

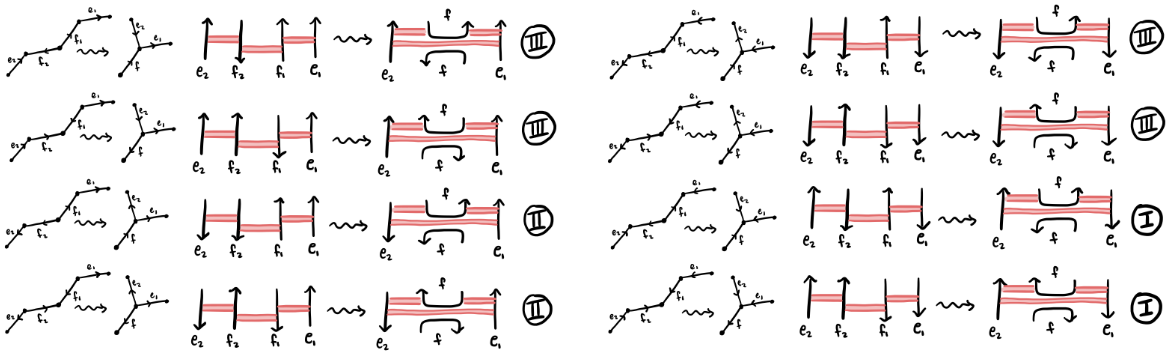


Figure 27: Orientation checks for the inductive step of the claim when the edges f_1 and f_2 have different labels. We abuse notation and use e_1, e_2, f_1, f_2 and f to also refer to the curves/arcs in the diagram that correspond to these edges in Γ. These figures show the possible orientations for e_1, e_2, f_1, f_2 and f relative to each other in Γ (on the left), and how in the corresponding diagrams (on the right), a disjoint orientation-preserving band between e_1 and e_2 can be found inductively when they are oriented as in cases (I) and (II), and a disjoint orientation-reversing band between e_1 and e_2 can be found inductively when they are oriented as in case (III). These pictures suffice for all cases in which e_1 and e_2 are connected to the nonshared vertex of f_1 and f_2, respectively, and are not incident before the fold.

taking the existing band from e_1 to f_1, following through the "tunnel" created by the band sum, and continuing along the existing band from f_2 to e_2. See Figure 27 for the orientation checks. □

Example 2.11 (more complicated band sums) In Figure 28 we present an example containing more complicated band sums (compared to our running example). Here our surface is S^2. In the top box we start with a preliminary diagram $\mathfrak{D}$ coming from a given ϕ (and a choice of cancellation), and from this we produce a graph Γ. The middle box shows a sequence of Stallings folds, which results in three band sums (corresponding to the highlighted folds). The bottom box shows the result of the band sums in the diagram $\mathfrak{D}$. Note that in this example there is a necessary order for the band sums: 1 and 3 cannot be banded together until 2 and 5, and then 6 and 4, are banded together.

3 Closed 3-manifolds and bridge split links

In this section and the next, we translate fundamental topological theorems into the algebraic setup we described in Section 2, and obtain correspondences between topology and algebra. In each subsection we take as input a topological theorem and show how this translates to algebra. Because we pass through diagrams in between topology and algebra and multiple notions of equivalence are involved, the proofs are rather technical. We include full details in Section 3A and omit some details further on, as each subsection builds from the previous and the proofs follow similarly.

As a warm-up, we begin in Section 3A with the case of closed 3-manifolds where our topological input theorem is the Reidemeister–Singer theorem. Then in Section 3B we show how link theory in 3-manifolds

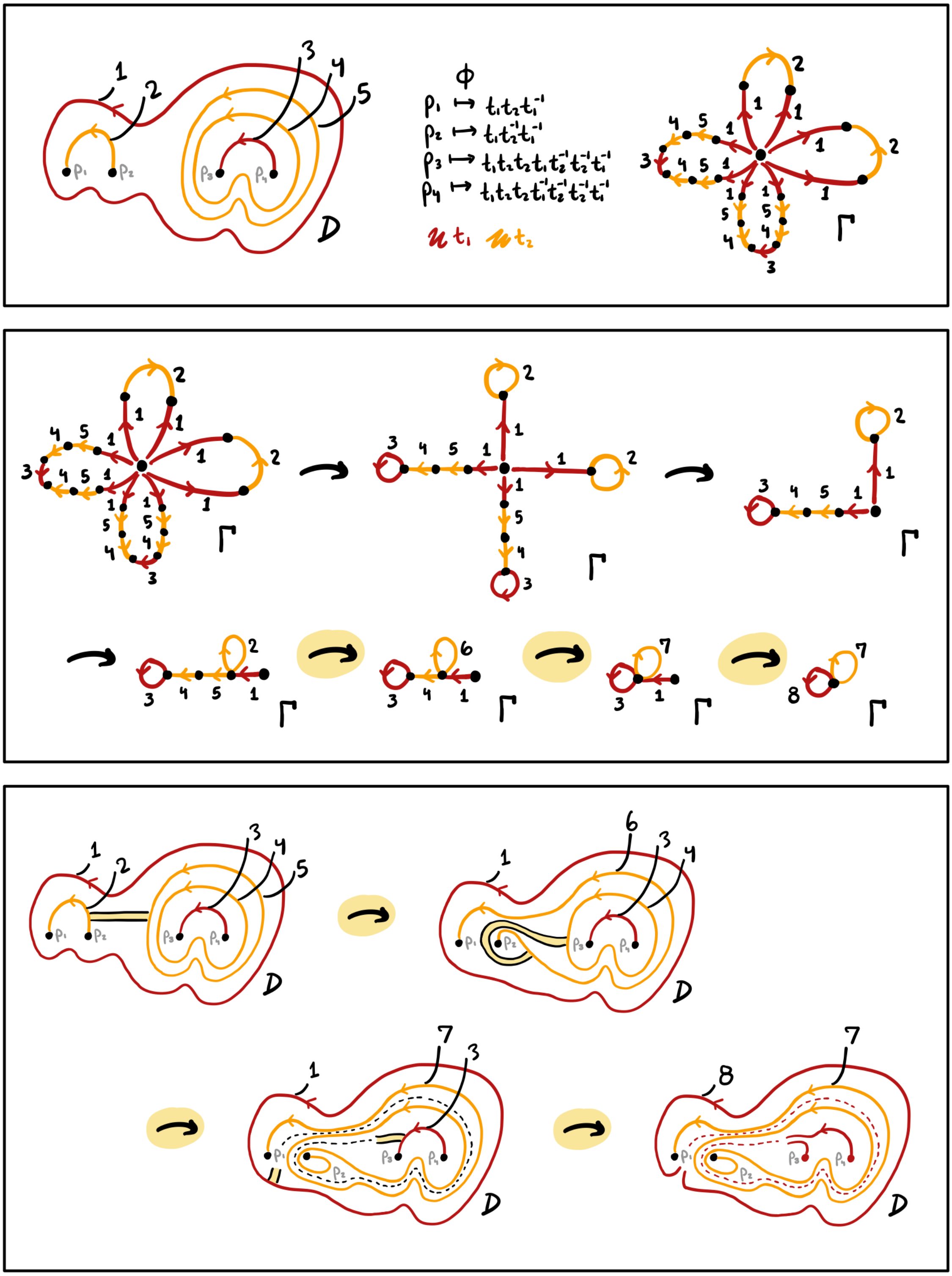

1
2
3
4
5
ρ_1
ρ_2
ρ_3
ρ_4
D
ϕ
$\rho_1 \mapsto t_1 t_2 t_1^{-1}$
$\rho_2 \mapsto t_1 t_2^{-1} t_1^{-1}$
$\rho_3 \mapsto t_1 t_2 t_2 t_1 t_2^{-1} t_2^{-1} t_1^{-1}$
$\rho_4 \mapsto t_1 t_2 t_2 t_1^{-1} t_2^{-1} t_2^{-1} t_1^{-1}$
t_1
t_2
Γ
6
7
8

Figure 28: An example containing more complicated band sums.

can be translated into the algebra of bounding homomorphisms up to stabilization, starting from the observation that a pair of bounding homomorphisms determines a link in bridge position in a Heegaard split 3-manifold. We then state a correspondence theorem in this setting.

In Section 4A we recall the 4-dimensional story of group trisections of closed 4-manifolds. Then in Section 4B we consider the case of surfaces inside 4-manifolds, where a triple of bounding homomorphisms with a pairwise freeness condition determines a bridge trisected surface in a trisected 4-manifold.

3A Closed 3-manifolds

This section is heavily inspired by Jaco's announcement [20] of a result similar to Theorem 3.1. Our topological input theorem here is the Reidemeister–Singer theorem.

Topological input theorem [36; 39] *Any two Heegaard splittings of a fixed* 3*-manifold become isotopic after some number of stabilizations.*

Let `Man`^{3} denote the set of all closed, connected, oriented 3-manifolds considered up to orientation-preserving diffeomorphism (or equivalently homeomorphism). Let `Alg`^{3} denote the set of pairs of homomorphisms (ϕ_1, ϕ_2) where $\phi_i : \pi_1(\Sigma_g, *) \twoheadrightarrow F_g$ for $i = 1, 2$ are surjections. We will call such pairs (ϕ_1, ϕ_2) *splitting homomorphisms* [40; 19]. Given a single such surjection ϕ, which is a bounding homomorphism for the special case where $b = 0$, using Theorem 2.10 we obtain a handlebody, which we will denote by $H(\phi)$, such that the following diagram commutes for an isomorphism ψ as in Lemma 2.4:

$$\begin{tikzcd}
 & \pi_1(H(\phi), *) \arrow[d, "\psi", "\cong"'] \\
\pi_1(\Sigma_g, *) \arrow[ur, two heads, "\iota"] \arrow[r, two heads, "\phi"'] & F_g
\end{tikzcd}$$

By Lemma 2.3, $H(\phi)$ is the unique handlebody bounding Σ_g with the property that there exists a vertical isomorphism ψ in this diagram making it commute.

Therefore, given $(\phi_1, \phi_2) \in$ `Alg`^{3} we can form two handlebodies $H(\phi_1)$, $H(\phi_2)$ with boundary Σ_g, and thus we obtain a compact 3-manifold $M(\phi_1, \phi_2) = H(\phi_1) \cup_{\Sigma_g} -H(\phi_2)$, which is given the orientation that naturally results from gluing the orientations of $H(\phi_1)$ and $-H(\phi_2)$. We thus have a map

$$M : \texttt{Alg}^3 \to \texttt{Man}^3, \quad (\phi_1, \phi_2) \mapsto M(\phi_1, \phi_2).$$

We will define three relations on `Alg`^{3}, denoted by $\sim_h$, $\sim_m$ and $\sim_s$, such that this map M descends to the quotient of `Alg`^{3} by these relations. Both of the relations $\sim_h$ and $\sim_m$ are actually equivalence relations on the set `Alg`^{3}, while $\sim_s$ is not. We will abuse notation and also denote by $\sim_s$ the equivalence relation on `Alg`^{3} generated by the relation $\sim_s$ (that is, the smallest equivalence relation containing $\sim_s$). The h here stands for "handleslide", the m for "mapping class", and the s for "stabilization". The proof of Theorem 3.2 motivates this choice of notation.

We say $(\phi_1,\phi_2)\sim_h(\phi_1',\phi_2')$ if, for $i=1,2$, there exist isomorphisms $h_i\colon F_g\to F_g$ such that the following diagram commutes:

$$\begin{tikzcd}
 & F_g \arrow[dd, "\cong"', "h_i"] \\
\pi_1(\Sigma_g,*) \arrow[ur, two heads, "\phi_i"] \arrow[dr, two heads, "\phi_i'"'] & \\
 & F_g
\end{tikzcd}$$

We call an automorphism $m\colon\pi_1(\Sigma_g,*)\to\pi_1(\Sigma_g,*)$ *orientation-preserving* if the induced automorphism $H_2(\pi_1(\Sigma_g,*);\mathbb{Z})\to H_2(\pi_1(\Sigma_g,*);\mathbb{Z})$ is the identity. We write $(\phi_1,\phi_2)\sim_m(\phi_1',\phi_2')$ if there exists an orientation-preserving isomorphism $m\colon\pi_1(\Sigma_g,*)\to\pi_1(\Sigma_g,*)$ such that, for $i=1,2$, the following diagram commutes:

$$\begin{tikzcd}
\pi_1(\Sigma_g,*) \arrow[dd, "\cong"', "m"] \arrow[dr, two heads, "\phi_i"] & \\
 & F_g \\
\pi_1(\Sigma_g,*) \arrow[ur, two heads, "\phi_i'"'] &
\end{tikzcd}$$

Next we define $\sim_s$. We note that ϕ_i' will be a map from $\pi_1(\Sigma_g,*)\twoheadrightarrow F_g$ while ϕ_i will be a map from $\pi_1(\Sigma_{g+1},*)\twoheadrightarrow F_{g+1}$. Let a_i, b_i be the generators of $\pi_1(\Sigma_g,*)$ (and, abusing notation, $\pi_1(\Sigma_{g+1},*)$), and h_i be the generators of F_g (and, abusing notation, F_{g+1}). We say $(\phi_1,\phi_2)\sim_s(\phi_1',\phi_2')$ if $\phi_i(a_j)=\phi_i'(a_j)$ and $\phi_i(b_j)=\phi_i'(b_j)$ for $i=1,2$ and $j=1,\dots,g$ (where we are identifying F_g naturally as a subset of F_{g+1}), and the rest of the generators are mapped as follows:

$$\phi_1(a_{g+1})=h_{g+1},\quad \phi_1(b_{g+1})=1,\quad \phi_2(a_{g+1})=1,\quad \phi_2(b_{g+1})=h_{g+1}.$$

Let $\sim$ denote the equivalence relation on $\mathtt{Alg}^3$ generated by $\sim_h$, $\sim_m$ and $\sim_s$. Now we proceed to the main result of this section. A result that is similar in spirit was announced in [20].

Theorem 3.1 *The map $M\colon\mathtt{Alg}^3\to\mathtt{Man}^3$ descends to $\mathtt{Alg}^3/{\sim}$ and the resulting map is a bijection.*

Proof We consider an intermediate set $\mathtt{Diag}^3$ whose elements are *Heegaard diagrams*, that is, tuples (Σ_g,α,β) where α and β are cut systems on Σ_g (which are only considered up to isotopy). We will refer to these simply as *diagrams*. Then the map M factors as

$$\begin{tikzcd}
\mathtt{Alg}^3 \arrow[rr, "M"] \arrow[dr, "D"] & & \mathtt{Man}^3 \\
 & \mathtt{Diag}^3 \arrow[ur, "R"] &
\end{tikzcd}$$

The map $R\colon\mathtt{Diag}^3\to\mathtt{Man}^3$ is the topological realization of a diagram (Σ_g,α,β), where we cross Σ_g with an interval, glue disks on the respective sides to α and β, and then glue 3-balls to the resulting sphere boundary components. The map $D\colon\mathtt{Alg}^3\to\mathtt{Diag}^3$ is the construction of M using Theorem 2.10, but where we stop at just a diagram (rather than realizing the manifold) with α corresponding to ϕ_1

$\sim_h$	an equivalence relation on $\mathtt{Alg}^3$ (as defined above)
$\sim_m$	an equivalence relation on $\mathtt{Alg}^3$ (as defined above)
$\sim_s$	an equivalence relation on $\mathtt{Alg}^3$ (as defined above)
(ϕ_1, ϕ_2)	an element of $\mathtt{Alg}^3$
$[\phi_1, \phi_2]$	an equivalence class in $\mathtt{Alg}^3/\sim_h$
$\llbracket \phi_1, \phi_2 \rrbracket$	an equivalence class in $\mathtt{Alg}^3/\sim_h, \sim_m$
$\sim_h$	an equivalence relation on $\mathtt{Diag}^3$ (generated by handleslides)
$\sim_m$	an equivalence relation on $\mathtt{Diag}^3$ (generated by mapping classes)
$\sim_s$	an equivalence relation on $\mathtt{Diag}^3/\sim_h$ (generated by stabilizations)
$(\Sigma_g, \alpha, \beta)$	an element of $\mathtt{Diag}^3$
$[\Sigma_g, \alpha, \beta]$	an equivalence class in $\mathtt{Diag}^3/\sim_h$
$\llbracket \Sigma_g, \alpha, \beta \rrbracket$	an equivalence class in $\mathtt{Diag}^3/\sim_h, \sim_m$

Table 1: A summary of the notation used throughout the proof of Theorem 3.1.

and β corresponding to ϕ_2. We use the notation $D(\phi_1)$ and $D(\phi_2)$ to denote these cut systems, so that $D\colon (\phi_1, \phi_2) \mapsto (\Sigma_g, D(\phi_1), D(\phi_2))$.

The following commutative diagram is a guide to the logic of the proof:

$$\begin{tikzcd}
\mathtt{Alg}^3 \arrow[rr, "M"] \arrow[dr, "D"] \arrow[dd] & & \mathtt{Man}^3 \\
& \mathtt{Diag}^3 \arrow[ur, "R"] \arrow[d, "p_1"] & \\
\mathtt{Alg}^3/\sim_h \arrow[r, "D_1"] \arrow[d] & \mathtt{Diag}^3/\sim_h \arrow[d, "p_2"] \arrow[uur, bend right, two heads] & \\
\mathtt{Alg}^3/\sim_h, \sim_m \arrow[r, "D_2"] \arrow[d] & \mathtt{Diag}^3/\sim_h, \sim_m \arrow[d, "p_3"] \arrow[uuur, bend right, two heads] & \\
\mathtt{Alg}^3/\sim_h, \sim_m, \sim_s \arrow[r, "D_3"] & (\mathtt{Diag}^3/\sim_h, \sim_m)/\sim_s \arrow[uuuur, bend right, two heads] &
\end{tikzcd}$$

The goal is to define a bijection $(\mathtt{Alg}^3/\sim_h, \sim_m, \sim_s) \to \mathtt{Man}^3$, so we must show that this map, which passes through an intermediate set of diagrams, is well defined, injective and surjective. We do this by descending by quotients on the algebraic and diagrammatic sides, and checking each time that the relevant map factors through and a bijection between the quotients is achieved.

In Table 1 we summarize the notation used throughout the proof, with precise definitions for the relations on the diagrammatic side following. We abuse notation and use the symbols $\sim_h$, $\sim_m$, $\sim_s$ to denote equivalence relations on both the algebraic and diagrammatic sides.

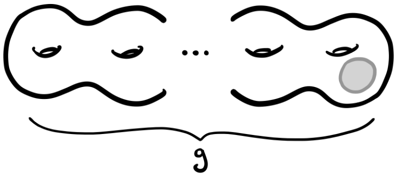


Figure 29: Our standard model for the closed genus g surface, with a fixed disk for stabilization indicated.

Given two diagrams $(\Sigma_g, \alpha, \beta)$ and $(\Sigma_g, \alpha', \beta')$, we write $(\Sigma_g, \alpha, \beta) \sim_h (\Sigma_g, \alpha', \beta')$ if there is a sequence of handleslides from the curves α to α' and similarly from β to β'. We write $(\Sigma_g, \alpha, \beta) \sim_m (\Sigma_g, \alpha', \beta')$ if there exists a single mapping class $\Sigma_g \to \Sigma_g$ taking α to α' and β to β' simultaneously. Let $[\![\Sigma_g, \alpha, \beta]\!]$ denote an equivalence class of a diagram $(\Sigma_g, \alpha, \beta)$ under the equivalence relation generated by $\sim_h$ and $\sim_m$.

Recall that stabilizing a Heegaard diagram entails performing a connect sum with the standard genus 1 diagram of S^3. In order to connect sum in a controlled manner, recall that we have fixed a standard model for the closed genus g surface, and we additionally fix a disk on this model where the connect sum will be performed. See Figure 29. Because we will mod out by handleslides and mapping classes first, we can assume our Heegaard diagram looks like the standard model. We write $[\![\Sigma_{g+1}, \alpha, \beta]\!] \sim_s [\![\Sigma_g, \alpha', \beta']\!]$ if we obtain α and β on Σ_{g+1} from α' and β' on Σ_g by

1. choosing an isotopy of α' and β' such that they do not intersect the connect sum disk, and
2. modifying Σ_g to be Σ_{g+1} (using the fixed disk for the connect sum) and adding the two new curves in the standard genus 1 diagram of S^3 to α' and β'.

(Note that this description incorporates both stabilization and destabilization, depending on which diagram is seen as the original and which as the modified one.) This operation is not well defined in $\mathtt{Diag}^3$ because of the choice of isotopy; for instance, see Figure 30. However once we quotient by $\sim_h$ this *is* well defined, as we are able to use handleslides to "move" the curve over the attached handle. See Figure 31. Thus we write $(\mathtt{Diag}^3/{\sim_h}, {\sim_m})/{\sim_s}$ rather than $\mathtt{Diag}^3/{\sim_h}, {\sim_m}, {\sim_s}$ because unique stabilizations only occur after modding out by handleslides, and additionally we wish to assume our Heegaard diagram looks like our standard model equipped with our fixed disk.

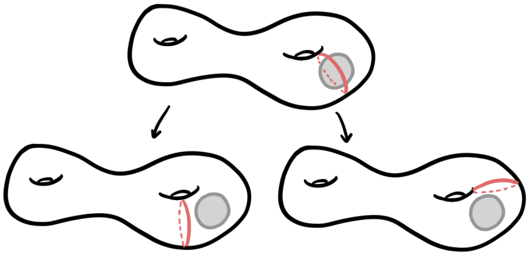


Figure 30: A curve which intersects the disk for stabilization, and two choices of isotopy for moving the curve off of the disk.

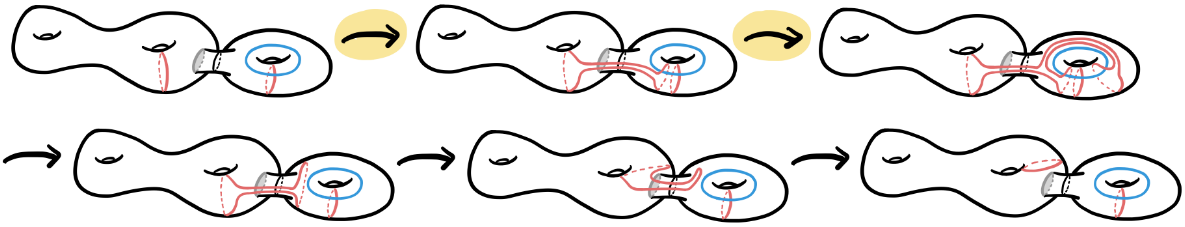


Figure 31: Using handleslides, denoted by the highlighted arrows, to move the curve over the attached handle.

Claim 1 *The map $p_1 \circ D$ factors through $\mathtt{Alg}^3/{\sim_h}$ and the resulting map D_1 is bijective.*

Recall that two handlebodies H_1 and H_2 bounding a given surface are equivalent if and only if their diagrams differ by handleslides [22]. Given two splitting homomorphisms $(\phi_1, \phi_2), (\phi_1', \phi_2') \in \mathtt{Alg}^3$ with $(\phi_1, \phi_2) \sim_h (\phi_1', \phi_2')$, by Lemma 2.3 we see that $H(\phi_1) = H(\phi_1')$ and $H(\phi_2) = H(\phi_2')$. Therefore $(\Sigma_g, D(\phi_1), D(\phi_2)) \sim_h (\Sigma_g, D(\phi_1'), D(\phi_2'))$ and thus the map factors through $\mathtt{Alg}^3/{\sim_h}$, as desired.

To see that the map $D_1 \colon \mathtt{Alg}^3/{\sim_h} \to \mathtt{Diag}^3/{\sim_h}$ is injective, suppose that $D_1([\phi_1, \phi_2]) = D_1([\phi_1', \phi_2'])$ with $[\phi_1, \phi_2], [\phi_1', \phi_2'] \in \mathtt{Alg}^3/{\sim_h}$. Then $H(\phi_1) = H(\phi_1')$ and $H(\phi_2) = H(\phi_2')$ and therefore, by the definition of equivalence of handlebodies, we have for $i = 1, 2$ the commutative diagram

$$\begin{tikzcd}
& \pi_1(H(\phi_i), *) \arrow[d, "h_i", "\cong"'] \\
\pi_1(\Sigma_g, *) \arrow[ur, two heads] \arrow[dr, two heads] & \\
& \pi_1(H(\phi_i'), *)
\end{tikzcd}$$

for some isomorphisms $h_i \colon \pi_1(H(\phi_i), *) \to \pi_1(H(\phi_i'), *)$, where the other maps are induced by inclusion. From this, it follows that $(\phi_1, \phi_2) \sim_h (\phi_1', \phi_2')$. Thus D_1 is injective.

To see that the map $D_1 \colon \mathtt{Alg}^3/{\sim_h} \to \mathtt{Diag}^3/{\sim_h}$ is surjective, we define a section

$$\sigma \colon \mathtt{Diag}^3/{\sim_h} \to \mathtt{Alg}^3/{\sim_h}.$$

Let $[\Sigma_g, \alpha, \beta]$ denote the equivalence class of $(\Sigma_g, \alpha, \beta)$ in $\mathtt{Diag}^3/{\sim_h}$. We define $\sigma([\Sigma_g, \alpha, \beta])$ by taking the diagram $(\Sigma_g, \alpha, \beta)$ with a particular choice of the curves α and β (so they are no longer isotopy classes but fixed curves). We then reverse the construction of the map D. That is, we consider each of the sets of curves α and β separately and apply the construction just as in Lemma 2.4 to obtain maps $\phi_1, \phi_2 \colon \pi_1(\Sigma_g, *) \to F_g$, where here we have chosen orientations for each of the curves in α and β. These maps ϕ_1, ϕ_2 are independent of the choice of representatives of the isotopy classes of the curves in α and β, as well as the choice of orientations, when we consider the result $[\phi_1, \phi_2]$ in $\mathtt{Alg}^3/{\sim_h}$, giving a map $\mathtt{Diag}^3 \to \mathtt{Alg}^3/{\sim_h}$. (Note that we have also implicitly chosen an ordering of the curves in each cut system in this construction; however because there are automorphisms of the free group permuting all of the canonical generators, this choice does not matter.) This map factors through to give a map

$\sigma\colon \mathtt{Diag}^3/{\sim_h} \to \mathtt{Alg}^3/{\sim_h}$ which is a section to D_1 by Lemma 2.3 together with the fact that two cut systems determine the same handlebody if and only if they differ by handleslides. Thus D_1 is surjective.

Claim 2 *The map $p_2 \circ D_1$ factors through $\mathtt{Alg}^3/{\sim_h}, {\sim_m}$ and the resulting map D_2 is bijective.*

For well-definedness, suppose that $(\phi_1, \phi_2) \sim \cdots \sim (\phi_1', \phi_2')$, where each $\sim$ is either $\sim_h$ or $\sim_m$. We must show that $(\Sigma_g, D(\phi_1), D(\phi_2)) \sim \cdots \sim (\Sigma_g, D(\phi_1'), D(\phi_2'))$ where each $\sim$ is either $\sim_h$ or $\sim_m$. By the previous step of the proof, we know that every $\sim_h$ equivalence of splitting homomorphisms produces diagrams that are equivalent with respect to $\sim_h$. Assume $(\phi_1, \phi_2) \sim_m (\phi_1', \phi_2')$. By the Dehn–Nielsen–Baer theorem, there exists an orientation-preserving diffeomorphism $\Sigma_g \to \Sigma_g$ that fixes the basepoint and realizes the isomorphism $\pi_1(\Sigma_g, *) \to \pi_1(\Sigma_g, *)$ that is contained in the assumption that $(\phi_1, \phi_2) \sim_m (\phi_1', \phi_2')$ [7]. This then implies that $(\Sigma_g, D(\phi_1), D(\phi_2))$ and $(\Sigma_g, D(\phi_1'), D(\phi_2'))$ are equivalent using $\sim_h$ and $\sim_m$. Therefore, the resulting map D_2 is well defined.

Assume $D_2(\llbracket \phi_1, \phi_2 \rrbracket) = D_2(\llbracket \phi_1', \phi_2' \rrbracket)$, where $\llbracket \phi_1, \phi_2 \rrbracket, \llbracket \phi_1', \phi_2' \rrbracket \in \mathtt{Alg}^3/{\sim_h}, {\sim_m}$. We must show that $(\phi_1, \phi_2) \sim \cdots \sim (\phi_1', \phi_2')$, where each $\sim$ is either $\sim_h$ or $\sim_m$. By assumption, $(\Sigma_g, D(\phi_1), D(\phi_2)) \sim \cdots \sim (\Sigma_g, D(\phi_1'), D(\phi_2'))$, where each $\sim$ is either $\sim_h$ or $\sim_m$. Let $(\Sigma_g, \alpha, \beta)$ and $(\Sigma_g, \alpha', \beta')$ be two diagrams in the above chain of relations such that $(\Sigma_g, \alpha, \beta) \sim_m (\Sigma_g, \alpha', \beta')$. Then there exists an orientation-preserving diffeomorphism

$$F\colon \Sigma_g \to \Sigma_g, \qquad \alpha \mapsto \alpha', \quad \beta \mapsto \beta',$$

which we can assume fixes the basepoint $*$. Let $(f_1, f_2) = \sigma(\Sigma_g, \alpha, \beta)$ and $(f_1', f_2') = \sigma(\Sigma_g, \alpha', \beta')$, where σ is the map from the preceding claim. (Note that σ is technically defined on $\mathtt{Diag}^3/{\sim_h}$, but we can similarly apply the same construction to any specific diagram with curves transverse to the generators $a_1, \dots, b_g$ which are not considered up to isotopy.) Then we have for $i = 1, 2$ the commutative diagram

$$\begin{tikzcd}
\pi_1(\Sigma_g, *) \arrow[dd, "\cong"', "\pi_1(F,*)"] \arrow[dr, two heads, "f_i"] & \\
& F_g \\
\pi_1(\Sigma_g, *) \arrow[ur, two heads, "f_i'"'] &
\end{tikzcd}$$

for $i = 1, 2$, where $\pi_1(F, *)$ is orientation-preserving. Therefore, $(f_1, f_2) \sim_m (f_1', f_2')$, so the chain of equivalences from $(\Sigma_g, D(\phi_1), D(\phi_2))$ to $(\Sigma_g, D(\phi_1'), D(\phi_2'))$ can be converted to a chain of equivalences from (ϕ_1, ϕ_2) to (ϕ_1', ϕ_2'), and the map D_2 is injective.

It follows similarly that

$$\sigma\colon \mathtt{Diag}^3/{\sim_h} \to \mathtt{Alg}^3/{\sim_h}$$

descends to a map

$$\sigma\colon \mathtt{Diag}^3/{\sim_h}, {\sim_m} \to \mathtt{Alg}^3/{\sim_h}, {\sim_m},$$

and that it is a section for D_2.

Claim 3 *The map $p_3 \circ D_2$ factors through $\mathtt{Alg}^3/{\sim_h}, {\sim_m}, {\sim_s}$ and the resulting map D_3 is bijective.*

For well-definedness, suppose $(\phi_1, \phi_2) \sim_s (\phi_1', \phi_2')$. Then, by construction of the map D, we will have $[\![\Sigma_{g+1}, D(\phi_1), D(\phi_2)]\!] \sim_s [\![\Sigma_g, D(\phi_1'), D(\phi_2')]\!]$. Well-definedness therefore follows.

Similarly, by construction of D, if

$$[\![\Sigma_{g+1}, D(\phi_1), D(\phi_2)]\!] \sim_s [\![\Sigma_g, D(\phi_1'), D(\phi_2')]\!],$$

then $[\![\phi_1, \phi_2]\!] \sim_s [\![\phi_1', \phi_2']\!]$, so D_3 is injective. If $[\![\Sigma_{g+1}, \alpha, \beta]\!] \sim_s [\![\Sigma_g, \alpha', \beta']\!]$, then, again by construction, $[\![\sigma(\Sigma_{g+1}, \alpha, \beta)]\!] \sim_s [\![\sigma(\Sigma_g, \alpha', \beta')]\!]$, so σ factors through to give a section of D_3.

We note that this claim follows more immediately than the previous ones since the definition of the algebraic relation $\sim_s$ is explicit in the sense that it does not involve any choices, as compared to the definitions of the algebraic relations $\sim_h$ and $\sim_m$. In later sections, we will define other notions of stabilization and they will be similarly explicit.

Claim 4 *The map R factors through $(\mathtt{Diag}^3/{\sim_h}, {\sim_m})/{\sim_s}$ and the resulting map is bijective.*

Every closed, orientable 3-manifold admits a Heegaard decomposition (for example, by taking a triangulation and taking the Heegaard splitting surface to be the boundary of a regular neighborhood of the 1-skeleton). Let Y be a closed, orientable 3-manifold and let $S \subset Y$ be a Heegaard splitting surface of genus g. Choose an identification of S with Σ_g. We have $Y = H_1 \cup_S H_2$ for two handlebodies H_1 and H_2, and by choosing collections of g disjoint properly embedded disks $\mathbb{D}_1$ and $\mathbb{D}_2$ in H_1 and H_2, respectively, that cut H_1 and H_2 into a 3-ball, then looking at $(S, \partial\mathbb{D}_1, \partial\mathbb{D}_2)$, we have a diagram whose topological realization is Y. (Here the topological realization is as before: thicken S, glue disks to $\mathbb{D}_1$ and $\mathbb{D}_2$ on their respective sides, and glue in 3-balls to the resulting spheres.) Using the identification of S with Σ_g, we obtain a diagram $(\Sigma_g, \alpha, \beta) \in \mathtt{Diag}^3$ where α and β are the respective images of $\partial\mathbb{D}_1$ and $\partial\mathbb{D}_2$, and the image of $(\Sigma_g, \alpha, \beta)$ in $\mathtt{Man}^3$ is Y. Therefore the map $R\colon \mathtt{Diag}^3 \to \mathtt{Man}^3$ is surjective.

The factored-through map $R\colon (\mathtt{Diag}^3/{\sim_h}, {\sim_m})/{\sim_s} \to \mathtt{Man}^3$ is injective by the Reidemeister–Singer theorem, and hence a bijection. By composing D_3 with this map, we obtain the theorem. □

3B Bridge split links in 3-manifolds

In Section 3A we used that any pair of Heegaard splittings of the same fixed 3-manifold become isotopic after some number of stabilization operations (which corresponds to connect summing with the genus 1 splitting of the 3-sphere). The goal of this section will be translating the corresponding uniqueness up to perturbation statement for bridge splittings of links in 3-manifolds into the algebra of bounding homomorphisms.

Topological input theorem [15; 44] *Let L be a link in a fixed Heegaard split 3-manifold. Then any two bridge splittings of L become isotopic after some number of perturbations.*

Let $\mathtt{Man}^{(3,1)}$ denote the set of closed, connected, oriented 3-manifolds M together with a link $L \subset M$ modulo orientation-preserving diffeomorphisms preserving the links. Let $\mathtt{Alg}^{(3,1)}$ denote the set of pairs (ϕ_1, ϕ_2) such that $\phi_1, \phi_2 \colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow F_{g+b}$ are bounding homomorphisms. Throughout this section, let a_j, b_j and p_k denote the generators of $\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$, where a_j, b_j are the surface generators (for $j = 1, \dots, g$) and p_k are the puncture generators (for $k = 1, \dots, 2b$), and let h_j, t_l denote the generators of F_{g+b} (for $l = 1, \dots, b$).

We define the map $(M, L) \colon \mathtt{Alg}^{(3,1)} \to \mathtt{Man}^{(3,1)}$ as follows. Given $(\phi_1, \phi_2) \in \mathtt{Alg}^{(3,1)}$, let $H(\phi_1)$ and $H(\phi_2)$ be the handlebodies bounding Σ_g that result from the application of Theorem 2.10, and further let $T(\phi_1) \subset H(\phi_1)$ and $T(\phi_2) \subset H(\phi_2)$ be the resulting trivial tangles in these handlebodies. We then define $(M, L)(\phi_1, \phi_2)$ to be the 3-manifold $H(\phi_1) \cup_{\Sigma_g} H(\phi_2)$ (with the orientation as in Section 3A) together with the link $T(\phi_1) \cup_{\{p_1, \dots, p_{2b}\}} T(\phi_2)$.

We now define the analogues of the equivalence relations $\sim_h$, $\sim_m$ and $\sim_s$ on $\mathtt{Alg}^3$ in this setting. Given $(\phi_1, \phi_2), (\phi'_1, \phi'_2) \in \mathtt{Alg}^{(3,1)}$, we write $(\phi_1, \phi_2) \sim_h (\phi'_1, \phi'_2)$ if, for $i = 1, 2$, there exist isomorphisms $h_i \colon F_{g+b} \to F_{g+b}$ such that the following diagram commutes:

$$\begin{tikzcd}
 & F_{g+b} \arrow[d, "h_i", "\cong"'] \\
\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \arrow[ur, two heads, "\phi_i"] \arrow[dr, two heads, "\phi'_i"'] & F_{g+b}
\end{tikzcd}$$

Let $m \colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \to \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$ be an automorphism that preserves the conjugacy classes of $p_1, \dots, p_{2b}$ setwise. Then m descends to an automorphism $\pi_1(\Sigma_g, *) \to \pi_1(\Sigma_g, *)$, by the surjective map $\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow \pi_1(\Sigma_g, *)$ which sends each p_i to the identity and is the identity on all of the elements $a_1, b_1, \dots, a_g, b_g$. We call m *orientation-preserving* if this corresponding automorphism $\pi_1(\Sigma_g, *) \to \pi_1(\Sigma_g, *)$ is orientation-preserving. Given $(\phi_1, \phi_2), (\phi'_1, \phi'_2) \in \mathtt{Alg}^{(3,1)}$, we write $(\phi_1, \phi_2) \sim_m (\phi'_1, \phi'_2)$ if there exists an orientation-preserving isomorphism

$$m \colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \to \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$$

such that, for $i = 1, 2$, the following diagram commutes:

$$\begin{tikzcd}
\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \arrow[d, "m", "\cong"'] \arrow[dr, two heads, "\phi_i"] & \\
\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \arrow[r, two heads, "\phi'_i"'] & F_{g+b}
\end{tikzcd}$$

While $\sim_h$ and $\sim_m$ as defined in this section are very similar to $\sim_h$ and $\sim_m$ as defined in Section 3A, the analogue of $\sim_s$ is a bit more complicated. We will have one such relation $\sim_{s_g}$, which is directly analogous to $\sim_s$ in Section 3A; namely, it captures the idea of increasing the genus of the Heegaard splitting while leaving everything else fixed. In addition, there are two relations $\sim_{s_b^1}$ and $\sim_{s_b^2}$ which will correspond to the idea of modifying a link in bridge position by perturbation.

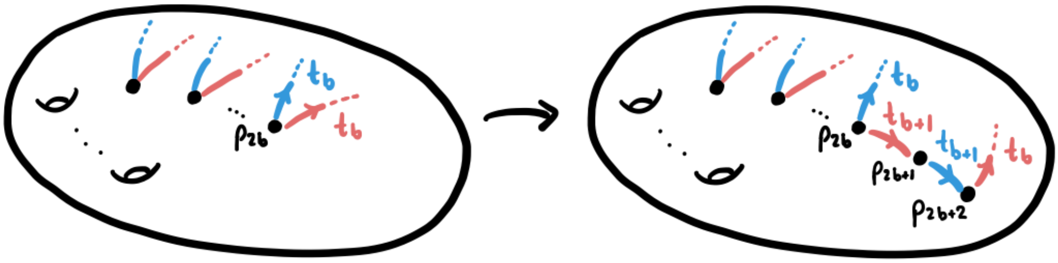


Figure 32: Performing a perturbation, where one of the pink tangle strands is being pulled through the surface. This adds a tangle strand to each side and increases the number of punctures by two. If ϕ_1 corresponds to pink and ϕ_2 corresponds to blue, then this is a picture of the $\sim_{s_b^1}$ version of perturbation. Switching the colors would give a picture for the $\sim_{s_b^2}$ version.

Given $(\phi_1, \phi_2), (\phi_1', \phi_2') \in \mathtt{Alg}^{(3,1)}$, we now define the relation $\sim_{s_g}$. We note that ϕ_i' will be a map from $\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow F_{g+b}$ while ϕ_i will be a map from $\pi_1(\Sigma_{g+1} - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow F_{g+1+b}$. We say $(\phi_1, \phi_2) \sim_{s_g} (\phi_1', \phi_2')$ if $\phi_i(a_j) = \phi_i'(a_j)$, $\phi_i(b_j) = \phi_i'(b_j)$ and $\phi_i(p_k) = \phi_i'(p_k)$ for $i = 1, 2$, $j = 1, \dots, g$ and $k = 1, \dots, 2b$ (where we are identifying F_{g+b} naturally as a subset of F_{g+1+b}, identifying the h_i generators in F_{g+b} with h_i in F_{g+1+b} and similarly with t_i), and the rest of the generators are mapped as follows:

$$\phi_1(a_{g+1}) = h_{g+1}, \quad \phi_1(b_{g+1}) = 1, \quad \phi_2(a_{g+1}) = 1, \quad \phi_2(b_{g+1}) = h_{g+1}.$$

Finally, we define $\sim_{s_b^1}$ and $\sim_{s_b^2}$, which correspond to perturbation of the tangle strands. See Figure 32. Suppose now that $b > 0$. There are two such operations since we can either push a tangle strand from the side corresponding to ϕ_1 across Σ_g, or we can push a tangle strand corresponding to ϕ_2. The motivation for this comes from investigating Figure 32 and imagining applying the operation σ from the proof of Theorem 3.1 to the before and after parts of the figure. Let $(\phi_1, \phi_2), (\phi_1', \phi_2') \in \mathtt{Alg}^{(3,1)}$. We note that ϕ_i' will be a map from $\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow F_{g+b}$ while ϕ_i will be a map from $\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}, p_{2b+1}, p_{2b+2}\}, *) \twoheadrightarrow F_{g+b+1}$. Assume without loss of generality that $\phi_1'(p_{2b}) = \phi_2'(p_{2b}) = t_b$. (We can do this because we first mod out by mapping class group elements; see the proof of Theorem 3.1, Claim 1, for details.) We write $(\phi_1, \phi_2) \sim_{s_b^1} (\phi_1', \phi_2')$ if

$$\begin{aligned} &\phi_1(p_{2b}) = t_{b+1}, \quad &&\phi_1(p_{2b+1}) = (t_{b+1})^{-1}, \quad &&\phi_1(p_{2b+2}) = t_b, \\ &\phi_2(p_{2b}) = t_b, &&\phi_2(p_{2b+1}) = t_{b+1}, &&\phi_2(p_{2b+2}) = (t_{b+1})^{-1}, \end{aligned}$$

and ϕ_i and ϕ_i' agree for all other elements in the generating sets (suitably identifying the groups) for $i = 1, 2$. We similarly define $\sim_{s_b^2}$ by swapping the roles of the indices 1 and 2.

Let $\sim$ denote the equivalence relation on $\mathtt{Alg}^{(3,1)}$ generated by $\sim_h$, $\sim_m$, $\sim_{s_g}$, $\sim_{s_b^1}$ and $\sim_{s_b^2}$.

Theorem 3.2 *The map $(M, L)\colon \mathtt{Alg}^{(3,1)} \to \mathtt{Man}^{(3,1)}$ descends to $\mathtt{Alg}^{(3,1)}/{\sim}$ and the resulting map is a bijection.*

Proof As in the proof of Theorem 3.1, we consider an intermediate set $\mathtt{Diag}^{(3,1)}$ whose elements are *diagrams*, that is, tuples $(\Sigma_g, \alpha, \beta, S_\alpha, S_\beta)$ where α, β are cut systems on Σ_g, and S_α, S_β are shadow diagrams for trivial tangles with endpoints $\{p_1, \dots, p_{2b}\}$ (which are all only considered up to isotopy). In other words, (α, S_α) is a curve-and-arc system for one tangle and handlebody, and (β, S_β) is a curve-and-arc system for the other. (Refer to the beginning of Section 2 for the definition of *curve-and-arc system*.) Then the map (M, L) factors as

$$\begin{tikzcd}
\mathtt{Alg}^{(3,1)} \arrow[rr, "{(M,L)}"] \arrow[dr, "D"'] & & \mathtt{Man}^{(3,1)} \\
& \mathtt{Diag}^{(3,1)} \arrow[ur, "R"'] &
\end{tikzcd}$$

The map $R\colon \mathtt{Diag}^{(3,1)} \to \mathtt{Man}^{(3,1)}$ is the topological realization of a diagram $(\Sigma_g, \alpha, \beta, S_\alpha, S_\beta)$, where we cross Σ_g with an interval, glue disks on the respective sides to α and β, glue three balls to the resulting sphere boundary components to obtain two handlebodies, and then push the interiors of the shadow arcs in S_α and S_β into their respective handlebody to obtain two tangles. The map $D\colon \mathtt{Alg}^{(3,1)} \to \mathtt{Diag}^{(3,1)}$ is the construction of (M, L) using Theorem 2.10, but where we stop at just a diagram with curves α and arcs S_α corresponding to ϕ_1 and curves β and arcs S_β corresponding to ϕ_2.

In the proof of Theorem 3.1 (see Claim 1) we used the fact that two handlebodies bounding a given surface are equivalent if and only if their diagrams differ by handleslides. In this proof the following fact will take its place: two tangles in a handlebody are isotopic (fixing their boundary points) if and only if their curve-and-arc systems are related by a sequence of isotopies and slides. This folklore fact appears, for instance, in [32, Proposition 3.1; 30, Proposition 5.2], both citing [21] for the proof idea. Our topological input theorem can now be translated into the following diagrammatic statement: two diagrams of the same link in a 3-manifold are related by a sequence of perturbations, deperturbations, and moves from the above fact. In other words, these are diagrammatic equivalence relations which have the algebraic counterparts $\sim_h$, $\sim_m$, $\sim_{s_g}$, $\sim_{s_b^1}$ and $\sim_{s_b^2}$ as described above.

The rest of the proof then follows in similar fashion as before; mod out the map D by these diagrammatic equivalence relations, and then show each time that the map factors through and a bijection between quotients is achieved. We leave the details to the reader. Then we use both our topological input theorem from this section and the Reidemeister–Singer theorem (as previously) to achieve the result. □

Remark 3.3 Theorem 3.2 is in fact a generalization of Theorem 3.1 where we take the links to be empty. Note also that the number of components of the link resulting from a given pair (ϕ_1, ϕ_2) can easily be read off from the maps ϕ_1 and ϕ_2. We can thereby describe the partition of the set $\mathtt{Alg}^{(3,1)}$ so that the different equivalence classes correspond in the bijection above to manifolds with links of $0, 1, 2 \dots$ components (in particular, $\sim$ respects this partition). In particular, Theorem 3.1 is recovered by restricting to the case where $b = 0$.

4 Closed 4-manifolds and bridge trisected surfaces

Here we continue the translation of topology into algebra that we started in the previous section, but moving up a dimension. The proofs in this section follow similarly to those in the previous, and thus we omit some of the details.

4A Closed 4-manifolds

In 2016 Gay and Kirby introduced *trisections* of 4-manifolds, which can be seen as 4-dimensional analogues of Heegaard splittings.

Definition 4.1 (trisection of a 4-manifold [8]) A $(g; k_1, k_2, k_3)$-*trisection* of a smooth, closed, connected, oriented 4-manifold X^4 is a decomposition $X^4 = X_1 \cup X_2 \cup X_3$ with the following properties:

(1) Each X_i is a 4-dimensional 1-handlebody, that is, diffeomorphic to $\natural^{k_i}(S^1 \times B^3)$. If $k_i = 0$, we interpret this boundary connected sum as B^4.

(2) The X_i's intersect pairwise in genus g handlebodies; that is, their pairwise intersections are diffeomorphic to $H_g := \natural^g(S^1 \times B^2)$.

(3) The triple intersection of the X_i's is a genus g oriented surface (denoted by Σ_g), called the *central surface*.

If $k_1 = k_2 = k_3 =: k$, we call the trisection *balanced*, and denote this by $(g; k)$.

A trisection is determined by its *spine*, namely the central surface along with the three handlebodies the surface bounds, as there is a unique way, up to diffeomorphism, to fill this in with 4-dimensional 1-handlebodies [28]. Every smooth, closed, connected, oriented 4-manifold admits a trisection, and a trisection of a given 4-manifold is unique up to stabilization (see [8] for a proof of this result and the precise definition of stabilization).

Topological input theorem [8] *Any pair of trisections of a fixed* 4*-manifold become isotopic after some number of stabilizations.*

Technically, there are two notions of stabilization: a balanced one (which increases the genus of the central surface by three) and an unbalanced one (which increases the genus by one). We will assume all trisections are balanced and work only with balanced stabilizations. We can make this assumption without loss of generality as any unbalanced trisection can be made balanced with some number of unbalanced stabilizations. Note that we could easily expand the algebraic relations included in this section to include those which would correspond to unbalanced stabilizations, but for the sake of simplicity of notation we have not done this.

Let `Man`^{4} be the set of closed, connected, oriented, smooth 4-manifolds up to orientation-preserving diffeomorphism. Let `Alg`^{4} denote the set of triples (ϕ_1, ϕ_2, ϕ_3) of surjective homomorphisms $\pi_1(\Sigma_g, *) \twoheadrightarrow F_g$ for some integer g such that $M(\phi_1, \phi_2)$, $M(\phi_2, \phi_3)$ and $M(\phi_3, \phi_1)$ are all diffeomorphic to $\#^k(S^1 \times S^2)$

for some integer k, where M is as in Section 3A. By the Poincaré conjecture [35], this is equivalent to the property that the pushout of ϕ_i and ϕ_j for $i \neq j$ are (necessarily finitely generated) free groups. Note that here, an individual ϕ is a bounding homomorphism for the special case $b = 0$ which satisfies the additional pushout property, and a triple $(\phi_1, \phi_2, \phi_3) \in \mathtt{Alg}^4$ is a *group trisection* of the fundamental group of a 4-manifold, as described in [1] and Section 1.

We have a map $X\colon \mathtt{Alg}^4 \to \mathtt{Man}^4$ just as in [1]. Namely, given $(\phi_1, \phi_2, \phi_3) \in \mathtt{Alg}^4$, identify the handlebodies $H(\phi_1), H(\phi_2), H(\phi_3)$ along their common boundary Σ_g. Now the three 3-manifolds $M(\phi_i, \phi_j) = H(\phi_i) \cup_{\Sigma_g} H(\phi_j)$ for $i \neq j$ are all diffeomorphic to some $\#^k(S^1 \times S^2)$ by the assumption that the pairwise pushouts of the ϕ_i and ϕ_j are free groups. Therefore we may glue in three 4-dimensional 1-handlebodies (uniquely by [28]) to obtain a smooth, closed 4-manifold $X(\phi_1, \phi_2, \phi_3)$. We orient $M(\phi_1, \phi_2) \subset X(\phi_1, \phi_2, \phi_3)$ as in Section 3A and we orient X so that the orientation restricted to the 4-dimensional 1-handlebody that is glued to $M(\phi_1, \phi_2)$ induces this orientation on $M(\phi_1, \phi_2)$.

We define the relations $\sim_h$ and $\sim_m$ in an analogous fashion as was done for $\mathtt{Alg}^3$ in Section 3A. The stabilization relation $\sim_s$ on $\mathtt{Alg}^4$ is defined as follows. Here ϕ'_i will be a map from $\pi_1(\Sigma_g, *) \twoheadrightarrow F_g$ while ϕ_i will be a map from $\pi_1(\Sigma_{g+3}, *) \twoheadrightarrow F_{g+3}$. Let a_i, b_i be the generators of $\pi_1(\Sigma_g, *)$ (and, abusing notation, $\pi_1(\Sigma_{g+3}, *)$), and h_i be the generators of F_g (and, abusing notation, F_{g+3}). We say $(\phi_1, \phi_2, \phi_3) \sim_s (\phi'_1, \phi'_2, \phi'_3)$ if $\phi_i(a_j) = \phi'_i(a_j)$ and $\phi_i(b_j) = \phi'_i(b_j)$ for $i = 1, 2, 3$ and $j = 1, \dots, g$ (where we are identifying F_g naturally as a subset of F_{g+3}), and the rest of the generators are mapped as follows:

$$\begin{array}{lll}
\phi_1(a_{g+1}) = h_{g+1}, & \phi_2(a_{g+1}) = h_{g+1}, & \phi_3(a_{g+1}) = 1, \\
\phi_1(b_{g+1}) = 1, & \phi_2(b_{g+1}) = 1, & \phi_3(b_{g+1}) = h_{g+1}, \\
\phi_1(a_{g+2}) = h_{g+2}, & \phi_2(a_{g+2}) = 1, & \phi_3(a_{g+2}) = h_{g+2}, \\
\phi_1(b_{g+2}) = 1, & \phi_2(b_{g+2}) = h_{g+2}, & \phi_3(b_{g+2}) = 1, \\
\phi_1(a_{g+3}) = 1, & \phi_2(a_{g+3}) = h_{g+3}, & \phi_3(a_{g+3}) = h_{g+3}, \\
\phi_1(b_{g+3}) = h_{g+3}, & \phi_2(b_{g+3}) = 1, & \phi_3(b_{g+3}) = 1.
\end{array}$$

This is an algebraic analogue of the topological operation of stabilizing a trisection, just as in Section 3A where we presented the analogous notion for stabilizations of Heegaard splittings. As before, we let the equivalence relation $\sim_s$ be the symmetrization of the relation $\sim_s$.

As we now have three maps, we must define one more relation, which does not have a counterpart in Section 3, corresponding to cyclically permuting the "colors" of the curves on a trisection diagram (that is, cyclically permuting the roles of the cut system curves α, β and γ). We say $(\phi_1, \phi_2, \phi_3) \sim_c (\phi'_1, \phi'_2, \phi'_3)$ if $\phi_1 = \phi'_2$, $\phi_2 = \phi'_3$ and $\phi_3 = \phi'_1$.

Let $\sim$ denote the equivalence relation on $\mathtt{Alg}^4$ generated by $\sim_h$, $\sim_m$, $\sim_s$ and $\sim_c$. A result very similar to the following theorem is stated in [1, Theorem 5].

Theorem 4.2 (compare with [1]) *The map $X\colon \mathtt{Alg}^4 \to \mathtt{Man}^4$ descends to $\mathtt{Alg}^4/{\sim}$ and the resulting map is a bijection.*

Proof The proof proceeds analogously to the discussion of the Heegaard splittings of closed 3-manifolds, where the role of the Reidemeister–Singer theorem is taken on by the uniqueness of trisections up to stabilization from [8]. □

Remark 4.3 We can algorithmically determine if the pushout of such a pair ϕ_i and ϕ_j is in fact a free group. Namely, we can construct the corresponding 3-manifold and algorithmically check if it is a (possibly empty) connected sum of copies of $S^1 \times S^2$ (see for example [27]). We are unaware if there is a more direct algebraic method to verify this condition.

4B Bridge trisected surfaces in 4-manifolds

Finally we turn to the setting of knotted surfaces in 4-manifolds. For us, a *knotted surface* is a closed surface smoothly embedded in a smooth, closed, connected, oriented 4-manifold. In particular, our surfaces are not necessarily orientable or connected.

In 2018 Meier and Zupan showed that knotted surfaces in trisected 4-manifolds can always be isotoped into a compatibly trisected surface. This inherited decomposition is called a *bridge trisection*.

Definition 4.4 (bridge trisection of a knotted surface [31; 32]) A $(g;k_1,k_2,k_3;b;c_1,c_2,c_3)$*-bridge trisection* of a knotted surface S in a 4-manifold X^4 is a decomposition

$$(X^4,S)=(X_1,\mathscr{D}_1)\cup(X_2,\mathscr{D}_2)\cup(X_3,\mathscr{D}_3)$$

with the following properties:

(1) The decomposition $X^4=X_1\cup X_2\cup X_3$ is a $(g;k_1,k_2,k_3)$-trisection of X^4.

(2) Each $\mathscr{D}_i$ is a boundary parallel collection of c_i disks in X_i.

(3) The $\mathscr{D}_i$'s intersect pairwise in trivial b-strand tangles (denoted by T_b) in the handlebodies which are the pairwise intersections of the X_i's.

Just as before, if $k_1=k_2=k_3=:k$, we replace these parameters with just one k and call the trisection *balanced*. We assume again that all trisections are balanced (with respect to the k_i parameter; the c_i may be different).

Pairwise, the tangles T_b form unlinks. A bridge trisection is determined by these three tangles, as there is a unique way to smoothly cap off unlinks in $\#^k(S^1\times S^2)$ with disks [32]. A knotted surface in a trisected 4-manifold is said to be in *bridge position* if its intersection with the trisection of the 4-manifold results in a bridge trisection. Knotted surfaces can always be isotoped to be in bridge position, and this is unique up to perturbation (for the proof of this result and the precise definition of pertubation, see [31; 32; 18]).

Topological input theorem [31; 32; 18] *Any pair of bridge trisections for a smoothly embedded surface in a fixed underlying trisection of a* 4*-manifold become isotopic after some number of perturbations.*

Throughout this section, let a_j, b_j and p_k denote the generators of $\pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *)$, where a_j, b_j are the surface generators (for $j = 1, \dots, g$) and p_k are the puncture generators (for $k = 1, \dots, 2b$), and let h_j, t_l denote the generators of F_{g+b} (for $l = 1, \dots, b$). Given a bounding homomorphism

$$\phi\colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow \langle t_1, \dots, t_b, h_1, \dots, h_g \rangle,$$

we have an associated homomorphism, which we call the *associated closed bounding homomorphism* for ϕ, denoted by

$$\overline{\phi}\colon \pi_1(\Sigma_g, *) \twoheadrightarrow \langle h_1, \dots, h_g \rangle,$$

which is given by postcomposing ϕ by the map quotienting out all of the t_l (and again calling the images of the h_j by h_j) and then sending all a_j and b_j (now thought of as in $\pi_1(\Sigma_g, *)$) to the resulting elements of the free group generated by the h_j. Note that since ϕ is a bounding homomorphism, $\overline{\phi}$ is also.

Let $\texttt{Man}^{(4,2)}$ denote the set of closed, connected, oriented, smooth 4-manifolds X together with a union of closed (potentially nonorientable or disconnected) surfaces $S \subset X$ modulo orientation-preserving diffeomorphisms preserving the surfaces setwise. Let $\texttt{Alg}^{(4,2)}$ denote the set of triples (ϕ_1, ϕ_2, ϕ_3) of bounding homomorphisms $\phi_1, \phi_2, \phi_3\colon \pi_1(\Sigma_g - \{p_1, \dots, p_{2b}\}, *) \twoheadrightarrow F_{g+b}$ such that

1. the pushouts of pairs of the associated closed bounding homomorphisms $\overline{\phi}_i, \overline{\phi}_j$ for $i \neq j$ are all free groups, and
2. the pushouts of pairs of the bounding homomorphisms ϕ_i, ϕ_j are all free groups of rank equal to the sum of the rank of the pushout of $\overline{\phi}_i$ and $\overline{\phi}_j$ plus the number of components of (ϕ_1, ϕ_2) as in Remark 3.3.

In other words, an element in $\texttt{Alg}^{(4,2)}$ is a *group trisection* of a knotted surface group, which could also be described with the following commutative diagram, where every face is a pushout and every homomorphism is surjective (analogous to that in Section 1):

$$\begin{tikzcd}
 & \pi_1(H_g - T_b, *) \arrow[rr, ->>] \arrow[dd, ->>] & & \pi_1(X_1 - \mathscr{D}_1, *) \arrow[dd, ->>] \\
\pi_1(\Sigma_g - \{2b \text{ pts}\}, *) \arrow[ur, ->>] \arrow[rr, ->>] \arrow[dd, ->>] & & \pi_1(H_g - T_b, *) \arrow[ur, ->>] \arrow[dd, ->>] & \\
 & \pi_1(X_3 - \mathscr{D}_3, *) \arrow[rr, ->>] & & \pi_1(X^4 - S, *) \\
\pi_1(H_g - T_b, *) \arrow[ur, ->>] \arrow[rr, ->>] & & \pi_1(X_2 - \mathscr{D}_2, *) \arrow[ur, ->>] &
\end{tikzcd}$$

We will need the following lemma in the construction of the map $(X, S)\colon \texttt{Alg}^{(4,2)} \to \texttt{Man}^{(4,2)}$.

Lemma 4.5 (the free group characterizes unlinks) *The fundamental group detects the unlink in the connected sum* $\#^k(S^1 \times S^2)$. *That is, if* $L = L_1 \sqcup \cdots \sqcup L_n \hookrightarrow \#^k(S^1 \times S^2)$ *is an* n*-component link with* $\pi_1(S^3 - L, *) \cong F_{n+k}$ *a free group on* $n+k$ *generators, then* L *is the* n*-component unlink.*

Proof This proof is a generalization of the argument in [17, Theorem 1] for unlinks in S^3. Any abelian subgroup of a free group is cyclic, so the meridian-longitude generators of the torus around a link component span an abelian group in the free $\pi_1(S^1 \times S^2 - L, *)$. Since the meridians generate the first homology of a link complement, we know that the longitudes have to be nullhomotopic in the link exterior. Now apply the loop theorem to obtain disjointly embedded disks recognizing the split unlink. □

Remark 4.6 There exist nontrivial links $L \subset \#^k(S^1 \times S^2)$ whose complement has free fundamental group (but not free of the same rank as the group of the unlink). For example, the core curve of one of the solid tori in the standard genus 1 Heegaard splitting of $S^1 \times S^2$ has complement homotopy equivalent to the other solid torus; that is, the fundamental group of its complement is free on one generator. On the other hand, the fundamental group of the complement of the unknot in $S^1 \times S^2$ is free on two generators (the meridian of the unknot and the generator of $\pi_1(S^1 \times S^2, *)$).

Remark 4.7 An analogous statement to Lemma 4.5 is false for higher dimensional links. Cochran [6] exhibited links of 2-spheres in 4-spheres whose groups are free (but not free on their meridians).

Now we define the map $(X, S)\colon \mathtt{Alg}^{(4,2)} \to \mathtt{Man}^{(4,2)}$ as follows. Given $(\phi_1, \phi_2, \phi_3) \in \mathtt{Alg}^{(4,2)}$, let $H(\phi_1)$, $H(\phi_2)$ and $H(\phi_3)$ be the handlebodies bounding Σ_g that result from the application of Theorem 2.10, and further let $T(\phi_i) \subset H(\phi_i)$ for $i = 1, 2, 3$ be the resulting trivial tangles in these handlebodies. We glue the handlebodies $H(\phi_1)$, $H(\phi_2)$ and $H(\phi_3)$ together along the common surface Σ_g and obtain a closed, oriented, smooth 4-manifold $X = X(\overline{\phi}_1, \overline{\phi}_2, \overline{\phi}_3)$ as in Section 4A. (Here we have used the condition that the pairwise pushouts of the associated closed bounding homomorphisms are free groups to ensure that we obtain $\#^k(S^1 \times S^2)$ as the result of gluing two of the handlebodies together, and thus we can cap off with 4-dimensional 1-handlebodies.) The surface S in X is obtained by considering the unions of the tangles $T(\phi_i) \cup_{\{p_1,\dots,p_{2b}\}} T(\phi_j) \subset H(\phi_i) \cup_{\Sigma_g} H(\phi_j)$ for $i \neq j$. Since the pushout of ϕ_i and ϕ_j is a free group of the appropriate rank, by Lemma 4.5 we know that the unions of these tangles are all unlinks. Therefore, we can take disjoint bridge disks bounding $T(\phi_i) \cup_{\{p_1,\dots,p_{2b}\}} T(\phi_j)$ in $H(\phi_i) \cup_{\Sigma_g} H(\phi_j)$ and push these into the 4-dimensional 1-handlebody bounding $H(\phi_i) \cup_{\Sigma_g} H(\phi_j)$ in X. Then the union of these three sets of disks (with one set in each 4-dimensional 1-handlebody) is the knotted surface S.

We now define the analogues of the various relations from Sections 3B and 4A in this setting. The definitions of $\sim_h$ and $\sim_m$ are exactly analogous to the definitions in Section 3B. Just as stabilization appeared in Section 3B as several different relations — namely, $\sim_{s_g}$ for changing the genus of the surface and $\sim_{s_b^1}$, $\sim_{s_b^2}$ for perturbing each of the two tangles — in our current setting, stabilization will also manifest as several different relations. Here we will have $\sim_{s_g}$ corresponding to changing the genus of the surface, and $\sim_{s_b^1}$, $\sim_{s_b^2}$ and $\sim_{s_b^3}$ corresponding to changing the number of strands in the tangles.

Given $(\phi_1,\phi_2,\phi_3),(\phi'_1,\phi'_2,\phi'_3)\in \texttt{Alg}^{(4,2)}$, we now define $(\phi_1,\phi_2,\phi_3)\sim_{s_g}(\phi'_1,\phi'_2,\phi'_3)$. We note that ϕ'_i will be a map from $\pi_1(\Sigma_g-\{p_1,\dots,p_{2b}\},*)\twoheadrightarrow F_{g+b}$ while ϕ_i will be a map from

$$\pi_1(\Sigma_{g+1}-\{p_1,\dots,p_{2b}\},*)\twoheadrightarrow F_{g+3+b}.$$

We say $(\phi_1,\phi_2,\phi_3)\sim_{s_g}(\phi'_1,\phi'_2,\phi'_3)$ if $\phi_i(a_j)=\phi'_i(a_j)$, $\phi_i(b_j)=\phi'_i(b_j)$ and $\phi_i(p_k)=\phi'_i(p_k)$ for $i=1,2,3$, $j=1,\dots,g$ and $k=1,\dots,2b$ (where we are identifying F_{g+b} naturally as a subset of F_{g+3+b}, identifying the h_i generators in F_{g+b} with h_i in F_{g+3+b} and similarly with t_i), and the rest of the generators are mapped just as in the definition of $\sim_s$ in Section 4A.

Now we discuss the algebraic version of perturbation in this setting, which is analogous to the definitions of $\sim_{s_b^1}$, $\sim_{s_b^2}$ in Section 3B. Suppose that $b>0$. This definition is motivated by the pictures of perturbation shown in Figures 33 and 34, where two arcs of different colors are banded together to create three new arcs, one of each color. There are three such operations to consider depending on how we cyclically permute the colors. For each operation, there are two cases under which the operation can occur: either the two arcs to be banded together share an endpoint, or they do not.

Let $(\phi_1,\phi_2,\phi_3),(\phi'_1,\phi'_2,\phi'_3)\in \texttt{Alg}^{(4,2)}$. In both cases ϕ'_i will be a map from $\pi_1(\Sigma_g-\{p_1,\dots,p_{2b}\},*)\twoheadrightarrow F_{g+b}$ while ϕ_i will be a map from $\pi_1(\Sigma_g-\{p_1,\dots,p_{2b},p_{2b+1},p_{2b+2}\},*)\twoheadrightarrow F_{g+b+1}$. If the arcs to be banded together share an endpoint, then assume without loss of generality that $\phi'_1(p_{2b})=\phi'_2(p_{2b})=t_b$. (We can do this because we first mod out by mapping class group elements; see the proof of Theorem 3.1, Claim 1, for details.) We write $(\phi_1,\phi_2,\phi_3)\sim_{s_b^1}(\phi'_1,\phi'_2,\phi'_3)$ if

$$\begin{aligned}
&\phi_1(p_{2b})=t_{b+1}, && \phi_1(p_{2b+1})=(t_{b+1})^{-1}, && \phi_1(p_{2b+2})=t_b,\\
&\phi_2(p_{2b})=t_{b+1}, && \phi_2(p_{2b+1})=(t_{b+1})^{-1}, && \phi_2(p_{2b+2})=t_b,\\
& && \phi_3(p_{2b+1})=t_{b+1}, && \phi_3(p_{2b+2})=(t_{b+1})^{-1},
\end{aligned}$$

and ϕ_i and ϕ'_i agree for all other elements in the generating sets (suitably identifying the groups) for $i=1,2,3$. See Figure 33. We similarly define $\sim_{s_b^2}$ and $\sim_{s_b^3}$ in this case by swapping the roles of the indices.

If the arcs to be banded together do not share an endpoint, then assume without loss of generality that $\phi'_1(p_{2b-1})=t_b$ and $\phi'_2(p_{2b})=t_b$. We write $(\phi_1,\phi_2,\phi_3)\sim_{s_b^1}(\phi'_1,\phi'_2,\phi'_3)$ if

$$\begin{aligned}
&\phi_1(p_{2b-1})=t_{b+1}, && && \phi_1(p_{2b+1})=(t_{b+1})^{-1}, && \phi_1(p_{2b+2})=t_b,\\
& && \phi_2(p_{2b})=t_{b+1}, && \phi_2(p_{2b+1})=(t_{b+1})^{-1}, && \phi_2(p_{2b+2})=t_b,\\
& && && \phi_3(p_{2b+1})=t_{b+1}, && \phi_3(p_{2b+2})=(t_{b+1})^{-1},
\end{aligned}$$

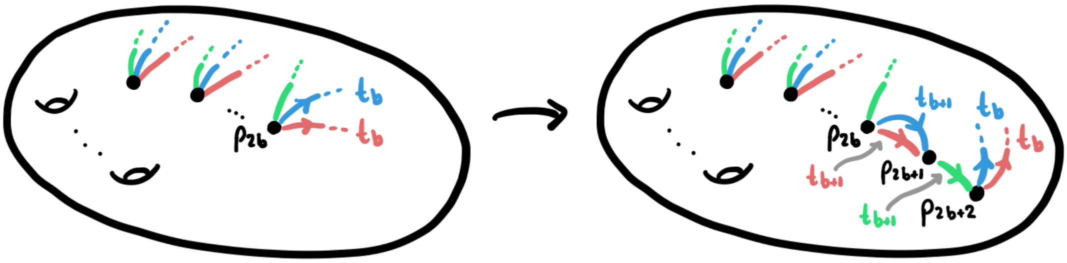


Figure 33: Performing a perturbation, where two arcs of different colors with the same endpoint are banded together. This creates three new tangle strands, one of each color, and increases the number of punctures by two. If ϕ_1 corresponds to pink, ϕ_2 corresponds to blue, and ϕ_3 corresponds to green, then this is a picture of the $\sim_{s_b^1}$ version of perturbation. Cyclically permuting the colors gives a picture for the $\sim_{s_b^2}$ and $\sim_{s_b^3}$ version.

and ϕ_i and ϕ_i' agree for all other elements in the generating sets (suitably identifying the groups) for $i = 1, 2, 3$. See Figure 34. We similarly define $\sim_{s_b^2}$ and $\sim_{s_b^3}$ in this case by swapping the roles of the indices.

Finally, we must again include a relation corresponding to cyclically permuting the "colors" of the curves and arcs on a trisection diagram (that is, cyclically permuting the roles of the three curve-and-arc systems). We say $(\phi_1, \phi_2, \phi_3) \sim_c (\phi_1', \phi_2', \phi_3')$ if $\phi_1 = \phi_2'$, $\phi_2 = \phi_3'$ and $\phi_3 = \phi_1'$. Then let $\sim$ denote the equivalence relation on $\mathtt{Alg}^{(4,2)}$ generated by $\sim_h$, $\sim_m$, $\sim_{s_g}$, $\sim_{s_b^1}$, $\sim_{s_b^2}$, $\sim_{s_b^3}$ and $\sim_c$.

Theorem 4.8 *The map $(X, S)\colon \mathtt{Alg}^{(4,2)} \to \mathtt{Man}^{(4,2)}$ descends to $\mathtt{Alg}^{(4,2)}/{\sim}$ and the resulting map is a bijection.*

Proof As in the proofs of Theorems 3.1, 3.2 and 4.2, we consider an intermediate set $\mathtt{Diag}^{(4,2)}$ whose elements are *trisection diagrams*, that is, tuples $(\Sigma_g, \alpha, \beta, \gamma, S_\alpha, S_\beta, S_\gamma)$ where α, β, γ are cut systems

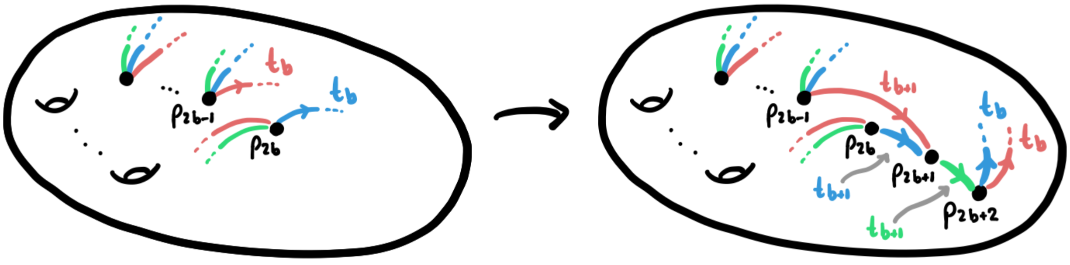


Figure 34: Performing a perturbation, where two arcs of different colors with different endpoints are banded together. This creates three new tangle strands, one of each color, and increases the number of punctures by two. If ϕ_1 corresponds to pink, ϕ_2 corresponds to blue, and ϕ_3 corresponds to green, then this is a picture of the $\sim_{s_b^1}$ version of perturbation. Cyclically permuting the colors gives a picture for the $\sim_{s_b^2}$ and $\sim_{s_b^3}$ version.

on Σ_g and S_α, S_β, S_γ are shadow diagrams for trivial tangles with endpoints $\{p_1, \dots, p_{2b}\}$ (which are all only considered up to isotopy). In other words, (α, S_α) is a curve-and-arc system for one tangle and handlebody, and similarly for the other two pairs. (Refer to the beginning of Section 2 for the definition of *curve-and-arc system.*) Then the map (X, S) factors as

$$\begin{tikzcd}
\mathtt{Alg}^{(4,2)} \arrow[rr, "{(X,S)}"] \arrow[dr, "D"'] & & \mathtt{Man}^{(4,2)} \\
& \mathtt{Diag}^{(4,2)} \arrow[ur, "R"'] &
\end{tikzcd}$$

The map $R\colon \mathtt{Diag}^{(4,2)} \to \mathtt{Man}^{(4,2)}$ is the topological realization of a diagram $(\Sigma_g, \alpha, \beta, \gamma, S_\alpha, S_\beta, S_\gamma)$, where we cross Σ_g with a disk, glue disks on the respective sides to α, β and γ, glue 3-balls to the resulting sphere boundary components to obtain three handlebodies, and then push the interiors of the shadow arcs in S_α, S_β and S_γ into their respective handlebody to obtain three tangles which are pairwise unlinks. Since the pairwise unions of the handlebodies are diffeomorphic to $\#^k (S^1 \times S^2)$, we can fill these in uniquely with 4-dimensional 1-handlebodies. Then cap off the unlinks with disks in the pairwise unions of the handlebodies, and then push these disks into the 4-dimensional 1-handlebodies bounded by these unions to create a knotted surface. The map $D\colon \mathtt{Alg}^{(4,2)} \to \mathtt{Diag}^{(4,2)}$ is the construction of (X, S) using Theorem 2.10, but where we stop at just a diagram with curves α and arcs S_α corresponding to ϕ_1, curves β and arcs S_β corresponding to ϕ_2, and curves γ and arcs S_γ corresponding to ϕ_3.

Our topological input theorem can now be translated into the following diagrammatic statement: two diagrams of the same knotted surface in a 4-manifold are related by a sequence of isotopies, slides, perturbations and depurtubations, as in the proof of Theorem 3.2, and, additionally, cyclically permuting the colors. In other words, these are diagrammatic equivalence relations which have the algebraic counterparts $\sim_h$, $\sim_m$, $\sim_{s_g}$, $\sim_{s_b^1}$, $\sim_{s_b^2}$, $\sim_{s_b^3}$ and $\sim_c$ as described above.

The rest of the proof then follows in similar fashion as before; mod out the map D by these diagrammatic equivalence relations, and show each time that the map factors through and a bijection between quotients is achieved. Then by our topological input theorem from this section, along with that from Section 4A, we achieve the result. □

One consequence of Theorem 4.8 is the following corollary. By the Gordon–Luecke theorem, (one-dimensional) knots in S^3 are determined by the oriented homeomorphism type of their complements [14], but the same is not true for knotted surfaces in S^4; see for instance [12; 42; 26]. However, the extra information contained in a group trisection *is* enough to distinguish knotted surfaces.

Corollary 4.9 *Although smoothly knotted surfaces in the* 4*-sphere cannot always be distinguished by their fundamental groups* (*or even their complements*), *they* ***can*** *always be distinguished by the group trisections of their fundamental groups.*

5 Examples and consequences

In this section we discuss some examples and consequences of our results, including examples of nonequivalent group trisections of the same group, an algebraic version of the smooth unknotting conjecture, a group-theoretic characterization of knot groups, and musings on algorithmic decidability. We will frequently make use of the following definition.

Definition 5.1 (fundamental group of pairs or triples of maps) The *fundamental group* of (ϕ_1, ϕ_2) (in $\mathtt{Alg}^3$ or $\mathtt{Alg}^{(3,1)}$) is their pushout. Similarly, the *fundamental group* of (ϕ_1, ϕ_2, ϕ_3) (in $\mathtt{Alg}^4$ or $\mathtt{Alg}^{(4,2)}$) is the group that results from pushing out the three homomorphisms.

Alternatively, the fundamental group of (ϕ_1, ϕ_2) or (ϕ_1, ϕ_2, ϕ_3) is the fundamental group of the corresponding topological space constructed from these maps.

5A Nonequivalent group trisections of the same group

Here we present a few examples of nonequivalent group trisections of the same group. Specifically, we provide nonequivalent triples (ϕ_1, ϕ_2, ϕ_3) in $\mathtt{Alg}^{(4,2)}/{\sim}$ with the same fundamental group, where the nonequivalence follows from known topological results about the knotted surfaces realizing these groups as their fundamental group. We leave many details of the calculations to the reader, but see [3; 37] for a more thorough treatment of the following examples, as well as [23, Section 4.1] for a description of how to calculate fundamental groups from tri-plane diagrams.

Note that in the case of closed 4-manifolds, there are as many (nonequivalent) group trisections of the trivial group as there are exotic smooth structures on simply connected 4-manifolds. As a given 4-manifold with an embedded surface can admit nonequivalent group trisections of the same group, which correspond to different surfaces, the theory of group trisections for (4-manifold, knotted surface) pairs is even richer than that for 4-manifolds alone.

5A.1 Unknotted $\mathbb{RP}^2$s in S^4 For our first example of nonequivalent group trisections of knotted surface groups, consider the following.

Corollary 5.2 *There exist two elements of $\mathtt{Alg}^{(4,2)}$ which are not equivalent in the quotient $\mathtt{Alg}^{(4,2)}/{\sim}$, but both push out to $\mathbb{Z}/2\mathbb{Z}$. Under topological realization these elements correspond to unknotted $\mathbb{RP}^2$s in S^4.*

Proof As shown in [31, Figure 15], we can represent two unknotted $\mathbb{RP}^2$s in S^4, with Euler numbers ±2, by the tri-plane diagrams in Figures 35 and 36. Both $\mathbb{RP}^2$s produce group trisections of $\mathbb{Z}/2\mathbb{Z}$, but by Theorem 4.8 these group trisections cannot be equivalent as the surfaces are not isotopic. Here we include some details of the construction of these group trisections for the sake of illustration, but see [3,

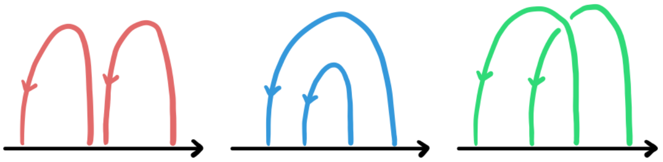


Figure 35: A tri-plane diagram for the unknotted $\mathbb{RP}^2$ in S^4 with Euler number -2.

Section 4.2.3] for a full treatment, including an example of the use of Theorem 2.10 to recover the γ (green) tangle from the group trisection.

As $\mathbb{RP}^2$ is nonorientable, it is not possible to consistently orient the tri-plane diagrams, but we choose arbitrary orientations for each tangle separately in order to write down the group trisection maps; the choice of orientations here will not matter. For both $\mathbb{RP}^2$s, presentations for the groups making up the initial three epimorphisms of the group trisection are as follows, where (abusing notation) T_2^α, T_2^β and T_2^γ are the trivial tangles as shown in Figures 35 and 36:

$$\begin{aligned}\pi_1(\Sigma_0-\{p_1,\dots,p_4\},*)&=\langle p_1,p_2,p_3,p_4\mid p_1p_2p_3p_4=1\rangle,\\ \pi_1(H_0-T_2^\alpha,*)&=\langle x_1,x_2\rangle,\\ \pi_1(H_0-T_2^\beta,*)&=\langle y_1,y_2\rangle,\\ \pi_1(H_0-T_2^\gamma,*)&=\langle z_1,z_2\rangle.\end{aligned}$$

Below we write down the initial three maps (in other words, the elements in $\mathtt{Alg}^{(4,2)}$) for the $\mathbb{RP}^2$ with Euler number -2, corresponding to the α tangle (left/red), β tangle (center/blue) and γ tangle (right/green):

$$\begin{aligned}\alpha\colon&\quad p_1\mapsto x_1,&\quad p_2\mapsto x_1^{-1},&\quad p_3\mapsto x_2,&\quad p_4\mapsto x_2^{-1},\\ \beta\colon&\quad p_1\mapsto y_1,&\quad p_2\mapsto y_2,&\quad p_3\mapsto y_2^{-1},&\quad p_4\mapsto y_1^{-1},\\ \gamma\colon&\quad p_1\mapsto z_1,&\quad p_2\mapsto z_2,&\quad p_3\mapsto z_1^{-1},&\quad p_4\mapsto z_1z_2^{-1}z_1^{-1}.\end{aligned}$$

The only difference between the initial three maps for the $\mathbb{RP}^2$ with Euler number $+2$ and those above is a slight change in the map for the γ tangle (the maps for the α and β tangles are unchanged), corresponding to the crossing change between the two tangles:

$$\gamma\colon\quad p_1\mapsto z_1,\quad p_2\mapsto z_2,\quad p_3\mapsto z_2z_1^{-1}z_2^{-1},\quad p_4\mapsto z_2^{-1}.$$

As the initial maps determine the entire pushout cube, it is sufficient to provide these maps; to recover the other maps in the cube, push out repeatedly until the entire cube is formed. □

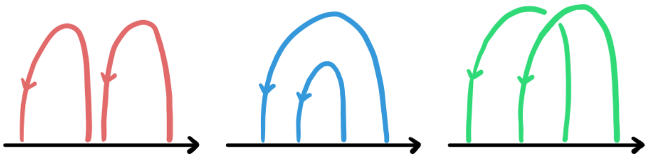


Figure 36: A tri-plane diagram for the unknotted $\mathbb{RP}^2$ in S^4 with Euler number $+2$.

5A.2 Twist-spun torus knots in S^4 A well-known family of twist-spun torus knots gives the following corollary of a combination of [11] and Theorem 4.8.

Corollary 5.3 *There exists a collection of three elements of* $\mathtt{Alg}^{(4,2)}$ *which are not equivalent in the quotient* $\mathtt{Alg}^{(4,2)}/{\sim}$*, but push out to the same group. Under topological realization these elements correspond to knotted spheres in* S^4 *which have isomorphic knotted surface groups.*

Proof Under topological realization, these elements of $\mathtt{Alg}^{(4,2)}$ correspond to the knotted spheres $S^2 \hookrightarrow S^4$ listed below:

- $\tau^2 T_{3,5}$, the 2-twist spin of the $(3,5)$-torus knot;
- $\tau^3 T_{2,5}$, the 3-twist spin of the $(2,5)$-torus knot;
- $\tau^5 T_{2,3}$, the 5-twist spin of the $(2,3)$-torus knot.

Presentations for the corresponding knotted surface groups are given by, for instance,

$$\pi_1(S^4 - \tau^2 T_{3,5}, *) \cong \langle x, y \mid x^3 = y^5, [a^2, x] = 1, [a^2, y] = 1 \rangle,$$

where $a = a(x, y)$ is a meridian of $T_{3,5}$ expressed as a word in the nonmeridional generators x, y coming from the genus 1 Heegaard decomposition of the 3-sphere. Recall that this Heegaard splitting leads to the presentation $\pi_1(S^3 - T_{3,5}, *) \cong \langle x, y \mid x^3 = y^5 \rangle$ of the group of the torus knot complement. If we write the first homology of the complement multiplicatively generated by $H_1(S^4 - \tau^2 T_{3,5}) \cong \langle t \rangle$, then under abelianization the generators of the knot group map to $x \mapsto t^5$, $y \mapsto t^3$, and we pick our orientations so that the meridian maps to $a \mapsto t$.

To write down the presentations for the other knot groups, we will abuse notation and reuse the letters x, y for the generators of the torus knot complement and a for a choice of meridian. With this,

$$\pi_1(S^4 - \tau^3 T_{2,5}, *) \cong \langle x, y \mid x^2 = y^5, [a^3, x] = 1, [a^3, y] = 1 \rangle,$$
$$\pi_1(S^4 - \tau^5 T_{2,3}, *) \cong \langle x, y \mid x^2 = y^3, [a^5, x] = 1, [a^5, y] = 1 \rangle.$$

However all three of these groups $\pi_1(S^4 - \tau^i T_{j,k}, *)$ are abstractly isomorphic to a direct product $\mathbb{Z} \times \mathrm{Dod}^*$ of the integers with the binary dodecahedral group Dod^*; see [43].

See [37, Section 16] for explicit constructions of group trisections of these groups which come from the tri-plane diagrams described in [31, Figure 20]. These group trisections are not equivalent, which follows from Theorem 4.8 together with Gordon's observation [11] that $\tau^2 T_{3,5}$, $\tau^3 T_{2,5}$ and $\tau^5 T_{2,3}$ are all distinct, nonisotopic 2-knots. Gordon shows that the minimal exponent of a meridian's power which is central in the groups $\pi_1(S^4 - \tau^i T_{j,k}, *)$ has to divide the twisting parameter. Pick a generator p of the punctured sphere group and follow it through the group trisection cube; we call its image in the final knotted sphere group p as well. Observe that the image of p is a meridional generator, and thus the smallest power of p that lands these elements in the center is an invariant distinguishing the group, and consequently the group trisections. □

Remark 5.4 Corollary 5.3 generalizes. Take coprime integers $p, q, r \geq 2$ and consider the following knotted 2-spheres:

- $\tau^p T_{q,r}$, the p-twist spin of the (q,r)-torus knot;
- $\tau^q T_{p,r}$, the q-twist spin of the (p,r)-torus knot;
- $\tau^r T_{p,q}$, the r-twist spin of the (p,q)-torus knot.

Since p, q, r are pairwise coprime, all three of these knotted spheres $\tau^i T_{j,k}\colon S^2 \hookrightarrow S^4$ for $\{i,j,k\} = \{p,q,r\}$ have the same fundamental group for their complement. By Zeeman [43], all of them are fibered by the punctured bounded Brieskorn sphere $\Sigma(p,q,r)$, but these fibrations have monodromies with different periods. If the fiber is a homology sphere, one can show that the resulting groups of the 2-knots are abstractly isomorphic. The example for $(p,q,r) = (2,3,5)$ is also discussed by Boyle [4], where he starts with Zeeman's observation that these twist spun 2-knots are fibered by a punctured Poincaré homology sphere $\Sigma(2,3,5)$. Even though they share the same group, these are nonisotopic 2-knots and one way to distinguish them is to look at the smallest power of a meridian which lies in the center of the knot group. This family of knots is further discussed in [11].

Remark 5.5 In [42], Suciu constructs an infinite family of ribbon 2-knots $R_i\colon S^2 \hookrightarrow S^4$ in the 4-sphere, each of which has $\pi_1(S^4 - R_i, *)$ isomorphic to the trefoil group. The 2-knots are pairwise nonisotopic, and can be distinguished by the $\mathbb{Z}\pi_1$-module structure of π_2 of their complements. By bridge trisecting these examples, one can construct an infinite family of elements in $\texttt{Alg}^{(4,2)}$ satisfying the same properties as in Corollary 5.3. See [37, Section 17].

5B Algebraic version of the smooth unknotting conjecture

The smooth unknotting conjecture posits that an embedded sphere in S^4 is smoothly unknotted if and only if the group of its complement is infinite cyclic. Here we use our correspondence in Theorem 4.8 to give an algebraic statement equivalent to the unknotting conjecture. This is in the spirit of a result in [1] which gives a similar group-theoretic conjecture that is equivalent to the smooth 4-dimensional Poincaré conjecture (following the tradition of [40]).

Note that given $(\phi_1, \phi_2, \phi_3) \in \texttt{Alg}^{(4,2)}$, we can tell just from the maps (ϕ_1, ϕ_2, ϕ_3) whether the resulting surface is connected. Similarly, we can determine the Euler characteristic of the resulting surface directly from (ϕ_1, ϕ_2, ϕ_3). We say (ϕ_1, ϕ_2, ϕ_3) is *spherical* if its corresponding surface is a sphere. Just as mentioned in [1], it follows from Theorem 4.2 that there is a purely group-theoretic condition on whether or not $X(\overline{\phi}_1, \overline{\phi}_2, \overline{\phi}_3)$, the 4-manifold built from the associated closed bounding homomorphisms $\overline{\phi}_i$, is orientation-preserving diffeomorphic to the 4-sphere with the standard orientation. We say in this case that (ϕ_1, ϕ_2, ϕ_3) *represents* S^4. Then from Theorem 4.8, it follows that the group-theoretic conjecture below is equivalent to the smooth unknotting conjecture.

Conjecture 5.6 *For every spherical* $(\phi_1,\phi_2,\phi_3)\in \mathtt{Alg}^{(4,2)}$ *which represents* S^4, *we have* $(\phi_1,\phi_2,\phi_3)\sim(s_1,s_2,s_3)$, *where* (s_1,s_2,s_3) *is the group trisection given by maps*

$$s_1,s_2,s_3\colon \pi_1(S^2-\{p_1,p_2\},*)\twoheadrightarrow \mathbb{Z},$$

with $s_1=s_2=s_3$ *and* $s_1(p_1)=1$ (*this corresponds to the unknotted sphere in* S^4).

5C A group-theoretic characterization of knot groups

All knots discussed here are smooth. This subsection is motivated by the question: Exactly which groups arise as fundamental groups of codimension-2 knot complements in S^n? An algebraic characterization of such groups was given by Kervaire; namely, such groups are exactly the finitely presentable groups G with $H_1(G)=\mathbb{Z}$, $H_2(G)=0$, and such that G can be normally generated by a single element [33]. When $n=3$, Artin gave a characterization for fundamental groups of knot complements in terms of group presentations [2]. When $n=4$, Kamada [24] and González-Acuña [9] have given similar characterizations also in terms of group presentations.

The methods of this paper give alternative group-theoretic characterizations of fundamental groups of knot complements in S^n for $n=3,4$, in a unified fashion. We say $(\phi_1,\phi_2)\in\mathtt{Alg}^{(3,1)}$ is *connected* if the resulting link is connected; note that this could instead be phrased entirely in terms of the maps ϕ_1 and ϕ_2. We say (ϕ_1,ϕ_2) *represents* S^3 if the resulting 3-manifold is diffeomorphic to the 3-sphere; note that this too can be phrased in an entirely algebraic way due to Theorem 3.1. (See the previous subsection for the corresponding definition when $n=4$.) Then directly from Theorems 3.2 and 4.8 we have the following characterizations.

Corollary 5.7 *A group G is the fundamental group of a knot complement in* S^3 *if and only if there exists some connected* $(\phi_1,\phi_2)\in\mathtt{Alg}^{(3,1)}$ *such that* (ϕ_1,ϕ_2) *represents* S^3, *and the fundamental group of* (ϕ_1,ϕ_2) (*recall Definition 5.1*) *is G.*

Similarly, a group G is the fundamental group of a knot complement in S^4 *if and only if there exists some spherical* $(\phi_1,\phi_2,\phi_3)\in\mathtt{Alg}^{(4,2)}$ *such that* (ϕ_1,ϕ_2,ϕ_3) *represents* S^4, *and the fundamental group of* (ϕ_1,ϕ_2,ϕ_3) (*recall Definition 5.1*) *is G.*

5D Algorithmic decidability

Recognizing whether a given finitely presented group is the group of the complement of a knotted surface in the 4-sphere is not algorithmic; this is the case $\mathcal{A}=\mathcal{G}$, $\mathcal{B}=\mathcal{K}_2$ in [10, Theorem 1.1]. It is undecidable whether a given finite presentation describes the fundamental group of the complement of a codimension-2 knot in S^n for $n\geq 3$ [13]. The following then is a corollary of a combination of [10, Theorem 1.1] and Theorem 4.8.

Corollary 5.8 *Given a group G, it is an undecidable problem to find a group trisection (ϕ_1, ϕ_2, ϕ_3) in* $\mathtt{Alg}^{(4,2)}$ *representing S^4 with $b > 0$ and fundamental group G (recall Definition 5.1), or to show that none exists (because this would decide whether G is a knotted surface group).*

Recall that there *is* an algorithm for deciding whether the pushouts appearing in group trisections are free groups (see Remark 4.3).

Question 5.9 There are many undecidability results in group theory (see for example [34]). Can the above techniques be used to show the existence or nonexistence of an algorithm to recognize if a closed smooth 4-manifold is diffeomorphic to S^4? Can the above techniques be used to show the existence or nonexistence of an algorithm to recognize if a smooth embedding of a 2-sphere in S^4 is unknotted? In a different direction, can our framework be useful for determining if there is a polynomial-time algorithm for deciding whether a knot in S^3 is the unknot?

Question 5.10 Are there extensions of the bijections between sets of the form $\mathtt{Alg}^n/\sim$ and sets of the form $\mathtt{Man}^n$ (as in Sections 3 and 4) for $n > 4$? A negative answer says that somehow the free groups and surface groups are responsible for the unique character of low-dimensional topology.

References

[1] **A Abrams**, **D T Gay**, **R Kirby**, *Group trisections and smooth 4-manifolds*, Geom. Topol. 22 (2018) 1537–1545 MR

[2] **E Artin**, *Theorie der Zöpfe*, Abh. Math. Sem. Univ. Hamburg 4 (1925) 47–72 MR

[3] **S Blackwell**, *Combinatorial and group theoretic approaches to trisected surfaces in 4-manifolds*, PhD thesis, University of Georgia (2022) MR Available at `https://www.proquest.com/docview/2709946407`

[4] **J Boyle**, *Classifying 1-handles attached to knotted surfaces*, Trans. Amer. Math. Soc. 306 (1988) 475–487 MR

[5] **M Clay**, *Free groups and folding*, from "Office hours with a geometric group theorist" (M Clay, D Margalit, editors), Princeton Univ. Press (2017) 66–84 MR

[6] **T Cochran**, *Ribbon knots in S^4*, J. London Math. Soc. 28 (1983) 563–576 MR

[7] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR

[8] **D Gay**, **R Kirby**, *Trisecting 4-manifolds*, Geom. Topol. 20 (2016) 3097–3132 MR

[9] **F González-Acuña**, *A characterization of 2-knot groups*, Rev. Mat. Iberoamericana 10 (1994) 221–228 MR

[10] **F González-Acuña**, **C M Gordon**, **J Simon**, *Unsolvable problems about higher-dimensional knots and related groups*, Enseign. Math. 56 (2010) 143–171 MR

[11] **C M Gordon**, *Some higher-dimensional knots with the same homotopy groups*, Quart. J. Math. Oxford Ser. 24 (1973) 411–422 MR

[12] **C M Gordon**, *Knots in the* 4*-sphere*, Comment. Math. Helv. 51 (1976) 585–596 MR

[13] **C M Gordon**, *Some embedding theorems and undecidability questions for groups*, from "Combinatorial and geometric group theory", London Math. Soc. Lecture Note Ser. 204, Cambridge Univ. Press (1995) 105–110 MR

[14] **C M Gordon**, **J Luecke**, *Knots are determined by their complements*, J. Amer. Math. Soc. 2 (1989) 371–415 MR

[15] **C Hayashi**, *Stable equivalence of Heegaard splittings of* 1*-submanifolds in* 3*-manifolds*, Kobe J. Math. 15 (1998) 147–156 MR

[16] **J Hempel**, 3*-manifolds*, Annals of Mathematics Studies 86, Princeton Univ. Press (1976) MR

[17] **J A Hillman**, *Alexander ideals of links*, Lecture Notes in Math. 895, Springer (1981) MR

[18] **M C Hughes**, **S Kim**, **M Miller**, *Isotopies of surfaces in* 4*-manifolds via banded unlink diagrams*, Geom. Topol. 24 (2020) 1519–1569 MR

[19] **W Jaco**, *Heegaard splittings and splitting homomorphisms*, Trans. Amer. Math. Soc. 144 (1969) 365–379 MR

[20] **W Jaco**, *Stable equivalence of splitting homomorphisms*, from "Topology of manifolds", Markham, Chicago, IL (1970) 153–156 MR

[21] **K Johannson**, *Topology and Combinatorics of* 3*-manifolds*, Lecture Notes in Math. 1599, Springer (1995) MR

[22] **J Johnson**, *Notes on Heegaard splittings*, unpublished (2006)

[23] **J Joseph**, **J Meier**, **M Miller**, **A Zupan**, *Bridge trisections and classical knotted surface theory*, Pacific J. Math. 319 (2022) 343–369 MR

[24] **S Kamada**, *A characterization of groups of closed orientable surfaces in* 4*-space*, Topology 33 (1994) 113–122 MR

[25] **S Kamada**, *Braid and knot theory in dimension four*, Mathematical Surveys and Monographs 95, Amer. Math. Soc., Providence, RI (2002) MR

[26] **T Kanenobu**, **K-i Kazama**, *The peripheral subgroup and the second homology of the group of a knotted torus in* S^4, Osaka J. Math. 31 (1994) 907–921 MR

[27] **G Kuperberg**, *Algorithmic homeomorphism of* 3*-manifolds as a corollary of geometrization*, Pacific J. Math. 301 (2019) 189–241 MR

[28] **F Laudenbach**, **V Poénaru**, *A note on* 4*-dimensional handlebodies*, Bull. Soc. Math. France 100 (1972) 337–344 MR

[29] **C J Leininger**, **A W Reid**, *The co-rank conjecture for* 3*-manifold groups*, Algebr. Geom. Topol. 2 (2002) 37–50 MR

[30] **J Meier**, *Filling braided links with trisected surfaces*, Algebr. Geom. Topol. 24 (2024) 803–895 MR

[31] **J Meier**, **A Zupan**, *Bridge trisections of knotted surfaces in* S^4, Trans. Amer. Math. Soc. 369 (2017) 7343–7386 MR

[32] **J Meier**, **A Zupan**, *Bridge trisections of knotted surfaces in* 4*-manifolds*, Proc. Natl. Acad. Sci. USA 115 (2018) 10880–10886 MR

[33] **F Michel**, **C Weber**, *Michel Kervaire work on knots in higher dimensions*, preprint (2014) arXiv 1409.0704

[34] **C F Miller, III**, *Decision problems for groups—survey and reflections*, from "Algorithms and classification in combinatorial group theory", Math. Sci. Res. Inst. Publ. 23, Springer (1992) 1–59 MR

[35] **G Perelman**, *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds*, preprint (2003) arXiv math/0307245

[36] **K Reidemeister**, *Zur dreidimensionalen Topologie*, Abh. Math. Sem. Univ. Hamburg 9 (1933) 189–194 MR

[37] **B M Ruppik**, *Casson–Whitney unknotting, deep slice knots and group trisections of knotted surface type*, PhD thesis, Max Planck Institute for Mathematics and the University of Bonn (2022) Available at `https://nbn-resolving.org/urn:nbn:de:hbz:5-66903`

[38] **H Schubert**, *Über eine numerische Knoteninvariante*, Math. Z. 61 (1954) 245–288 MR

[39] **J Singer**, *Three-dimensional manifolds and their Heegaard diagrams*, Trans. Amer. Math. Soc. 35 (1933) 88–111 MR

[40] **J Stallings**, *How not to prove the Poincaré conjecture*, from "Topology seminar", Ann. of Math. Stud. 60, Princeton Univ. Press (1966) 83–88 MR

[41] **J R Stallings**, *Topology of finite graphs*, Invent. Math. 71 (1983) 551–565 MR

[42] **A I Suciu**, *Infinitely many ribbon knots with the same fundamental group*, Math. Proc. Cambridge Philos. Soc. 98 (1985) 481–492 MR

[43] **E C Zeeman**, *Twisting spun knots*, Trans. Amer. Math. Soc. 115 (1965) 471–495 MR

[44] **A Zupan**, *Bridge and pants complexities of knots*, J. Lond. Math. Soc. 87 (2013) 43–68 MR

SB: *Department of Mathematics, University of Virginia*
Charlottesville, VA, United States

RK: *Department of Mathematics, University of California, Berkeley*
Berkeley, CA, United States

MK: *Department of Mathematics, University of Chicago*
Chicago, IL, United States

VL: *Department of Mathematics, University of Connecticut*
Storrs, CT, United States

BR: *Heinrich-Heine-Universität*
Düsseldorf, Germany

`blackwell@virginia.edu`, `kirby@math.berkeley.edu`, `michael.r.klug@gmail.com`,
`vincent.2.longo@uconn.edu`, `benjamin.ruppik@hhu.de`

`https://seblackwell.com`, `https://math.berkeley.edu/~kirby`,
`https://mathematics.uchicago.edu/people/profile/michael-klug`,
`https://math.uconn.edu/person/vincent-longo`, `https://bruppik.de`

Received: 9 August 2023 Revised: 24 April 2024

Algebraic & Geometric Topology **25**:8 (2025) 4719–4785
DOI: 10.2140/agt.2025.25.4719
Published: 20 November 2025

Classification of genus-two surfaces in S^3

FILIPPO BARONI

We describe an algorithm to decide whether two genus-two surfaces embedded in the 3-sphere are isotopic or not. The algorithm employs well-known techniques in 3-manifolds topology, as well as a new algorithmic solution to a problem on free groups.

57-08, 57K30

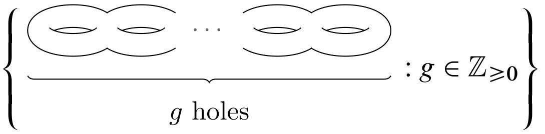
1 Introduction

1.1 The classification problem

In mathematics, by “classification” we usually mean a list, finite or infinite, of all objects of a given type, up to a given equivalence relation. For instance, the list

$$\{\mathbb{Z}/4\mathbb{Z}, \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}\}$$

is a classification of groups of order 4 up to isomorphism, and the list



is a classification of closed orientable surfaces up to homeomorphism. Naturally, we want the description of such a list to be somewhat explicit: the reader will surely agree that

$$\{\text{finitely generated Abelian groups up to isomorphism}\}$$

is not a classification of finitely generated Abelian groups up to isomorphism, but

$$\{\mathbb{Z}^k \times \mathbb{Z}/a_1\mathbb{Z} \times \cdots \times \mathbb{Z}/a_n\mathbb{Z} : k \in \mathbb{Z}_{\geqslant 0},\ n \in \mathbb{Z}_{\geqslant 0},\ a_1, \ldots, a_n \in \mathbb{Z}_{\geqslant 2},\ a_1 \mid \cdots \mid a_n\}$$

is. This example raises the question of what degree of explicitness is required for something to be considered a classification.

For instance, let us consider the set of knots in the 3-sphere: can we classify them up to isotopy?

- A task that is surely within our grasp is deciding whether two given knots in S^3 are isotopic. In fact, two knots being isotopic is equivalent to their two complements being homeomorphic via a homeomorphism which is meridian-preserving; this can be decided algorithmically thanks to Theorem 1.4.
- Moreover, it is not hard to devise an algorithm which produces an infinite list $\mathcal{L}_0$ of knot diagrams, so that every knot is isotopic to one element (and possibly more) of our list; this follows from the fact that there are only finitely many knot diagrams with n crossings up to isotopy for every $n \geqslant 0$, and they can be enumerated algorithmically.

We can combine these two ingredients to produce

(1) a computable list $\mathcal{L}$ of knots containing exactly one representative for each isotopy class, by taking knots in $\mathcal{L}_0$ which are not isotopic to elements appearing earlier in the list;

(2) an algorithm which takes a knot as input and returns the unique knot in $\mathcal{L}$ which is isotopic to it.

Given how complex the world of knots is, we cannot expect to find a classification which is as neat as the ones given above for surfaces or Abelian groups. This algorithmic answer is probably the best we can hope for, and we believe it is explicit enough to be considered a classification.

We hope that, in light of the above discussion, the reader will accept the following definition as a sensible one. A *classification* of a set $\mathcal{X}$ up to an equivalence relation $\sim$ is the datum of

(1) an algorithm producing a list $\mathcal{L}$ containing exactly one representative for each equivalence class of $\mathcal{X}/\sim$, and

(2) an algorithm which takes an element of $\mathcal{X}$ as input and returns the unique element of $\mathcal{L}$ equivalent to it.

Like in the knots example, in order to have a classification of $\mathcal{X}$ up to $\sim$, it is enough to provide a computable list $\mathcal{L}_0$ containing *at least* one representative for each equivalence class of $\mathcal{X}/\sim$, and an algorithm to decide equivalence; since the second ingredient is usually the hardest to find, we call it the *classification algorithm*.

Going back to the case of knots, a routine topological argument shows that the knot classification problem is equivalent to the classification of tori embedded in S^3 up to isotopy: the one-to-one correspondence is given by associating each knot to the boundary of its regular neighbourhood. It is then only natural to try and generalise this question to higher-genus surfaces.[1] In fact, the following appears as Problem 3.11 in Kirby's problem list.

---

[1] It should be noted that the classification of genus-zero surfaces in S^3 is trivial, since by Alexander's theorem there is only one isotopy class of embedded 2-spheres (see [1]).

Problem [13, Problem 3.11] *Classify embeddings of orientable surfaces in S^3.*

This is precisely the aim of this article, which solves a special case of the above problem by providing a classification of genus-two surfaces in S^3. Producing a redundant list $\mathcal{L}_0$ is very easy, by simply enumerating all simplicial genus-two surfaces in all subdivisions of S^3 (see Section 1.3). Therefore, the only goal of this article can be summarised as follows.

Theorem *There is an algorithm to decide whether two genus-two surfaces in S^3 are isotopic.*

The article is structured as follows. In the rest of Section 1, we introduce the notation and terminology we will use in our exposition, and we briefly discuss some foundational results in low-dimensional algorithmic topology. We also give a few examples of phenomena which make the genus-two classification problem intrinsically harder than the knot classification problem. Section 2 is devoted to the computation of mapping class groups of 3-manifolds. The exposition in this section closely follows the work of Johannson in [12], revisited to provide constructive and effective proofs. The arguments here are somewhat technical and involved, and may be skipped on a first reading; the main result of this section, that is, Theorem 2.23, can be assumed as a black box without compromising the understanding of the rest of the article. In Section 3, we present a solution to an algorithmic problem on free groups, as well as an application to a topological decision problem. Finally, in Section 4, we focus on the genus-two classification problem. A careful case analysis and elementary topological arguments, combined with Theorem 2.23 and Corollary 3.5, allow us to prove the theorem stated above; a more detailed outline of the inner workings of the classification algorithm is given at the end of Section 4.1.

1.2 Notation

Manifolds

- Throughout this article, we will always be working in the PL category: all manifolds will have a PL structure, and functions between manifolds will be assumed to be PL.
- All 3-manifolds will be compact, connected and oriented unless otherwise stated. Codimension-zero submanifolds and boundaries of 3-manifolds will be implicitly oriented accordingly.
- All surfaces (2-manifolds) will be compact, but we make no assumptions about connectedness or orientability.
- For integers $g \geqslant 0$ and $k \geqslant 0$, we denote by $\Sigma_{g,k}$ the connected orientable surface of genus g with k holes (also referred to as "punctures"). If $g \geqslant 1$, we denote by $U_{g,k}$ the surface obtained by taking the connected sum of g projective planes and then removing k open discs. In both cases, the integer k may be omitted when it is equal to 0. We will occasionally refer to the torus Σ_2 by the symbol T^2.
- If X is a manifold and $Z \subseteq X$ is a subspace, we denote by $\operatorname{clos}(Z)$ and $\operatorname{int}(Z)$ its closure and interior in X, respectively; the ambient space X will always be clear from the context.

Fibre bundles

- For a fibre bundle $p\colon M \to B$ with fibre F, we denote by $\partial_h M$ its horizontal boundary, that is, the induced ∂F-bundle on B; we will refer to the vertical boundary $p^{-1}(\partial B)$ as $\partial_v M$.

- We denote by I the 1-manifold $[0, 1]$. For every surface B, there is exactly one (orientable) I-bundle over B; we denote it by $B \times I$ (the "product I-bundle") if B is orientable, and by $B \mathbin{\tilde{\times}} I$ (the "twisted I-bundle") if it is not.

- If X is a manifold and $Z \subseteq X$ is a properly embedded submanifold, we denote by $\mathcal{N}(Z)$ and $\mathring{\mathcal{N}}(Z)$ the closed and open regular neighbourhoods of Z in X, respectively. The closed neighbourhood $\mathcal{N}(Z)$ is naturally endowed with a bundle structure over Z with fibre the $(\operatorname{codim} Z)$-dimensional disc.

Surfaces in 3-manifolds

- If X is a manifold and $Z \subseteq X$ is a properly embedded codimension-one submanifold, we define $X \mathbin{\backslash\backslash} Z = X \setminus \mathring{\mathcal{N}}(Z)$ to be the result of cutting X along Z; if Z is separating, the manifold $X \mathbin{\backslash\backslash} Z$ will be disconnected.

- As far as the notions of irreducible, boundary irreducible, incompressible, boundary incompressible, and sufficiently large are concerned, we adopt the same definitions as [20]. Unless otherwise stated, by "(boundary) compression disc" we will always mean "nontrivial (boundary) compression disc".

Homeomorphisms

- Unless otherwise stated, homeomorphisms between 3-manifolds will be orientation-preserving.

- If X and Y are manifolds, and $A_1, \dots, A_n$ and $B_1, \dots, B_n$ are subspaces, a function

$$f : (X, A_1, \dots, A_n) \to (Y, B_1, \dots, B_n)$$

is a function $X \to Y$ such that $f(A_i) \subseteq B_i$ for $1 \leqslant i \leqslant n$. We say that f is a homeomorphism if it is bijective and $f(A_i) = f(B_i)$ for $1 \leqslant i \leqslant n$. When the number n is clear from the context or irrelevant, we will write $(X, \boldsymbol{A})$ as a shorthand for $(X, A_1, \dots, A_n)$.

- If X and Y are oriented, we denote by $\boldsymbol{H}((X, \boldsymbol{A}); (Y, \boldsymbol{B}))$ the set of homeomorphisms from $(X, \boldsymbol{A})$ to $(Y, \boldsymbol{B})$, modulo isotopies through homeomorphisms of the same kind. Note that $\boldsymbol{H}((X, \boldsymbol{A}); (X, \boldsymbol{A}))$ has a natural group structure; we use the shorthand $\boldsymbol{H}(X, \boldsymbol{A})$ to refer to it. Moreover, if $S \subseteq X$ is a subspace, we denote by $\boldsymbol{H}_S(X, \boldsymbol{A})$ the group of self-homeomorphisms of $(X, \boldsymbol{A})$ fixing S pointwise, modulo isotopies through homeomorphisms of the same kind.

- If $f\colon (X, \boldsymbol{A}) \to (Y, \boldsymbol{B})$ is a function, we define its *trace*
$$f|_\partial\colon (\partial X, \boldsymbol{A} \cap \partial X) \to (\partial Y, \boldsymbol{B} \cap \partial Y)$$
to be its restriction to the boundary. The trace behaves well with isotopies — that is, isotopic homeomorphisms have isotopic traces. As a consequence, when X and Y are oriented, we have a well-defined *trace map*
$$(-)|_\partial\colon \boldsymbol{H}((X, \boldsymbol{A}); (Y, \boldsymbol{B})) \to \boldsymbol{H}((\partial X, \boldsymbol{A} \cap \partial X); (\partial Y, \boldsymbol{B} \cap \partial Y)),$$
which is functorial in the sense that it preserves composition. For the sake of convenience, if W is a set of (isotopy classes of) homeomorphisms, we define $W|_\partial = \{f|_\partial : f \in W\}$.

Dehn twists

- In the most general sense, a Dehn twist of a manifold X about a properly embedded two-sided codimension-one submanifold Z is a self-homeomorphism of X which is the identity outside $\mathcal{N}(Z)$.
- When Z is an essential two-sided curve in a surface X, the group of Dehn twists about Z modulo isotopies is isomorphic to $\mathbb{Z}$. If X is oriented there is a way to select a preferred generator of this group, which we call *the* Dehn twist about Z and denote by τ_Z. If X is orientable without a preferred orientation, fixing one arbitrarily allows us to pick the generating Dehn twists about different curves consistently. When X is nonorientable, instead, we arbitrarily call one of the generating Dehn twists τ_Z, so that the other one is τ_Z^{-1}.
- Similarly, if Z is an incompressible boundary incompressible annulus in an (oriented) 3-manifold X the group of Dehn twists about Z modulo isotopies is isomorphic to $\mathbb{Z}$. In this case, there is no preferred generator of this group, so we arbitrarily pick one to call τ_Z. We remark that, if z_1 and z_2 are the two boundary curves of Z, then $\tau_Z|_\partial\colon \partial X \to \partial X$ is equal to either $\tau_{z_1}\tau_{z_2}^{-1}$ or $\tau_{z_1}^{-1}\tau_{z_2}$, depending on the choice of τ_Z.

1.3 Classical algorithms on 3-manifolds

Triangulations of 3-manifolds We are, of course, interested in algorithms on 3-manifolds. The first issue we should address is, perhaps, what a 3-manifold *is* from an algorithmic point of view. The simplest way to think about a 3-manifold M in this setting is seeing it as a collection of 3-simplices (or "tetrahedra"), with some pairs of faces identified by simplicial isomorphisms. Such a description is called a *triangulation* of M, and it fully encodes the topology of M in a discrete and combinatorial fashion. Of course, from the perspective of algorithmic implementation, the name "tetrahedra" is purely suggestive: 3-simplices are encoded as ordered 4-tuples of "vertices" — represented, for example, by integers — and an identification between faces can be specified by two ordered triples of vertices, each describing a face of a tetrahedron.

Some care must be taken in order to ensure that the topological space M obtained by gluing the tetrahedra is actually a manifold. We deal with orientation by orienting the tetrahedra, and requiring that gluing maps are orientation-reversing. Then M is a 3-manifold if and only if the link of each vertex is a disc — for vertices on the boundary — or a sphere. Note that both conditions can be easily checked algorithmically.

Simplicial subdivisions Ideally, we would want all the "objects" we have to deal with algorithmically (namely, maps and submanifolds) to be simplicial; in other words, we want to describe them discretely at the level of simplices. Fortunately, in the world of 3-manifolds, everything can be made simplicial, up to subdivision and isotopy.

Definition 1.1 Let $\mathcal{T}$ be a triangulation of a 3-manifold M. A *subdivision* of $\mathcal{T}$ is a triangulation $\mathcal{T}'$ of a 3-manifold M' together with a choice, for each vertex v of $\mathcal{T}'$, of a tetrahedron $\Delta(v)$ of $\mathcal{T}$ and four nonnegative rational numbers $q_1(v)$, $q_2(v)$, $q_3(v)$ and $q_4(v)$, satisfying the following constraints.

(i) For every vertex v of $\mathcal{T}'$, the equality $q_1(v)+q_2(v)+q_3(v)+q_4(v)=1$ holds.

(ii) Fix a vertex v of $\mathcal{T}'$, and denote by $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$ and $\boldsymbol{u}_4$ the vertices of $\Delta(v)$, which we interpret as the four canonical base vectors in $\mathbb{R}^4$. If we think of $\Delta(v)$ as the convex hull of its vertices, then there is a unique point

$$p(v)=\sum_{i=1}^{4} q_i(v)\boldsymbol{u}_i \in \Delta(v) \subseteq M.$$

We require that, if v_1, v_2, v_3 and v_4 are vertices of a tetrahedron of $\mathcal{T}'$, then there is some tetrahedron of $\mathcal{T}$ containing $p(v_1)$, $p(v_2)$, $p(v_3)$ and $p(v_4)$.

(iii) Define a function $f\colon M'\to M$ by extending the map $v\mapsto p(v)$ linearly on the tetrahedra of $\mathcal{T}'$; the previous constraint guarantees that f is well defined. We require that f is a homeomorphism.

With slight abuse of notation, we will omit the choices of $\Delta(v)$ and $q_i(v)$, and simply call the triangulation $\mathcal{T}'$ a subdivision of $\mathcal{T}$. With this definition, subdivisions can be described in a fully combinatorial fashion. Note that, with some work, the statement "$\mathcal{T}'$ is a subdivision of $\mathcal{T}$" can be decided algorithmically.

The following two facts are crucial.

- Let $\mathcal{T}$ be a triangulation of a 3-manifold M. Then every submanifold $N\subseteq M$ (of dimension 0, 1, 2, or 3) is isotopic to a simplicial submanifold of a subdivision of $\mathcal{T}$.
- Let $\mathcal{T}_M$ and $\mathcal{T}_N$ be triangulations of the 3-manifolds M and N, respectively. Then every homeomorphism $f\colon M\to N$ is isotopic to a simplicial isomorphism between a subdivision of $\mathcal{T}_M$ and a subdivision of $\mathcal{T}_N$.

Therefore, in every step of our algorithms, we will always assume that submanifolds of 3-manifolds are simplicial. More precisely, a submanifold of a 3-manifold M with a triangulation $\mathcal{T}$ will be represented as a subdivision $\mathcal{T}'$ of $\mathcal{T}$, together with a set of simplices of $\mathcal{T}'$ (of the appropriate dimension). Similarly, a homeomorphism between 3-manifolds M and N with triangulations $\mathcal{T}_M$ and $\mathcal{T}_N$, respectively, will be represented as subdivisions $\mathcal{T}'_M$ and $\mathcal{T}'_N$ of $\mathcal{T}_M$ and $\mathcal{T}_N$, respectively, together with a simplicial isomorphism between $\mathcal{T}'_M$ and $\mathcal{T}'_N$.

Algorithmic operations A useful property of subdivisions is that every two subdivisions $\mathcal{T}_1$ and $\mathcal{T}_2$ of the same triangulation $\mathcal{T}$ have a common subdivision $\mathcal{T}_{12}$, which can be computed algorithmically. Without going into too much detail, let us simply remark that this implies that we can — and will — always assume that different "objects" in the same 3-manifold are simplicial with respect to the same subdivision.

All the natural operations one may want to perform on 3-manifolds can be done algorithmically; here is a long — but by no means complete — list of examples:

- composing and inverting homeomorphisms;
- finding images of submanifolds under homeomorphisms;
- cutting along a properly embedded codimension-one submanifold;
- finding (simplicial) regular neighbourhoods of properly embedded submanifolds;
- isotoping properly embedded submanifolds so that they are in general position;
- finding intersections of properly embedded submanifolds in general position.

Some of these constructions are trivial, while some require a lot of work. For more details on triangulations and subdivisions of PL manifolds, we refer the reader to the excellent introductory book [22] by Rourke and Sanderson.

Infinite search template The infinite search template algorithm can be informally described as follows: if the elements of a set $\mathcal{S}$ can be enumerated algorithmically and $\mathcal{S}$ is guaranteed to be nonempty, then there is an algorithm to construct an element of $\mathcal{S}$. As obvious as it may sound, this template will often allow us to blur the distinction between "proving that something exists" and "being able to algorithmically construct it".

As an example, consider two 3-manifolds M and N, with triangulations $\mathcal{T}_M$ and $\mathcal{T}_N$, respectively. We can algorithmically enumerate all triples $(\mathcal{T}'_M, \mathcal{T}'_N, f)$ where $\mathcal{T}'_M$ is a subdivision of $\mathcal{T}_M$, $\mathcal{T}'_N$ is a subdivision of $\mathcal{T}_N$, and f is a simplicial isomorphism between $\mathcal{T}'_M$ and $\mathcal{T}'_M$. If we know that M and N are homeomorphic, then the algorithm enumerating all the triples will eventually find one, effectively constructing a homeomorphism from M to N. As a consequence, after proving that two 3-manifolds are homeomorphic, we will always assume to have a homeomorphism available for our algorithmic purposes.

Let us remark that what we have described above is *not* an algorithm to solve the homeomorphism problem — that is, to decide whether M and N are homeomorphic. In fact, the algorithm will terminate if M and N are guaranteed to be homeomorphic, but will run forever if they are not; at any given point in time, there is no way to know if the algorithm has not found a homeomorphism because we have not waited long enough, or because there is none.

Solved algorithmic problems In the realm of surfaces, all the elementary questions can be settled algorithmically.

Theorem 1.2 *Given a surface F, the following operations can be carried out algorithmically*:

(1) *deciding whether F is orientable or not*;

(2) *computing the unique integers $g \geqslant 0$ and $k \geqslant 0$ such that F is homeomorphic to $\Sigma_{g,k}$ if it is orientable, and to $U_{g,k}$ if it is not*;

(3) *given another surface F', deciding if F and F' are homeomorphic*;

(4) *given a curve in F, deciding if it is trivial* (*that is*, *it bounds a disc in F*) *and if it is boundary parallel*;

(5) *given two* (*multi*)*curves in F, deciding if they are isotopic or not*;

(6) *given another surface F' and two homeomorphisms $F \to F'$, deciding whether they are isotopic or not.*

Of course, the situation for 3-manifolds is substantially more complicated. The homeomorphism problem has only been recently settled, thanks to the contributions of many authors, the last piece needed being the geometrisation theorem (see [2; 16; 23]). The isotopy problem for codimension-1 submanifolds has also proved to be quite difficult to investigate. The sole purpose of this article, after all, is to provide an algorithm solving the isotopy problem for one specific surface — namely, Σ_2 — embedded in one specific 3-manifold — namely, S^3. There exist, however, classical algorithms to answer some basic questions about 3-manifolds.

Theorem 1.3 *Given a* 3*-manifold M, the following operations can be carried out algorithmically*:

(1) *deciding whether M is irreducible*;

(2) *if M is irreducible*, *deciding whether M is boundary irreducible*;

(3) *if M is irreducible*, *deciding whether ∂M admits a nonseparating compression disc in M*;

(4) *deciding whether M is a handlebody* (*possibly a* 3*-ball*); *if it is*, *computing its genus*;

(5) *given a surface properly embedded in M, deciding if it is boundary parallel*;

(6) *given an orientable surface which is either properly embedded in M or embedded in ∂M, deciding whether it is incompressible*;

(7) *if M is irreducible and boundary irreducible*, *given an orientable incompressible surface properly embedded in M, deciding whether it is boundary incompressible*;

(8) *if M is irreducible*, *given two connected incompressible surfaces properly embedded in M, deciding whether they are isotopic.*

Proof Statements (1), (2), (6), and (7) are proved by Matveev in Theorems 4.1.12, 4.1.13, 4.1.15, and 4.1.19 of [20], respectively. Statement (4) is solved by [11, Algorithm 9.3]. Concerning statement (5), recall that a surface $F \subseteq M$ is boundary parallel if and only if it cobounds a product $F \times I$ with some surface in ∂M, where $F \subseteq M$ is identified with $F \times \{0\}$. It is then enough to check whether F is separating in M and, if it is, whether there is a component N of $M \setminus F$ such that $(\mathrm{clos}(N), \partial F)$ is homeomorphic to $(F \times I, \partial F \times \{0\})$; we can decide this thanks to Theorem 1.4 below.

In order to address statement (8), let us loosely rephrase a result of Waldhausen, namely [27, Proposition 5.4]: if F_1 and F_2 are isotopic incompressible surfaces properly embedded in M intersecting transversely with $\partial F_1 = \partial F_2$, then a component of $F_1 \setminus F_2$ is isotopic to a component of $F_2 \setminus F_1$. First of all, we check if ∂F_1 and ∂F_2 are isotopic in ∂M. If they are not, then F_1 and F_2 are not isotopic. Otherwise, we can assume that $\partial F_1 = \partial F_2$. Then we iterate over all components of $M \setminus\setminus (F_1 \cup F_2)$ and check whether any of these is a product $G \times I$ with $G \times \{0\} \subseteq F_1$ and $G \times \{1\} \subseteq F_2$, once again by means of Theorem 1.4. If we find such a region, then either $G \times \{0\} = F_1$ and $G \times \{1\} = F_2$ — thus showing that F_1 and F_2 are isotopic — or we use the product $G \times I$ to isotope $G \times \{1\} \subseteq F_2$ across $G \times \{0\} \subseteq F_1$, so as to reduce the number of components of $F_1 \cap F_2$. Hence, after finitely many steps, we either prove that F_1 and F_2 are isotopic, or we cannot find any more product regions; in this case, the two surfaces are not isotopic.

Finally, we prove statement (3). The crucial claim is the following: let $D \subseteq M$ be a separating compression disc for ∂M; then M admits a nonseparating compression disc for its boundary if and only if the same holds for one component of $M \setminus\setminus D$. In fact, let $E \subseteq M$ be a nonseparating compression disc for ∂M. Since M is irreducible, we can arrange for D and E to intersect transversely in a collection of arcs. Denote by $E_1, \dots, E_k$ the closures of the components of $E \setminus D$; they are all discs, and we can think of them as being properly embedded in $M \setminus\setminus D$. Restriction induces an isomorphism $H^1(M) \cong H^1(M \setminus\setminus D)$ on cohomology groups, sending $[E]$ — the cohomology class represented by E — to $[E_1] + \dots + [E_k]$. Since E is nonseparating in M, it is nontrivial in cohomology. Therefore, the same must hold for one of the discs $E_1, \dots, E_k$, thus showing that at least one component of $M \setminus\setminus D$ admits a nonseparating compression disc for the boundary. The algorithm then proceeds as follows. If M is boundary irreducible, then clearly the answer is no. Otherwise, let D be a compression disc for the boundary; if it is nonseparating we are done. Otherwise, we run the algorithm on the components of $M \setminus\setminus D$, and return yes if and only if we find a nonseparating disc in at least one of them. The algorithm will eventually terminate, since ∂M can only be compressed finitely many times. □

Concerning statement (8), let us remark that surfaces in S^3 are quite the opposite of incompressible. This is what makes classifying them a challenging task, even for surfaces of genus two. We refer the reader to Section 1.4 for a survey of the many different topological phenomena which can arise when embedding a genus-two surface in S^3, despite the trivial topology of the ambient space.

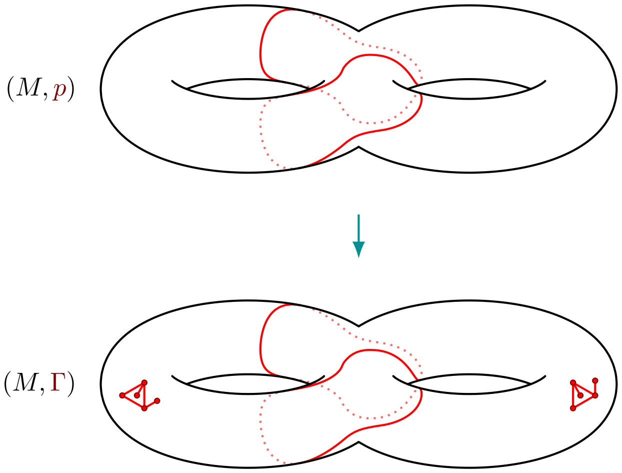


Figure 1: The homeomorphism problem can be reduced to the unoriented homeomorphism problem by adding chiral graphs on the boundary.

Homeomorphism problem for manifolds with boundary pattern Most importantly, the homeomorphism problem has been solved for what Matveev calls “Haken 3-manifolds with boundary pattern” in [20]. It is impossible to overstate how crucial this result is for our classification algorithm. In fact, the reader will probably find us referring to the following theorem more frequently then any other, often implicitly.

Theorem 1.4 *Let M and M' be irreducible 3-manifolds with nonempty boundary, and let $p \subseteq \partial M$ and $p' \subseteq \partial M'$ be (possibly empty) unions of curves such that $\partial M \mathbin{\backslash\backslash} p$ and $\partial M' \mathbin{\backslash\backslash} p'$ are incompressible in M and M', respectively. There is an algorithm which, given as input (M, p) and (M', p'), decides whether they are homeomorphic.*

Theorem 1.4 is a restricted version of [20, Theorem 6.1.6] which deals with the more general settings in which p and p' are allowed to be arbitrary graphs (what Matveev calls “boundary patterns”). There is, however, a technical point that needs to be addressed. Matveev’s theorem only guarantees an algorithmic solution to the question “is there a possibly orientation-reversing homeomorphism $(M, p) \to (M', p')$?”. One could, in theory, go through the proofs in [20, Chapter 6] and convince themselves that only trivial modifications are needed to answer the oriented version of the homeomorphism question. There is, however, a trick which allows us to derive Theorem 1.4 directly from [20, Theorem 6.1.6]. Simply augment the boundary pattern p by adding the same small chiral graph to each component of $M \mathbin{\backslash\backslash} p$, obtaining the boundary pattern $\Gamma \subseteq \partial M$; carry out the same procedure for p' to construct the boundary pattern $\Gamma' \subseteq M'$. Then (M, p) and (M', p') are orientation-preservingly homeomorphic if and only if (M, Γ) and (M, Γ') are possibly orientation-reversingly homeomorphic; a graphical explanation is provided in Figure 1.

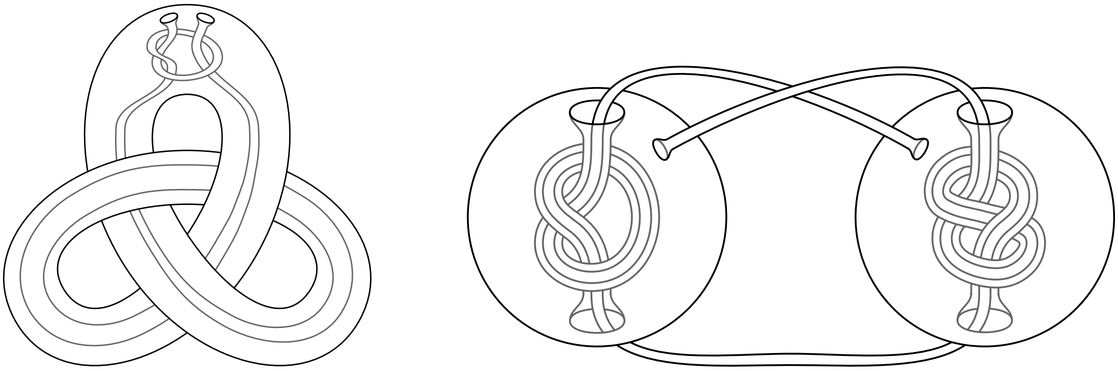


Figure 2: Left: a genus-two surface in S^3 which does not bound a handlebody on either side. Right: another genus-two surface in S^3 which does not bound a handlebody on either side.

1.4 Examples

Classification of genus-two surfaces in S^3 is significantly harder than the same task for tori. In fact, increasing the genus by one is enough for a set of "wild" topological phenomena to appear; this is in contrast with the relative "tameness" of tori embedded in S^3.

- First of all, it is not always true that a genus-two surface $S \subseteq S^3$ bounds a handlebody. A class of examples, as seen in Figure 2, left, can be constructed by starting with a knotted torus T and adding a tube inside the solid torus component of $S^3 \setminus T$. Alternatively, one can start with two knotted tori and join them with a tube; see Figure 2, right, for an example.
- Even if one component of $S^3 \setminus S$ is a handlebody, the other is not necessarily boundary irreducible; an example is described in Figure 3, left, where two knotted tori are connected by a "trivial" tube.
- Moreover, even in the case where one component of $S^3 \setminus S$ is a handlebody and the other is boundary irreducible (like in Figure 3), right, it is not always easy to identify a canonical meridian curve on S; we invite the reader to compare this with the existence of a standard meridian in tori embedded in S^3, which makes the classification problem significantly easier.

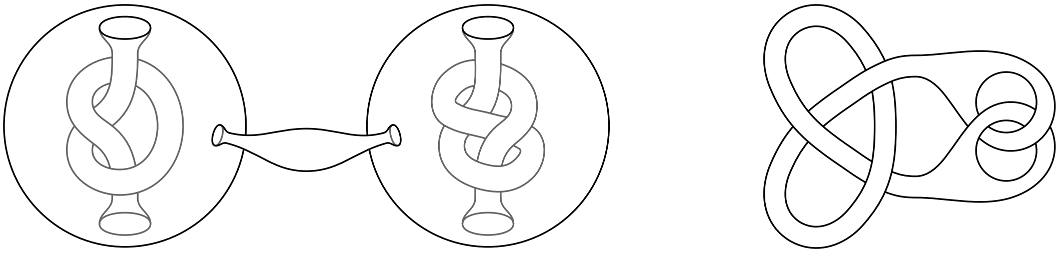


Figure 3: Left: a genus-two surface splitting S^3 into two components: one is a handlebody, but the other is not boundary irreducible. Right: a genus-two surface splitting S^3 into two components: one is a handlebody, and the other is boundary irreducible.

- Finally, for the genus-one case, one can completely bypass the matter of meridian curves by exploiting the fact that tori in S^3 are isotopic if and only if their complementary regions are homeomorphic. This result, which follows from the work of Gordon and Luecke (see [7]), does not hold for genus-two surfaces, even when they bound a handlebody on one side; an example of this phenomenon is given in [17].

2 Homeomorphisms of 3-manifolds

2.1 Rationale

Section 2 is dedicated to the study of the mapping class groups of 3-manifolds from an algorithmic point of view. These groups, of course, are often infinite, but they can be described with finite amount of data. To be more specific, it has been known for a long time (see [12, Corollary 27.6]) that the mapping class group of an irreducible sufficiently large 3-manifold contains a finite-index subgroup H which is generated by Dehn twists about annuli and tori. By going through the proofs in Johannson's book, one realises that, in fact, only finitely many Dehn twists are required to generate H. Therefore, theoretically, the mapping class group of a 3-manifold can be fully described by providing finitely many Dehn twists about annuli and tori, together with representatives of the cosets of H in the mapping class group.

There is, however, a substantial gap between acknowledging the existence of H and being able to algorithmically find generators of H and representatives of its cosets. The proofs by Johannson are for the most part constructive, and one could carefully convert them into algorithms for finding the generating Dehn twists (whereas representatives of the cosets of H are somewhat trickier to extract from said proofs). This is exactly the path we will follow in Section 2, albeit with a few *caveats*.

- Instead of working in the full generality of what Johannson calls "3-manifolds with complete and useful boundary pattern", we restrict our attention to *irreducible* 3*-manifold pairs*, following the work of Johannson in [10]. We admit that our approach is perhaps less elegant and powerful than Johannson's, but the decrease in flexibility is compensated by the (relative) conciseness of some of the algorithmic procedures we will describe.

- We make use of the geometrisation theorem (see [21; 24]) and Kuperberg's excellent exposition (see [16]) in order to compute the (finite) mapping class group of simple 3-manifolds. This could probably be avoided, following Johannson's inductive argument which relies on hierarchies. We decided to go for the shorter solution, rather than the one which was more faithful to Johannson's original work.

- As will become clear in Section 4, we are only interested in the trace of the mapping class group, and care little about what happens in the interior of the 3-manifold. This is why Theorem 2.23 is stated the way it is. It is true that little additional effort would have been required to give a complete description of the mapping class group, but once again we decided to avoid exceeding in generality, in order to keep this section reasonably short.

In Section 2.2, we start by recalling the definition of JSJ system for an irreducible sufficiently large 3-manifold pair, following [10] as closely as possible. Sections 2.3, 2.4, and 2.5 address the computation of the mapping class groups and the homeomorphism problem for Seifert fibred-spaces, I-bundles, and simple manifolds, respectively. In Section 2.6, we show how to actually compute the JSJ decomposition of a 3-manifold pair (with some additional assumptions). Finally, in Section 2.7, we put the results of the previous section together to deliver, as anticipated, a description of the mapping class group in the form of Theorem 2.23.

2.2 JSJ decomposition

This section is little more than a restatement of the definitions and the main theorem of [10, Chapter V, Section 6].

Definition 2.1 For $n \geqslant 2$, a 3*-manifold* n*-tuple* is an n-tuple $(M, R_1, \dots, R_{n-1})$ where M is a 3-manifold, and $R_1, \dots, R_{n-1}$ are surfaces in ∂M such that $R_i \cap R_j$ is a collection of curves for $1 \leqslant i < j \leqslant n-1$. We say *pair* and *triple* instead of 2-tuple and 3-tuple, respectively.

For a 3-manifold pair (M, R), we give the following definitions.

- We say that (M, R) is *irreducible* if M is irreducible and R is incompressible in M.
- We say that (M, R) is sufficiently large if M is.
- A surface $F \subseteq M$ is *properly embedded* in (M, R) if it is properly embedded in M and ∂F lies in R.
- Two disjoint surfaces F and F' in M with ∂F and $\partial F'$ lying in R are *parallel* in (M, R) if there is a 3-manifold $W \subseteq M$ homeomorphic to $F \times I$ such that $\partial_h W = F \cup F'$ and $\partial W \setminus \partial_h W \subseteq R$.
- A surface F in M is *boundary parallel* in (M, R) if it is parallel in (M, R) to a surface in R.

Our aim is to develop an algorithmic theory for the JSJ decomposition of irreducible sufficiently large 3-manifold pairs. Loosely speaking, as is the case for ordinary 3-manifolds, the decomposition for a pair (M, R) comes from an incompressible surface F properly embedded in (M, R) such that cutting M along it yields pieces which are in some sense "simpler" than the original pair. Note that these pieces immediately inherit a 3-manifold pair structure by considering the remnants of R as a surface in the boundary of $M \setminus\setminus F$. By doing so, however, we lose information about what parts of ∂M come from F. This is why, whenever we have a 3-manifold pair (M, R) with a properly embedded surface F, we naturally think of $M \setminus\setminus F$ as a 3-manifold triple. More precisely, we define $(M, R) \setminus\setminus F$ to be the 3-manifold triple (M', R', F'), where

- $M' = M \setminus\setminus F = M \setminus \mathring{\mathcal{N}}(F)$;
- $R' = R \cap M'$;
- $F' = \mathcal{N}(F) \cap M'$.

We now proceed to define the types of pieces we allow in the JSJ decomposition of a 3-manifold pair.

Definition 2.2 A 3-manifold n-tuple $(M, \boldsymbol{R})$ is a *Seifert n-tuple* if $\partial M = R_1 \cup \cdots \cup R_{n-1}$ and M admits a Seifert fibration such that $R_1, \dots, R_{n-1}$ are unions of fibres. Such a fibration is called a *Seifert fibration* for $(M, \boldsymbol{R})$. A 3-manifold n-tuple equipped with a Seifert fibration is called a *Seifert-fibred n-tuple*.

Definition 2.3 A 3-manifold triple is an *I-bundle triple* if it is homeomorphic to $(X, \partial_h X, \partial_v X)$ for some I-bundle X over a surface.

Definition 2.4 A 3-manifold triple (M, R, F) is a *simple triple* if

- every incompressible torus properly embedded in M is parallel in M to a component of R or of F;
- every incompressible annulus A properly embedded in M with $\partial A \subseteq \operatorname{int}(R)$ is parallel in (M, R) to an annulus in R or in F.

We can finally define the JSJ system for a 3-manifold pair. As expected, the JSJ system is unique up to isotopy.

Definition 2.5 Let (M, R) be an irreducible sufficiently large 3-manifold pair. A *JSJ system* of (M, R) is a surface F properly embedded in (M, R), possibly empty, satisfying the following properties.

(i) Each component of F is an incompressible torus or annulus, and it is not boundary parallel in (M, R).

(ii) Each component of $(M, R) \mathbin{\backslash\backslash} F$ is a Seifert triple, an I-bundle triple, or a simple triple.

(iii) The surface F is minimal with respect to inclusion among all surfaces properly embedded in (M, R) satisfying properties (i) and (ii).

Theorem 2.6 [10, Chapter V, Section 6, generalised splitting theorem] *Let (M, R) be an irreducible sufficiently large 3-manifold pair. Then (M, R) admits a JSJ system, which is unique up to isotopy in M fixing $\partial M \setminus \operatorname{int}(R)$.*

We now embark on an in-depth study of the pieces of the JSJ decomposition, with the aim of achieving a better understanding of the original 3-manifold pair. Specifically, for a piece (N, R', F') of $(M, R) \mathbin{\backslash\backslash} F$, we will be interested in the following algorithmic problems:

- solving the homeomorphism problem for (N, R', F');
- deciding whether a homeomorphism $F' \to F'$ extends to a self-homeomorphism of (N, R', F');
- describing the group $\boldsymbol{H}_{F'}(N, R')$ in a computationally feasible way (that is, with a finite amount of data).

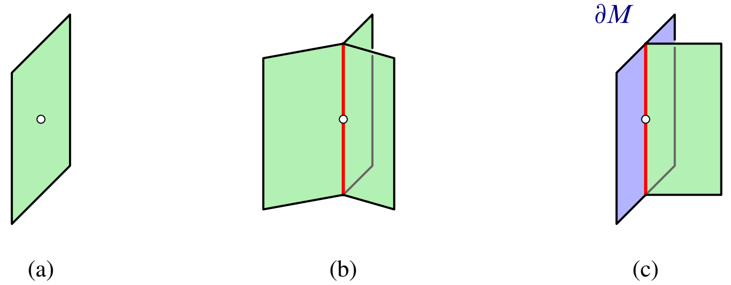


Figure 4: Local models for a 2-complex defining a Seifert fibration for a 3-manifold M. The 2-complex is coloured green, while the fibres are highlighted in red.

2.3 Seifert pieces

We start with an analysis of Seifert pieces. In fact, we will work with general Seifert n-tuples instead of just triples. The reason will become apparent in Section 2.7, when we will need to deal with Seifert 4-tuples as well. We follow the convention that a *Seifert manifold* is a 3-manifold which admits a Seifert fibration, whereas a *Seifert-fibred manifold* is a 3-manifold endowed with a fixed Seifert fibration.

Now is a good time to briefly discuss a "computationally friendly" definition of Seifert fibration. For our algorithmic purposes, a Seifert fibration of a triangulated 3-manifold is given by a 2-subcomplex Z of (some subdivision of) M satisfying the following properties.

(i) Every point of Z has a neighbourhood which looks like one of the local models in Figure 4; the set of points whose local models are of type (b) or (c) is a union of circles called *fibres*.

(ii) The closure of each component of $M \setminus Z$ is a solid torus, and has at least one fibre on its boundary.

(iii) No fibre bounds a disc in $M \setminus Z$.

On one hand, if M is equipped with a Seifert fibration in the usual sense and $p\colon M \to B$ is the projection to the base surface, we can obtain a 2-complex $Z \subseteq M$ by taking the preimage under p of a suitable 1-complex cutting B into discs which contain at most one singular point. Conversely, given a 2-complex Z as above, we can recover the Seifert fibration by fibring each complementary solid torus compatibly with the fibres of Z. Note that properties (i), (ii), and (iii) are easy to check algorithmically. As a consequence, given a Seifert manifold, we can find a Seifert fibration for it. This discussion naturally extends to a Seifert n-tuple $(M, \boldsymbol{R})$; in this case, we additionally require that ∂R_i is a union of fibres of Z for $1 \leqslant i \leqslant n-1$.

With the following we provide a version of [9, Lemma VI.19] which better suits our needs.

Lemma 2.7 *Let M and M' be Seifert-fibred manifolds. Let $f\colon M \to M'$ be a homeomorphism, and suppose that there is a nonempty finite union of fibres, fibred annuli, and fibred tori $X \subseteq \partial M$ such that the restriction of f to X is fibre-preserving. Then f can be isotoped, fixing X pointwise, to a fibre-preserving homeomorphism.*

Proof The proof is essentially identical to that of Jaco, and we present it here for the sake of completeness. Let $p' \colon M' \to B'$ be the projection to the base surface of the Seifert fibration, and let M' have n exceptional fibres. The proof proceeds by induction on $-\chi(B') + n$.

We can easily isotope f, fixing X, to be fibre-preserving on each component of ∂M which intersects X. Therefore, if $Y \subseteq \partial M$ is the union of all such components, then we can assume that $f|_Y \colon Y \to \partial M'$ is fibre-preserving, and prove that f can be isotoped, fixing Y, to a fibre-preserving homeomorphism.

If B' is a disc and $n \leqslant 1$, then M and M' are fibred solid tori of the same type, and it is easy to construct the desired isotopy. If instead B' is not a disc or $n \geqslant 2$, then let $y' \in \partial B'$ the image under p' of one of the fibres in $f(Y)$. There exists an arc a' properly embedded in B' connecting y' to a different point in $\partial B'$ and avoiding exceptional points, such that a' is not trivial in $\pi_1(B'_\bullet, \partial B')$, where $B'_\bullet$ is the punctured surface obtained by removing the exceptional points from B'. Then $A' = (p')^{-1}(a')$ is an incompressible boundary incompressible vertical annulus properly embedded in M'. The annulus $A = f^{-1}(A')$ is incompressible and boundary incompressible in M. It need not be vertical, but at least one of the components of ∂A is a fibre, since it lies in Y. We can isotope A to be vertical in M, and by [12, Proposition 5.6][2] we can assume that the isotopy fixes $A \cap Y$ pointwise. In fact, it is not hard to see that we can even take the isotopy to fix all of Y. As a consequence, we find that f restricts to a homeomorphism $M \mathbin{\backslash\backslash} A \to M' \mathbin{\backslash\backslash} A'$ which is fibre-preserving on $Y \cap (M \mathbin{\backslash\backslash} A)$ as well as on the two annuli $\operatorname{clos}(\partial(M \mathbin{\backslash\backslash} A) \setminus \partial M)$. The conclusion now follows by induction. □

The next two propositions generalise well-known facts about Seifert spaces — namely uniqueness of the Seifert fibration and solution to the homeomorphism problem — to the setting of Seifert n-tuples.

Proposition 2.8 *Let $(M, \boldsymbol{R})$ be a Seifert $(n+1)$-tuple with nonempty boundary. Then $(M, \boldsymbol{R})$ admits exactly one Seifert fibration up to isotopy in $(M, \boldsymbol{R})$, unless the surfaces $R_1, \dots, R_n$ are unions of tori and one of the following holds.*

1. *The 3-manifold M is a solid torus; in this case, every fibration of ∂M which does not contain a meridian of M extends to a unique Seifert fibration of M with at most one exceptional fibre; the fibration of ∂M containing a meridian of M does not extend to a Seifert fibration of M.*
2. *The 3-manifold M is homeomorphic to $T^2 \times I$; in this case, every fibration of $T^2 \times \{0\}$ extends to a unique fibration of M.*
3. *The 3-manifold M is homeomorphic to $U_2 \mathbin{\tilde{\times}} I$; in this case, M admits two Seifert fibrations, one over a Möbius band and one over a disc.*

Proof First of all, let us remark that as soon as one of the surfaces $R_1, \dots, R_n$ has an annulus component, the Seifert fibration for $(M, \boldsymbol{R})$ is necessarily unique: given two such fibrations, the homeomorphism

---

[2] It is immediate to see that exceptions 5.1.1–5.1.5 listed by Johannson do not occur, since M has nonempty boundary and the case of a solid torus with a single exceptional fibre has already been addressed.

id: $M \to M$ is fibre-preserving on $X = \partial R_1 \cup \cdots \cup \partial R_n$; by Lemma 2.7, we can isotope one fibration to the other in $(M, R_1, \dots, R_n)$.

When instead all the components of $R_1, \dots, R_n$ are tori, it is clear that two Seifert fibrations for $(M, \boldsymbol{R})$ are isotopic in $(M, \boldsymbol{R})$ if and only if they are isotopic in M. The conclusion then follows from [9, Theorem VI.18]. □

Remark 2.9 The two Seifert fibrations of $U_2 \mathbin{\tilde{\times}} I$ are *inequivalent*, by which we mean that no self-homeomorphism of $U_2 \mathbin{\tilde{\times}} I$ can send one to the other. In other words, if we fix a Seifert fibration for $U_2 \mathbin{\tilde{\times}} I$, then every self-homeomorphism of this 3-manifold can be isotoped to be fibre-preserving.

Proposition 2.10 *There is an algorithm which, given as input two Seifert $(n+1)$-tuples $(M, \boldsymbol{R})$ and $(M', \boldsymbol{R}')$, decides whether they are homeomorphic or not.*

Proof Fix two Seifert fibrations for $(M, \boldsymbol{R})$ and $(M', \boldsymbol{R}')$. It is very well known (for a thorough and complete discussion, see [19, Theorem 10.4.19]) that computing the Seifert invariants for the fibrations of M and M' leads to a straightforward algorithm for deciding whether M and M' are homeomorphic. If they are not, then clearly neither are $(M, \boldsymbol{R})$ and $(M', \boldsymbol{R}')$. If M and M' are closed and homeomorphic, then we are done. On the other hand, if $(M', \boldsymbol{R}')$ is one of the exceptions described in (1) and (2) of Proposition 2.8, then answering the homeomorphism problem is easy.

Therefore, let us assume that $(M', \boldsymbol{R}')$ has nonempty boundary and is not one of these exceptions, and let $f\colon M \to M'$ be a homeomorphism. We can find a finite set $\mathcal{F}_0$ of fibre-preserving self-homeomorphisms of M' such that every permutation of the boundary components of M' is induced by an element of $\mathcal{F}_0$. Let $\iota\colon M' \to M'$ be a fibre-preserving homeomorphism such that $\iota|_\partial$ acts as $-\mathrm{id}$ on $H_1(\partial M')$, and define

$$\mathcal{F} = \bigcup_{g \in \mathcal{F}} \{g, \iota g\}.$$

We claim that $(M, \boldsymbol{R})$ and $(M', \boldsymbol{R}')$ are homeomorphic if and only if $g^{-1} f\colon M \to M'$ can be isotoped to a homeomorphism $(M, \boldsymbol{R}) \to (M', \boldsymbol{R}')$ for some $g \in \mathcal{F}$ — this condition is easy to check algorithmically. The reverse implication is trivial. Conversely, assume that there is a homeomorphism

$$f'\colon (M, \boldsymbol{R}) \to (M', \boldsymbol{R}').$$

We can isotope f in M so that $f' f^{-1}\colon M' \to M'$ is fibre-preserving. There is a homeomorphism $g \in \mathcal{F}$ such that $f' f^{-1} g$ preserves each boundary component of M', as well as the orientation of the fibres on $\partial M'$. But then the homeomorphism $f' f^{-1} g$ acts as a power of the Dehn twist about a fibre on each boundary component of M', and hence it can be isotoped to a self-homeomorphism of $(M', \boldsymbol{R}')$. We immediately conclude that $g^{-1} f$ can be isotoped to a homeomorphism $(M, \boldsymbol{R}) \to (M', \boldsymbol{R}')$. □

We are now ready to give an algorithmic description of the mapping class group of (most) Seifert n-tuples with boundary.

Proposition 2.11 *There is an algorithm which, given as input a Seifert-fibred $(n+2)$-tuple $(M, \boldsymbol{R}, F)$ with nonempty boundary and not homeomorphic to one of the exceptions described in (1) and (2) of Proposition 2.8, returns*

- *a finite collection $\mathcal{F}$ of self-homeomorphisms of $(\partial M, \boldsymbol{R})$ fixing F pointwise, and*
- *a finite collection $\mathcal{C} = \{(a_1, b_1), \dots, (a_m, b_m)\}$, where $a_1, \dots, a_m, b_1, \dots, b_m$ are pairwise disjoint fibres in $\operatorname{int}(R_1) \cup \dots \cup \operatorname{int}(R_n)$,*

such that

$$\boldsymbol{H}_F(M, \boldsymbol{R})|_\partial = \bigcup_{f \in \mathcal{F}} \langle \tau_{a_1}\tau_{b_1}^{-1}, \dots, \tau_{a_m}\tau_{b_m}^{-1} \rangle f$$

as a subgroup of $\boldsymbol{H}_F(\partial M, \boldsymbol{R})$.

Proof Let $p\colon (M, \boldsymbol{R}, F) \to (B, \overline{\boldsymbol{R}}, \overline{F})$ be the projection to the base surface of the Seifert fibration, and consider a homeomorphism $f \in \boldsymbol{H}_F(M, \boldsymbol{R})$. We can isotope f to be fibre-preserving fixing F pointwise, using Proposition 2.8 and Remark 2.9 if F is empty and Lemma 2.7 otherwise. Denote by $\bar{f}$ the induced self-homeomorphism of $(B, \overline{\boldsymbol{R}}, \overline{F})$, which is not necessarily orientation-preserving.

Identity on $\boldsymbol{B}$ As we will see, the role of $\mathcal{F}$ is to encode the action of $\boldsymbol{H}_F(M, \boldsymbol{R})$ on the space $(\partial B, \overline{\boldsymbol{R}}, \overline{F})$. First of all, note that we can easily list all permutations of the components of ∂B which are induced by elements of $\boldsymbol{H}_F(M, \boldsymbol{R})$. Therefore, we can compute a set of representatives $\mathcal{F}_0 \subseteq \boldsymbol{H}_F(M, \boldsymbol{R})$ such that, up to replacing f with ff_0 for some $f_0 \in \mathcal{F}_0$, we can assume that f sends each component of ∂M to itself.

The homeomorphism f induces a self-homeomorphism of $(\partial B, \overline{\boldsymbol{R}}, \overline{F})$ which fixes $\overline{F}$ pointwise and sends each component of ∂B to itself. If F is nonempty, the homeomorphism $\bar{f}$ acts orientation-preservingly on each component of ∂B; otherwise, it will consistently preserve or reverse the orientations of these components. Either way, there are only finitely many such self-homeomorphisms of $(\partial B, \overline{\boldsymbol{R}}, \overline{F})$, and they are all induced by elements of $\boldsymbol{H}_F(M, \boldsymbol{R})$. We can then compute a set $\mathcal{F}_1 \subseteq \boldsymbol{H}_F(M, \boldsymbol{R})$ such that, up to replacing f with ff_1 for some $f_1 \in \mathcal{F}_1$, we can assume that $\bar{f}$ is the identity on ∂B.

In fact, up to further composing f with a self-homeomorphism of M fixing ∂M pointwise, we can assume that $\bar{f}$ is the identity on all of B; this follows from the fact that if a self-homeomorphism of B fixing ∂B lifts to a self-homeomorphism of M, then it lifts to a self-homeomorphism of M fixing ∂M.

Identity on almost all of $\boldsymbol{\partial M}$ Let $\mathcal{X}$ be the set of components of $\partial M \setminus \operatorname{int}(F)$. For each pair X, Y of different components in $\mathcal{X}$, let $A_{X,Y}$ be a vertical annulus connecting X and Y. We can of course assume that the boundary curves of all these annuli lie in $\operatorname{int}(R_1) \cup \dots \cup \operatorname{int}(R_n)$ and are pairwise disjoint. Moreover, for each component $X \in \mathcal{X}$, we can define k_X to be the unique integer such that $f|_X$ is isotopic to the composition of k_X Dehn twists of X about a fibre, through an isotopy which fixes ∂X pointwise. Up to composing with powers of Dehn twists about the annuli $A_{X,Y}$ — which induce the identity on B — we can assume that $k_X = 0$ for all components $X \in \mathcal{X}$ but one. In other words, we may assume that the restriction of f to $\partial M \setminus \operatorname{int}(X)$ is the identity for some component $X \in \mathcal{X}$.

Identity on ∂M In fact, we now prove that f restricts to the identity on all of ∂M. Let $N_1, \dots, N_m$ be fibred regular neighbourhoods of the exceptional fibres of M. Without loss of generality, we may assume that f restricts to the identity on $N_1 \cup \dots \cup N_m$. Define $M' = M \setminus \operatorname{int}(N_1 \cup \dots \cup N_m)$, which is an S^1-bundle over the surface $B' = B \setminus \operatorname{int}(p(N_1 \cup \dots \cup N_m))$; since B' has nonempty boundary, the bundle is uniquely determined by B'. The homeomorphism f restricts to $f' \colon M' \to M'$ such that f' is the identity on $\partial M' \setminus \operatorname{int}(X)$.

- Let us first deal with the case where B is orientable. Let S be a horizontal copy of B' in M'. At the level of homology, we have that

$$0 = f'_*[\partial S] = [\partial S] + k_X \cdot [c] = k_X \cdot [c] \in H_1(M'),$$

where $[c]$ is the homology class of a fibre. We conclude that $k_X = 0$ and that f does indeed restrict to the identity on ∂M.

- Assume now that B — and hence B' — is nonorientable. Let $\widetilde{M}'$ be the double covering of M' which is a product S^1-bundle over the orientable double covering $\widetilde{B}'$ of B'. Denote by $q \colon \widetilde{M}' \to M'$ the covering map, and by $\iota \colon \widetilde{M}' \to \widetilde{M}'$ the nontrivial deck transformation of q. The homeomorphism f' lifts to $\widetilde{f}' \colon \widetilde{M}' \to \widetilde{M}'$, which is the identity on $\partial \widetilde{M}' \setminus q^{-1}(\operatorname{int}(X))$; denote by $\widetilde{X}_1$ and $\widetilde{X}_2$ the two components of $q^{-1}(X)$, and by k_1 and k_2 the integers such that $\widetilde{f}'|_{\widetilde{X}_i}$ is the $(k_i)^{\text{th}}$ power of the Dehn twist about a fibre for $i \in \{1,2\}$. Let $\widetilde{S}$ be a horizontal copy of $\widetilde{B}'$ in $\widetilde{M}'$. Fixing an orientation of $\widetilde{S}$ allows us to define canonically oriented horizontal curves a_1 and a_2, which are the boundary components of $\widetilde{S}$ intersecting $\widetilde{X}_1$ and $\widetilde{X}_2$, respectively. Moreover, let b_1 and b_2 be fibres of $\widetilde{X}_1$ and $\widetilde{X}_2$, oriented so that they intersect a_1 and a_2 with positive sign. We can then define k_i as the integer such that $[\widetilde{f}'(a_i)] = [a_i] + k_i \cdot [b_i]$ in $H_1(\partial \widetilde{M}')$ for $i \in \{1,2\}$. The homological argument described in the orientable case shows that $k_1 + k_2 = 0$. On the other hand, note that $[\iota(a_1)] = -[a_2]$ and $[\iota(b_1)] = -[b_2]$. Then the relation $\widetilde{f}'\iota = \iota \widetilde{f}'$ readily implies that $k_1 = k_2$, since

$$\begin{aligned}
[\widetilde{f}'\iota(a_1)] &= -[\widetilde{f}'(a_2)] = -[a_2] - k_2 \cdot [b_2], \\
[\iota \widetilde{f}'(a_1)] &= [\iota(a_1)] + k_1 \cdot [\iota(b_2)] = -[a_2] - k_1 \cdot [b_2].
\end{aligned}$$

We conclude that $\widetilde{f}'$ restricts to the identity on $\partial \widetilde{M}'$, and hence f restricts to the identity on ∂M.

Conclusion We have decomposed f as a product

$$t \circ h \circ f_1 \circ f_0 \in \boldsymbol{H}_F(M, \boldsymbol{R}),$$

where $f_0 \in \mathcal{F}_0$, $f_1 \in \mathcal{F}_1$, $h \in \boldsymbol{H}_{\partial M}(M)$, and t is a product of powers of Dehn twists of M about the annuli $A_{X,Y}$ for $X, Y \in \mathcal{X}$. The algorithm will then return

$$\mathcal{F} = \{(f_0 f_1)|_\partial : f_0 \in \mathcal{F}_0,\ f_1 \in \mathcal{F}_1\} \quad \text{and} \quad \mathcal{C} = \{\partial A_{X,Y} : X, Y \in \mathcal{X} \text{ and } X \neq Y\}. \qquad \square$$

As an easy consequence, we can also solve the extension problem for Seifert n-tuples.

Corollary 2.12 *Let $(M, \boldsymbol{R}, F)$ and $(M', \boldsymbol{R}', F')$ be Seifert-fibred $(n{+}2)$-tuples not homeomorphic to one of the exceptions described in (1) and (2) of Proposition 2.8, and let $f : F \to F'$ be a fibre-preserving homeomorphism. There is an algorithm which, given as input the two $(n{+}2)$-tuples and f, decides whether f extends to a (not necessarily fibre-preserving) homeomorphism*

$$(M, \boldsymbol{R}, F) \to (M', \boldsymbol{R}', F').$$

Proof It suffices to be able to answer the following two questions algorithmically.

(1) Are the $(n{+}2)$-tuples $(M, \boldsymbol{R}, F)$ and $(M', \boldsymbol{R}', F')$ homeomorphic? Note that if they are, then we can choose the fibration on $(M, \boldsymbol{R}, F)$ so that they are fibre-preservingly so.

(2) Given a fibre-preserving homeomorphism $g\colon F \to F$, does it extend to a self-homeomorphism of $(M, \boldsymbol{R}, F)$?

The first question is addressed in Proposition 2.10. Thanks to Proposition 2.11, we can reduce the second to the following problem (if $\partial M = \varnothing$, we are already done).

(2') Given fibres $a_1, \dots, a_m, b_1, \dots, b_m$ in ∂M, does g extend to a self-homeomorphism of $(\partial M, \boldsymbol{R}, F)$ belonging to the subgroup of $\boldsymbol{H}(\partial M, \boldsymbol{R}, F)$ generated by $\tau_{a_1}\tau_{b_1}^{-1}, \dots, \tau_{a_m}\tau_{b_m}^{-1}$?

If g permutes the components of F in a nontrivial way or it reverses the orientation of the fibres on some component of F then the answer is no, since Dehn twists along fibres preserve components of F and the orientation of the fibres. Otherwise, the homeomorphism g acts like the identity on each annulus component of F, and is an integer power of the Dehn twist about a fibre on each torus component of F. Moreover, for each torus component T of F, we can compute the integer k_T such that $g|_T$ is the composition of k_T Dehn twists about a fibre. Clearly, the answer to the extension problem for g only depends on the values k_T, and is surely yes if they are all equal to zero.

Let us define a graph Γ, whose vertices are in one-to-one correspondence with components of ∂M. Each pair (a_i, b_i) defines an edge of Γ, joining the two vertices corresponding to the two (not necessarily distinct) boundary tori of M containing a_i and b_i. It is not hard to see that, if T is a torus component of F, the pair (a_i, b_i) represents an edge of Γ with $a_i \subseteq T$, and X is the component of ∂M which contains b_i, then the following operations do not change the answer to the extension problem for g:

- if $X \subseteq F$, then increase k_T by 1 and decrease k_X by 1, or vice versa;
- otherwise, increase or decrease k_T by 1.

By the proof of Proposition 2.11, the graph Γ is connected. Therefore, we immediately conclude that the answer to the extension problem is yes if and only if one of the following condition holds:

(i) the boundary of M is not completely covered by F;

(ii) the boundary of M is completely covered by F, and $\sum_T k_T = 0$. □

2.4 I-bundle pieces

We now move on to the — somewhat easier — study of I-bundle pieces. We first show how to solve the extension problem, and then address the mapping class group in Proposition 2.14.

Proposition 2.13 *Let (M, R, F) and (M', R', F') be I-bundle triples, and let $f\colon F \to F'$ be a homeomorphism. There is an algorithm which, given as input the two triples, decides whether f extends to a homeomorphism*

$$(M, R, F) \to (M', R', F').$$

Proof It suffices to be able to solve the following two questions algorithmically.

(1) Are the triples (M, R, F) and (M', R', F') homeomorphic?

(2) Given a homeomorphism $g\colon F \to F$, does it extend to a self-homeomorphism of (M, R, F)?

The first question can be easily answered, since we can compute the base surfaces of the I-bundles (M, R, F) and (M', R', F') and check if they are the same. As far as the second is concerned, we may assume that g is fibre-preserving. Moreover, since the group $\boldsymbol{H}(M, R, F)$ acts transitively on the components of F, we can further assume that g maps each component of F to itself. Now, the restriction of g to an annulus component A of F can either preserve or swap the boundary components of A. If g consistently preserves or swaps the boundary components of every component of F, then it extends to a self-homeomorphism of (M, R, F); otherwise, it does not. □

Proposition 2.14 *There is an algorithm which, given as input an I-bundle triple (M, R, F), returns*

- *a finite collection $\mathcal{F}$ of self-homeomorphisms of $(\partial M, R, F)$ fixing F pointwise, and*
- *a finite collection $\mathcal{C} = \{(a_1, b_1), \dots, (a_m, b_m)\}$, where $a_1, \dots, a_m, b_1, \dots, b_m$ are curves in $\operatorname{int}(R)$ with $a_1 \cap b_1 = \dots = a_m \cap b_m = \varnothing$,*

such that

$$\boldsymbol{H}_F(M, R)|_\partial = \bigcup_{f \in \mathcal{F}} \langle \tau_{a_1}\tau_{b_1}^{-1}, \dots, \tau_{a_m}\tau_{b_m}^{-1} \rangle f$$

as a subgroup of $\boldsymbol{H}_F(\partial M, R)$. Moreover, if (M, R, F) is the product I-bundle over Σ_0, $\Sigma_{0,1}$, $\Sigma_{0,2}$, or $\Sigma_{0,3}$, or the twisted I-bundle over U_1, $U_{1,1}$, or $U_{1,2}$, then the curves in $\mathcal{C}$ are pairwise disjoint.

Proof Let $p\colon (M, F) \to (B, \partial B)$ be the projection to the base surface of the I-bundle, and consider a homeomorphism $f \in \boldsymbol{H}_F(M, R)$. We can isotope f to be fibre-preserving fixing F pointwise. Denote by $\bar{f}$ the induced self-homeomorphism of B, which is not necessarily orientation-preserving. We now distinguish two cases, depending on whether B is orientable or not.

- If B is orientable, then $\bar{f}$ is necessarily orientation-preserving, unless $F = \varnothing$. If this is the case, let us choose a homeomorphism $\iota \in \boldsymbol{H}(M, R)$ such that the induced homeomorphism $\bar{\iota}\colon B \to B$ is orientation-reversing; otherwise, we set $\iota = \mathrm{id}\colon M \to M$. Up to composing f with ι, we can assume that $\bar{f}$ is orientation-preserving. There is a well-known explicit set of curves $c_1, \dots, c_m \subseteq B$ such that $\tau_{c_1}, \dots \tau_{c_m}$ generate $\boldsymbol{H}_{\partial B}(B)$ — see, for instance, [5, Section 4.4.4]. As a consequence, the homeomorphism f belongs to the subgroup of $\boldsymbol{H}_F(M, R)$ generated by $\tau_{A_1}, \dots, \tau_{A_m}$, where $A_i = p^{-1}(c_i)$ for $1 \leqslant i \leqslant m$. The algorithm will then return

$$\mathcal{F} = \{\mathrm{id}, \iota|_\partial\} \quad \text{and} \quad \mathcal{C} = \{\partial A_i : 1 \leqslant i \leqslant m\}.$$

Note that, if B is one of Σ_0, $\Sigma_{0,1}$, $\Sigma_{0,2}$, or $\Sigma_{0,3}$, then the curves $c_1, \dots, c_m$ can be chosen to be disjoint.

- If B is nonorientable, denote by $\boldsymbol{H}^{\pm}_{\partial B}(B)$ the group of self-homeomorphisms of B fixing ∂B pointwise, modulo isotopies through homeomorphisms of the same kind. Chillingworth (in [3, Section 3]) and Kobayashi and Omori (in [14, Proposition 3.2]), respectively, for the cases $\partial B = \varnothing$ and $\partial B \neq \varnothing$, provide an explicit set of two-sided curves $c_1, \dots, c_m \subseteq B$ and a homeomorphism $y \in \boldsymbol{H}^{\pm}_{\partial B}(B)$ such that

$$\boldsymbol{H}^{\pm}_{\partial B}(B) = \langle \tau_{c_1}, \dots, \tau_{c_m}, y \rangle.$$

The homeomorphism y is a *Y-homeomorphism*, a description of which can be found in [18, Section 2]; in particular, we have that y^2 is a Dehn twist about a curve c_0 which can be computed explicitly. Easy algebraic manipulations show that

$$\boldsymbol{H}^{\pm}_{\partial B}(B) = \bigcup_{g \in \{\mathrm{id}, y\}} \langle \tau_{c_1}, \dots, \tau_{c_m}, y\tau_{c_1}y^{-1}, \dots, y\tau_{c_m}y^{-1}, y^2 \rangle g.$$

Define $c_{m+i} = y(c_i)$ for $1 \leqslant i \leqslant m$, so that $y\tau_{c_i}y^{-1} = \tau_{c_{m+i}}$; let $h \in \boldsymbol{H}_F(M, R)$ be a homeomorphism such that $\bar{h} = y$. The algorithm will then return

$$\mathcal{F} = \{\mathrm{id}, h|_\partial\} \quad \text{and} \quad \mathcal{C} = \{\partial p^{-1}(c_i) : 0 \leqslant i \leqslant 2m\}.$$

Note that, if B is one of $U_{1,0}$, $U_{1,1}$, or $U_{1,2}$, then the curves $c_1, \dots, c_m$ can be chosen to be disjoint. □

2.5 Simple pieces

We finally turn our attention to simple pieces. As anticipated, instead of using Johannson's original approach with hierarchies, we will prove the finiteness of the mapping class group of these pieces by means of hyperbolic geometry and the geometrisation theorem. While sacrificing generality, this strategy will allow us to bypass the algorithmically challenging induction procedure of Johannson.

In Sections 2.3 and 2.4 we dealt with Seifert n-tuples and I-bundle triples in full generality (save for a few exceptions). For simple triples, instead, we will restrict our attention to a subclass satisfying additional conditions. This will, of course, impose further restrictions on the type of 3-manifold pairs which our final algorithm will accept as input. However, we will maintain a level of generality which will be sufficient for our purposes. The additional constraints are listed in the following definition.

Definition 2.15 A 3-manifold triple (M, R, F) is *reasonable* if it satisfies the following properties:

- the pair (M, R) is irreducible;
- no component of R is a disc;
- the surface F is incompressible in M;
- each component of F is a torus or an annulus, and is not boundary parallel in (M, R);
- each component of $\partial M \setminus \operatorname{int}(R \cup F)$ is an annulus.

Part of the appeal of simple reasonable triples is that the two surfaces R and F cover the whole boundary of the 3-manifold M, save for a few explicit exceptions.

Lemma 2.16 *Let (M, R, F) be a reasonable 3-manifold triple. Then one of the following holds.*

(i) *The triple (M, R, F) is not simple.*

(ii) *The surfaces R and F cover the boundary of M.*

(iii) *The 3-manifold M is a solid torus, and $\partial M \setminus \operatorname{int}(R \cup F)$ is a union of annuli which are incompressible in M.*

(iv) *The 3-manifold M is homeomorphic to $T^2 \times I$ in such a way that $T^2 \times \{1\}$ is a component of R or F.*

Proof Suppose that (M, R, F) is simple, and that $\partial M \setminus \operatorname{int}(R \cup F)$ is nonempty. Let A be a component of this surface; by assumption, A is an annulus. Let T be the boundary component of M containing A; we show that T is a torus. Let A' be the component of $\partial M \setminus \operatorname{int}(R)$ containing A. If A' is a torus, then $A' = T$ and we are done. Otherwise, the surface A' is an annulus. If A' were compressible in M, then one component of $\partial A' \subseteq \partial R$ would bound a disc in M; since R is incompressible and has no disc components, this is impossible. Therefore, A' is incompressible. Since the triple (M, R, F) is simple, it follows that T is indeed a torus.

- If T is compressible, then M is in fact a solid torus, and the components of R, F, and $\partial M \setminus \operatorname{int}(R \cup F)$ are parallel annuli in ∂M; by assumption, these annuli are incompressible in M.
- If T is incompressible, then it must be parallel in (M, R) to a torus in R or in F. This immediately implies that M is homeomorphic to $T^2 \times I$, where $T^2 \times \{0\}$ is identified with T and $T^2 \times \{1\}$ is a component of either R or F. □

As a final restriction, we only want to deal with simple triples which do not belong to the classes we have already analysed — namely, I-bundles and Seifert pieces.

Definition 2.17 A 3-manifold triple (M, R, F) is *strongly simple* if it is simple, it is not an I-bundle triple, and $(M, R, F, \partial M \setminus \operatorname{int}(R \cup F))$ is not a Seifert 4-tuple.

Under the additional assumptions we have listed, we can now prove that the mapping class group of a simple triple is finite, and present an algorithm to compute it. At the same time, we will describe a solution to the homeomorphism problem for simple triples.

Proposition 2.18 *The following hold for reasonable strongly simple* 3*-manifold triples* (M, R, F) *and* (M', R', F')*.*

1. *There is an algorithm which, given as input the triples* (M, R, F) *and* (M', R', F')*, decides whether they are homeomorphic.*
2. *The group* $\boldsymbol{H}(M, R, F)$ *is finite. Moreover, there is an algorithm which, given as input the triple* (M, R, F)*, returns representatives of* $\boldsymbol{H}(M, R, F)$*.*

Proof By Lemma 2.16, we immediately see that, if R and F do not cover the boundary of M, the 4-tuple $(M, R, F, \partial M \setminus \operatorname{int}(R \cup F))$ admits a Seifert fibration. Therefore, we have that $\partial M = R \cup F$.

Boundary incompressibility of $\boldsymbol{F}$ We claim that F is boundary incompressible in (M, R) in the following sense: for every disc D in M whose boundary can be written as the union of two arcs a and b with $a \cap b = \partial a = \partial b$, $D \cap F = a$, and $D \cap R = b$, we have that a cuts a disc off of F.

In fact, suppose by contradiction that D is a disc as described above, with a an essential arc in an annulus component A of F. Let A' be the disc obtained by boundary compressing A along D, that is, $A' = (A \setminus \partial_v \mathcal{N}(D)) \cup \partial_h \mathcal{N}(D)$. By irreducibility of (M, R), the disc A' cobounds a 3-ball B with some disc in R. If B is disjoint from D then A is boundary parallel. If instead B contains D then A is compressible, contradicting incompressibility of F.

The easy case Let us deal first with the case where every component of R is a torus; as a consequence, the same holds for the components of F. By assumption, we have that M is irreducible, boundary irreducible, atoroidal,[3] and not a Seifert manifold; additionally, every boundary component of M is a torus. By geometrisation, the interior of M admits a complete finite-volume hyperbolic structure. We can now invoke [16, Theorem 6.1] and conclude that

- if (M', R', F') is another reasonable strongly simple 3-manifold triple such that each component of R' and F' is a torus, then we can algorithmically decide whether M and M' are homeomorphic or not;
- the group $\boldsymbol{H}(M)$ is finite and can be computed algorithmically.

---

[3] There are several slightly different definitions of "atoroidal" in the literature. In this article, by "X is atoroidal" we mean that every incompressible torus in X is boundary parallel in X or, equivalently, that the triple $(X, \varnothing, \partial X)$ is simple. Other definitions could be inequivalent in general, but they all agree with ours when X is not a Seifert manifold.

Note that $\boldsymbol{H}(M, R, F)$ is a subgroup of $\boldsymbol{H}(M)$, and checking if an element of $\boldsymbol{H}(M)$ belongs to $\boldsymbol{H}(M, R, F)$ is trivial; this is enough to prove statement (2). As far as statement (1) is concerned, we simply check whether M and M' are homeomorphic. If they are, we find a homeomorphism $f\colon M \to M'$ and consider the (finitely many) elements of $\boldsymbol{H}(M')f$: if any of these gives a homeomorphism $(M, R, F) \to (M', R', F')$ then the two triples are homeomorphic, otherwise they are not.

Doubling M We now turn to the more involved case where not all components of R are tori. Let R_0 be the union of the components of R which are not tori. Let $\overline{M}$ be the result of gluing two copies of M along R_0. More precisely, we define

$$\overline{M} = M \times \{0, 1\}/\{(x, 0) \sim (x, 1) : x \in R_0\}.$$

By assumption, the surface R_0 is nonempty, so the 3-manifold $\overline{M}$ is connected. Note that every boundary component of $\overline{M}$ is either a copy of a torus component of R, a copy of a torus component of F, or the union of two copies of an annulus component of F. As a consequence, we see that every component of $\partial\overline{M}$ is a torus.

The surface R_0 has a canonical copy $\overline{R_0}$ in $\overline{M}$ — specifically, the image in the quotient of $R_0 \times \{0\}$; this surface is incompressible in $\overline{M}$. More generally, if S is a surface in M, we let $\overline{S}$ be the "doubled surface", that is, the image in the quotient of the surface $S \times \{0, 1\} \subseteq M \times \{0, 1\}$. Similarly, if $f\colon (M, R_0) \to (M, R_0)$ is a function, we let $\bar{f}\colon \overline{M} \to \overline{M}$ be the function defined by $\bar{f}(x, i) = (f(x), i)$ for $(x, i) \in M \times \{0, 1\}$.

Note that $\overline{M}$ comes equipped with a natural orientation-reversing involution $\iota\colon \overline{M} \to \overline{M}$ defined by $\iota(x, i) = (x, 1-i)$; for every homeomorphism $f\colon (M, R, F) \to (M, R, F)$, we have the relation $\bar{f}\iota = \iota\bar{f}$. Finally, let us remark that, since $\overline{R_0}$ is incompressible, the 3-manifold $\overline{M}$ is sufficiently large.

The double is irreducible and boundary irreducible Irreducibility of $\overline{M}$ follows immediately from the fact that M is irreducible and $\overline{R_0}$ is incompressible in $\overline{M}$.

Let now D be a disc properly embedded in $\overline{M}$. Up to isotopy, we can assume that D is in general position with respect to $\overline{R_0}$. Additionally, let us isotope D so that the number of components of $D \cap \overline{R_0}$ is as small as possible. A standard innermost circle argument on D shows that we can assume that all these components are arcs. If $D \cap \overline{R_0}$ is empty, then D can be interpreted as a compression disc for either R or F in M, and hence it must be trivial in M and in $\overline{M}$ as well. Otherwise, let a be an arc in $D \cap \overline{R_0}$ which is outermost in D; the arc a cuts off a disc E from D such that $E \cap \overline{R_0} = a$. Note that $a \subseteq \overline{R_0}$, and $\partial D \setminus \operatorname{int}(a)$ is an arc in $\overline{F}$. The disc E can then be interpreted as a boundary compression disc for F in (M, R). By boundary incompressibility of F, it is easy to see that D can be isotoped to remove a from the intersection $D \cap \overline{R_0}$, thus contradicting our minimality assumption.

The double is strongly simple More precisely, we prove that the triple $(\overline{M}, \overline{R_1}, \overline{F})$ is strongly simple, where $R_1 = R \setminus R_0$ is the union of the torus components of R. First, let us show that $\overline{M}$ is not a Seifert manifold. If by contradiction $\overline{M}$ admits a Seifert fibration, then up to isotopy we can assume that $\overline{R_0}$ is either horizontal or vertical. Since M is homeomorphic to either component of $\overline{M} \setminus\setminus \overline{R_0}$, it is easy to see that

- if $\overline{R_0}$ is horizontal, then M inherits an I-bundle structure; the surfaces R and F have no torus components, and in fact they are the horizontal and vertical boundary, respectively, of the fibration inherited by M;
- if $\overline{R_0}$ is vertical, then M is itself Seifert-fibred, in such a way that R and F are unions of fibres.

Both cases are incompatible with (M, R, F) being strongly simple.

We now address the "torus" condition for simple triples. Let T be an incompressible torus in $\overline{M}$. By the equivariant torus theorem (see [8, Corollary 4.6]), we can assume that T is either disjoint from or equal to $\iota(T)$.

- If T is disjoint from $\iota(T)$, then we can think of T as a torus in M; incompressibility of $\overline{R_0}$ in $\overline{M}$ implies that T is incompressible in M. Since (M, R, F) is a simple triple, T must be parallel in M to a component of R or F. But then T is boundary parallel in $\overline{M}$ as well.
- If $T = \iota(T)$, then $T = \bar{A}$ for some annulus A properly embedded in (M, R). The annulus A is incompressible in M since T is incompressible in $\overline{M}$. Then A must be parallel in (M, R) to an annulus in R or in F; the first case is impossible, for otherwise T would be compressible in $\overline{M}$. Therefore, A is parallel in (M, R) to an annulus component A' of F. We conclude that $T = \bar{A}$ is parallel to the boundary component $\overline{A'}$ of $\overline{M}$.

Finally, let us consider the "annulus" condition. Let A be an incompressible annulus properly embedded in $(\overline{M}, \overline{R_1})$. If A is boundary compressible then it is boundary parallel. Otherwise, by the equivariant annulus theorem (see [15]), we can assume that A is either disjoint from or equal to $\iota(A)$.

- If A is disjoint from $\iota(A)$, then we can think of A as an incompressible annulus properly embedded in (M, R). By simplicity of (M, R, F), this implies that A is boundary parallel in M and, hence, in $\overline{M}$ as well.
- If $A = \iota(A)$, then $A = \overline{A_0}$ for some incompressible annulus $A_0 \subseteq M$ whose boundary intersects two boundary components of M, which contradicts simplicity of (M, R, F).

The double is hyperbolic It is not hard to verify that the triple $(\overline{M}, \overline{R_1}, \overline{F})$ is reasonable. Since, by definition, every component of $\overline{R_1}$ is a torus, we have already proved statements (1) and (2) for this triple. In the process, we also noted that the interior of $\overline{M}$ admits a complete finite-volume hyperbolic metric. In fact, for our purposes, it will be more convenient to think of $\overline{M}$ as a compact hyperbolic 3-manifold whose boundary components are flat tori — in other words, as the result of truncating the cusps of a complete finite-volume noncompact hyperbolic 3-manifold. By combining Mostow–Prasad rigidity with [27, Theorem 7.1], we have the following.

(∗) Every self-homeomorphism of $\overline{M}$ is isotopic to a unique isometry $\overline{M} \to \overline{M}$.

As a consequence, let us remark that we can choose the hyperbolic metric on $\overline{M}$ in such a way that ι is an isometry. In fact, the involution ι is isotopic to an isometry $\hat{\iota}\colon \overline{M} \to \overline{M}$. Note that $\hat{\iota}$ is itself an involution: the isometry $\hat{\iota}\circ\hat{\iota}$ is isotopic to $\iota\circ\iota=\mathrm{id}$ and, therefore, equal to the identity. We now apply [25, Theorem B][4] to deduce that $\hat{\iota}=h\iota h^{-1}$ for some homeomorphism $h\colon \overline{M}\to\overline{M}$. Then ι is an isometry with respect to the pull-back of the hyperbolic metric on $\overline{M}$ by h.

Solving the homeomorphism problem Consider another reasonable strongly simple 3-manifold triple (M',R',F'); suppose that, additionally, there is at least one component of R' which is not a torus — otherwise (M,R,F) and (M',R',F') could not possibly be homeomorphic. Define $\overline{M'}$, ι', and R_1' like we did with $\overline{M}$, ι, and R_1. Recall that $\overline{M'}$ can be endowed with a finite-volume hyperbolic metric with flat boundary, with respect to which ι' is an isometry. Naturally, $(*)$ also holds for homeomorphisms $\overline{M}\to\overline{M'}$.

If the triples $(\overline{M},\overline{R_1},\overline{F})$ and $(\overline{M'},\overline{R_1'},\overline{F'})$ are not homeomorphic — which we can algorithmically decide thanks to statement (1) — then clearly neither are (M,R,F) and (M',R',F'). Otherwise, let

$$f\colon (\overline{M},\overline{R_1},\overline{F})\to(\overline{M'},\overline{R_1'},\overline{F'})$$

be a homeomorphism. We claim that the following are equivalent:

(i) There is a homeomorphism $g\colon (M,R,F)\to(M',R',F')$ such that

$$\overline{g}\colon (\overline{M},\overline{R_1},\overline{F})\to(\overline{M'},\overline{R_1'},\overline{F'})$$

is isotopic to f.

(ii) The homeomorphism $f^{-1}\iota' f\iota\colon \overline{M}\to\overline{M}$ is isotopic to the identity.

It is clear that (i) implies (ii). Conversely, if (ii) holds, we can assume up to isotopy that f restricts to an isometry on the interior of $\overline{M}$. But then $f^{-1}\iota' f\iota$ is an isometry of $\overline{M}$ which is isotopic to the identity and, hence, equal to it. In other words, the following diagram commutes:

$$\begin{tikzcd}
(\overline{M},\overline{R_1},\overline{F}) \arrow[r, "f"] \arrow[d, "\iota"] & (\overline{M'},\overline{R_1'},\overline{F'}) \arrow[d, "\iota'"] \\
(\overline{M},\overline{R_1},\overline{F}) \arrow[r, "f"] & (\overline{M'},\overline{R_1'},\overline{F'})
\end{tikzcd}$$

We then conclude that f is induced by a homeomorphism $g\colon M\to M'$; it is easy to see that g must send F to F' and, therefore, R to R'.

In order to decide whether (M,R,F) and (M',R',F') are homeomorphic, we can then carry out the following procedure. Let f_0 be any homeomorphism from $(\overline{M},\overline{R_1},\overline{F})$ to $(\overline{M'},\overline{R_1'},\overline{F'})$. Thanks to statement (1), we can compute representatives for the finite group $\boldsymbol{H}(\overline{M'},\overline{R_1'},\overline{F'})$. For each element f of $\boldsymbol{H}(\overline{M'},\overline{R_1'},\overline{F'})f_0$, we check whether $f^{-1}\iota' f\iota$ is isotopic to the identity of $\overline{M}$ (we can do this algorithmically since $\boldsymbol{H}(\overline{M},\overline{R_1},\overline{F})$ is finite and we can compute representatives for it). If this happens for some f, then (M,R,F) and (M',R',F') are homeomorphic, otherwise they are not.

---

[4] Using Tollefson's notation, the case where $\beta\neq\mathrm{id}$ does not apply here, since the fundamental group of complete finite-volume hyperbolic 3-manifolds has trivial centre.

Computing the mapping class group By the very same argument as the one presented in the previous step, we can compute representatives of the subgroup

$$\{f \in \boldsymbol{H}(\overline{M}, \overline{R_1}, \overline{F}) : f \text{ is isotopic to } \overline{g} \text{ for some } g \in \boldsymbol{H}(M, R, F)\} \leqslant \boldsymbol{H}(\overline{M}, \overline{R_1}, \overline{F}).$$

We claim that the group homomorphism $\boldsymbol{H}(M, R, F) \to \boldsymbol{H}(\overline{M}, \overline{R_1}, \overline{F})$ sending g to $\overline{g}$ is injective; in other words, if $\overline{g}$ is isotopic to the identity in $\overline{M}$ then g is isotopic to the identity in (M, R, F).

- First of all, if $\overline{g}$ is isotopic to the identity, then we can isotope g in (M, R, F) so that it restricts to the identity on $R_1 \cup F$. As a consequence, the homeomorphism $\overline{g}$ restricts to the identity on $\partial\overline{M}$.

- Let $\pi\colon \widetilde{M} \to \overline{M}$ be the universal covering; we will think of $\widetilde{M}$ as a closed subset of hyperbolic 3-space $\mathbb{H}^3$, whose complement is a collection of open horoballs. We define a retraction $r\colon \mathbb{H}^3 \to \widetilde{M}$ as follows. If x is a point of $\widetilde{M}$ we set $r(x) = x$. If instead x is contained in a horoball B, we set $r(x) = \gamma \cap \partial B$, where γ is the geodesic emanating from the centre of B and passing through x. Finally, if x and y are two (not necessarily distinct) points of $\widetilde{M}$, we define $\gamma_{x,y}\colon [0,1] \to \mathbb{H}^3$ to be the geodesic such that $\gamma_x(0) = x$ and $\gamma_{x,y}(1) = y$.

- Let $\widetilde{g}\colon \widetilde{M} \to \widetilde{M}$ be a homeomorphism that is a lift of $\overline{g}$. Since $\overline{g}$ is isotopic to the identity in $\overline{M}$ and restricts to the identity on $\partial\overline{M}$, we can choose $\widetilde{g}$ so that it restricts to the identity on $\partial\widetilde{M}$. Let us define a homotopy $\widetilde{g}_t\colon \widetilde{M} \to \widetilde{M}$ as $\widetilde{g}_t(x) = r(\gamma_{x,\widetilde{g}(x)}(t))$, so that $\widetilde{g}_0 = \mathrm{id}$ and $\widetilde{g}_1 = \widetilde{g}$. It is easy to check that $\widetilde{g}_t$ is invariant under deck transformations of π, and hence it descends to a homotopy $\overline{g}_t\colon \overline{M} \to \overline{M}$ between the identity and $\overline{g}$. By definition, for each $t \in [0,1]$ the homeomorphism $\overline{g}_t$ fixes $\partial\overline{M}$ pointwise.

- Since $\iota\overline{g} = \overline{g}$ and the homotopy $\overline{g}_t$ has been defined only in terms of the metric, we must have that $\iota\overline{g}_t = \overline{g}_t$ for every $t \in [0,1]$. In particular, we find that $\overline{g}_t(\overline{R_0}) \subseteq \overline{R_0}$, from which we deduce that $g|_{R_0}$ is homotopic to the identity in R_0, through a homotopy that fixes ∂R_0. By a classical result of Epstein (namely, [4, Theorem 6.4]), this implies that $g|_{R_0}$ is actually isotopic to the identity in R_0, and we can therefore assume that $g|_\partial\colon \partial M \to \partial M$ is the identity.

- The homotopy $\overline{g}_t$ preserves the two components of $\overline{M} \mathbin{\backslash\backslash} \overline{R_0}$ for each $t \in [0,1]$. Therefore, there is an induced homotopy $g_t\colon M \to M$ between g and the identity, which restricts to the identity on ∂M for each $t \in [0,1]$. By [27, Theorem 7.1], this implies that g is isotopic to the identity in (M, R, F).

As a final remark, given $f \in \boldsymbol{H}(\overline{M}, \overline{R_1}, \overline{F})$, we can algorithmically find $g \in \boldsymbol{H}(M, R, F)$ such that f is isotopic to $\overline{g}$, provided that one exists. Therefore, we have presented a complete algorithm to compute representatives of $\boldsymbol{H}(M, R, F)$. □

2.6 Computing the JSJ decomposition

After analysing the types of pieces individually, we now show how to actually compute the JSJ decomposition. The next three propositions explain how to recognise Seifert pieces, I-bundle pieces and (strongly) simple pieces, respectively. Once we have these three algorithms, computing the JSJ decomposition is straightforward, as the proof of Theorem 2.22 witnesses.

Proposition 2.19 *There is an algorithm which, given a* 3*-manifold* $(n+1)$*-tuple* $(M, \boldsymbol{R})$ *as input, decides whether it is a Seifert* $(n+1)$*-tuple or not.*

Proof If $R_1 \cup \cdots \cup R_n \neq \partial M$, or one of the surfaces $R_1, \dots, R_n$ has a component which is not a torus or an annulus, then the answer is no. We can use [11, Algorithm 8.2] to decide whether M is a Seifert manifold or not. If it is not then we are done. If M is homeomorphic to a solid torus, a product $T^2 \times I$, or a twisted I-bundle $U_2 \tilde{\times} I$, then we can easily decide if $(M, \boldsymbol{R})$ is a Seifert $(n+1)$-tuple or not. Otherwise, find the unique Seifert fibration for M; if it is compatible with the surfaces $R_1, \dots, R_n$ then the answer is yes; otherwise, the answer is no. □

Proposition 2.20 *There is an algorithm which, given a* 3*-manifold triple* (M, R, F) *as input, decides whether it is an* I*-bundle triple or not.*

Proof If any of the following hold, then (M, R, F) is not an I-bundle triple:

- at least one component of F is not an annulus;
- R has zero or more than two components;
- R has two nonhomeomorphic components;
- there are one component of R and one component of F which are disjoint.

Otherwise, let B be one component of R if R has two components, or the nonorientable surface doubly covered by R if R has only one component; if (M, R, F) is an I-bundle triple, then it is the (product or twisted) I-bundle over B. Let $p \subseteq F$ be the union of the core curves of the annulus components of F. Let X be the I-bundle over B, and denote by $q \subseteq \partial X$ the union of the core curves of the annuli making up the vertical boundary of X. Then (M, R, F) is an I-bundle triple if and only if the 3-manifolds with boundary pattern (M, p) and (X, q) are homeomorphic. □

Proposition 2.21 *There is an algorithm which, given a reasonable* 3*-manifold triple* (M, R, F) *as input, decides whether it is a simple triple or not.*

Proof Let us first deal with the cases where M is homeomorphic to a solid torus or to $T^2 \times I$.

- In a solid torus, every incompressible annulus is boundary parallel. Therefore, there are only finitely many incompressible annuli properly embedded in (M, R) up to isotopy. We can enumerate them and check whether any of them is not parallel to a component of R or F.

- When M is homeomorphic to $T^2 \times I$, the triple (M, R, F) is not simple if and only if one of the following happens:
  - $T^2 \times \{i\} \subseteq R$ and $T^2 \times \{i\} \neq R$ for some $i \in \partial I$;
  - there are isotopic curves $c_0, c_1 \subseteq T^2$ such that $c_i \times \{i\} \subseteq \partial R$ for each $i \in \partial I$;
  - there are two annulus components R_1 and R_2 of $R \cap (T^2 \times \{i\})$ for some $i \in \partial I$, such that neither component of $(T^2 \times \{i\}) \setminus \mathrm{int}(R_1 \cup R_2)$ is contained in F.

From now on, we will assume that M is neither a solid torus nor homeomorphic to $T^2 \times I$. Moreover, thanks to Lemma 2.16, we can assume that $\partial M = R \cup F$. Let $p = R \cap F$; using the terminology of [20], we have that the 3-manifold with boundary pattern (M, p) is boundary irreducible.

Detecting essential tori We can use [20, Lemma 6.4.7] to decide if M contains an incompressible torus which is not parallel to a boundary component of M. If it does, then (M, R, F) is not simple. Likewise, if M has a torus boundary component which is not a component of R or F, then the triple (M, R, F) is not simple. If none of these two condition is satisfied, then (M, R, F) satisfies the "torus" condition for being simple, and we only need to check for incompressible annuli.

Detecting essential annuli We can use [20, Lemma 6.4.8] to decide if M contains an incompressible annulus A with the following properties:

- the boundary of A lies in $\operatorname{int}(R)$;
- A is not parallel to a union of annulus components of R and F;
- every nontrivial boundary compression disc for A intersects F.

If there is such an annulus, then (M, R, F) is not simple. If there is an annulus component of R whose two boundary curves lie in different components of F, then (M, R, F) is not simple. If none of these two conditions holds, then (M, R, F) satisfies the "annulus" condition for being simple, and is therefore simple. □

Theorem 2.22 *Let (M, R) be an irreducible sufficiently large 3-manifold pair. Suppose that no component of R is a disc and that $\partial M \setminus \operatorname{int}(R)$ is a (possibly empty) collection of annuli. There is an algorithm which, given as input the pair (M, R), returns the JSJ system F of (M, R). Moreover, each component of $(M, R) \setminus\!\!\setminus F$ is reasonable.*

Proof It is easy to check that, under the assumptions made in the statement, the pieces of the JSJ decomposition of (M, R) are reasonable. We can search through all surfaces F' properly embedded in (M, R) whose components are annuli and tori, until we find one satisfying properties (i) and (ii) of Definition 2.5 and, additionally, the property that each component of $(M, R) \setminus\!\!\setminus F'$ is reasonable. All these conditions can be checked algorithmically, and we are guaranteed to find at least one such surface F'. Then, for every union of components of F', we check whether it still satisfies the above properties and, among those that do, we return one that is minimal. □

Let us remark that the additional constraints we put on the pair (M, R) — namely the fact that R has no disc components and $\partial M \setminus \operatorname{int}(R)$ is union of annuli — are a consequence of the assumption of reasonableness we require for simple pieces.

2.7 Piecing things together

For the sake of convenience, we will slightly deviate from the statement of Theorem 2.6 and classify pieces of the JSJ decomposition as follows. If (M, R) is an irreducible sufficiently large 3-manifold pair and F is its JSJ system, we say that a component (N, R', F') of $(M, R) \mathbin{\backslash\backslash} F$ is

- a *Seifert piece* if $(N, R', F', \partial N \setminus \operatorname{int}(R' \cup F'))$ is a Seifert 4-tuple;
- an *I-bundle piece* if (N, R', F') is an I-bundle triple;
- a *strongly simple piece* if (N, R', F') is a strongly simple triple.

It is clear that every component of $(M, R) \mathbin{\backslash\backslash} F$ is either a Seifert piece, an I-bundle piece, or a strongly simple piece. Note that the three possibilities are not mutually exclusive: it is possible for a component to be a Seifert piece and an I-bundle piece at the same time.

Theorem 2.23 *Let (M, R) be an irreducible sufficiently large 3-manifold pair. Suppose that no component of R is a disc and that $\partial M \setminus \operatorname{int}(R)$ is a (possibly empty) collection of annuli. There is an algorithm which, given as input the pair (M, R), returns*

- *a finite collection $\mathcal{F}$ of self-homeomorphisms of $(\partial M, R)$, and*
- *a finite collection $\mathcal{C} = \{(a_1, b_1), \dots, (a_m, b_m)\}$, where $a_1, \dots, a_m, b_1, \dots, b_m$ are curves in R with $a_1 \cap b_1 = \dots = a_m \cap b_m = \varnothing$,*

such that

$$\boldsymbol{H}(M, R)|_\partial = \bigcup_{f \in \mathcal{F}} \langle \tau_{a_1} \tau_{b_1}^{-1}, \dots, \tau_{a_m} \tau_{b_m}^{-1} \rangle f$$

as a subgroup of $\boldsymbol{H}(\partial M, R)$. Moreover, if every I-bundle piece in the JSJ decomposition of (M, R) is a product I-bundle over Σ_0, $\Sigma_{0,1}$, $\Sigma_{0,2}$, or $\Sigma_{0,3}$, or the twisted I-bundle over U_1, $U_{1,1}$, or $U_{1,2}$, then the curves in $\mathcal{C}$ are pairwise disjoint.

Proof We use Theorem 2.22 to compute the JSJ system F of (M, R), and note that all the pieces are reasonable.

Exceptional cases First of all, let us deal with the cases where the JSJ decomposition of (M, R) is trivial or contains exceptional Seifert pieces.

- Suppose that F is empty. Then $(M, R, \varnothing)$ is either a strongly simple piece, an I-bundle piece, or a Seifert piece. If it is strongly simple, then we can just return $\mathcal{F} = \boldsymbol{H}(M, R)|_\partial$ and $\mathcal{C} = \varnothing$ thanks to Proposition 2.18. If $(M, R, \partial M \setminus \operatorname{int}(R))$ is a Seifert triple, then we conclude immediately by Proposition 2.11. Finally, if $(M, R, \varnothing)$ is an I-bundle triple, then we apply Proposition 2.14.
- Suppose that a component (N, R', F') is homeomorphic to $(T^2 \times I, \varnothing, T^2 \times \partial I)$. The two boundary components of N must come from the same torus in F, otherwise the surface F would not be minimal. We then have that M is homeomorphic to the mapping torus of some homeomorphism $T^2 \to T^2$, and we simply return $\mathcal{F} = \{\mathrm{id}\}$ and $\mathcal{C} = \varnothing$.

From now on, we can therefore assume that F is nonempty and that there are no $(T^2 \times I, \varnothing, T^2 \times \partial I)$ components in $(M, R) \mathbin{\backslash\backslash} F$. As a consequence of Proposition 2.8, it is easy to see that every Seifert piece either has a unique Seifert fibration, or it has two and it is homeomorphic to $(U_2 \mathbin{\tilde{\times}} I, \varnothing, \partial(U_2 \mathbin{\tilde{\times}} I))$. Since the two fibrations of $U_2 \mathbin{\tilde{\times}} I$ are inequivalent, let us arbitrarily and consistently pick one for every $(U_2 \mathbin{\tilde{\times}} I, \varnothing, \partial(U_2 \mathbin{\tilde{\times}} I))$ component of $(M, R) \mathbin{\backslash\backslash} F$ — say the one over the Möbius band. Then every Seifert piece in the JSJ decomposition of (M, R) has a distinguished fibration, either the unique one — for pieces which are not homeomorphic to the twisted I-bundle over a Klein bottle — or the one we have selected. Moreover, every homeomorphism between Seifert pieces can be isotoped to be fibre-preserving.

JSJ graph We closely follow the construction of the JSJ graph given by Kuperberg in [16, Section 6.3], with slight changes to accommodate the presence of boundary. More precisely, we define a graph Γ as follows.

- There is one vertex for each component of $(M, R) \mathbin{\backslash\backslash} F$.
- There is one edge for each component of F. Each edge joins the two (not necessarily distinct) vertices corresponding to the JSJ pieces whose boundaries this component lies on.

An *automorphism* of the JSJ graph Γ is an automorphism $\varphi\colon \Gamma \to \Gamma$ of the underlying graph, together with some additional data:

- for each vertex (N, R', F') representing a strongly simple piece, a homeomorphism $\varphi|_N$ from (N, R', F') to the component of $(M, R) \mathbin{\backslash\backslash} F$ associated to the vertex $\varphi(N, R', F')$, defined up to isotopy in $\varphi(N, R', F')$;
- for each edge X, a homeomorphism $\varphi|_X$ from X to the component of F associated to the edge $\varphi(X)$, defined up to isotopy in $\varphi(X)$; note that if X is an annulus, there are only two possible values for $\varphi|_X$.

Moreover, we enforce the following consistency condition for each strongly simple piece (N, R', F'): for each edge X having (N, R', F') as an endpoint, we ask that the restriction of $\varphi|_N$ to X is isotopic to $\varphi|_X$ in $\varphi(X)$. If both endpoints of X are equal to (N, R', F'), then we ask that this condition holds for both copies of X in F'.

Finitely many automorphisms Every homeomorphism $f\colon (M, R) \to (M, R)$ preserving F induces an automorphism φ_f of Γ in a natural way. We claim that the set

$$\Phi = \{\varphi_f : f \text{ is a self-homeomorphism of } (M, R) \text{ preserving } F\}$$

is finite. In fact, the only issue that needs addressing is the potentially infinite number of homeomorphisms between JSJ tori connecting two Seifert pieces. Let T be such a torus, connecting Seifert pieces (N_1, R_1, S_1) and (N_2, R_2, S_2). Let c_1 and c_2 be curves in T which are fibres of (N_1, R_1, S_1) and (N_2, R_2, S_2), respectively; note that c_1 and c_2 are not isotopic in T, otherwise we could remove T

from the JSJ system F. Consider a homeomorphism $f\colon (M,R)\to (M,R)$ preserving F. Necessarily, the curves c_1 and c_2 will be sent by f to curves in $\varphi_f(T)$ which are isotopic to fibres c_1' and c_2' of $\varphi_f(N_1,R_1,S_1)$ and $\varphi_f(N_2,R_2,S_2)$, respectively. Since the curves c_1, c_2, c_1', and c_2' do not depend on f, and there are only finitely many isotopy classes of homeomorphisms $T\to\varphi_f(T)$ sending c_1 to c_1' and c_2 to c_2', we conclude that there are only finitely many choices for $\varphi_f|_T$.

We remark that this argument is still valid if (N_1,R_1,S_1) and (N_2,R_2,S_2) are the same piece; in this case, we simply have to consider the possibility that f sends c_1 to c_2' and c_2 to c_1'.

Computable automorphisms Our argument for the finiteness of Φ actually provides an algorithm to compute a finite superset Φ' of Φ. As a consequence, in order to compute Φ, we simply need an algorithm to decide whether an automorphism of Γ is induced by a self-homeomorphism of (M,R) preserving F.

Let φ be an automorphism of Γ. If (N,R',F') is not homeomorphic to $\varphi(N,R',F')$ for some piece (N,R',F') then clearly φ is not induced by a homeomorphism. Otherwise, for each piece (N,R',F') which is not strongly simple, we ask the following question: is there a homeomorphism $f_N\colon (N,R',F')\to \varphi(N,R',F')$ such that, for each component X of F', the homeomorphisms $f_N|_X$ and $\varphi|_X$ are isotopic in $\varphi(X)$? If the answer — which we can compute thanks to Corollary 2.12 and Proposition 2.13 — is no, then φ cannot be induced by a homeomorphism. Otherwise, we easily see that φ belongs to Φ. In fact, we can isotope the homeomorphisms f_N so that they agree on their shared boundary F, and then combine them to produce a homeomorphism $f\colon (M,R)\to(M,R)$ preserving F; here, for ease of notation, we have used f_N to refer to the homeomorphism $\varphi|_N$ for strongly simple pieces (N,R',F').

We can therefore algorithmically construct a finite set $\mathcal{F}_0$ by picking, for each $\varphi\in\Phi$, a self-homeomorphism of (M,R) preserving F and inducing φ. This finite set has the property that every homeomorphism $f\colon (M,R)\to(M,R)$ can be isotoped so that, for some $f_0\in\mathcal{F}_0$, the homeomorphism ff_0 is the identity on F and on all the strongly simple pieces of the JSJ decomposition. In fact, it is enough to isotope f so to preserve F and pick $f_0\in\mathcal{F}_0$ such that $\varphi_{f_0}=\varphi_{f^{-1}}$, so that ff_0 induces the trivial automorphism of Γ. Up to further isotoping f, we can then assume that ff_0 is the identity on F and on strongly simple pieces, as desired.

Finitely many Dehn twists Finally, let us show how to compute the sets $\mathcal{F}$ and $\mathcal{C}$. For each Seifert or I-bundle piece (N,R',F'), we apply Propositions 2.11 and 2.14 to find

- a finite collection $\mathcal{F}_N$ of self-homeomorphisms of $(\partial N,R')$ fixing F' pointwise, and
- a finite collection $\mathcal{C}_N=\{(a_{N,1},b_{N,1}),\dots,(a_{N,m_N},b_{N,m_N})\}$ where $a_{N,1},\dots,a_{N,m_N},b_{N,1},\dots,b_{N,m_N}$ are curves in $\partial N\setminus(F\cup\partial R)$ with $a_{N,1}\cap b_{N,1}=\dots=a_{N,m_N}\cap b_{N,m_N}=\varnothing$,

such that

$$\boldsymbol{H}_{F'}(N,R')|_\partial=\bigcup_{f\in\mathcal{F}_N}\langle \tau_{a_{N,i}}\tau_{b_{N,i}}^{-1} : 1\leqslant i\leqslant m_N\rangle f$$

as a subgroup of $\boldsymbol{H}_{F'}(\partial N, R')$. Moreover, if the piece (N, R', F') is a product I-bundle over Σ_0, $\Sigma_{0,1}$, $\Sigma_{0,2}$, or $\Sigma_{0,3}$, or a twisted I-bundle over U_1, $U_{1,1}$, or $U_{1,2}$, then the curves in $\mathcal{C}_N$ are pairwise disjoint. Let us extend the homeomorphisms in $\mathcal{F}_N$ to self-homeomorphisms of $(\partial M, R)$ by setting them equal to the identity on $\partial M \setminus N$. Finally, if some curve in $\mathcal{C}_N$ lies in $\partial M \setminus R$, we can replace it with a parallel curve in R without changing the isotopy class of the corresponding Dehn twist. It is then easy to see that the algorithm can return the collections

$$\mathcal{F} = \{f_0|_\partial : f_0 \in \mathcal{F}_0\} \cup \bigcup_N \mathcal{F}_N \quad \text{and} \quad \mathcal{C} = \bigcup_N \mathcal{C}_N. \qquad \square$$

Remark 2.24 Translated in the language of Matveev's boundary patterns, Theorem 2.23 can be used to compute the mapping class group of a Haken 3-manifold (M, p) with boundary pattern such that p is a collection of simple closed curves and no component of $\partial M \setminus\!\!\setminus p$ is a disc.

As a consequence of Theorem 2.23 and the work of Gordon and Luecke (see [7]), we have the following. Note that statement (3) can also be proved using an oriented version of Matveev's boundary patterns.

Corollary 2.25 *Let $M \subseteq S^3$ be the complement of a nontrivial knot.*

(1) *Every self-homeomorphism of M sends the meridian curve of ∂M to itself.*

(2) *The group $\boldsymbol{H}(M)|_\partial$ is either trivial or generated by the homeomorphism*

$$(-\mathrm{id})\colon \partial M \to \partial M$$

inducing multiplication by -1 on $H_1(\partial M)$.

(3) *There is an algorithm which, given M as input, returns representatives of $\boldsymbol{H}(M)|_\partial$.*

Proof Statement (1) follows from [7]; statement (2) is an immediate consequence. As far as statement (3) is concerned, we can apply Theorem 2.23 to the pair $(M, \partial M)$. Since homeomorphisms of the form $\tau_a \tau_b^{-1}$ cannot act as $(-\mathrm{id})$ on $H_1(\partial M)$, it is enough to check whether the collection $\mathcal{F}$ returned by the algorithm contains any nontrivial homeomorphism. $\square$

3 An algorithmic problem on free groups

3.1 Introducing band systems

We now take a break from topology and enter the realm of combinatorics and free groups. The bulk of Section 3 will be devoted to proving Theorem 3.4. The reader should however not be surprised by how seemingly distant the statement of the theorem is to anything concerning isotopies of surfaces. In fact, an easy consequence of the theorem — namely Corollary 3.5 — will allow us to answer an algorithmic question about extensions of homeomorphisms to the interior of handlebodies.

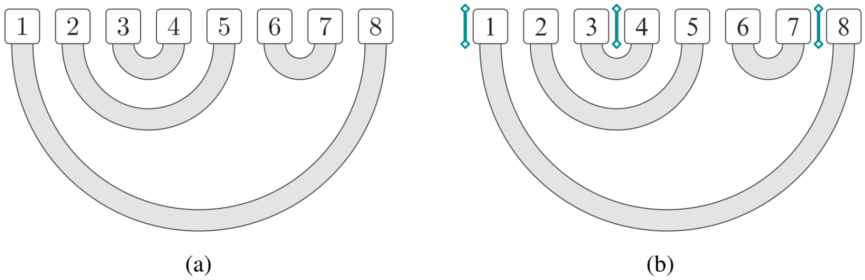


Figure 5: A band system (a) and a marked band system (b).

We start by introducing the notation we will need to talk about cancellation of words in free groups, namely (marked) band systems and bundling maps. Section 3.2 is entirely dedicated to the proof of Theorem 3.4, while the topological corollary is presented in Section 3.3.

Definition 3.1 Let n be a positive integer. A *band system*[5] of length $2n$ is a collection

$$B = \{(a_1, b_1), \dots, (a_n, b_n)\}$$

such that

(i) $\{a_1, b_1, \dots, a_n, b_n\} = \{1, \dots, 2n\}$;

(ii) for each $1 \leqslant i \leqslant n$, the inequality $a_i < b_i$ holds;

(iii) there is no pair of indices i, j such that $a_i < a_j < b_i < b_j$.

The individual pairs (a_i, b_i) are called *bands*, with a_i being the *left endpoint* and b_i being the *right endpoint*.

Figure 5(a) shows a graphical depiction of the band system

$$B = \{(1,8), (2,5), (3,4), (6,7)\} \text{ of length } 2n = 8,$$

hopefully justifying the name. If we think of a pair $(a, b) \in B$ as a physical band connecting the elements a and b, then (iii) ensures that these bands do not intersect.

Definition 3.2 A *marked band system* is a pair (B, p) where B is a band system of length — say — $2n$, and $p = (p_1, \dots, p_r)$ is a weakly increasing sequence of numbers with

$$p_i \in \left\{\tfrac{1}{2}, 1 + \tfrac{1}{2}, \dots, 2n + \tfrac{1}{2}\right\} \quad \text{for } 1 \leqslant i \leqslant r.$$

---

[5] It was brought to our attention by an anonymous referee that what we call "band systems" here were already known in the literature under the name "crossingless matchings". However, to the best of our knowledge, the theory and results developed in this section are still novel.

Figure 5(b) depicts the same band system as Figure 5(a), equipped with the marking

$$p = \left(\tfrac{1}{2}, 3 + \tfrac{1}{2}, 7 + \tfrac{1}{2}\right).$$

We think of an element $p_i = a + \frac{1}{2}$ as a "separator" between the integers a and $a+1$.

Let (B, p) be a marked band system. The *maximal bundle*[6] of (B, p) is, loosely speaking, the marked band system (B', p') constructed from (B, p) by bundling together parallel bands as much as possible, while making sure that the bundles don't cross any marked spot. Formally, let $2n$ be the length of B. Consider the equivalence relation $\sim$ on $\{1, \dots, 2n\}$ generated by

$$a_1 \sim a_2 \text{ and } b_1 \sim b_2 \quad \text{whenever} \quad \begin{cases} (a_1, b_1), (a_2, b_2) \in B, \\ a_2 = a_1 + 1, \, b_2 = b_1 - 1, \\ a_1 + \frac{1}{2}, b_1 - \frac{1}{2} \notin p. \end{cases}$$

Let $E = \{1, \dots, 2n\}/\sim$ be the quotient set, and let $2n'$ be its cardinality (which is easily seen to be even). Since equivalence classes consist of consecutive integers, there is a natural bijection between E and $\{1, \dots, 2n'\}$ such that classes containing smaller integers are mapped to smaller integers. Let

$$\iota \colon \{1, \dots, 2n\} \to \{1, \dots, 2n'\}$$

be the composition of the projection map $\{1, \dots, 2n\} \to E$ with said bijection; we will call ι the *bundling map*. We can then define

$$B' = \{(\iota(a), \iota(b)) : (a, b) \in B\}.$$

As far as the marking is concerned, if $p = (p_1, \dots, p_r)$, for each $1 \leqslant i \leqslant r$ we define

$$p'_i = \begin{cases} \frac{1}{2} & \text{if } p_i = \frac{1}{2}, \\ \iota\left(p_i - \frac{1}{2}\right) + \frac{1}{2} & \text{otherwise,} \end{cases}$$

and set $p' = (p'_1, \dots, p'_r)$. It is not hard to verify that (B', p') is a marked band system.

As an example, let us look at Figure 6. The marked band system on the top is defined by

$$\begin{aligned} &B = \{(1, 12), (2, 11), (3, 8), (4, 7), (5, 6), (9, 10)\} \text{ of length } 2n = 12, \\ &p = \left(2 + \tfrac{1}{2}, 6 + \tfrac{1}{2}, 9 + \tfrac{1}{2}, 12 + \tfrac{1}{2}\right). \end{aligned}$$

In order to construct the maximal bundle, we group together the bands $(1, 12)$ and $(3, 4)$, since they are parallel and do not cross any marked spot. We do the same for bands $(3, 8)$ and $(4, 7)$; note that the band $(5, 6)$ does not belong to the same bundle, since the spot $6 + \frac{1}{2}$ is marked. The maximal bundle of (B, p) is then (B', p'), where

$$\begin{aligned} &B' = \{(1, 8), (2, 5), (3, 4), (6, 7)\} \text{ of length } 2n' = 8, \\ &p' = \left(1 + \tfrac{1}{2}, 4 + \tfrac{1}{2}, 6 + \tfrac{1}{2}, 8 + \tfrac{1}{2}\right). \end{aligned}$$

Let us now remark a few properties of the bundling map ι.

---

[6] This notion of "bundle" has nothing to do with that of "fibre bundle" used elsewhere in the article.

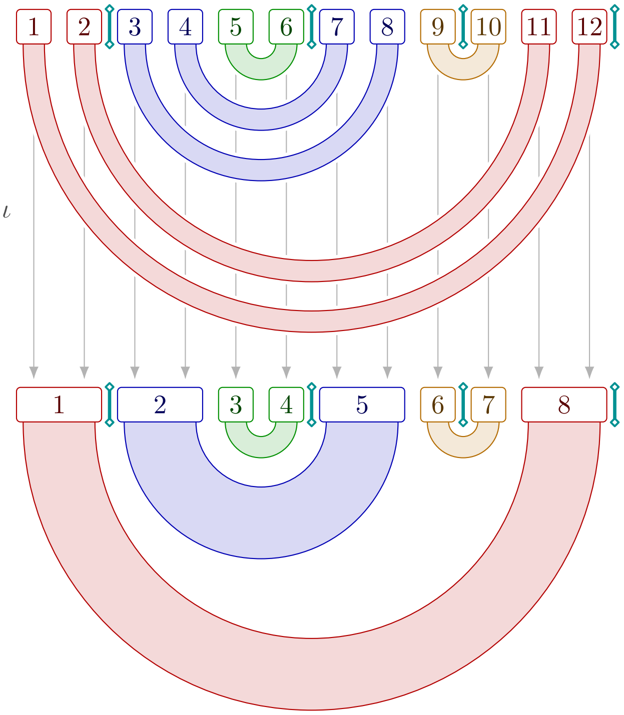


Figure 6: A marked band system (on the top) and its maximal bundle (on the bottom). Bands of the same colour get bundled together; vertical arrows represent the bundling map ι.

- The bundling map sends bands of B to bands of B'. From now on, with slight abuse of notation, for each band $(a,b) \in B$ we will denote by $\iota(a,b)$ the band $(\iota(a),\iota(b)) \in B'$.
- If $\iota(a_1,b_1) = \iota(a_2,b_2)$ where (a_1,b_1) and (a_2,b_2) are bands of B with $a_1 < a_2 < b_2 < b_1$, then $(a_1+i, b_1-i) \in B$ and $\iota(a_1+i,b_1-i) = \iota(a_1,b_1)$ for each $0 \leqslant i \leqslant a_2 - a_1$. Moreover, no element of p lies between a_1 and a_2 or between b_2 and b_1.

The following facts are completely elementary, but we prefer to state them here so as not to clutter the proof of Theorem 3.4.

Lemma 3.3 *Let (B, p) be a marked band system.*

(1) *If (a,b) is a band of B, then there is a band $(c,c+1) \in B$ for some $a \leqslant c < b$.*

(2) *Let (B', p') be the maximal bundle of (B, p), with bundling map ι. If (a_1,b_1) and (a_2,b_2) are bands of B with $a_1 < a_2 < b_2 < b_1$, then either*

- $\iota(a_1,b_1) = \iota(a_2,b_2)$,
- *there is an element of p lying between a_1 and a_2 or between b_2 and b_1, or*
- *there is a band $(c,c+1) \in B$ for some*

$$c \in \{a_1,\dots,a_2-1\} \cup \{b_2,\dots,b_1-1\}.$$

Proof In order to prove statement (1), simply consider the narrowest band amongst those having left endpoint in $\{a,\dots,b-1\}$. We now turn to statement (2). If there is a band $(a,b)\in B$ with $a_1<a<b<a_2$ or $b_2<a<b<b_1$ then we can apply statement (1) and immediately conclude. Otherwise, for every band $(a,b)\in B$ we have that $a_1\leqslant a\leqslant a_2$ if and only if $b_2\leqslant b\leqslant b_1$. But then, using the notation from the definition of maximal bundle, it is clear that either

$$(a_1,b_1)\sim(a_1+1,b_1-1)\sim\cdots\sim(a_2,b_2),$$

or there is marked spot (in the marking p) lying between a_1 and a_2 or between b_2 and b_1. □

Let (B',p') be a marked band system of length $2n'$. We say that (B',p') is *maximal* if it is its own maximal bundle. Of course, if (B',p') is maximal, there are infinitely many marked band systems (B,p) having (B',p') as their maximal bundle; in fact, these marked band systems can be parametrised quite nicely. An *unbundling map* for (B',p') is a function $\varphi\colon\{1,\dots,2n'\}\to\mathbb{Z}_{>0}$ such that $\varphi(a')=\varphi(b')$ for each band $(a',b')\in B'$. There is a one-to-one correspondence between unbundling maps for (B',p') and marked band systems having (B',p') as their maximal bundle.

- Given a marked band system (B,p) whose maximal bundle is (B',p'), we can define an unbundling map φ as $\varphi(a')=\left|\iota^{-1}(a')\right|$, where ι is the bundling map for (B,p).
- Conversely, let $\varphi\colon\{1,\dots,2n'\}\to\mathbb{Z}_{>0}$ be an unbundling map for (B',p'). Let $2n=\varphi(1)+\cdots+\varphi(2n')$; for each $1\leqslant a\leqslant 2n$, define $\iota(a)$ to be the unique integer in $\{1,\dots,2n'\}$ such that

$$\varphi(1)+\cdots+\varphi(\iota(a)-1)<a\leqslant\varphi(1)+\cdots+\varphi(\iota(a)).$$

It is now easy to construct the unique marked band system (B,p) having (B',p') as maximal bundle and ι as bundling map. In fact, we can set

$$B=\bigcup_{(a',b')\in B'}\bigl\{(a+i,b-i):\{a+1,\dots,a+\varphi(a')\}=\iota^{-1}(a'),\{b-\varphi(a'),\dots,b-1\}=\iota^{-1}(b'),1\leqslant i\leqslant\varphi(a')\bigr\}$$

and $p=(p_1,\dots,p_r)$, where $p'=(p'_1,\dots,p'_r)$ and

$$p_i=\begin{cases}\frac{1}{2} & \text{if } p'_i=\frac{1}{2},\\ \max\left(\iota^{-1}\left(p'_i-\frac{1}{2}\right)\right)+\frac{1}{2} & \text{otherwise}\end{cases}\qquad\text{for } 1\leqslant i\leqslant r.$$

Figure 7 shows an instance of this construction. The marked band system of length $2n'=4$ on the top is defined by $B'=\{(1,4),(2,3)\}$ and $p'=\left(1+\frac{1}{2}\right)$, and is clearly maximal. We choose the unbundling map φ such that $\varphi(1)=\varphi(4)=4$ and $\varphi(2)=\varphi(3)=2$. If we replace each band $(a,b)\in B'$ with $\varphi(a)$ parallel bands, we obtain the marked band system (B,p) displayed on the bottom, where

$$B=\{(1,12),(2,11),(3,10),(4,9),(5,8),(6,7)\}\text{ of length }2n=12,$$
$$p=\left(4+\tfrac{1}{2}\right).$$

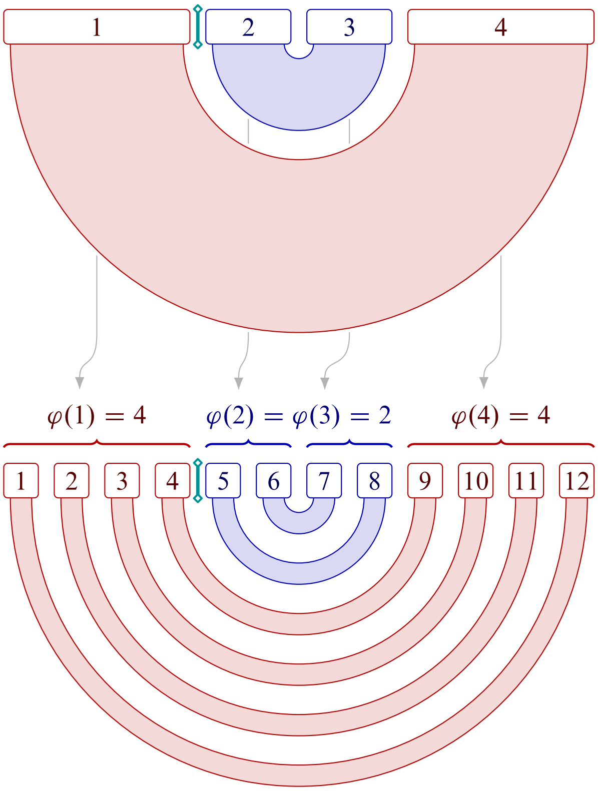


Figure 7: A maximal marked band system (on the top) and the marked band system associated to an unbundling map φ (on the bottom).

Given an unbundling map φ for (B', p'), it is sometimes convenient to define the corresponding *cumulative unbundling map* $\widehat{\varphi}\colon \{1, \dots, 2n'+1\} \to \mathbb{Z}_{\geqslant 0}$ as

$$\widehat{\varphi}(a') = \varphi(1) + \cdots + \varphi(a'-1).$$

With this definition, we can compactly write $\iota^{-1}(a')$ as $\{\widehat{\varphi}(a')+1, \dots, \widehat{\varphi}(a'+1)\}$, without having to refer to the band system (B, p) associated to φ.

3.2 Band systems and free groups

Let us introduce some notation about free groups which we will employ in this section. Let g be a positive integer. Fix g symbols $\{x_1, \dots, x_g\}$ and denote by $\mathbb{F}_g$ the free group over them. If s is a word of length m in the symbols $x_1, \dots, x_g$ and their inverses, we use the following notation:

- s_i is the i^{th} symbol of s, where $1 \leqslant i \leqslant m$;
- $s_{[i\,:\,j]}$ is the subword $s_i s_{i+1} \dots s_j$, where $1 \leqslant i \leqslant j \leqslant m$;
- s^{-1} is the word such that $(s^{-1})_i = (s_{m+1-i})^{-1}$ for $1 \leqslant i \leqslant m$;

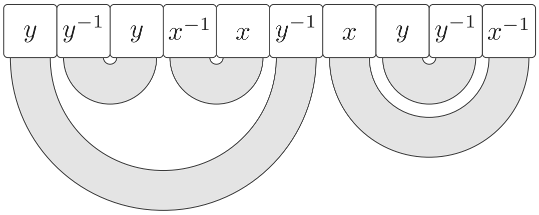


Figure 8: A cancellation band system for the word $yy^{-1}yx^{-1}xy^{-1}xyy^{-1}x^{-1}$.

- s^k is the concatenation of k copies of s, where $k \geqslant 0$; if $k < 0$, we define $s^k = (s^{-1})^{-k}$;
- s^∞ the infinite word obtained by concatenating infinitely many copies of s; this does not correspond to an element of $\mathbb{F}_g$, and we will only ever make use of finite subwords of infinite words.

Let $s = s_1 s_2 \cdots s_{2n}$ be a word which reduces to 1 in $\mathbb{F}_g$. Then it must do so via a sequence of cancellations of adjacent symbols of the form xx^{-1} or $x^{-1}x$ (by "adjacent" we mean "adjacent after performing the previous cancellations in the sequence"). For any such sequence, we can define a band system

$$B = \{(a,b) : 1 \leqslant a < b \leqslant 2n,\ s_a \text{ cancels with } s_b\}.$$

We say that B is a *cancellation band system* for s; cancellation band systems need not be unique, but each word reducing to 1 has at least one of them. For example, Figure 8 shows how the band system

$$B = \{(1,6),(2,3),(4,5),(7,10),(8,9)\}$$

is a cancellation band system for the word $yy^{-1}yx^{-1}xy^{-1}xyy^{-1}x^{-1}$ in the free group generated by the symbols x and y.

The main technical result of this section is the following.

Theorem 3.4 *Let $w_0,\dots,w_l,t_1,\dots,t_l \in \mathbb{F}_g$. Consider the set*

$$\mathcal{A} = \left\{(k_1,\dots,k_l) \in \mathbb{Z}^l : w_0 t_1^{k_1} w_1 t_2^{k_2} w_2 \cdots t_l^{k_l} w_l = 1\right\}.$$

There is an algorithm which, given as input $w_0,\dots,w_l$ and $t_1,\dots,t_l$, returns a collection of sets $A_1,\dots,A_N$ such that $\mathcal{A} = A_1 \cup \dots \cup A_N$, and each A_j is of the form

$$A_j = \left\{\boldsymbol{z}_j + \boldsymbol{M}_j \boldsymbol{v} : \boldsymbol{v} \in \mathbb{Z}_{\geqslant 0}^{d_j}\right\}$$

for some vector $\boldsymbol{z}_j \in \mathbb{Z}^l$ and some matrix $\boldsymbol{M}_j \in \mathbb{Z}^{l \times d_j}$.

Proof Let $w_0,\dots,w_l$ and $t_1,\dots,t_l$ be given as reduced words over $\{x_1,\dots,x_g\}$. For a vector $\boldsymbol{k} \in \mathbb{Z}^l$, denote by $w(\boldsymbol{k})$ the word $w_0 t_1^{k_1} w_1 \cdots t_l^{k_l} w_l$.

Cyclically reduced words For each $1 \leqslant i \leqslant l$ we can write $t_i = u_i t_i' u_i^{-1}$, where t_i' is a cyclically reduced word and u_i is a reduced word. Define

- $w_0' = w_0 u_1$;
- $w_i' = u_i^{-1} w_i u_{i+1}$ for $1 \leqslant i < l$;
- $w_l' = u_l^{-1} w_l$.

Then

$$\mathcal{A} = \left\{ (k_1, \dots, k_l) \in \mathbb{Z}^l : w_0' (t_1')^{k_1} w_1' \cdots (t_l')^{k_l} w_l' = 1 \right\}.$$

Therefore, up to replacing each w_i with w_i' and each t_i with t_i', we can assume that each t_i is cyclically reduced.

Only nonnegative powers For each $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_l) \in \{-1, 1\}^l$ define

$$\mathcal{A}^{\boldsymbol{\epsilon}} = \left\{ (\epsilon_1 k_1, \dots, \epsilon_l k_l) : (k_1, \dots, k_l) \in \mathbb{Z}^l_{\geqslant 0},\ w_0 t_1^{\epsilon_1 k_1} w_1 \cdots t_l^{\epsilon_l k_l} w_l = 1 \right\},$$

so that

$$\mathcal{A} = \bigcup_{\boldsymbol{\epsilon} \in \{-1,1\}^l} \mathcal{A}^{\boldsymbol{\epsilon}}.$$

Denote by $\mathcal{A}^+$ the set

$$\mathcal{A}^{(1,\dots,1)} = \left\{ (k_1, \dots, k_l) \in \mathbb{Z}_{\geqslant 0} : w_0 t_1^{k_1} w_1 \cdots t_l^{k_l} w_l = 1 \right\}.$$

Suppose we have an algorithm to decompose $\mathcal{A}^+$ as described in the statement, and fix a vector $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_l) \in \{-1, 1\}^l$. Then we can algorithmically decompose

$$\left\{ (k_1, \dots, k_l) \in \mathbb{Z}^l_{\geqslant 0} : w_0 (t_1^{\epsilon_1})^{k_1} w_1 \cdots (t_l^{\epsilon_l})^{k_l} w_l = 1 \right\} = A_1 \cup \cdots \cup A_N,$$

where each A_j is of the form

$$A_j = \left\{ \boldsymbol{z}_j + \boldsymbol{M}_j \boldsymbol{v} : \boldsymbol{v} \in \mathbb{Z}^{d_j}_{\geqslant 0} \right\}$$

for some vector $\boldsymbol{z}_j \in \mathbb{Z}^l$ and some matrix $\boldsymbol{M}_j \in \mathbb{Z}^{l \times d_j}$. But then we have the decomposition

$$\mathcal{A}^{\boldsymbol{\epsilon}} = A_1' \cup \cdots \cup A_N',$$

where, for each $1 \leqslant j \leqslant N$,

- $A_j = \left\{ \boldsymbol{z}_j' + \boldsymbol{M}_j' \boldsymbol{v} : \boldsymbol{v} \in \mathbb{Z}^{d_j}_{\geqslant 0} \right\}$;
- $\boldsymbol{z}_j'$ is obtained from $\boldsymbol{z}_j$ by multiplying the i^{th} coordinate by ϵ_i for $1 \leqslant i \leqslant l$;
- $\boldsymbol{M}_j'$ is obtained from $\boldsymbol{M}_j$ by multiplying the i^{th} row by ϵ_i for $1 \leqslant i \leqslant l$.

Since there are only finitely many choices of $\boldsymbol{\epsilon} \in \{-1, 1\}^l$, it is enough to describe an algorithm to compute the set $\mathcal{A}^+$.

Nonempty words Suppose we have an algorithm which works when every t_i is a nonempty word (that is, $t_i \neq 1$ as an element of $\mathbb{F}_g$). If t_i is the empty word for some $1 \leqslant i \leqslant l$, by induction on l we can assume that the set

$$\left\{(k_1, \dots, k_{i-1}, k_{i+1}, \dots, k_l) \in \mathbb{Z}_{\geqslant 0}^{l-1} : w_0 t_1^{k_1} w_1 \cdots t_{i-1}^{k_{i-1}} (w_{i-1} w_i) t_{i+1}^{k_{i+1}} w_{i+1} \cdots t_l^{k_l} w_l = 1\right\}$$

can be algorithmically decomposed as $A_1 \cup \dots \cup A_N$, where each A_j is of the form

$$A_j = \left\{\boldsymbol{z}_j + \boldsymbol{M}_j \boldsymbol{v} : \boldsymbol{v} \in \mathbb{Z}_{\geqslant 0}^{d_j}\right\}$$

for some vector $\boldsymbol{z}_j \in \mathbb{Z}^{l-1}$ and some matrix $\boldsymbol{M}_j \in \mathbb{Z}^{(l-1)\times d_j}$. For each $1 \leqslant j \leqslant N$, write

$$\boldsymbol{z}_j = \begin{pmatrix} \boldsymbol{z}_j^{(1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \quad \text{and} \quad \boldsymbol{M}_j = \begin{pmatrix} \boldsymbol{M}_j^{(1)} \\ \boldsymbol{M}_j^{(2)} \end{pmatrix},$$

where $\boldsymbol{z}_j^{(1)} \in \mathbb{Z}^{i-1}$, $\boldsymbol{z}_j^{(2)} \in \mathbb{Z}^{l-i}$, and similarly $\boldsymbol{M}_j^{(1)} \in \mathbb{Z}^{(i-1)\times d_j}$, $\boldsymbol{M}_j^{(2)} \in \mathbb{Z}^{(l-i)\times d_j}$. Define

$$\boldsymbol{z}_j' = \begin{pmatrix} \boldsymbol{z}_j^{(1)} \\ 0 \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \in \mathbb{Z}^l \quad \text{and} \quad \boldsymbol{M}_j' = \begin{pmatrix} \boldsymbol{M}_j^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \\ \boldsymbol{M}_j^{(2)} & \boldsymbol{0} \end{pmatrix} \in \mathbb{Z}^{l\times(d_j+1)}.$$

We get the required decomposition

$$\mathcal{A}^+ = \bigcup_{j=1}^{N} \left\{\boldsymbol{z}_j' + \boldsymbol{M}_j' \boldsymbol{v} : \boldsymbol{v} \in \mathbb{Z}_{\geqslant 0}^{d_j+1}\right\}.$$

Therefore, we can assume that each t_i is nonempty.

Fixed cancellation bundle Let $\boldsymbol{k} = (k_1, \dots, k_l)$ be a vector in $\mathcal{A}^+$. Consider a cancellation band system B for the word $w(\boldsymbol{k})$. Moreover, define the *block marking* $p = (p_1, \dots, p_{2l+2})$ to separate different "blocks" of $w(\boldsymbol{k})$, by setting

$$\begin{aligned} p_{2i-1} &= |w_0| + k_1\,|t_1| + |w_1| + \dots + k_{i-1}\,|t_{i-1}| + \tfrac{1}{2} \quad &&\text{for } 1 \leqslant i \leqslant l+1, \\ p_{2i} &= |w_0| + k_1\,|t_1| + |w_1| + \dots + |w_{i-1}| + \tfrac{1}{2} &&\text{for } 1 \leqslant i \leqslant l+1. \end{aligned}$$

Denote by (B', p') the maximal bundle of the marked band system (B, p); we say that (B', p') is a *cancellation bundle* for $\boldsymbol{k}$ if it can be obtained in this way (that is, as the maximal bundle of a cancellation band system for $w(\boldsymbol{k})$ endowed with the block marking described above). Let

$$\iota\colon \{1, \dots, 2n\} \to \{1, \dots, 2n'\}$$

be the bundling map, where $2n$ and $2n'$ are the lengths of B and B', respectively. Let us label the integers $\{1, \dots, 2n'\}$ with the symbols $W_0, \dots, W_l, T_1, \dots, T_l$ depending on which "block" of $w(\boldsymbol{k})$ they come from. More precisely, we label

- the integers between p'_{2i+1} and p'_{2i+2} with the symbol W_i for $0 \leqslant i \leqslant l$, and
- the integers between p'_{2i} and p'_{2i+1} with the symbol T_i for $1 \leqslant i \leqslant l$.

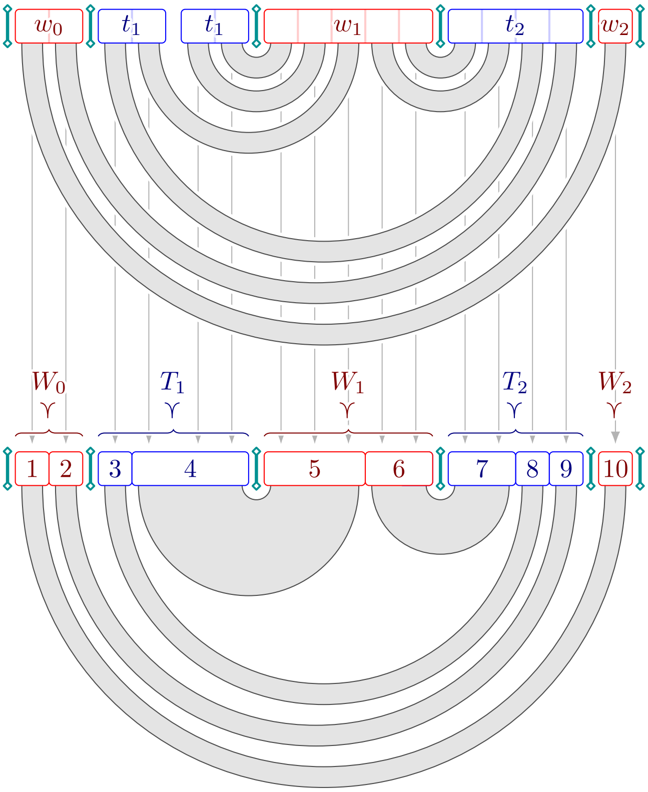


Figure 9: The cancellation bundle (on the bottom) induced by a cancellation band system (on the top). In this specific example, we have that $|w_0| = 2$, $|t_1| = 2$, $|w_1| = 5$, $|t_2| = 4$, and $|w_2| = 1$. The word $w(\boldsymbol{k})$ on the top corresponds to the vector $\boldsymbol{k} = (2, 1)$, and reduces to 1 via the cancellation band system represented in the picture — although this is of course impossible to verify without knowing the actual the words involved. The marked spots separate the different "blocks" of $w(\boldsymbol{k})$, namely w_0, t_1^2, w_1, t_2, and w_2. The maximal bundling procedure yields the marked band system depicted on the bottom, with vertical arrows representing the bundling map. Here, the integers in $\{1, \dots, 10\}$ are labelled with W_0, T_1, W_1, T_2, or W_2 according to the block they come from.

We will use the notation $i \prec X$ to signify that $i \in \{1, \dots, 2n'\}$ has label $X \in \{W_0, \dots, W_l, T_1, \dots, T_l\}$. See Figure 9 for a graphical representation of a cancellation bundle.

The crucial fact we seek to prove now is that $2n'$ is bounded above by a constant which does not depend on $\boldsymbol{k}$. This is a consequence of the following two statements.

(1) For each $1 \leqslant i \leqslant l$, there is no band of B' with both endpoints labelled T_i. In fact, suppose that $(a', b') \in B'$ is such a band, and let (a, b) be a band of B with $\iota(a, b) = (a', b')$. By statement (1) of Lemma 3.3, there is a band $(c, c + 1) \in B$ with $a \leqslant c < b$. But this is impossible, since positions a and b (and therefore c and $c + 1$) belong to the block $t_i^{k_i}$; the word t_i is cyclically reduced, therefore there can be no cancellation between the symbols in positions c and $c + 1$.

(2) For each $1 \leqslant i < j \leqslant l$, there is at most one band of $(a', b') \in B'$ with $a' \prec T_i$ and $b' \prec T_j$. In fact, suppose that (a'_1, b'_1) and (a'_2, b'_2) are two such bands, with $a'_1 < a'_2 < b'_2 < b'_1$. Let (a_1, b_1) and (a_2, b_2) be bands of B with $\iota(a_1, b_1) = (a'_1, b'_1)$ and $\iota(a_2, b_2) = (a'_2, b'_2)$. Consider statement (2) of Lemma 3.3. Clearly,

$$\iota(a_1, b_1) = (a'_1, b'_1) \neq (a'_2, b'_2) = \iota(a_2, b_2).$$

Moreover, since $a'_1, a'_2 \prec T_i$ and $b'_1, b'_2 \prec T_j$, we have that

$$p_{2i} < a_1 < a_2 < p_{2i+1} \leqslant p_{2j} < b_2 < b_1 < p_{2j+1}.$$

As a consequence, there must be a band $(c, c+1) \in B$ for some

$$c \in \{a_1, \ldots, a_2 - 1\} \cup \{b_2, \ldots, b_1 - 1\},$$

but this is once again impossible since t_i and t_j are cyclically reduced words.

As anticipated, these two facts imply the existence of an upper bound C for the length $2n'$ of B' which is independent of $\boldsymbol{k}$. An estimate for C can be found as follows. Say that a band $(a', b') \in B'$ is

- of *type* 1 if $a' \prec W_i$ or $b' \prec W_i$ for some $0 \leqslant i \leqslant l$;
- of *type* 2 if it is not of type 1, that is, if $a' \prec T_i$ and $b' \prec T_j$ for some $1 \leqslant i, j \leqslant l$.

An easy counting argument shows that there are at most $|w_0| + \cdots + |w_l|$ bands of type 1 and at most $l(l-1)/2$ bands of type 2. Therefore, a suitable upper bound for $2n'$ is

$$2n' \leqslant C = 2(|w_0| + \cdots + |w_l|) + l(l-1).$$

Let us now forget our choice of $\boldsymbol{k}$, and fix any maximal marked band system (B', p') such that B' has length $2n' \leqslant C$ and $p' = (p'_1, \ldots, p'_{2l+2})$ with $p'_1 = \frac{1}{2}$ and $p'_{2l+2} = 2n' + \frac{1}{2}$. Since there are only finitely many such marked band systems, and since every $\boldsymbol{k} \in \mathcal{A}^+$ has a cancellation bundle of this form, it is enough to describe an algorithm to compute the set

$$\left\{\boldsymbol{k} = (k_1, \ldots, k_l) \in \mathbb{Z}^l_{\geqslant 0} : w_0 t_1^{k_1} w_1 t_2^{k_2} w_2 \cdots t_l^{k_l} w_l = 1,\ (B', p') \text{ is a cancellation bundle for } \boldsymbol{k}\right\},$$

which will be henceforth referred to as $\mathcal{A}^+_{(B', p')}$.

Let us remark that the labelling procedure described above can be carried out without any reference to $\boldsymbol{k}$; in fact, only the integer $2n'$ and the marking p' are required. Therefore, we will from now on assume that the integers in $\{1, \ldots, 2n'\}$ are labelled with symbols $W_0, \ldots, W_l, T_1, \ldots, T_l$ according to the marking p'.

Fixed unbundling class Let us dwell some more on what it means for a vector $\boldsymbol{k} \in \mathbb{Z}^l_{\geqslant 0}$ to have (B', p') as a cancellation bundle. Recall that band systems having (B', p') as their maximal bundle are parametrised by unbundling maps for (B', p'). It is easy to see that an unbundling map φ for (B', p') describes a cancellation band system for $\boldsymbol{k}$ if and only if

(i) $\sum_{a' \prec W_i} \varphi(a') = |w_i|$ for $0 \leqslant i \leqslant l$;

(ii) $\sum_{a' \prec T_i} \varphi(a') = k_i\, |t_i|$ for $1 \leqslant i \leqslant l$;

(iii) $w(\boldsymbol{k})_{[\hat{\varphi}(a')+1 : \hat{\varphi}(a'+1)]} = \left(w(\boldsymbol{k})_{[\hat{\varphi}(b')+1 : \hat{\varphi}(b'+1)]}\right)^{-1}$ for each $(a', b') \in B'$.

Moreover, (ii) implies that φ describes a cancellation band system for at most one vector $\boldsymbol{k}(\varphi) \in \mathbb{Z}^l_{\geqslant 0}$, whose coordinates can be easily computed as

$$k_i(\varphi) = \frac{\sum_{a' \prec T_i} \varphi(a')}{|t_i|} \quad \text{for } 1 \leqslant i \leqslant l \tag{1}$$

(recall that $|t_i| \neq 0$ for each i).

Let us say that an unbundling map φ for (B', p') is *compatible with the labelling* if

$$\sum_{a' \prec W_i} \varphi(a') = |w_i| \quad \text{for } 0 \leqslant i \leqslant l \qquad \text{and} \qquad \sum_{a' \prec T_i} \varphi(a') \equiv 0 \pmod{|t_i|} \quad \text{for } 1 \leqslant i \leqslant l.$$

If, additionally, we have

$$w(\boldsymbol{k}(\varphi))_{[\hat{\varphi}(a')+1\,:\,\hat{\varphi}(a'+1)]} = \bigl(w(\boldsymbol{k}(\varphi))_{[\hat{\varphi}(b')+1\,:\,\hat{\varphi}(b'+1)]}\bigr)^{-1} \quad \text{for each } (a', b') \in B',$$

then we say that φ is *cancelling*. We can then equivalently define $\mathcal{A}^+_{(B',p')}$ as the set

$$\bigl\{\boldsymbol{k}(\varphi) : \varphi \text{ is an unbundling map for } (B', p') \text{ which is compatible with the labelling and cancelling}\bigr\}.$$

Let us define an equivalence relation on the set of unbundling maps for (B', p') which are compatible with the labelling. Given two unbundling maps φ and ψ, we say that $\varphi \sim \psi$ if the following properties hold for each $1 \leqslant a' \leqslant 2n'$:

- $\varphi(a') = \psi(a')$ whenever a' is an endpoint of a band of type 1;
- $\varphi(a') \equiv \psi(a') \pmod{|t_i|}$ whenever $a' \prec T_i$ for some $1 \leqslant i \leqslant l$.

Fix an equivalence class Ξ. After choosing a representative $\varphi \in \Xi$, with slight abuse of notation, for each $1 \leqslant a' \leqslant 2n'$ let us denote by $\Xi(a')$

- the integer $\varphi(a')$ if $a' \prec W_i$ for some $0 \leqslant i \leqslant l$;
- the residue class of $\varphi(a')$ modulo $|t_i|$ if $a' \prec T_i$ for some $1 \leqslant i \leqslant l$.

In both cases, the value of $\Xi(a')$ does not depend on the choice of φ. The values of $\Xi(a')$ for $1 \leqslant a' \leqslant 2n'$ uniquely identify Ξ, and hence there are only finitely many equivalence classes of unbundling maps. As a consequence, it is enough to describe an algorithm to compute the set

$$\{\boldsymbol{k}(\varphi) : \varphi \in \Xi \text{ is cancelling}\},$$

which will be henceforth referred to as $\mathcal{A}^+_{(B',p'),\Xi}$.

Final computation The reason why fixing the unbundling class Ξ is beneficial is that conditions of the form

$$w(\boldsymbol{k}(\varphi))_{[\hat{\varphi}(a')+1\,:\,\hat{\varphi}(a'+1)]} = \bigl(w(\boldsymbol{k}(\varphi))_{[\hat{\varphi}(b')+1\,:\,\hat{\varphi}(b'+1)]}\bigr)^{-1} \tag{2}$$

for bands $(a', b') \in B'$ become mutually independent. More precisely, whether condition (2) is satisfied for a band (a', b') or not only depends on the value of $\varphi(a')$, as long as $\varphi \in \Xi$; we will now clarify and justify this claim.

Let $a' \in \{1, \dots, 2n'\}$ be an integer, and let $X \in \{W_0, \dots, W_l, T_1, \dots, T_l\}$ be its label. Define

$$h(a') = \sum_{\substack{b' \prec X \\ b' < a'}} \Xi(b').$$

When $X = T_i$ for some $1 \leqslant i \leqslant l$, the integer $h(a')$ is only well defined modulo $|t_i|$; by convention, we will assume that $0 \leqslant h(a') < |t_i|$. We can now rewrite the subwords involved in (2): for each $\varphi \in \Xi$ we have

$$w(\boldsymbol{k}(\varphi))_{[\hat{\varphi}(a')+1\,:\,\hat{\varphi}(a'+1)]} = \begin{cases} (w_i)_{[h(a')+1\,:\,h(a')+\varphi(a')]} & \text{if } a' \prec W_i, \\ (t_i^{\infty})_{[h(a')+1\,:\,h(a')+\varphi(a')]} & \text{if } a' \prec T_i. \end{cases}$$

Consider a band $(a', b') \in B'$. It is now clear that whether (a', b') satisfies (2) or not only depends on the value of $\varphi(a') = \varphi(b')$. Our next task is to show how to compute the set

$$S(a', b') = \{s \in \mathbb{Z}_{>0} : (a', b') \text{ satisfies (2) if and only if } \varphi(a') = s\},$$

and we do so by analysing two cases.

- Suppose that (a', b') is a band of type 1, with $a' \prec W_i$ for some $0 \leqslant i \leqslant l$ (the case where $b' \prec W_i$ is identical). Then $\varphi(a') = \Xi(a')$ is fixed, and condition (2) reads

$$(w_i)_{[h(a')+1\,:\,h(a')+\Xi(a')]} = \big((w_j)_{[h(b')+1\,:\,h(b')+\Xi(a')]}\big)^{-1} \quad \text{if } b' \prec W_j, \text{ or}$$
$$(w_i)_{[h(a')+1\,:\,h(a')+\Xi(a')]} = \Big((t_j^{\infty})_{[h(b')+1\,:\,h(b')+\Xi(a')]}\Big)^{-1} \quad \text{if } b' \prec T_j.$$

Either way, whether the equality of words holds or not only depends on Ξ. We have that $S(a', b') = \{\Xi(a')\}$ if the two words are equal, and $S(a', b') = \varnothing$ if they are not.

- Suppose that (a', b') is a band of type 2, with $a' \prec T_i$ and $b' \prec T_j$ for some $1 \leqslant i < j \leqslant l$. Then all the possible values for $\varphi(a') = \varphi(b')$ are of the form $q + Lv$, where
  - $L = \operatorname{lcm}(|t_i|, |t_j|)$,
  - v is a nonnegative integer, and
  - q is the only integer in $\{1, \dots, L\}$ such that $q \equiv \Xi(a') \pmod{|t_i|}$ and $q \equiv \Xi(b') \pmod{|t_j|}$;

if no such q exists then the class Ξ is empty, and we conclude that $\mathcal{A}^{+}_{(B',p'),\Xi} = \varnothing$. Otherwise, all that is left to do is determining for which values of $v \geqslant 0$ the band (a', b') satisfies (2). We can rewrite

$$(t_i^{\infty})_{[h(a')+1\,:\,h(a')+\varphi(a')]} = \big((t_i^{\infty})_{[h(a')+1\,:\,h(a')+L]}\big)^{v}\, (t_i^{\infty})_{[h(a')+1\,:\,h(a')+q]},$$

and similarly for $(t_j^{\infty})_{[h(b')+1\,:\,h(b')+\varphi(b')]}$. Therefore, condition (2) reads

$$\begin{aligned}\big((t_i^{\infty})_{[h(a')+1\,:\,h(a')+L]}\big)^{v}\, &(t_i^{\infty})_{[h(a')+1\,:\,h(a')+q]} \\ &= \Big((t_j^{\infty})_{[h(b')+q+1\,:\,h(b')+q+L]}\Big)^{-v} \Big((t_j^{\infty})_{[h(b')+1\,:\,h(b')+q]}\Big)^{-1}.\end{aligned}$$

There are three cases.

- If
$$(t_i^\infty)_{[h(a')+1\,:\,h(a')+q]} \neq \left((t_j^\infty)_{[h(b')+1\,:\,h(b')+q]}\right)^{-1},$$
then (2) is not satisfied for any value of v, and $S(a',b') = \varnothing$.
- If
$$(t_i^\infty)_{[h(a')+1\,:\,h(a')+q]} = \left((t_j^\infty)_{[h(b')+1\,:\,h(b')+q]}\right)^{-1}$$
but
$$(t_i^\infty)_{[h(a')+1\,:\,h(a')+L]} \neq \left((t_j^\infty)_{[h(b')+q+1\,:\,h(b')+q+L]}\right)^{-1},$$
then (2) is satisfied only for $v = 0$, leading to $S(a',b') = \{q\}$.
- If
$$(t_i^\infty)_{[h(a')+1\,:\,h(a')+L]} = \left((t_j^\infty)_{[h(b')+q+1\,:\,h(b')+q+L]}\right)^{-1},$$
then (2) is satisfied for every value of $v \geqslant 0$, so $S(a',b') = \{q + Lv : v \geqslant 0\}$.

In conclusion, we have shown how to compute a set $S(a',b')$ for each band $(a',b') \in B'$ with the following property: a function $\varphi\colon \{1,\dots,2n'\} \to \mathbb{Z}_{>0}$ is a cancelling representative of Ξ if and only if $\varphi(a') = \varphi(b') \in S(a',b')$ for every $(a',b') \in B'$. If $S(a',b') = \varnothing$ for some band (a',b') then clearly $\mathcal{A}^+_{(B',p'),\Xi} = \varnothing$. Otherwise, each set $S(a',b')$ can be written in the form
$$S(a',b') = \{q(a',b') + L(a',b')v : v \in \mathbb{Z}_{\geqslant 0}\}$$
for some integers $q(a',b') > 0$ and $L(a',b') \geqslant 0$. Therefore, every cancelling unbundling map $\varphi \in \Xi$ is described by a (not necessarily unique) vector $\boldsymbol{v} \in \mathbb{Z}^{n'}_{\geqslant 0}$ whose coordinates $v_{(a',b')}$ are indexed by bands $(a',b') \in B'$, such that
$$\varphi(a') = \varphi(b') = q(a',b') + L(a',b')v_{(a',b')} \quad \text{for each } (a',b') \in B'.$$
Conversely, each vector $\boldsymbol{v} \in \mathbb{Z}^{n'}_{\geqslant 0}$ describes a cancelling unbundling map $\varphi(\boldsymbol{v}) \in \Xi$, defined by the previous equations. As a consequence, we find that
$$\mathcal{A}^+_{(B',p'),\Xi} = \left\{\boldsymbol{k}(\varphi(\boldsymbol{v})) : \boldsymbol{v} \in \mathbb{Z}^{n'}_{\geqslant 0}\right\}.$$
Finally, recalling formula (1), it is easy to see that $\boldsymbol{k}(\varphi(\boldsymbol{v}))$ can be computed as $\boldsymbol{k}(\varphi(\boldsymbol{v})) = \boldsymbol{z} + \boldsymbol{M}\boldsymbol{v}$, where:

- $\boldsymbol{z} \in \mathbb{Z}^{l}_{\geqslant 0}$ is the vector whose coordinates are defined by
$$z_i = \frac{\sum_{a' \prec T_i} q(a')}{|t_i|} \quad \text{for } 1 \leqslant i \leqslant l,$$
where for ease of notation we have defined $q(a') = q(b') = q(a',b')$ for each band $(a',b') \in B'$;
- $\boldsymbol{M} \in \mathbb{Z}^{l \times n'}$ is the matrix whose entries are indexed by $\{1,\dots,l\} \times B'$ and defined by
$$M_{i,(a',b')} = \begin{cases} \dfrac{L(a',b')}{|t_i|} & \text{if } a' \prec T_i \text{ or } b' \prec T_i, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } 1 \leqslant i \leqslant l,\ (a',b') \in B'.$$

Therefore, we find that
$$\mathcal{A}^+_{(B',p'),\Xi} = \left\{\boldsymbol{z} + \boldsymbol{M}\boldsymbol{v} : \boldsymbol{v} \in \mathbb{Z}^{n'}_{\geqslant 0}\right\}. \qquad \square$$

3.3 Topological applications

As anticipated, the reason why we are interested in Theorem 3.4 is the following topological consequence. In Section 4.6, we will make use of Corollary 3.5 to solve the isotopy problem for surfaces which bound a genus-two handlebody on one side, while the other side is such that the trace of its mapping class group can be described in terms of Dehn twists about disjoint curves. Even though we will only apply this corollary to handlebodies of genus $g = 2$, we state and prove it in greater generality, at no additional cost in terms of effort or simplicity.[7]

Corollary 3.5 *Let V be a handlebody of genus $g \geqslant 1$. Let m be a nonnegative integer, and let $a_1, \dots, a_{2m}$ be pairwise disjoint curves in ∂V. Let $f : \partial V \to \partial V$ be a homeomorphism. There is an algorithm which, given as input $a_1, \dots, a_{2m}$ and f, decides whether there exist integers $h_1, \dots, h_m$ such that*

$$\tau_{a_1}^{h_1} \cdots \tau_{a_m}^{h_m} \tau_{a_{m+1}}^{-h_1} \cdots \tau_{a_{2m}}^{-h_m} f : \partial V \to \partial V$$

extends to a homeomorphism $V \to V$.

Proof Fix a basepoint $x_0 \in \partial V$ once and for all; additionally, let us arbitrarily pick orientations for $a_1, \dots, a_{2m}$. Consider a complete system of oriented meridians $b_1, \dots, b_g \subseteq \partial V$. For each b_r, we give a few definitions. Let $f(b_r)$ intersect $a_1 \cup \dots \cup a_{2m}$ transversely at points $p_{r,1}, \dots, p_{r,l_r}$, numbered in the order they appear on $f(b_r)$. For $j = 1, \dots, l_r$, let (when $j = l_r$, we take $j + 1$ to mean 1)

- $\beta_{r,j}$ be the subarc of $f(b_r)$ going from $p_{r,j}$ to $p_{r,j+1}$;
- $\gamma_{r,j} \subseteq \partial V$ be an arc joining x_0 to $p_{r,j}$;
- $i(r, j)$ be the only integer in $\{1, \dots, 2m\}$ such that $p_{r,j} \in a_{i(r,j)}$;
- $w_{r,j}$ be the element of $\pi_1(V, x_0)$ represented by $\gamma_{r,j} * \beta_{r,j} * \gamma_{r,j+1}^{-1}$, where "$*$" denotes the concatenation of paths;
- $t_{r,j}$ be the element of $\pi_1(V, x_0)$ represented by $\gamma_{r,j} * a_{i(r,j)}^{\epsilon} * \gamma_{r,j}^{-1}$, where we consider $a_{i(r,j)}$ as a loop based at $p_{r,j}$, and ϵ is 1 or -1 depending on the "sign" of the intersection between $f(b_r)$ and $a_{i(r,j)}$ at $p_{r,j}$, as depicted in Figure 10.

It is not hard to see that, given integers $h_1, \dots, h_{2m}$, the conjugacy class of the curve $\tau_{a_1}^{h_1} \cdots \tau_{a_{2m}}^{h_{2m}} f(b_r)$ in $\pi_1(V, x_0)$ is represented by the element

$$t_{r,1}^{h_{i(r,1)}} w_{r,1} \cdots t_{r,l_r}^{h_{i(r,l_r)}} w_{r,l_r};$$

we hope that Figure 11 will be convincing enough for the reader. Consider the set

$$\mathcal{A}_r = \left\{ (k_1, \dots, k_{l_r}) \in \mathbb{Z}^{l_r} : t_{r,1}^{k_1} w_{r,1} \cdots t_{r,l_r}^{k_{l_r}} w_{r,l_r} = 1 \right\}.$$

---

[7] In fact, the same proof works for a statement which is even more general. Given pairwise disjoint curves $a_1, \dots, a_m \subseteq \partial V$, a homeomorphism $f : \partial V \to \partial V$, and a matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$ with integer entries of suitable sizes, we can decide whether there exists a vector $\boldsymbol{h} \in \mathbb{Z}^m$ such that $\boldsymbol{A}\boldsymbol{h} \leqslant \boldsymbol{b}$ and $\tau_{a_1}^{h_1} \cdots \tau_{a_m}^{h_m} f$ extends to a homeomorphism $V \to V$.

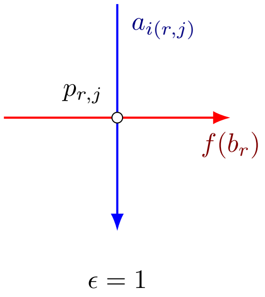


Figure 10: Choice of ϵ at the intersection $p_{r,j}$ between $f(b_r)$ and $a_{i(r,j)}$.

By the previous remark, we have that $\tau_{a_1}^{h_1}\cdots\tau_{a_{2m}}^{h_{2m}} f(b_r)$ is a meridian if and only if $\big(h_{i(r,j)}\big)_{j=1}^{l_r} \in \mathcal{A}_r$. Consequently, $\tau_{a_1}^{h_1}\cdots\tau_{a_{2m}}^{h_{2m}} f$ extends to a homeomorphism of V if and only if $\big(h_{i(r,j)}\big)_{j=1}^{l_r} \in \mathcal{A}_1\cap\ldots\cap\mathcal{A}_g$.

By Theorem 3.4, the sets $\mathcal{A}_r$ can be algorithmically decomposed as

$$\mathcal{A}_r = A_{r,1}\cup\cdots\cup A_{r,N_r},$$

where

$$A_{r,q} = \left\{\boldsymbol{z}_{r,q} + \boldsymbol{M}_{r,q}\boldsymbol{v} : \boldsymbol{v}\in\mathbb{Z}_{\geqslant 0}^{d_{r,q}}\right\}$$

for some vector $\boldsymbol{z}_{r,q}\in\mathbb{Z}^{l_r}$ and some matrix $\boldsymbol{M}_{r,q}\in\mathbb{Z}^{l_r\times d_{r,q}}$. Therefore, we finally find that there exist integers $h_1,\ldots,h_m$ such that

$$\tau_{a_1}^{h_1}\cdots\tau_{a_m}^{h_m}\tau_{a_{m+1}}^{-h_1}\cdots\tau_{a_{2m}}^{-h_m} f$$

extends to a homeomorphism of V if and only if, for some choice of indices

$$q_1\in\{1,\ldots,N_1\},\ \ldots, q_g\in\{1,\ldots,N_g\},$$

this system of linear equations has at least one solution in the variables $\boldsymbol{v}_1\in\mathbb{Z}_{\geqslant 0}^{d_{1,q_1}},\ldots,\boldsymbol{v}_g\in\mathbb{Z}_{\geqslant 0}^{d_{g,q_g}}$, $h_1,\ldots,h_m\in\mathbb{Z}$:

$$\begin{cases} \boldsymbol{z}_{r,q_r} + \boldsymbol{M}_{r,q_r}\boldsymbol{v}_r = \big(h_{i(r,j)}\big)_{j=1}^{l_r} & \text{for } r=1,\ldots,g,\\ h_r = -h_{m+r} & \text{for } r=1,\ldots,m. \end{cases}$$

This condition can be checked algorithmically (a classical result proved in [6] by Gathen and Sieveking) for every choice of indices $q_1,\ldots,q_g$, providing an algorithmic solution to the question in the statement. □

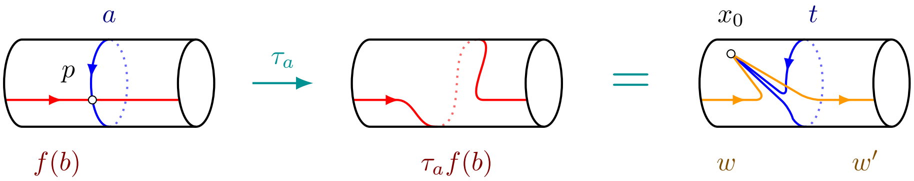


Figure 11: The effect of applying a Dehn twist about a to the curve $f(b)$; indices were omitted for clarity.

4 The classification algorithm

4.1 Outline

Section 4 will be entirely devoted to providing a proof of the following.

Theorem 4.1 *There is an algorithm to decide whether two genus-two oriented surfaces embedded in* S^3 *are isotopic or not.*

More precisely, let S_1 and S_2 be genus-two surfaces embedded in S^3, and let us fix an orientation for S_1 and for S_2. We say that S_1 and S_2 are isotopic (as oriented surfaces) if there is an isotopy $f_t\colon S^3 \to S^3$ such that f_0 is the identity and f_1 maps S_1 to S_2 orientation-preservingly. Since every self-homeomorphism of S^3 is isotopic to the identity, the oriented surfaces S_1 and S_2 are isotopic if and only if there exists a homeomorphism $f\colon S^3 \to S^3$ sending S_1 to S_2 orientation-preservingly.

Another equivalent definition is the following. Let M_1 be the closure of the component of $S^3 \setminus S_1$ which lies on the positive side of S_1, and let N_1 be the closure of the other component; we will sometimes call M_1 and N_1 the *sides* of S_1. Define M_2 and N_2 in a similar way. Then S_1 and S_2 are isotopic if and only if there exist homeomorphisms $f\colon M_1 \to M_2$ and $g\colon N_1 \to N_2$ such that $f|_\partial$ and $g|_\partial$ are isotopic as homeomorphisms $S_1 \to S_2$.

Remark 4.2 A nonoriented version of Theorem 4.1 states that there is an algorithm to decide whether two genus-two surfaces embedded in S^3 are isotopic or not. Of course, the nonoriented version is an immediate corollary of the oriented one. In fact, in order to decide whether the surfaces S_1 and S_2 are isotopic without any constraint on orientation, we can simply check if they are orientation-preservingly isotopic for at least one choice of orientations.

From now on, we will take S_1, S_2, M_1, N_1, M_2, and N_2 to refer to the objects we have already introduced; in particular, we will assume orientations for S_1 and S_2 have been fixed. Our aim will be to provide a proof of Theorem 4.1. The full algorithm is quite involved, and for the sake of convenience we will split it into several cases. Before we start describing the actual algorithm, let us briefly outline the structure of the following sections.

- Several cases of the proof of Theorem 4.1 will be addressed by finding *canonical* compression discs for S_1 and S_2, then compressing S_1 and S_2 along these discs and — carefully — reducing the question to the isotopy problem of tori in S^3. In Section 4.2 we provide a general framework for dealing with these canonical compression discs; by abstracting the repetitive parts of the algorithm away, we will later be able to focus on what is meaningfully different in each of these cases.
- In Section 4.3 we deal with the problem of finding such canonical compression discs. The recipe we present relies on showing the existence of discs satisfying properties which are *stable under boundary compressions*; if these properties are suitably chosen, uniqueness will easily follow.

- After the introductory sections, we finally start describing the classification algorithm. Sections 4.4 and 4.5 deal with the case where S_1 and S_2 are *partially compressible on one side*. Specifically, the case where one side has at least one — and, a posteriori, only one — nonseparating compression disc is addressed in Section 4.4; Section 4.5 explains the strategy to follow when one side only has separating compression discs. In both cases, we heavily rely on the tools developed in Sections 4.2 and 4.3.

- We are then left to address the case where one side of S_1 is a handlebody and the other is boundary irreducible. The algorithm we use here will depend on the JSJ decomposition of the boundary irreducible component. If all the I-bundle pieces are "small", then the algorithmic problem on free groups discussed in Section 3 will quickly lead to a solution of the isotopy problem, as described in Section 4.6. Otherwise, only two cases can arise: either a product I-bundle over a punctured torus appears in the JSJ decomposition — and we deal with this in Section 4.7 — or one of the pieces is a twisted I-bundle over a punctured Klein bottle; the latter situation is addressed in Section 4.8. In both cases, an ad hoc discussion is enough to settle the isotopy problem.

4.2 General strategy for canonical compression discs

We say that nontrivial compression discs D_1 for S_1 and D_2 for S_2 are *canonical* if every homeomorphism $S^3 \to S^3$ sending S_1 to S_2 orientation-preservingly sends D_1 to D_2, up to isotopy preserving S_2. As anticipated, we will often rely on compressing the surfaces along canonical compression discs in order to reduce the isotopy question to the more manageable problem for tori in S^3. Let us begin our discussion with a definition.

Definition 4.3 Let X_1 and X_2 be 3-manifolds, and let $p_1 \subseteq \partial X_1$ and $p_2 \subseteq \partial X_2$ be simple closed curves. Let X_1' and X_2' be the 3-manifolds obtained by attaching a 2-handle, respectively, to X_1 along p_1 and to X_2 along p_2. Define the *handle extension map*

$$\mathrm{HExt}_{p_1,p_2}\colon \boldsymbol{H}((X_1,p_1);(X_2,p_2)) \to \boldsymbol{H}(X_1';X_2')$$

by extending homeomorphisms $(X_1,p_1) \to (X_2,p_2)$ to the 2-handles.

It is easy to see that this map is well defined, in that its output does not depend on the choice of representative for the isotopy class nor on the specific extension to the 2-handles. Moreover, the handle extension map is functorial in the following sense: for 3-manifolds X_1, X_2, X_3, curves $p_1 \subseteq \partial X_1$, $p_2 \subseteq \partial X_2$, $p_3 \subseteq \partial X_3$, and homeomorphisms

$$f\colon (X_1,p_1) \to (X_2,p_2), \quad g\colon (X_2,p_2) \to (X_3,p_3),$$

we have

$$\mathrm{HExt}_{p_1,p_3}(gf) = \mathrm{HExt}_{p_2,p_3}(g) \circ \mathrm{HExt}_{p_1,p_2}(f).$$

Suppose now that we can find canonical compression discs D_1 and D_2 for S_1 and S_2, respectively. Up to flipping the orientations of S_1 and S_2, we can assume that D_1 and D_2 lie in M_1 and M_2, respectively.

Let $p_1 \subseteq S_1$ and $p_2 \subseteq S_2$ be the boundary curves of D_1 and D_2, respectively. Let $P_1 = M_1 \mathbin{\setminus\!\!\setminus} D_1$ (which may be a disconnected 3-manifold) and $Q_1 = \operatorname{clos}(S^3 \setminus P_1)$. Note that Q_1 can also be obtained by attaching the 2-handle $\operatorname{clos}(Q_1 \setminus N_1)$ to N_1 along the curve p_1. The 3-manifolds P_1 and Q_1 have the same boundary T_1, which is either a torus or the union of two tori. Similarly, define $P_2 = M_2 \mathbin{\setminus\!\!\setminus} D_2$, $Q_2 = \operatorname{clos}(S^3 \setminus P_2)$, and $T_2 = \partial P_2 = \partial Q_2$.

Proposition 4.4 *In the situation described in the previous paragraph, the oriented surfaces S_1 and S_2 are isotopic if and only if there is a homeomorphism $f\colon (N_1, p_1) \to (N_2, p_2)$ such that $\operatorname{HExt}_{p_1,p_2}(f)|_\partial$, seen as a map from T_1 to T_2, extends to a homeomorphism $P_1 \to P_2$.*

Proof If S_1 and S_2 are isotopic, let $f\colon S^3 \to S^3$ be a homeomorphism sending S_1 to S_2 orientation-preservingly. Since D_1 and D_2 are canonical, we can isotope f so that it sends D_1 to D_2 (and, hence, p_1 to p_2). Then $f|_{N_1}\colon (N_1, p_1) \to (N_2, p_2)$ is a homeomorphism such that $\operatorname{HExt}_{p_1,p_2}(f|_{N_1})|_\partial$ extends to $f|_{P_1}\colon P_1 \to P_2$.

Conversely, let $f\colon (N_1, p_1) \to (N_2, p_2)$ and $g\colon P_1 \to P_2$ be homeomorphisms such that $\operatorname{HExt}_{p_1,p_2}(f)$ and g have the same trace $T_1 \to T_2$. Then $\operatorname{HExt}_{p_1,p_2}(f)$ and g can be combined to construct a self-homeomorphism of S^3 sending S_1 to S_2 orientation-preservingly. □

Proposition 4.4 will occasionally be useful on its own. However, for most of our applications, we will refer to the following result. The proof exploits the functoriality of the handle extension map — together with the output of Theorem 2.23 — to translate the isotopy problem into a group-theoretic question which, while unsolvable in general, can be easily answered in the special cases we need.

Proposition 4.5 *With the same notation as above, we can algorithmically decide whether the oriented surfaces S_1 and S_2 are isotopic provided that*

(i) *the 3-manifold pair $(N_1, \partial N_1 \mathbin{\setminus\!\!\setminus} p_1)$ is irreducible;*

(ii) *P_1 is either a (possibly trivial) knot complement, or the union of two nontrivial knot complements.*

Proof Since $(N_1, \partial N_1 \mathbin{\setminus\!\!\setminus} p_1)$ is irreducible, we can decide whether (N_1, p_1) is homeomorphic to (N_2, p_2) or not. If it is not, then clearly S_1 and S_2 are not isotopic. Otherwise, let $f_0\colon (N_1, p_1) \to (N_2, p_2)$ be a homeomorphism. Similarly, since P_1 is a union of possibly trivial knot complements, we can decide if P_1 and P_2 are homeomorphic. If they are not, then once again S_1 and S_2 are not isotopic. Otherwise, let $g_0\colon P_1 \to P_2$ be a homeomorphism. By Proposition 4.4, we can solve the isotopy problem if we can answer the following question: is there a homeomorphism $f \in \boldsymbol{H}(N_2, p_2)$ such that

$$\operatorname{HExt}_{p_2,p_2}(f)|_\partial \circ (\operatorname{HExt}_{p_1,p_2}(f_0)|_\partial \circ g_0|_\partial^{-1}) \in \boldsymbol{H}(P_2)|_\partial?$$

By applying Theorem 2.23, we can further reduce the question to the following: given homeomorphisms $f_1, \dots, f_n \in \boldsymbol{H}(Q_2)|_\partial$ which are products of powers of Dehn twists of T_2, and $h \in \boldsymbol{H}(T_2)$, does there exist $f \in \langle f_1, \dots, f_n\rangle$ such that $fh \in \boldsymbol{H}(P_2)|_\partial$?

When P_2 and Q_2 are solid tori Let $\ell, m \subseteq T_2$ be an oriented longitude and an oriented meridian of P_2, respectively; denote by $[\ell]$ and $[m]$ the corresponding homology classes in $H_1(T_2)$. The Dehn twists of T_2 which extend to Q_2 generate a subgroup of $\boldsymbol{H}(T_2)$ which is isomorphic to $\mathbb{Z}$ — namely, the homeomorphisms that fix $[\ell]$. We can therefore compute an integer k_0 such that[8] the group $\langle f_1, \dots, f_n \rangle$ is generated by the self-homeomorphism of T_2 sending $[\ell]$ to $[\ell]$ and $[m]$ to $k_0[\ell] + [m]$. It is immediate to check that the answer to the isotopy question is positive if and only if h sends $[m]$ to $k[\ell] \pm [m]$ where k is a multiple of k_0.

When P_2 is a solid torus but Q_2 is not In this case, the group $\langle f_1, \dots, f_n \rangle$ is actually trivial, since $\boldsymbol{H}(Q_2)|_\partial \leqslant \langle -\mathrm{id} \rangle$ and Dehn twists cannot act as $(-\mathrm{id})$ on $H_1(T_2)$. Therefore, it is sufficient to check whether $h \in \boldsymbol{H}(P_2)|_\partial$ or not.

When P_2 is a nontrivial knot complement By Corollary 2.25, the group $\boldsymbol{H}(P_2)|_\partial$ is finite and can be computed. Therefore, it is sufficient to be able to decide if $h \in \langle f_1, \dots, f_n \rangle$. Note that, in this case, Q_2 is a solid torus. As argued above, the homeomorphisms $f_1, \dots, f_n$ belong to the subgroup of $\boldsymbol{H}(T_2)$ fixing the homology class of the meridian of Q_2. This group is isomorphic to $\mathbb{Z}$, and we can compute a generator of $\langle f_1, \dots, f_n \rangle$. By studying the action of h on $H_1(T_2)$, we can easily decide whether h belongs to $\langle f_1, \dots, f_n \rangle$ or not.

When P_2 is the union of two nontrivial knot complements Like before, it is enough to be able to decide if $h \in \langle f_1, \dots, f_n \rangle$. In this case, the 3-manifold Q_2 is homeomorphic to the complement of the 2-component unlink in S^3. Denote by T and T' the two components of T_2. Let $m \subseteq T$ and $m' \subseteq T'$ be the two unique curves which are homologically trivial in Q_2 but not in T_2, and fix an orientation for each of them. Clearly, every self-homeomorphism of Q_2 preserves $m \cup m'$ as a set up to isotopy. Moreover, Dehn twists of T_2 cannot swap T and T' or invert the orientation of m or m'. Therefore, the homeomorphisms $f_1, \dots, f_n$ belong to the subgroup of $\boldsymbol{H}(T_2)$ fixing the homology classes $[m]$ and $[m']$ in $H_1(T_2)$. This group is isomorphic to $\mathbb{Z} \times \mathbb{Z}$; hence, by studying the action of h on $H_1(T_2)$, we can decide whether h belongs to $\langle f_1, \dots, f_n \rangle$ or not. □

4.3 Compression discs for the boundary

The reader should by now reasonably believe that canonical compression discs will play a crucial role in our classification algorithm. A good strategy to prove that two discs — say — in M_1 and M_2 are canonical consists in characterising them as the unique discs satisfying some property which only depends on the intrinsic topologies of M_1 and M_2. The results in this section will provide us with useful tools to prove such a characterisation.

---

[8] More explicitly, if f_i sends $[m]$ to $k_i[\ell] + [m]$ for $1 \leqslant i \leqslant n$, then $k_0 = \gcd(k_1, \dots, k_n)$. In fact, since the homeomorphisms $f_1, \dots, f_n$ are actually products of two Dehn twists about disjoint curves, the integers k_i will always be equal to -1, 0, or 1.

Let (M, R) be a 3-manifold pair. Consider a property $\mathcal{P}$ for discs properly embedded in M which is invariant under isotopies in (M, R), such as "being separating". We say that $\mathcal{P}$ is *stable under boundary compressions* if

- for every disc D properly embedded in (M, R), and
- for every possibly trivial boundary compression disc $E \subseteq M$ for D with $\partial E \subseteq D \cup R$,

we have that at least one of the two discs obtained by boundary compressing D along E satisfies $\mathcal{P}$. The crucial fact is that, (very) loosely speaking, discs satisfying a property which is stable under boundary compressions can be made disjoint. More precisely, we have the following.

Proposition 4.6 *Let M be an irreducible 3-manifold, and let $R \subseteq \partial M$ be a surface. Let F be a disc properly embedded in (M, R). Suppose that there is a disc properly embedded in (M, R) not isotopic to F in M and satisfying some property $\mathcal{P}$ which is stable under boundary compressions. Then there is a disc properly embedded in (M, R) which is not isotopic to F in M, is disjoint from F, and satisfies $\mathcal{P}$.*

Proof Among all discs properly embedded in (M, R) which are in general position with respect to F, are not isotopic to F, and satisfy $\mathcal{P}$, pick D to minimise the number of components of $F \cap D$; we will show that D is in fact disjoint from F. Since M is irreducible, a standard innermost circle argument shows that $F \cap D$ is a collection of arcs. Suppose that this intersection is nonempty, and let $a \subseteq F \cap D$ be an outermost arc in F. The arc a cuts a disc E off of F, such that E is a (possibly trivial) boundary compression disc for D and $\partial E \subseteq D \cup R$. Let D_1 and D_2 be the two discs obtained by boundary compressing D along E, with D_1 satisfying $\mathcal{P}$.

If D_1 is isotopic to F, then D can in fact be isotoped to be disjoint from F, contradicting our minimality assumption. Otherwise, D_1 is a disc satisfying $\mathcal{P}$ which is not isotopic to F, and its intersection with F has strictly fewer components than $F \cap D$. Again, this contradicts the minimality of D. It follows that D must be disjoint from F, as required. □

The next proposition shows that some of the properties we care about are, in fact, stable under boundary compressions.

Proposition 4.7 *Let M be an irreducible 3-manifold, and let $R \subseteq \partial M$ be a surface. The following properties for discs properly embedded in (M, R) are stable under boundary compressions*:

(1) *being nonseparating*;

(2) *having boundary which is a nontrivial curve in R.*

Proof Let D be a disc properly embedded in (M, R), and let $E \subseteq M$ be a possibly trivial boundary compression disc for D with $\partial E \subseteq D \cup \mathrm{int}(R)$. Denote by D_1 and D_2 the two discs obtained by boundary compressing D along E.

(1) The disc D is nonseparating if and only if $[D] \in H^1(M)$ is nontrivial. But

$$[D] = [D_1] + [D_2] \in H^1(M),$$

and therefore if D is nonseparating then one of D_1 and D_2 must be as well.

(2) Suppose that D_1 cobounds a ball $B \subseteq M$ with some disc in R. If B is disjoint from E, then D is actually isotopic to D_2, so ∂D_2 is trivial in R if and only if ∂D is. If instead B contains E then it also contains D_2, and it is easy to see that ∂D is trivial in R. □

This is a good time to remark that, when we say that a disc properly embedded in a 3-manifold is separating, we mean that it splits the 3-manifold itself, and not only the boundary, in two connected components. However, as the following proposition shows, there is no difference for 3-manifolds which can be embedded in S^3. Since all the 3-manifolds we will work with have this property — and, usually, already lie inside S^3 — we will freely use the term "separating" to denote discs which separate the 3-manifold they lie in and its boundary.

Proposition 4.8 *Let M be a 3-manifold embedded in S^3 with connected boundary, and let $D \subseteq M$ be a properly embedded disc. Then D separates M if and only if ∂D separates ∂M.*

Proof Clearly if D separates M then ∂D separates ∂M. Conversely, suppose that ∂D separates ∂M but D does not separate M. Let $P = M \setminus\setminus D$ be the result of cutting M along D; by assumption, we have that P is connected, but ∂P is the union of two components. We can interpret $\operatorname{clos}(M \setminus P)$ as a 1-handle which, when attached to P, yields the original 3-manifold M. Let $a \subseteq M$ be the cocore arc of this 1-handle, and let b be an arc properly embedded in P connecting the two endpoints of a. Then $a \cup b$ is a closed curve in M which intersects each boundary component of P exactly once. This provides a contradiction, since M is embedded in S^3. □

4.4 One nonseparating compression disc

We finally start presenting the actual classification algorithm. As anticipated, this section addresses the case where one side of S_1 — say M_1 — has a properly embedded nonseparating disc D_1 but is not a handlebody; see Figure 12 for an example. Define p_1, P_1, Q_1, and T_1 as described in Section 4.2 (even though we don't know that D_1 belongs to a pair of canonical discs yet). The torus T_1 bounds a solid torus in S^3. But P_1 cannot be a solid torus, for otherwise M_1 would be a handlebody. Hence, the solid torus must be the other component of $S^3 \setminus\setminus T_1$, namely Q_1. In other words, P_1 is the complement of a nontrivial knot and, as such, is boundary irreducible. As described in Figure 13, the 3-manifold M_1 is obtained by attaching a 1-handle to P_1, while drilling an arc from the solid torus Q_1 yields N_1. The following lemma applied to $(P_1, \partial P_1)$ implies, a posteriori, that no arbitrary choice was made when selecting the disc D_1. We remark that the lemma is stated in greater generality then we actually need here, but we will need this more general version later.

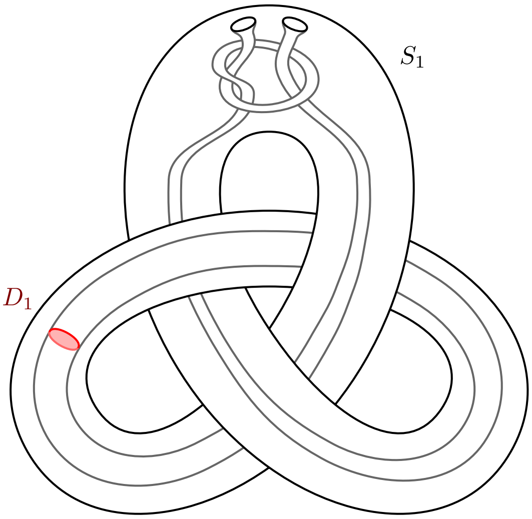
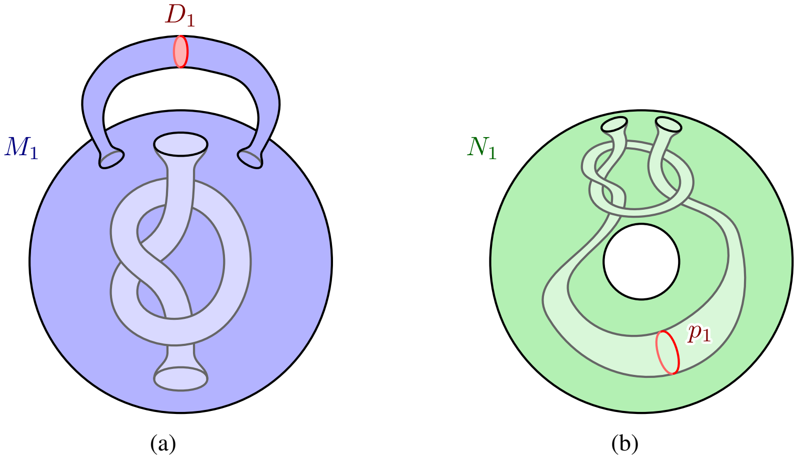


Figure 12: The surface S_1 admits a nonseparating compression disc D_1.

Lemma 4.9 *Let (P, R) be an irreducible 3-manifold pair. Let M be the result of attaching a 1-handle H to P along two discs in $\operatorname{int}(R)$, and define $R' = \partial M \setminus (\partial P \setminus R)$. Then the cocore of H is the unique compression disc for R' which does not separate M.*

Proof Let F be the cocore of H, and suppose there is another nonseparating compression disc D for R' which is not isotopic to F. By Propositions 4.6 and 4.7, we may assume that D is disjoint from F and, therefore, contained in P. Since the pair (P, R) is irreducible, D must cobound a ball in P with some disc $D' \subseteq R$. But D is nonseparating in M, and this is only possible if D' contains exactly one of the attaching discs of H. This would imply that D is isotopic to F, contrary to our assumption. □



(a) (b)

Figure 13: The 3-manifold M_1 (a) is homeomorphic to a knot complement P_1 with a 1-handle attached. The 3-manifold N_1 (b) is homeomorphic to a solid torus Q_1 with an arc drilled out.

Now, if M_2 is a handlebody or S_2 has no nonseparating compression discs in M_2 then clearly S_1 and S_2 are not isotopic. Otherwise, by imitating the above procedure for S_2 instead of S_1, we define p_2, D_2, P_2, Q_2, and T_2. Since D_1 (respectively, D_2) can be intrinsically defined as the unique nonseparating compression disc for S_1 (respectively, S_2) in M_1 (respectively, M_2), every homeomorphism $M_1 \to M_2$ must send D_1 to a disc isotopic to D_2. If the 3-manifold pair $(N_1, \partial N_1 \setminus\setminus p_1)$ is irreducible, we immediately conclude by applying Proposition 4.5; note that this is the case for the surface depicted in Figure 12, since the pair $(N_1, \partial N_1 \setminus\setminus p_1)$ shown in Figure 13 is irreducible.

Otherwise, the surface $\partial N_1 \setminus\setminus p_1$ is compressible in N_1, and we can find a compression disc $E_0 \subseteq N_1$ for it. The disc E_0 is properly embedded in the solid torus Q_1; if it is a meridian disc of Q_1, then let us define $E_1 = E_0$. Otherwise, E_0 must cut a ball off of Q_1; this ball necessarily contains the curve p_1, for otherwise E_0 would be a trivial compression disc for $\partial N_1 \setminus\setminus p_1$. We can therefore find a meridian disc E_1 of Q_1 which is disjoint from B (and, hence, from p_1).

Either way, we have found a disc E_1 properly embedded in $(N_1, \partial N_1 \setminus\setminus p_1)$ which is nonseparating in Q_1. Similarly, let E_2 be a disc properly embedded in $(N_2, \partial N_2 \setminus\setminus p_2)$ which is nonseparating in Q_2; see the top left of Figure 14 for an example. We claim that S_1 and S_2 are isotopic if and only if P_1 is homeomorphic to P_2 and (N_1, p_1) is homeomorphic to (N_2, p_2). The forward implication is trivial. Conversely, suppose that the 3-manifolds with boundary pattern (N_1, p_1) and (N_2, p_2) are homeomorphic. Note that ∂E_2 does not separate $\partial N_2 \setminus\setminus p_2$; hence, as shown in Figure 14, it is easy to construct a self-homeomorphism h of (N_2, p_2) which maps E_2 to itself with the opposite orientation, and is the identity in a neighbourhood of p_2. Then the Dehn twists about E_2, together with $\mathrm{HExt}_{p_2,p_2}(h)$, generate the whole mapping class group of the solid torus Q_2. As a consequence, the map

$$\mathrm{HExt}_{p_1,p_2}\colon \boldsymbol{H}((N_1, p_1); (N_2, p_2)) \to \boldsymbol{H}(Q_1; Q_2)$$

is surjective. If, moreover, the knot complements P_1 and P_2 are homeomorphic, then every homeomorphism $P_1 \to P_2$ extends to a homeomorphism $Q_1 \to Q_2$ which, in turn, is the image under HExt_{p_1,p_2} of some homeomorphism $(N_1, p_1) \to (N_2, p_2)$. By Proposition 4.4, this shows that S_1 and S_2 are isotopic.

Finally, we explain how to algorithmically decide whether (N_1, p_1) and (N_2, p_2) are homeomorphic or not. Note that $N_1 \setminus\setminus E_1$ is a knot complement, since E_1 is a meridian disc for Q_1; the same holds for $N_2 \setminus\setminus E_2$.

- If N_1 and N_2 are not handlebodies, then $N_1 \setminus\setminus E_1$ and $N_2 \setminus\setminus E_2$ are nontrivial knot complements, and p_1 and p_2 are meridian curves of $N_1 \setminus\setminus E_1$ and $N_2 \setminus\setminus E_2$, respectively. Lemma 4.9 then implies that E_1 and E_2 are the unique nonseparating compression discs for $\partial N_1 \setminus\setminus p_1$ and $\partial N_2 \setminus\setminus p_2$ in N_1 and N_2, respectively. As a consequence, every homeomorphism $(N_1, p_1) \to (N_2, p_2)$ maps E_1 to E_2; we deduce that (N_1, p_1) and (N_2, p_2) are homeomorphic if and only if $N_1 \setminus\setminus E_1$ and $N_2 \setminus\setminus E_2$ are.

- If N_1 and N_2 are handlebodies, we claim that (N_1, p_1) and (N_2, p_2) are always homeomorphic. Note that attaching a 2-handle to $N_1 \setminus\setminus E_1$ along p_1 yields a 3-ball. In other words, the curve p_1 is a longitude of the solid torus $N_1 \setminus\setminus E_1$. As a consequence, there exists a homeomorphism $(N_1 \setminus\setminus E_1, p_1) \to (N_2 \setminus\setminus E_2, p_2)$, and every such homeomorphism extends to a homeomorphism $(N_1, p_1) \to (N_2, p_2)$.

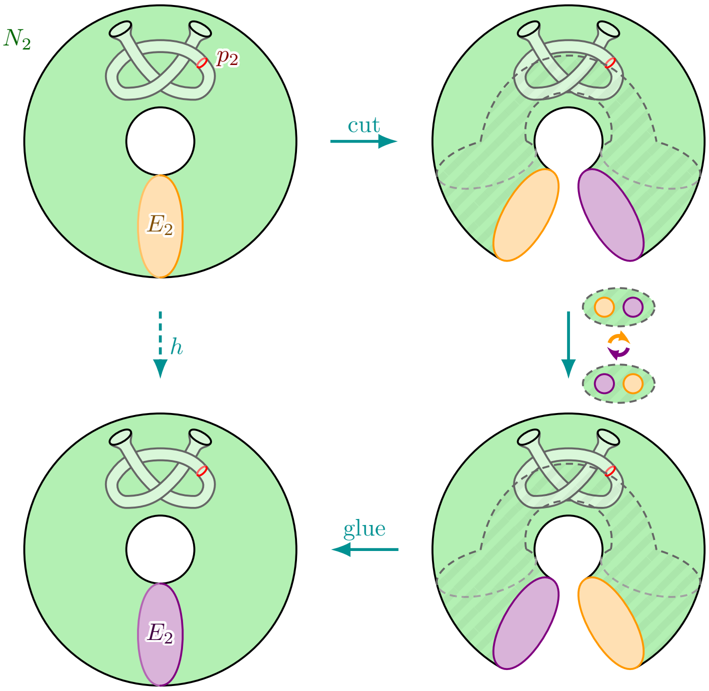


Figure 14: There exists a self-homeomorphism h of (N_2, p_2) which flips the disc E_2 and restricts to the identity in a neighbourhood of p_2. This homeomorphism can be constructed by cutting N_2 along E_2, swapping the two sides of E_2 by sliding in the shaded region of ∂N_2, and finally gluing the two sides back together.

4.5 One separating compression disc

We now deal with the case where one of the sides of S_1 — again, say M_1 — has compressible boundary, but all the compression discs are separating; an example of this situation is depicted in Figure 15. Let D_1 be a compression disc for S_1 in M_1. Define p_1, P_1, Q_1, and T_1 as described in Section 4.2. In this case, the 3-manifold P_1 is the union of two connected components $P_{1,1}$ and $P_{1,2}$, with boundary tori $T_{1,1}$ and $T_{1,2}$, respectively, so that $T_1 = T_{1,1} \cup T_{1,2}$.

The torus $T_{1,1}$ bounds a solid torus $\mathbb{T}$ in S^3. But $P_{1,1}$ cannot be a solid torus, since S_1 has no nonseparating compression discs in M_1. Hence, the solid torus $\mathbb{T}$ must be the closure of the other component of $S^3 \setminus T_{1,1}$. In particular, we have that $P_{1,2} \subseteq \mathbb{T}$. Now, the torus $T_{1,2}$ must be compressible in $\mathbb{T}$, since it is not π_1-injective. It cannot be compressible in $P_{1,2}$ though, because S_1 has no nonseparating compression discs in M_1. Therefore, it is easy to see that $P_{1,2}$ is contained in a ball and, moreover, it is a (nontrivial) knot complement. Clearly, the same holds for $P_{1,1}$. From this discussion, it follows that Q_1 is homeomorphic to the complement in S^3 of the 2-component unlink. As described in Figure 16, the 3-manifold M_1 is obtained by joining $P_{1,1}$ and $P_{1,2}$ by a 1-handle, while drilling an arc from Q_1 yields N_1.

The following lemma implies, a posteriori, that no arbitrary choice was made when selecting the disc D_1.

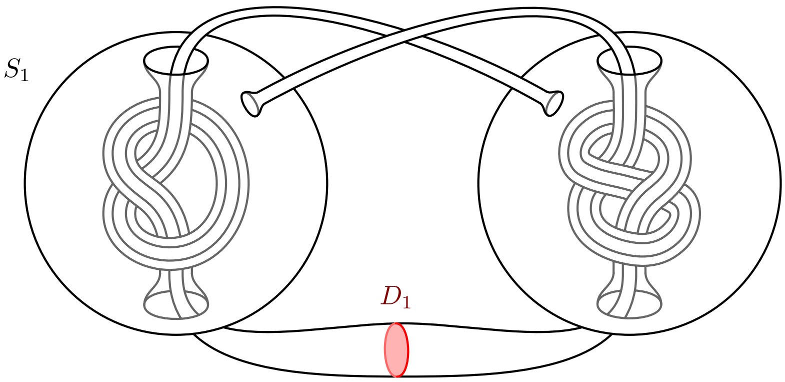


Figure 15: The surface S_1 admits a separating compression disc D_1.

Lemma 4.10 *Let P_1 and P_2 be irreducible boundary irreducible 3-manifolds, and let M be the result of joining them by a 1-handle H. Then the cocore of H is the unique nontrivial compression disc for ∂M in M.*

Proof Let F be the cocore of H, and suppose there is another (nontrivial) compression disc D for ∂M which is not isotopic to F; in particular, D is not boundary-parallel. By Propositions 4.6 and 4.7 applied

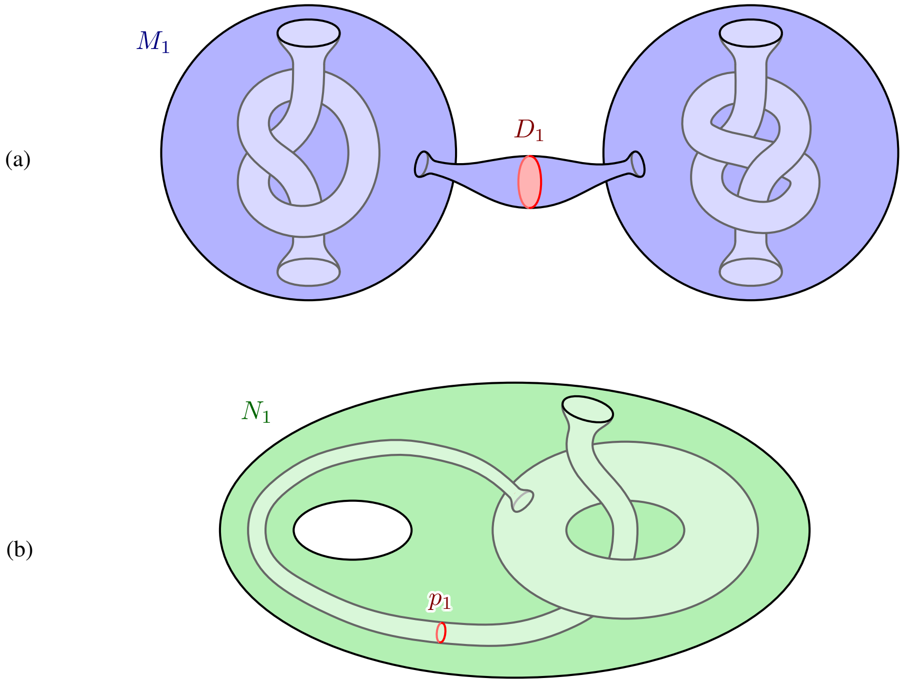


Figure 16: The 3-manifold M_1 (a) is homeomorphic to two knot complements $P_{1,1}$ and $P_{1,2}$ joined by a 1-handle. The 3-manifold N_1 (b) is homeomorphic to the complement Q_1 of a 2-unlink, from which an arc connecting the two boundary components has been drilled out.

to the pair $(M, \partial M)$, we may assume that D is disjoint from F and therefore, without loss of generality, contained in P_1. Since P_1 is irreducible and boundary irreducible, D must cobound a ball in P_1 with some disc $D' \subseteq \partial P_1$. But D is not boundary parallel in M, and this is only possible if D' contains one of the attaching discs of H. This would imply that D is isotopic to F, contrary to our assumption. □

Now, if M_2 is boundary irreducible or S_2 has a nonseparating compression disc in M_2, then clearly S_1 and S_2 are not isotopic. Otherwise, by imitating the above procedure for S_2 instead of S_1, we define p_2, D_2, P_2, Q_2, T_2, $P_{2,1}$, $P_{2,2}$, $T_{2,1}$, and $T_{2,2}$. Since D_1 (respectively, D_2) can be intrinsically defined as the unique compression disc for S_1 (respectively, S_2) in M_1 (respectively, M_2), every homeomorphism $M_1 \to M_2$ must send D_1 to a disc isotopic to D_2. If the 3-manifold pair (N_1, p_1) is irreducible, we immediately conclude by applying Proposition 4.5.

Otherwise, we can assume that the surfaces $\partial N_1 \setminus\!\!\setminus p_1$ and $\partial N_2 \setminus\!\!\setminus p_2$ are compressible in N_1 and N_2, respectively. We will show that, in this case, the surfaces S_1 and S_2 are isotopic if and only if P_1 and P_2 are homeomorphic. The forward implication is trivial. Conversely, suppose that P_1 and P_2 are homeomorphic. Let E be a nontrivial compression disc for $\partial N_1 \setminus\!\!\setminus p_1$ in N_1. The disc E is properly embedded in Q_1, and its boundary lies in a component of T_1, say $T_{1,1}$.

- If ∂E is nontrivial in $T_{1,1}$, then $N_1 \setminus\!\!\setminus E$ is homeomorphic to a solid torus $\mathbb{T}$ via a homeomorphism sending p_1 to a trivial curve in $\partial\mathbb{T}$. Therefore, N_1 is a handlebody and p_1 is a meridian curve of N_1.
- If ∂E is trivial in $T_{1,1}$, up to isotoping E, we can assume that $\partial E = p_1$. The sphere $S = E \cup D_1$ is embedded in Q_1, and does not bound a ball, hence it separates Q_1 into two components, each homeomorphic to a punctured solid torus. As a consequence, the disc E separates N_1 into two components, each homeomorphic to a solid torus. We conclude that N_1 is a handlebody, and p_1 is a meridian curve of N_1.

Either way, we have shown that (N_1, p_1) and (N_2, p_2) are both homeomorphic to a handlebody endowed with a separating meridian curve as a boundary pattern. Let $m_{1,1} \subseteq T_{1,1}$ and $m_{1,2} \subseteq T_{1,2}$ be the unique curves which are homologically trivial in Q_1 but not in T_1; define $m_{2,1} \subseteq T_{2,1}$ and $m_{2,2} \subseteq T_{2,2}$ analogously. By construction, the curves $m_{1,1}$, $m_{1,2}$, $m_{2,1}$ and $m_{2,2}$ are meridians of the knot complements $P_{1,1}$, $P_{1,2}$, $P_{2,1}$ and $P_{2,2}$, respectively. By Corollary 2.25, every homeomorphism $P_1 \to P_2$ sends $m_{1,1} \cup m_{1,2}$ to $m_{2,1} \cup m_{2,2}$ up to isotopy. Moreover, it is not hard to see that every homeomorphism $T_1 \to T_2$ with this property is induced by $\mathrm{HExt}_{p_1,p_2}(f)$ for some $f \in \boldsymbol{H}((N_1, p_1); (N_2, p_2))$; see Figure 17 for an explanation. Therefore, the trace of every homeomorphism $P_1 \to P_2$ extends to a homeomorphism $Q_1 \to Q_2$ which is the image under HExt_{p_1,p_2} of some homeomorphism $(N_1, p_1) \to (N_2, p_2)$. Thanks to Proposition 4.4, this is enough to conclude that S_1 and S_2 are isotopic.

4.6 Small I-bundles in the JSJ decomposition

Finally, we deal with the case where S_1 is not partially compressible on one side; in other words, we assume that each component of $S^3 \setminus\!\!\setminus S_1$ is either boundary irreducible or a handlebody. Since S_1 is

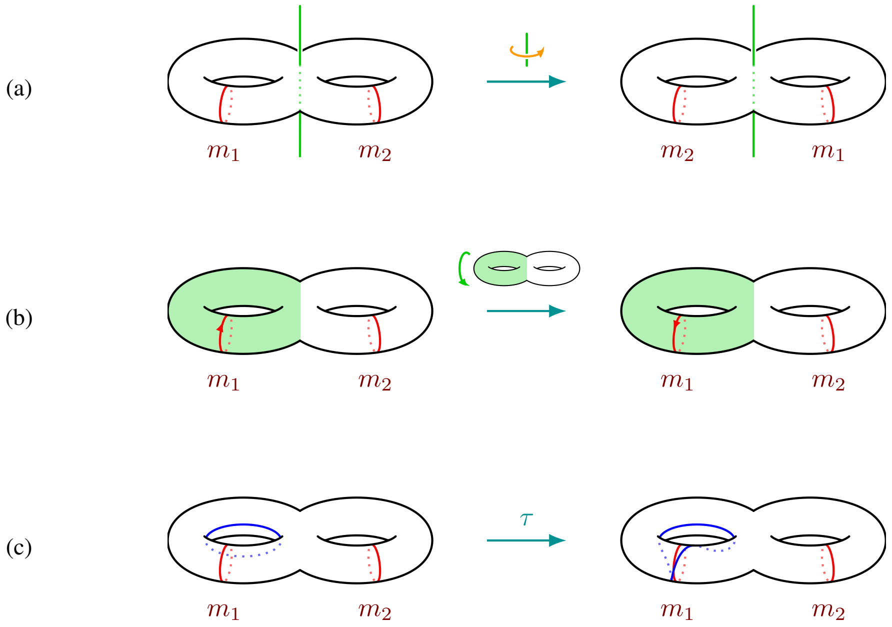


Figure 17: Three self-homeomorphisms of a handlebody N endowed with a separating meridian curve p as a boundary pattern (the boundary of the green region in (b)): (a) swaps the two meridian curves m_1 and m_2; (b) inverts the orientation of m_1 and leaves m_2 unchanged; (c) is a Dehn twist about the meridian disc bounded by m_1. If Q denotes the 3-manifold obtained by attaching a 2-handle to N along p and $T = \partial Q$, then the images of (a), (b), and (c) under $(-)|_\partial \circ \mathrm{HExt}_{p,p}$ generate the subgroup of $\boldsymbol{H}(T)$ of homeomorphisms preserving the isotopy class of $m_1 \cup m_2$.

compressible in S^3, at least one of the sides — say N_1 — must be a handlebody. If M_1 is a handlebody too, then $S^3 = M_1 \cup N_1$ is a Heegaard splitting of S^3. By a theorem of Waldhausen (see [26]), all genus-two Heegaard surfaces of S^3 are isotopic, and hence S_1 and S_2 are isotopic if and only if M_2 and N_2 are handlebodies.

We can therefore assume that M_1 is boundary irreducible. If M_2 is not homeomorphic to M_1 or N_2 is not a handlebody, then clearly S_1 and S_2 are not isotopic. Otherwise, M_2 is boundary irreducible too. Let F be the JSJ system of M_2. The boundary curves of the annuli components of F split the surface S_2 into subsurfaces $R_1, \dots, R_r$ (namely, the components of $S_2 \setminus\!\setminus \partial F$) such that $\chi(R_1) + \cdots + \chi(R_r) = -2$. A simple analysis shows that there are only a handful of possible surface types for each component R_i:

- the annulus $\Sigma_{0,2}$, with $\chi(\Sigma_{0,2}) = 0$;
- the pair of pants $\Sigma_{0,3}$, with $\chi(\Sigma_{0,3}) = -1$;
- the sphere with four punctures $\Sigma_{0,4}$, with $\chi(\Sigma_{0,4}) = -2$;
- the punctured torus $\Sigma_{1,1}$, with $\chi(\Sigma_{1,1}) = -1$;
- the torus with two punctures $\Sigma_{1,2}$, with $\chi(\Sigma_{1,2}) = -2$;
- the closed surface of genus two Σ_2, with $\chi(\Sigma_2) = -2$.

We are interested in the types of I-bundles which can occur in the JSJ decomposition of M_2.

(1) If one of the components is a product bundle $\Sigma_{1,1} \times I$, then $S_2 \setminus\!\!\setminus \partial F$ is the union of two surfaces homeomorphic to $\Sigma_{1,1}$ (that is, the horizontal boundary of the I-bundle) and a positive number of annuli.

(2) If one of the components is a twisted bundle $U_{2,1} \tilde{\times} I$, then $S_2 \setminus\!\!\setminus \partial F$ is the union of a surface homeomorphic to $\Sigma_{1,2}$ (that is, the horizontal boundary of the I-bundle) and a positive number of annuli.

(3) Otherwise, each I-bundle component of the JSJ decomposition of M_2 is either a product I-bundle over $\Sigma_{0,2}$ or $\Sigma_{0,3}$, or a twisted I-bundle over $U_{1,1}$ or $U_{1,2}$.

We will now provide a solution to the isotopy problem for the last case, deferring the analysis of the first two to Sections 4.7 and 4.8.

Let $f_0 \colon M_1 \to M_2$ be a homeomorphism. By Theorem 2.23, we can algorithmically compute

- a finite collection $\mathcal{F}$ of self-homeomorphisms of ∂M_2, and
- a finite collection $a_1, \dots, a_m, b_1, \dots, b_m$ of pairwise disjoint curves in ∂M_2,

such that

$$\boldsymbol{H}(M_1; M_2)|_\partial = \bigcup_{f \in \mathcal{F}} \langle \tau_{a_1} \tau_{b_1}^{-1}, \dots, \tau_{a_m} \tau_{b_m}^{-1} \rangle f f_0|_\partial.$$

Let $g_0 \colon N_1 \to N_2$ be a homeomorphism. We have that S_1 and S_2 are isotopic if and only if some element of $\boldsymbol{H}(M_1; M_2)|_\partial g_0|_\partial^{-1}$, seen as a self-homeomorphism of ∂N_2, extends to a homeomorphism $N_2 \to N_2$. For each $f \in \mathcal{F}$, thanks to Corollary 3.5, we can algorithmically decide whether there is an element of

$$\langle \tau_{a_1} \tau_{b_1}^{-1}, \dots, \tau_{a_m} \tau_{b_m}^{-1} \rangle f (f_0 g_0^{-1})|_\partial,$$

which extends to a self-homeomorphism of N_2. Since $\mathcal{F}$ is finite, this is enough to solve the isotopy problem for S_1 and S_2.

4.7 Product I-bundle over punctured torus

Only two very special cases are left — namely, those where the JSJ decomposition of M_2 contains a "large" I-bundle piece. In this section, we address the case where a component Z of $M_2 \setminus\!\!\setminus F$ is homeomorphic to $\Sigma_{1,1} \times I$, where the horizontal boundary $\partial_h Z$ consists of — say — the surfaces R_1 and R_2; see Figure 18 for an example. Due to the Euler characteristic constraint, the complement $\operatorname{clos}(S_2 \setminus \partial_h Z)$ is an annulus. By gluing this annulus to the vertical boundary of Z, we get a torus $T_2 = (S_2 \setminus \partial_h Z) \cup \partial_v Z$ embedded in S^3. This torus separates S^3 into two components: let P_2 be the one lying inside M_2, and $Q_2 = N_2 \cup Z$ be the other one. If P_2 were a solid torus, then $\partial R_1 \subseteq \partial P_2$ would be a longitude of P_2, since it bounds a surface in the complement of P_2. But then the meridian disc of P_2 would be a boundary compression disc for $\partial_v Z$ in M_2, contradicting the fact that $\partial_v Z$ is an annulus in a JSJ system and, hence, not boundary

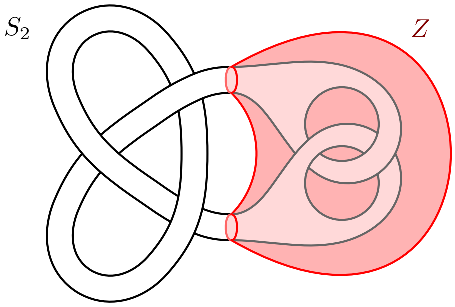


Figure 18: The surface S_2 splits S^3 into two components: one is a handlebody, and the other has a JSJ piece Z which is an I-bundle over a punctured torus.

parallel. Therefore, Q_2 must be a solid torus, and P_2 is the complement of a nontrivial knot. The situation is described in Figure 19.

Let K_2 be a section of the projection $Z \to \Sigma_{1,1}$ lying in $Z \setminus \partial_h Z$ (in other words, a surface of the form $\Sigma_{1,1} \times \{\frac{1}{2}\} \subseteq \Sigma_{1,1} \times I$). Note that K_2 is properly embedded in Q_2, and $Q_2 \setminus\!\setminus K_2$ is the handlebody N_2.

Lemma 4.11 *Let K be a punctured torus properly embedded in a solid torus $\mathbb{T}$ such that $\mathbb{T} \setminus\!\setminus K$ is a handlebody. Then*

1. *the curve $\partial K \subseteq \partial \mathbb{T}$ is a meridian*;
2. *every punctured torus K' properly embedded in $\mathbb{T}$ such that $\mathbb{T} \setminus\!\setminus K'$ is a handlebody is isotopic to K.*

Proof Statement (1) follows immediately from the fact that the meridian in $\partial \mathbb{T}$ is the only nontrivial curve which is trivial in $H_1(\mathbb{T})$. As far as statement (2) is concerned, there is a "standard" embedding of a punctured torus in $\mathbb{T}$, as depicted in Figure 20. We aim to show that K is isotopic to this standard punctured torus. Since K is not π_1-injective, it is compressible in $\mathbb{T}$. Let $D \subseteq \mathbb{T}$ be a compression disc

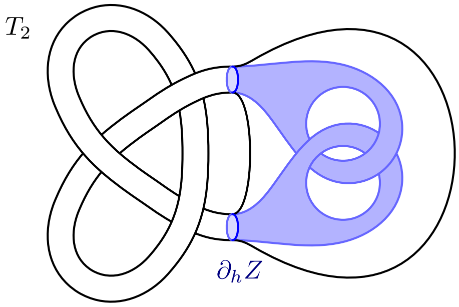


Figure 19: The torus T_2 splits S^3 into two components. One of them, namely P_2 (the "outside" in this picture), is a nontrivial knot complement; the other, namely Q_2, is a solid torus. The surface $\partial_h Z$ is a union of two punctured tori, it lies inside Q_2, and splits it into two components: the handlebody N_2 and the product bundle Z.

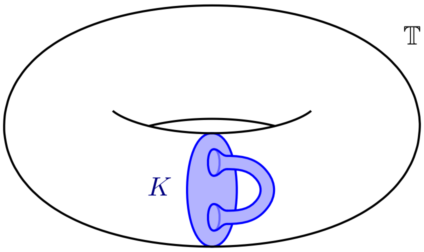


Figure 20: A “standard” punctured torus K embedded in a solid torus $\mathbb{T}$.

for K. We now analyse two cases, depending on whether ∂D separates K or not (although, a posteriori, we could always pick ∂D to be nonseparating).

- If ∂D is nonseparating in K, then compressing K along D yields a meridian disc $E \subseteq \mathbb{T}$. The punctured torus K can be recovered by removing $\partial_v \mathcal{N}(a)$ and adding $\partial_h \mathcal{N}(a)$ to E, where a is a suitable arc in $\mathbb{T}$ such that $a \cap E = \partial a$. Note that $\mathbb{T} \setminus\setminus E$ is a 3-ball, and $P = (\mathbb{T} \setminus\setminus E) \setminus \mathring{\mathcal{N}}(a)$ is a knot complement. Moreover, if we attach a 1-handle to P, we obtain a 3-manifold which is homeomorphic to the handlebody $\mathbb{T} \setminus\setminus K$. Looking at fundamental groups, this implies that $\pi_1(P) * \mathbb{Z} = \mathbb{Z} * \mathbb{Z}$, and hence the knot complement P must in fact be a solid torus. This readily implies that the arc a must be trivial, by which we mean that it must cobound a disc D' with an arc in E, such that $\operatorname{int}(D') \cap E = \varnothing$. There is only one meridian disc E up to isotopy, and there are two trivial arcs, one on each side of E. But K is completely determined by E and a, and it is easy to see that both choices of a yield standard punctured tori in $\mathbb{T}$.
- If ∂D is separating in K, then compressing K along D yields the union of a meridian disc E and a torus T. The punctured torus K can be recovered by removing $\partial_v \mathcal{N}(a)$ and adding $\partial_h \mathcal{N}(a)$ to $E \cup T$, where a is a suitable arc in $\mathbb{T}$ which has one endpoint in E, one in T, and is otherwise disjoint from $E \cup T$. Let Q be the closure of the component of $\mathbb{T} \setminus T$ which is disjoint from E. If Q is a solid torus, then K admits a compression disc whose boundary does not separate it; by our analysis of the first case, it follows that K is the standard punctured torus. Otherwise, Q is a nontrivial knot complement, and $P = (\mathbb{T} \setminus\setminus E) \setminus \operatorname{int}(Q)$ is a punctured solid torus. Note that if we join the knot complement Q and the solid torus $P \setminus \mathring{\mathcal{N}}(a)$ with a 1-handle, we obtain a 3-manifold which is homeomorphic to the handlebody $\mathbb{T} \setminus\setminus K$. Looking at fundamental groups, this implies that $\pi_1(Q) * \mathbb{Z} = \mathbb{Z} * \mathbb{Z}$, contradicting the fact that Q is a nontrivial knot complement. □

We know that M_1 is homeomorphic to M_2; in particular, it has the same JSJ decomposition. Therefore, we can define T_1, P_1, Q_1 and K_1 like we did with T_2, P_2, Q_2 and K_2. These definitions are all canonical up to isotopy (there is exactly one component in the JSJ decomposition of M_2 which is an I-bundle $\Sigma_{1,1} \times I$). Consequently, every homeomorphism $S^3 \to S^3$ sending S_1 to S_2 orientation-preservingly can be isotoped so that it sends P_1 to P_2, Q_1 to Q_2, and K_1 to K_2. Clearly, if S_1 and S_2 are isotopic then the 3-manifolds with boundary pattern $(P_1, \partial K_1)$ and $(P_2, \partial K_2)$ are homeomorphic; we claim that the converse is also true. In fact, let $f \colon (P_1, \partial K_1) \to (P_2, \partial K_2)$ be a homeomorphism. Since ∂K_1 and ∂K_2 are meridians of the solid tori Q_1 and Q_2, respectively, f extends to a homeomorphism $f\colon S^3 \to S^3$ sending Q_1 to Q_2. By Lemma 4.11, we can isotope f so that it sends K_1 to K_2. Since $N_1 = Q_1 \setminus \mathring{\mathcal{N}}(K_1)$ and $N_2 = Q_2 \setminus \mathring{\mathcal{N}}(K_2)$, the homeomorphism f sends N_1 to N_2 and, hence, S_1 to S_2 orientation-preservingly.

The knot complements P_1 and P_2 are irreducible and boundary irreducible; therefore we can algorithmically decide whether $(P_1, \partial K_1)$ and $(P_2, \partial K_2)$ are homeomorphic or not. As we have shown, this provides a solution to the isotopy problem for S_1 and S_2.

4.8 Twisted I-bundle over punctured Klein bottle

Finally, suppose there is a component Z of $M_2 \mathbin{\backslash\backslash} F$ homeomorphic to $U_{2,1} \mathbin{\tilde{\times}} I$, where the horizontal boundary $\partial_h Z$ is a torus with two punctures K_2 embedded in S_2. Due to the Euler characteristic constraint, the complement $\operatorname{clos}(S_2 \setminus K_2)$ is an annulus. We know that M_1 is homeomorphic to M_2; in particular, it has the same JSJ decomposition. Therefore, we can define K_1 like we did with K_2. There are two cases.

When K_2 is incompressible in N_2 By gluing the annulus $\operatorname{clos}(S_2 \setminus K_2)$ to the vertical boundary of Z, we get a torus $T_2 = (S_2 \setminus K_2) \cup \partial_v Z$ embedded in S^3. This torus separates S^3 into two components: let P_2 be the one lying inside M_2, and $Q_2 = N_2 \cup Z$ be the other one. Note that K_2 is properly embedded in Q_2, and moreover it is incompressible in Q_2, since it is incompressible on both sides (that is, in N_2 and in Z). This implies that Q_2 cannot be a solid torus, because $\pi_1(K_2) = \mathbb{Z} * \mathbb{Z} * \mathbb{Z}$ does not embed in $\mathbb{Z}$. Therefore, Q_2 is a nontrivial knot complement and P_2 is a solid torus.

If K_1 is compressible in N_1, then clearly S_1 and S_2 are not isotopic. Otherwise, let us define T_1, P_1 and Q_1 like we did with T_2, P_2 and Q_2. These definitions are all canonical up to isotopy (there is exactly one component in the JSJ decomposition of M_2 which is an I-bundle $U_{2,1} \mathbin{\tilde{\times}} I$). It is then easy to see that S_1 and S_2 are isotopic if and only if there is a homeomorphism $S^3 \to S^3$ sending Q_1 to Q_2 and K_1 to K_2. If Q_1 is not homeomorphic to Q_2 then once again S_1 and S_2 are not isotopic. Otherwise, by Corollary 2.25, we can algorithmically produce a list $\mathcal{F}$ of representatives of isotopy classes of homeomorphisms $Q_1 \to Q_2$. Note that every $f \in \mathcal{F}$ extends to a homeomorphism $S^3 \to S^3$ by Corollary 2.25. If for some $f \in \mathcal{F}$ the surfaces $f(K_1)$ and K_2 are isotopic in Q_2 (which we can check, since the two surfaces are incompressible), then S_1 and S_2 are isotopic. Otherwise, they are not.

When K_2 is compressible in N_2 Let $D \subseteq N_2$ be a compression disc for K_2. Let $N_2' = N_2 \mathbin{\backslash\backslash} D$, and let $K_2' = \partial N_2' \setminus (\partial N_2 \setminus K_2)$ be the result of compressing K_2 along D. If D is separating in the handlebody N_2, then one of the two components of N_2' is a solid torus whose boundary is fully contained in K_2'. Therefore, up to replacing D with the meridian disc of this solid torus, we can assume that D is nonseparating.

- Let us first assume that K_2' is incompressible in N_2'. Lemma 4.9 implies that D is the unique nonseparating compression disc for K_2 in N_2, up to isotopy in N_2; let $D_2 = D$. By repeating the above procedure for K_1 in N_1, we can check whether K_1 admits a unique nonseparating compression disc in N_1. If this is not the case, then clearly S_1 and S_2 are not isotopic. Otherwise, let $D_1 \subseteq N_1$ be this unique nonseparating compression disc. By construction, the discs D_1 and D_2 are canonical.

- Assume now that K_2' is compressible in N_2'. Since K_2' is an annulus, this immediately implies that the annulus $\mathrm{clos}(\partial N_2 \setminus K_2)$ is compressible in N_2 or, equivalently, that the boundary components of K_2 are meridian curves of N_2. Let D_2 be a disc properly embedded in N_2 whose boundary ∂D_2 is a boundary component of K_2; obviously, this disc is unique up to isotopy. If the boundary components of K_1 do not bound discs in N_1, then clearly S_1 and S_2 are not isotopic. Otherwise, let D_1 be a disc properly embedded in N_1 whose boundary is a boundary component of K_1. As we said before, the discs D_1 and D_2 are canonical.

Either way, we have found two canonical nonseparating compression discs D_1 and D_2 for S_1 and S_2, respectively. Moreover, M_1 is boundary irreducible and the 3-manifold $N_1 \setminus\setminus D_1$ is a solid torus. By Proposition 4.5, we can algorithmically decide whether S_1 and S_2 are isotopic.

References

[1] **J W Alexander**, *On the subdivision of 3-space by a polyhedron*, Proc. Natl. Acad. Sci. USA 10 (1924) 6–8

[2] **M Aschenbrenner**, **S Friedl**, **H Wilton**, *Decision problems for 3-manifolds and their fundamental groups*, from "Interactions between low-dimensional topology and mapping class groups" (R I Baykur, J Etnyre, U Hamenstädt, editors), Geom. Topol. Monogr. 19, Geom. Topol. Publ., Coventry (2015) 201–236 MR

[3] **D R J Chillingworth**, *A finite set of generators for the homeotopy group of a non-orientable surface*, Proc. Cambridge Philos. Soc. 65 (1969) 409–430 MR

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

[5] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR

[6] **J von zur Gathen**, **M Sieveking**, *A bound on solutions of linear integer equalities and inequalities*, Proc. Amer. Math. Soc. 72 (1978) 155–158 MR

[7] **C M Gordon**, **J Luecke**, *Knots are determined by their complements*, J. Amer. Math. Soc. 2 (1989) 371–415 MR

[8] **W H Holzmann**, *An equivariant torus theorem for involutions*, Trans. Amer. Math. Soc. 326 (1991) 887–906 MR

[9] **W Jaco**, *Lectures on three-manifold topology*, CBMS Regional Conference Series in Mathematics 43, Amer. Math. Soc., Providence, RI (1980) MR

[10] **W Jaco**, **P B Shalen**, *Seifert fibered spaces in 3-manifolds*, from "Geometric topology" (J C Cantrell, editor), Academic Press, New York (1979) 91–99 MR

[11] **W Jaco**, **J L Tollefson**, *Algorithms for the complete decomposition of a closed 3-manifold*, Illinois J. Math. 39 (1995)

[12] **K Johannson**, *Homotopy equivalences of 3-manifolds with boundaries*, Lecture Notes in Mathematics 761, Springer (1979) MR

[13] **R Kirby**, *Problems in low-dimensional topology*, from "Geometric topology" (W H Kazez, editor), AMS/IP Stud. Adv. Math. 2.2, Amer. Math. Soc., Providence, RI (1997) 35–473 MR

[14] **R Kobayashi**, **G Omori**, *An infinite presentation for the mapping class group of a non-orientable surface with boundary*, Osaka J. Math. 59 (2022) 269–314 MR

[15] **T Kobayashi**, *Equivariant annulus theorem for* 3*-manifolds*, Proc. Japan Acad. Ser. A Math. Sci. 59 (1983) 403–406 MR

[16] **G Kuperberg**, *Algorithmic homeomorphism of 3-manifolds as a corollary of geometrization*, Pacific J. Math. 301 (2019) 189–241 MR

[17] **J H Lee**, **S Lee**, *Inequivalent handlebody-knots with homeomorphic complements*, Algebr. Geom. Topol. 12 (2012) 1059–1079 MR

[18] **W B R Lickorish**, *Homeomorphisms of non-orientable two-manifolds*, Proc. Cambridge Philos. Soc. 59 (1963) 307–317 MR

[19] **B Martelli**, *An introduction to geometric topology* (2016)

[20] **S Matveev**, *Algorithmic topology and classification of* 3*-manifolds*, 2nd edition, Algorithms and Computation in Mathematics 9, Springer (2007) MR

[21] **J W Morgan**, **F T-H Fong**, *Ricci flow and geometrization of* 3*-manifolds*, University Lecture Series 53, Amer. Math. Soc., Providence, RI (2010) MR

[22] **C P Rourke**, **B J Sanderson**, *Introduction to piecewise-linear topology*, Springer (1982) MR

[23] **P Scott**, **H Short**, *The homeomorphism problem for closed* 3*-manifolds*, Algebr. Geom. Topol. 14 (2014) 2431–2444 MR

[24] **W P Thurston**, *Three-dimensional manifolds, Kleinian groups and hyperbolic geometry*, Bull. Amer. Math. Soc. 6 (1982) 357–381 MR

[25] **J L Tollefson**, *Involutions of sufficiently large* 3*-manifolds*, Topology 20 (1981) 323–352 MR

[26] **F Waldhausen**, *Heegaard–Zerlegungen der* 3*-Sphäre*, Topology 7 (1968) 195–203 MR

[27] **F Waldhausen**, *On irreducible* 3*-manifolds which are sufficiently large*, Ann. of Math. 87 (1968) 56–88 MR

School of Mathematics, Trinity College Dublin
Dublin, Ireland

`baronif@tcd.ie`

Received: 17 November 2023 Revised: 8 August 2024

Algebraic & Geometric Topology **25**:8 (2025) 4787–4826
DOI: 10.2140/agt.2025.25.4787
Published: 20 November 2025

Meromorphic projective structures: signed spaces, grafting and monodromy

SPANDAN GHOSH
SUBHOJOY GUPTA

A meromorphic quadratic differential on a compact Riemann surface defines a complex projective structure away from the poles via the Schwarzian equation. In this article we first prove the analogue of Thurston's grafting theorem for the space of such structures with signings at regular singularities. This extends previous work of Gupta–Mj which only considered irregular singularities. We also define a framed monodromy map from the signed space extending work of Allegretti–Bridgeland, and we characterize the $\mathrm{PSL}_2(\mathbb{C})$-representations that arise as holonomy, generalizing results of Gupta–Mj and Faraco–Gupta. As an application of our grafting theorem, we also show that the monodromy map to the moduli space of framed representations (as introduced by Fock–Goncharov) is a local biholomorphism, proving a conjectured analogue of a result of Hejhal.

30F30, 57M50

1 Introduction

A marked and bordered surface is a pair $(\mathbb{S},\mathbb{M})$ where $\mathbb{S}$ is a compact oriented surface of genus g and k boundary components, together with a nonempty set $\mathbb{M}$ of finitely many marked points, where each boundary component has at least one marked point. We shall assume that if $g=0$ then $|\mathbb{M}|\geq 3$.

Let $\mathfrak{n}=(n_1,n_2,\dots,n_k)$ be a tuple of positive integers, each $n_i\geq 3$, such that n_i-2 is the number of marked points on the i^{th} boundary. The set of marked points in the interior of the surface, which can be considered as punctures, is denoted by $\mathbb{P}$. This paper concerns the space of (signed) meromorphic projective structures $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ which is a $2^{|\mathbb{P}|}$-fold branched cover over the space $\mathcal{P}(\mathbb{S},\mathbb{M})$ (see Section 2.3 for details).

Recall that such a meromorphic projective structure is determined by a meromorphic quadratic differential q on a compact Riemann surface X of genus g, with k poles having orders $(n_1,n_2,\dots,n_k)$ and $|\mathbb{P}|$ poles having order at most 2, via the Schwarzian equation

$$u''+\tfrac{1}{2}qu=0. \tag{1}$$

Namely, the ratio of solutions of (1) determines a *developing map* $f\colon \widetilde{X}\to\mathbb{C}\mathrm{P}^1$ that is equivariant with respect to a *holonomy/monodromy representation* $\rho\colon \pi_1(X)\to \mathrm{PSL}_2(\mathbb{C})$. Historically, these arose in the context of uniformizing a punctured sphere (see, for example, Chapters VIII and IX of [43]).

We recover $(\mathbb{S},\mathbb{M})$ as the underlying topological surface via a real blow-up of the k poles of orders at least 3; the horizontal directions of q at such a pole determine the marked points on the corresponding boundary circle. (Recall here that a pole of order n has $n-2$ horizontal directions, ie tangent directions $\pm v$ where $q(v)\in\mathbb{R}^+$.)

Our first result in this paper is a more geometric parametrization of $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ which is an analogue of Thurston's grafting theorem, which we now briefly recall. For a *closed* oriented surface S of genus $g\geq 2$, Thurston had introduced a "grafting" deformation of any hyperbolic structure on S that results in a new complex projective structure. This construction starts with a hyperbolic surface X with developing map $f\colon \widetilde{X}\to\mathbb{H}\subset\mathbb{C}\mathrm{P}^1$, and a measured geodesic lamination λ on X. The new complex projective structure is obtained by equivariantly inserting "lunes" along the images of the leaves of the lift of λ to the universal cover. Here a lune is a region of $\mathbb{C}\mathrm{P}^1$ bounded by two circular arcs, and its "angle" is determined by the transverse measure on λ; for details of this operation see [16; 46]. Thurston's grafting theorem (see [7; 33]) asserts that the space $\mathcal{P}_g$ of complex projective structures on S is *parametrized* by such deformations, ie the *grafting map*

$$\mathrm{Gr}\colon \mathcal{T}_g\times\mathcal{ML}\to\mathcal{P}_g \tag{2}$$

is a homeomorphism. (Here $\mathcal{T}_g$ is the *Teichmüller space* of marked hyperbolic structures and $\mathcal{ML}$ is the space of measured laminations on S.)

In our context, the analogues of the spaces in the left-hand side of (2) are enhanced Teichmüller space $\mathcal{T}^{\pm}(\mathbb{S},\mathbb{M})$ and the space of signed measured laminations $\mathcal{ML}^{\pm}(\mathbb{S},\mathbb{M})$, where in both cases the signing is associated with the punctures in $\mathbb{P}$ (see Section 2 for details). The former space already appears in previous literature in the context of marked and bordered surfaces (see, for example, [2]). The latter is a space that we introduce, and should be related to the spaces of real $\mathcal{A}$- and $\mathcal{X}$-laminations that Fock–Goncharov introduced in [19, Section 12] (see also [20]).

We shall prove:

Theorem 1.1 *There is a grafting map*

$$\widehat{\mathrm{Gr}}\colon \mathcal{T}^{\pm}(\mathbb{S},\mathbb{M})\times\mathcal{ML}^{\pm}(\mathbb{S},\mathbb{M})\to\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M}),$$

which is a homeomorphism.

In the case when the set $\mathbb{P}=\varnothing$, this was proved in [27]; the key technical step in the proof of Theorem 1.1 is to determine how the signings at the points of $\mathbb{P}$ play a role. In particular, we shall crucially use the relation between the signed grafting map and the Schwarzian derivative at these poles (see Lemma 3.7).

Along the way, we shall provide a "grafting description" of the projective structures corresponding to quadratic differentials with *simple* poles — see Corollary 3.4 — which could be of independent interest.

Next, we introduce a monodromy map

$$\widehat{\Phi}\colon \mathcal{P}^{\pm}(\mathbb{S},\mathbb{M}) \to \widehat{\chi}(\mathbb{S},\mathbb{M}), \tag{3}$$

where the target is the space of *framed* $\mathrm{PSL}_2(\mathbb{C})$-representations of the fundamental group $\pi_1(\mathbb{S}\setminus\mathbb{P})$ (see [25, Section 5] for more on this space). We briefly recall here that a framed representation is a pair of a $\mathrm{PSL}_2(\mathbb{C})$-representation ρ together with a framing $\beta\colon F_\infty \to \mathbb{C}\mathrm{P}^1$ which is a ρ-equivariant map defined on the lift of $\mathbb{M}$ to the ideal boundary of the universal cover (this set of ideal boundary points is the *Farey set* F_∞, following [19, Section 1.3]). Such a (framed) monodromy map had been previously defined by Allegretti–Bridgeland [3] for a subspace $\mathcal{P}^*(\mathbb{S},\mathbb{M})$ corresponding to meromorphic projective structures with *no apparent singularities*. At a regular singularity, they had defined the framing using the signing to choose a fixed point of the monodromy around it, and at an irregular singularity, they had defined the framing in terms of the asymptotics of the solutions of the Schwarzian equation in Stokes sectors. The latter description was shown to be equivalent to considering the asymptotic values of the developing map in [27] (see Section 4.1 of that paper). We extend this here to the case of regular singularities, to provide a more geometric definition of the framing in terms of the asymptotic behavior of the developing map, that also applies at apparent singularities.

Our next theorem then characterizes the image of this framed monodromy map. Here, $\widehat{\chi}^*(\mathbb{S},\mathbb{M})$ denotes the subset of nondegenerate framed representations (see Definition 4.2) that had been introduced in [3].

Theorem 1.2 *The image of* $\widehat{\Phi}$ *in* (3) *is precisely the space* $\widehat{\chi}^*(\mathbb{S},\mathbb{M})$ *of nondegenerate framed representations.*

Allegretti–Bridgeland had shown that the image of $\mathcal{P}^*(\mathbb{S},\mathbb{M})$ is *contained* in $\widehat{\chi}^*(\mathbb{S},\mathbb{M})$ (see Theorem 6.1 of their paper). We use completely different techniques, applying our (grafting) Theorem 1.1, to prove this inclusion for the monodromy map $\widehat{\Phi}$. In particular, this provides an alternative proof of [3, Theorem 6.1]. The opposite inclusion, needed to complete the proof of Theorem 1.2 is already known as it follows from the constructions in [18; 27] together with [3, Theorem 9.1] which implies that nondegenerate framed representations have Fock–Goncharov coordinates with respect to some ideal triangulation.

Once again, Theorem 1.2 had been proved in the case that $\mathbb{P}=\varnothing$ in [26], and in the "opposite" case when $\mathbb{M}=\mathbb{P}$ in [25] under the additional assumption of no apparent singularities. See also the recent paper of Nascimento [41] which discusses the case of projective structures with Fuchsian-type singularities, and Le Fils [39] which handles the case of projective structures with branch points. For a closed surface, the image of the monodromy map was characterized by Gallo–Kapovich–Marden in [23]; the case of a punctured surface was left as an open question in that paper.

As an immediate corollary we obtain:

Corollary 1.3 *A representation* $\rho\colon \pi_1(\mathbb{S}\setminus\mathbb{P}) \to \mathrm{PSL}_2(\mathbb{C})$ *arises as the monodromy of a meromorphic projective structure in* $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ (*forgetting the framing*) *if and only if there exists a framing* β *such that the pair* (ρ,β) *is nondegenerate.*

In Section 4 we provide an alternative characterization of the monodromy representations that appear, without involving framings — see Corollary 4.6. In the case that $\mathbb{S}$ has no boundary components (ie $\mathbb{M} = \mathbb{P}$), this coincides with the representations described in Theorem A of [18]. This paper thus provides an alternative proof of that result, sans the construction of affine structures that is discussed in detail in [18].

Finally, we also prove:

Theorem 1.4 *The monodromy map* $\widehat{\Phi}$ *in* (3) *is a local homeomorphism.*

For a closed surface S, the fact that the monodromy map from the space of projective structures to the $\mathrm{PSL}_2(\mathbb{C})$-representation variety

$$\Phi\colon \mathcal{P}_g \to \chi(S) \tag{4}$$

is a local homeomorphism was a classical result of Hejhal in [30]. This can be considered a special case of the Ehresmann–Thurston principle concerning the holonomy map from the deformation space of geometric structures on a closed manifold to the corresponding representation variety (see, for example, [24]). This local homeomorphism result was also proved in the case of projective structures with regular singularities having loxodromic monodromy by Luo [40] (see the discussion in Section 2.3 of [25]). Recently, it was also proved for projective structures with "cusps" (that allow apparent singularities) in [8, Theorem 5.6], and those without apparent singularities and with fixed residues at the poles in [44]. The work in [26] had proved the above theorem in the case that $\mathbb{P} = \varnothing$ (ie only irregular singularities); our proof here follows their strategy and reduces to an application of a relative version of the Ehresmann–Thurston principle.

Using [3], we can conclude:

Corollary 1.5 *The monodromy map* $\widehat{\Phi}$ *is a local biholomorphism.*

Theorem 1.4 verifies a conjecture by Allegretti in [1], and Corollary 1.5 answers a question raised in [3].

In the context of meromorphic projective structures, it remains to *"explore the nonuniqueness of projective structures with given monodromy"*, quoting from [23, Problem 12.2.1]. For a closed surface, the fibers of the monodromy map (4) are discrete by Hejhal's result, and have been studied by Baba in [5; 6]; it is conceivable that some of the techniques developed there can be generalized to the case of the framed monodromy map $\widehat{\Phi}$.

Recall that the Schwarzian equation (1) is a first-order linear ODE on the Riemann surface. In a broader context, linear differential equations of *higher* order $n \geq 2$ can be considered (see, for example, [31]). These determine what are called $\mathrm{SL}_n(\mathbb{C})$-opers in the literature (see [21, Section 15.6]); indeed, projective structures form the special case of $\mathrm{SL}_2(\mathbb{C})$-opers. Another direction to pursue would be to characterize the surface-group representations into $\mathrm{PSL}_n(\mathbb{C})$ that arise as the monodromy of SL_n-opers. In contrast with the $n = 2$ case, it follows from a dimension count that for $n > 2$ such representations form a subset of positive codimension in the $\mathrm{PSL}_n(\mathbb{C})$-character variety, so the analogue of Theorem 1.4 does not hold. The monodromy map has been studied recently by Alley in [4], in a special case of meromorphic cyclic $\mathrm{SL}_n(\mathbb{C})$-opers on $\mathbb{C}\mathrm{P}^1$.

The case of *first-order* linear ODEs on a Riemann surface is also interesting, as they relate to the theory of translation structures on surfaces associated with abelian differentials. In that context, the monodromy is determined by the periods of the differential, and the analogue of Theorem 1.2 for closed surfaces was classical work of Haupt in [29] that was reproved by Kapovich in [34], and extended recently by Le Fils [38] and Bainbridge–Johnson–Judge–Park [9]. For punctured surfaces and meromorphic differentials, an analogue of Theorem 1.2 was proved recently by Chenakkod–Faraco–Gupta [14], and was extended by Chen–Faraco [13].

Acknowledgements

Ghosh carried out most of this work while an undergraduate student at the Indian Institute of Science, and is grateful to Kishore Vaigyanik Protsahan Yojana (KVPY) for their fellowship and contingency grant. Gupta is grateful to the Department of Science and Technology, Govt of India for its support via grant no CRG/2022/ 001822. He also thanks Gianluca Faraco and Mahan Mj for their interest and their collaboration in previous papers that led up to this. Both authors would like to thank Dylan Allegretti and Lorenzo Ruffoni for helpful correspondences, and the anonymous referee for helpful comments. This work is also supported by the DST FIST program - 2021 [TPN - 700661].

2 Signed parameter spaces

As mentioned in the Introduction, let $(\mathbb{S},\mathbb{M})$ be a marked and bordered surface of genus g and k boundary components, and let $\mathfrak{n}=\{n_1,n_2,\dots,n_k\}$ with $n_i\geq 3$ be the associated integer-tuple. Here recall that n_i-2 is the number of marked points on the i^{th} boundary component, for each $1\leq i\leq k$. We shall also denote by $m=|\mathbb{P}|$ the number of marked points in the interior of $\mathbb{S}$. We define

$$\chi = 6g-6+\sum_{i=1}^{k}(n_i+1)+3m \tag{5}$$

and we shall assume this is positive, throughout this article. This equivalent to requiring that if $g=0$, then $|\mathbb{M}|\geq 3$.

2.1 The enhanced Teichmüller space

We shall define the space of hyperbolic structures on the marked and bordered surface $(\mathbb{S},\mathbb{M})$. Such a hyperbolic structure X will have either a cusp or a geodesic boundary component at each interior puncture $\mathbb{P}\subset\mathbb{M}$, and each boundary component of $\mathbb{S}$ with marked points will be a "crown end" (see [27, Section 3.2]) where each boundary arc between marked points is a bi-infinite geodesic (a *side* of the crown).

Definition 2.1 A marking of a hyperbolic surface X as above by the surface $(\mathbb{S}, \mathbb{M})$ is a homeomorphism $f\colon \mathbb{S} \setminus \mathbb{M} \to X^\circ$, where X° is the surface obtained from X by removing the boundary components homeomorphic to S^1. Such a marking must take the set of interior punctures $\mathbb{P}$ to geodesic ends or cusps, and boundary arcs between marked points to bi-infinite geodesics of the crown ends. Two such markings (f_1, X_1) and (f_2, X_2) are defined to be equivalent if there is an isometry $g\colon X_1 \to X_2$ such that $g \circ f_1$ is isotopic to f_2 relative to the marked points $\mathbb{M}$. Then, the set of such markings (f, X) under this equivalence relation forms the Teichmüller space $\mathcal{T}(\mathbb{S}, \mathbb{M})$.

Now, we want to provide an additional signing parameter to our hyperbolic surfaces, namely an orientation to the boundary components homeomorphic to S^1. For each component, this parameter is a choice of an element in $\{\pm 1\}$, depending on whether it agrees with the orientation induced from that on the surface. Recall that there are at most m such boundary components. Thus, we get:

Definition 2.2 (Definition 3.4 of [2]) The enhanced Teichmüller space $\mathcal{T}^{\pm}(\mathbb{S}, \mathbb{M})$ is a 2^m-fold branched cover of $\mathcal{T}(\mathbb{S}, \mathbb{M})$ (branched on the set of surfaces with at least one cusp end), with branching data given by a choice of a sign at each geodesic end.

Parametrizing this space with shear coordinates, we have the following:

Theorem 2.3 (Proposition 3.5 of [2]) *For a pair* $(\mathbb{S}, \mathbb{M})$ *with* $|\mathbb{M}| \geq 3$ *if* $g(\mathbb{S}) = 0$, *we have that* $\mathcal{T}^{\pm}(\mathbb{S}, \mathbb{M})$ *is homeomorphic to* $\mathbb{R}^{\chi}$.

Here the expression for χ is given in (5).

We denote an element of this set by a tuple (X, f, σ), where X is the hyperbolic surface, $f\colon \mathbb{S} \setminus \mathbb{M} \to X^\circ$ is the marking, and σ is the signing of each geodesic boundary component of X.

2.2 The enhanced space of measured laminations

Let (X, f, σ) be an element of $\mathcal{T}^{\pm}(\mathbb{S}, \mathbb{M})$ as above. Now, we define the space of measured geodesic laminations on this element. First, we define a geodesic lamination on the surface:

Definition 2.4 A geodesic lamination on X is a closed subset L in X that is a disjoint union of simple closed geodesics or a bi-infinite geodesics on X, including the boundary geodesics of X.

Since we always include the boundary geodesics, it will be helpful to consider L as the disjoint union of two subsets, $L^\circ \in X^\circ$ and ∂X. Since the geodesics in a lamination cannot meet the boundary geodesics, it follows that the leaves of the lamination must satisfy the following at the neighborhood of an end of X:

- If the end is a cusp, the leaves of the geodesic exiting the end must go inside the cusp, without any spiraling. This can be seen by lifting to the universal cover $\mathbb{H}^2$; if we assume that the cusp end C lifts to a horodisk $H = \{z \mid \operatorname{Im}(z) > h\}$ in the upper half-plane model of $\mathbb{H}^2$, such that $C = H/\langle z \mapsto z+1 \rangle$, then any such geodesic leaf lifts to a vertical line that remains in a single lift of the fundamental domain.

- If the end is a geodesic end, then each leaf entering a neighborhood of the end must spiral in a direction around the geodesic and accumulate on it. Since the leaves are disjoint, it follows that (if there is more than one leaf) they must all spiral in the same direction.
- If the end is a crown end, similar to a cusp, each leaf entering the end must go into one of the cusps of the crown.

We recall another useful lemma regarding geodesic laminations:

Lemma 2.5 *Given a geodesic lamination L on a hyperbolic surface X, there is a $\pi_1(X)$-invariant ideal triangulation of the universal cover $\widetilde{X}$ such that no leaf of the lamination intersects the interior of a triangle.*

Proof The completion of the complement of the geodesic lamination L comprises finitely many connected hyperbolic surfaces that have crowns, cusps or geodesic boundary (see, for example, [11, page 7]). One can equip each of these components with an ideal triangulation, where the (finitely many) edges of the triangulation are geodesic lines between cusps or crown boundary cusps or spiraling onto geodesic boundary components. The lift of these geodesic lines, together with the lift of L, defines a $\pi_1(X)$-invariant ideal triangulation of the universal cover. □

Lemma 2.6 *Given a geodesic lamination L, for each end of X there is a neighborhood of that end such that only finitely many leaves of the lamination enter it. For a crown end, we can take the neighborhood to be the whole crown.*

Proof Consider a triangulation of the universal cover of X as before. Then, the triangulation descends to the surface itself, due to $\pi_1(X)$-equivariance. From the Gauss–Bonnet formula, the number of triangles on the surface is exactly $-2\chi(X)$. Now, each triangle can contribute at most 3 geodesics going into the cusps or crown cusps of X. Since each geodesic going into such an end is adjacent to two triangles, it follows that the total number of geodesics going into cusps and crown cusps is at most $-2\chi(X) \times 3 = -6\chi(X)$. This proves the first statement. For the second, any leaf of the lamination intersecting a crown must be a geodesic going to one or two of its boundary cusps (ie crown tips), so the statement follows. □

We briefly recall the notion of a measured lamination, which is a geodesic lamination L as above, together with a transverse measure on it (for details see, for example, [42, Section 1.8]):

Definition 2.7 Given a geodesic lamination L, let $\Lambda(L)$ denote the collection of all compact 1-manifolds embedded in X which are transverse to L and such that their boundary (if it exists) lies in $X \setminus L$. Then, a measure on L refers to a function $\mu\colon \Lambda(L) \to \mathbb{R}_{\geq 0}$ such that it is transverse to L (ie if α and β are 1-manifolds that are homotopic via 1-manifolds with boundary in $X \setminus L$, then $\mu(\alpha) = \mu(\beta)$), σ-additive (ie if $\alpha = \bigcup_{i\in\mathbb{N}} \alpha_i$ with $\alpha_i \cap \alpha_j = \partial\alpha_i \cap \partial\alpha_j$, then $\mu(\alpha) = \sum_i \mu(\alpha_i)$), and its support is L°. This pair (L, μ) defines a measured lamination on X.

Given a measured lamination, every isolated curve in L obtains a weight. We can define a topology on the set of measured laminations as follows: First, given a lamination, we can lift the lamination to the universal cover to get a $\pi_1(X)$-invariant set of geodesics in $\mathbb{H}^2$. Thus, it gives us a subset of the space $M_\infty = (\partial\mathbb{H}^2 \times \partial\mathbb{H}^2 \setminus \Delta)/\sim$, where $\sim$ is the equivalence relation $(x, y) \sim (y, x)$. Then, a measure on the lamination gives us a Borel measure on the space M_∞. We define the topology on the set of measured laminations to be the topology it inherits from the weak-$*$ topology of measures on M_∞. Thus, we have:

Definition 2.8 The space of measured laminations on X, denoted by $\mathcal{ML}(X)$, is the set of all pairs (L, μ) defined above, endowed with the topology it inherits from being a subset of the space of measures on M_∞, endowed with the weak-$*$ topology.

Now, we can parametrize the space $\mathcal{ML}(X)$ via the use of Dehn–Thurston coordinates as described in the unpublished notes of Dylan Thurston [47], to get the following (see Proposition 1.5 of [28]):

Proposition 2.9 *The space $\mathcal{ML}(X)$ is homeomorphic to $\mathbb{R}^{\chi-p} \times \mathbb{R}^p_{\geq 0}$, where p is the number of cusp ends of X.*

Proof Recall here that χ is given by (5). For simplicity, let us first assume that there are no crown ends on X, ie $k = 0$ or $\mathbb{M} = \mathbb{P}$. In this case $\chi = 6g - 6 + 3m$ where $m = |\mathbb{M}|$.

We construct a pants decomposition of the topological surface underlying X; an Euler characteristic count this gives us a total of $\#P = 2g + m - 2$ pairs of pants. Hence, there are $t = \frac{1}{2}(3 \cdot \#P - m)$ simple closed curves in the interior of X that gives the decomposition. To this set of curves, we also add in the geodesic boundaries and peripheral loops around the cusps of X to obtain a total of $t + m$ curves. Let these curves be given by $(P_1, P_2, \dots, P_t, P_{t+1}, \dots, P_{t+m})$. Also, let us choose a collection of dual curves D_i for each P_i. Then, we have that associated to each pants curve P_i, there is a pair of parameters for the measured lamination, namely the length or intersection parameter $i(\mu, P_i)$ and the twist parameter $\theta(\mu, P_i)$, the latter measured using D_i. It follows from the discussion in [47] that the space of measured laminations is homeomorphic to this space of intersection and twist parameters. This gives a factor of $\mathbb{R}^{2t}$.

For a pants curve in the interior of X, the length parameter can take values in $\mathbb{R}_{\geq 0}$ while the twist parameter takes values in $\mathbb{R}$. Considering the pairs of parameter values $(i, \theta) \in \mathbb{R}_{\geq 0} \times \mathbb{R}$, observe that in the case of zero length, the twist parameter does not matter, leading to the identification $(0, \theta) \sim (0, -\theta)$. Hence, the parameter space is in fact homeomorphic to $\mathbb{R}^2$ for an interior curve. For a pants curve that is a boundary geodesic, the twist parameter can only take $\pm\infty$ as values if the length parameter is not zero. If the length parameter is zero, the twist parameter is also necessarily zero. Hence, it follows that the space parametrizing pairs (i, θ) for a geodesic boundary is homeomorphic to $\mathbb{R}$. This gives a factor of $\mathbb{R}^b$, where b is the number of geodesic boundary ends of X. For a pants curve that corresponds to a cusp end, hence there is no twist parameter. Thus the corresponding space of parameters is homeomorphic to $\mathbb{R}_{\geq 0}$. This gives a factor of $\mathbb{R}^p_{\geq 0}$. Note that $p + b = m$; collecting the factors, we get a space homeomorphic to $\mathbb{R}^{2t} \times \mathbb{R}^b \times \mathbb{R}^p_{\geq 0} \equiv \mathbb{R}^{\chi-p} \times \mathbb{R}^p_{\geq 0}$, as desired. This handles the case when there are no crown ends.

The case of (only) crown ends was established in [27, Proposition 3.8], and we now follow the arguments there to extend the above proof to include crown ends. As in that proof, we can first remove the crown ends to obtain a surface with k additional boundary components. Following the argument above, the space of measured laminations on the resulting surface X' with $b+k$ boundary components and p cusps is homeomorphic to $\mathbb{R}^{2t} \times \mathbb{R}^{b+k} \times \mathbb{R}^{p}_{\geq 0}$. By [27, Proposition 3.7], the space of measured laminations on each crown is homeomorphic to $\mathbb{R}^{n_i-1}$ where n_i-2 is the number of boundary cusps of the i^{th} crown. Moreover, these measured laminations can be glued to a measured lamination on X' as long as one parameter — the transverse measure of each boundary component that is glued — matches. Thus, the space of measured laminations on X is homeomorphic to $\mathbb{R}^{2t} \times \mathbb{R}^{b+k} \times \mathbb{R}^{p}_{\geq 0} \times \prod_{i=1}^{k} \mathbb{R}^{n_i-2} \equiv \mathbb{R}^{\chi-p} \times \mathbb{R}^{p}_{\geq 0}$, as desired. □

Now, we consider an additional signing parameter for elements of the space $\mathcal{ML}(X)$: At each end of X that is a cusp, we assign a sign ± to the weight of measured lamination entering the end. By this marking, we get a branched cover

$$\mathcal{ML}^{\pm}(X) \to \mathcal{ML}(X)$$

branched over those measured laminations which have at least one cusp end with no weight of the lamination entering it.

Lemma 2.10 *The space $\mathcal{ML}^{\pm}(X)$ is homeomorphic to the cell $\mathbb{R}^{\chi}$.*

Proof This follows from the proof of Proposition 2.9, by observing that at each cusp end, the nonnegative parameter, together with a sign, can be thought of as taking values in $\mathbb{R}$. □

Now, the spaces $\mathcal{ML}^{\pm}(X)$ are canonically homeomorphic to each other as X ranges over $\mathcal{T}(\mathbb{S}, \mathbb{M})$. This motivates us to define a space parametrizing signed measured laminations on the marked surface $(\mathbb{S}, \mathbb{M})$. The space parametrizing measured laminations for the case $\mathbb{P} = \varnothing$ was introduced in Section 3.3 of [27].

Definition 2.11 The space of signed measured laminations, denoted by $\mathcal{ML}^{\pm}(\mathbb{S}, \mathbb{M})$, is a space parametrizing measured laminations on $(\mathbb{S}, \mathbb{M})$ along with a choice of signings for the total incident weights at the set of punctures $\mathbb{P} \subset \mathbb{M}$. It has a topology induced from the transverse measures on finitely many closed curves on the surface, along with the measures of arcs crossing into the cusp ends for the interior punctures and the crowns.

Then the previous lemma can be interpreted as proving:

Proposition 2.12 *The space $\mathcal{ML}^{\pm}(\mathbb{S}, \mathbb{M})$ is homeomorphic to the cell $\mathbb{R}^{\chi}$.*

2.3 The space of signed projective structures

For a surface with a marking $(\mathbb{S}, \mathbb{M})$ as before, Allegretti–Bridgeland introduced the space $\mathcal{P}(\mathbb{S}, \mathbb{M})$ in [3]. This space parametrizes meromorphic projective structures on Σ_g with k poles of orders given by the tuple $\mathfrak{n}$ and m poles of order less than or equal to 2, which are marked by the pair $(\mathbb{S}, \mathbb{M})$. We have the following additional structure on this space:

Theorem 2.13 (Proposition 8.2 of [3]) *The space $\mathcal{P}(\mathbb{S},\mathbb{M})$ has the natural structure of a complex manifold of complex dimension χ, hence of real dimension 2χ.*

Similar to the definition of the space of signed measured laminations, we shall define a signed version of the space $\mathcal{P}(\mathbb{S},\mathbb{M})$, with the signing being given at each puncture $\mathbb{P}$ by the exponent of the projective structure at the puncture (defined below). The discussion here follows that in Section 8.2 of [3], though we shall provide an alternative description in terms of a fiber product (Lemma 2.18).

2.3.1 The exponent at a regular singularity Given a meromorphic projective structure, we can define the *leading coefficient* at a regular singularity p as

$$a_p := \lim_{z \to p} q(z) \cdot z^2,$$

where $q(z)dz^2$ is the Schwarzian derivative of the projective structure in a uniformizing neighborhood of the puncture. This is independent of the choice of the uniformizing chart, hence well defined for the projective structure. In other words, in any coordinate z around the regular singularity, the quadratic differential q has the form

$$q(z) = \left(\frac{a_p}{z^2} + \cdots\right) dz^2.$$

We then define:

Definition 2.14 (exponent) The exponent at a regular singularity $p \in \mathbb{P}$ of a meromorphic projective structure is the complex number

$$r_p := \pm 2\pi i \sqrt{1 - 2a_p} \tag{6}$$

defined up to sign, where a_p is the leading coefficient at p.

Remark Our definition of exponent slightly differs from the one in [3], in that they have $1 + 4a_p$ under the square root in place of $1 - 2a_p$. This is only matter of convention, arising from the fact that their Schwarzian equation has a constant factor of -1 of the zeroth-order term while ours (in (1)) has constant $\frac{1}{2}$.

2.3.2 The signed space of projective structures First, we associate a signing to our marked projective structures, as in Section 3.5 of [3]:

Definition 2.15 A signed marked projective structure is a marked projective structure on the pair $(\mathbb{S},\mathbb{M})$ along with a signing at each of the regular singularities, ie a choice of sign for the exponent at each regular singularity having nonzero exponent.

Now, we define the space of signed marked projective structures as a cover of $\mathcal{P}(\mathbb{S},\mathbb{M})$, following Proposition 8.4 of [3]:

Theorem 2.16 *There exists a complex manifold $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ that parametrizes signed, marked projective structures on $(\mathbb{S},\mathbb{M})$, along with a finite branched covering map*

$$\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M}) \to \mathcal{P}(\mathbb{S},\mathbb{M})$$

of degree $2^{|\mathbb{P}|}$, obtained by forgetting the signing.

Proof The proof is essentially the same as that in [3]. First, we define the map

$$a\colon \mathcal{P}(\mathbb{S},\mathbb{M}) \to \mathbb{C}^{|\mathbb{P}|}$$

sending each projective structure to the collection of leading coefficients at each of its regular singularities. We have that a is a holomorphic map because of the way the complex structure on $\mathcal{P}(\mathbb{S},\mathbb{M})$ is constructed (see, for example, [3, Proposition 7.3]). Moreover, a is a submersion, by a similar argument as in the proof of Lemma 6.1 in [12] — it suffices to construct a quadratic differential whose leading coefficient is nonzero at a given regular singularity and zero at all other regular singularities, which is clear from an application of Riemann–Roch.

Due to these properties of a, we can construct a $2^{|\mathbb{P}|}$-branched cover $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ of $\mathcal{P}(\mathbb{S},\mathbb{M})$, branched over the zero loci of $\{1-2a_p\}_{p\in\mathbb{P}}$, and such that the points in a fiber of the branching represent a choice of a sign (ie an element of $\{\pm 1\}$) for the nonzero exponents at the regular singularities. □

Remark Allegretti–Bridgeland [3] do the same construction as above, the only difference being that they define the covering space over the subset of projective structures without any apparent singularities.

2.3.3 Describing the signed space by fiber products

We provide another description of the space $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ via fiber products, that will be helpful in defining the grafting map later. Recall the definition of a fiber product (for example, see Chapter 1, Section 11 of [37]):

Definition 2.17 Let $\mathcal{C}$ be a category. Suppose we are given two morphisms $f\colon A\to C$ and $g\colon B\to C$ among objects A, B, C. Then, the fiber product of A and B with respect to the given morphisms is an object $A\times_C B$ in $\mathcal{C}$, along with morphisms $g'\colon A\times_C B\to A$ and $f'\colon A\times_C B\to B$, such that $f\circ g' = g\circ f'$, and the following universal property holds: given any object D with morphisms $f''\colon D\to B$ and $g''\colon D\to A$ such that $f\circ g'' = g\circ f''$, there is a unique morphism $i\colon D\to A\times_C B$ such that $f'' = f'\circ i$ and $g'' = g'\circ i$, ie the maps f'', g'' factor through i.

Now, we have the following:

Lemma 2.18 *Let $m = |\mathbb{P}|$ and consider the map*

$$r^2\colon \mathcal{P}(\mathbb{S},\mathbb{M}) \to \mathbb{C}^m$$

that assigns to a meromorphic projective structure the set of squares of exponents (see (6)) at its regular singularities $\mathbb{P}$, and

$$\mathrm{sq}\colon \mathbb{C}^m \to \mathbb{C}^m$$

$$\begin{tikzcd}
\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M}) \arrow[r, "r"'] \arrow[d, "\pi"] & \mathbb{C}^{|\mathbb{P}|} \arrow[d, "\mathrm{sq}"'] \\
\mathcal{P}(\mathbb{S},\mathbb{M}) \arrow[r, "r^2"] & \mathbb{C}^{|\mathbb{P}|}
\end{tikzcd}$$

Figure 1: Fiber product defining the space $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$. Here, π is the forgetful projection map.

that squares each coordinate of $\mathbb{C}^m$, *ie* $\mathrm{sq}(z_1, z_2, \dots, z_m) = (z_1^2, z_2^2, \dots, z_m^2)$. *Then,* $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ *is isomorphic to the fiber product* $\mathcal{P}(\mathbb{S},\mathbb{M}) \times_{\mathbb{C}^m} \mathbb{C}^m$, *in the category of complex manifolds as well as in the category of topological spaces.*

Although this is just a restatement of the way the branched cover is defined, and is in fact implicit in the construction described in [3, Section 8.2], it is convenient as we shall use the universal property of fiber products in Section 3.3 while defining the signed grafting map.

Remark The exponent at regular singularities was defined only up to sign on the space of unsigned projective structures. The construction of the signed space allows us to uniquely define the exponent at each regular singularity of a signed projective structure. Thus, we get a well-defined holomorphic map $r : \mathcal{P}^{\pm}(\mathbb{S},\mathbb{M}) \to \mathbb{C}^m$ which sends a signed projective structure to the exponents at each of its regular singularities. This is the map on the upper side of the commuting square in Figure 1.

3 The grafting theorem

Our first result concerns a grafting parametrization for the space of signed projective structures $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$. If one ignores signings, the grafting operation was briefly described in the Introduction; for details, see the references mentioned there, and for the present context of marked and bordered hyperbolic surfaces, we refer the reader to [25; 27].

We begin by stating the following general result concerning simply connected projective surfaces, essentially due to Kulkarni–Pinkall; see [27, Theorem 2.1; 35, Theorem 10.6].

Theorem 3.1 *Let* $\widetilde{X}$ *be a simply connected projective surface that is not projectively isomorphic to* $\mathbb{C}$, *or the universal cover of* $\mathbb{CP}^1 \setminus \{0, \infty\}$. *Then there exists a unique measured lamination* L *on the hyperbolic plane* $\mathbb{H}^2$ *such that* $\widetilde{X}$ *is obtained by grafting* $\mathbb{H}^2$ *along* L. *The map associating* L *to* $\widetilde{X}$ *is equivariant, ie if* $\widetilde{X}$ *is the universal cover of a projective surface* X, *and the developing map* $\widetilde{X} \to \mathbb{CP}^1$ *is* $\rho_{\mathbb{C}}$*-equivariant for a representation* $\rho_{\mathbb{C}} : \pi_1(X) \to \mathbb{PSL}_2(\mathbb{C})$, *then* L *is invariant under the image of a naturally associated representation* $\rho_{\mathbb{R}} : \pi_1(X) \to \mathbb{PSL}_2(\mathbb{R})$. *Moreover, the image* Γ *of* $\rho_{\mathbb{R}}$ *is discrete, and the quotient* $\mathbb{H}^2 / \Gamma$ *is homeomorphic to* X. *Finally, the mapping* $\widetilde{X} \mapsto L$ *is continuous.*

We refer the reader to the sketch of the proof provided in [27]. However, it will help to keep in mind the broad strategy of the proof: Each point x in the universal cover $\widetilde{X}$ is contained in "maximal disk" U_x whose image under the developing map of the projective structure is a round disk V_x in $\mathbb{C}\mathrm{P}^1$. The convex hull $C(V_x)$ in $\mathbb{H}^3$ is bounded by a totally geodesic hyperbolic plane. The envelope of these convex hulls, as x varies in $\widetilde{X}$, then defines a $\rho_{\mathbb{C}}$-equivariant "pleated plane" in $\mathbb{H}^3$, bent along an equivariant collection of geodesic pleating lines. In fact, each convex hull $C(\overline{U_x} \cap \partial_\infty \widetilde{X})$ maps to a totally geodesic face or "plaque" of the pleated plane, or a pleating line. "Straightening" the pleated plane then yields a totally geodesic copy of $\mathbb{H}^2$, on which the pleating lines define the measured lamination L. This equivariant pleated plane in $\mathbb{H}^3$ determined by the projective structure is a key intermediate object, interesting in its own right, that we shall refer to later.

The above result will be crucial in the proof of Theorem 1.1; in particular, we shall apply it to the lift of a given projective structure on $\mathbb{S}$ to its universal cover.

3.1 Grafting a marked surface along a measured lamination

Given an element $X \in \mathcal{T}^{\pm}(\mathbb{S}, \mathbb{M})$ and a measured geodesic lamination λ on X, we can perform the operation of grafting the surface along this measured lamination λ, to obtain a projective structure on $\mathbb{S}$. In this section we first show that this projective structure is in fact in $\mathcal{P}(\mathbb{S}, \mathbb{M})$, ie the Schwarzian derivative of the developing map descends to a quadratic differential with poles of prescribed orders at the punctures of the underlying Riemann surface. (For brevity we shall often abbreviate this by saying that the projective structure has poles of prescribed orders.) It suffices to verify this for each end of X; recall that there are only finitely many leaves of the lamination λ going into a cusp, geodesic end or crown end.

For a boundary component of $\mathbb{S}$ (which defines a crown end of X), we have the following from [27]:

Lemma 3.2 [27, Proposition 4.2] *The operation of grafting along a measured geodesic lamination exiting a crown end produces a meromorphic projective structure with a pole of order n_i on the underlying punctured Riemann surface.*

For the cusps and geodesic boundary components of X, we have the following lemma. For the computations in the proof, it will help to recall the definition of the Schwarzian derivative:

$$S(f) = \left(\frac{f''}{f'}\right)' - \frac{1}{2}\left(\frac{f''}{f'}\right)^2. \tag{7}$$

Although this is essentially Proposition 3.5 of [25], we provide a complete proof here that clarifies some of the arguments there; we shall refer to some of the computations throughout this paper.

Lemma 3.3 *The operation of grafting at cusps and geodesic ends produces a projective structure such that the Schwarzian derivative of the developing map on the underlying Riemann surface has a pole of order at most 2.*

Proof We shall consider the two cases of a cusp end and a geodesic end separately:

- For a cusp, we can assume that it has a punctured-disk neighborhood $\mathbb{D}^*$ that lifts to $H=\{z:\mathrm{Im}(z)>a\}$, with $\mathbb{D}^*$ biholomorphic to $H/\langle z\to z+1\rangle$. Let the lifts of the finitely many geodesics entering the cusp be given by $\{z:\mathrm{Re}(z)=a_j\}$ with corresponding weights α_j for $j=1,2,\dots,r$ (assuming without loss of generality that $0\le a_1<a_2<\cdots<a_r<1$). Let us define $\omega_j=e^{i\alpha_j}$. Define by $E_i(w)$ the elliptic element $z\mapsto \omega_i^{-1}(z-w)+w$ (ie clockwise rotation by α_j about the endpoints ∞ and w), and let T denote the translation $z\mapsto z+1$.

Now, referring to [16, Lemma 5.5], the monodromy after bending will be conjugate to the element $E_1(a_1)\circ E_2(a_2)\circ\cdots\circ E_r(a_r)\circ T$, which is easily seen to be

$$z\mapsto \omega_1^{-1}\omega_2^{-1}\dots\omega_r^{-1}z+c, \tag{8}$$

where

$$c=a_1+\sum_{i=1}^{r}\omega_1^{-1}\omega_2^{-1}\dots\omega_i^{-1}(a_{i+1}-a_i), \tag{9}$$

where we set $a_{r+1}=1$. This element is elliptic if $\alpha:=\sum\alpha_i$ is not a multiple of 2π. Otherwise, this is either a parabolic element or the identity.

To determine the developing map and compute the Schwarzian derivative, we divide into the following cases. The possible developing maps at a regular singularity are classified by studying solutions of the Schwarzian equation (1); see Section 4.1.1 for a discussion.

(i) **α is not an integer multiple of 2π** Recall that the operation of grafting introduces "lunes" (regions in $\mathbb{C}\mathrm{P}^1$ bounded by circular arcs) at every lift of a geodesic leaf entering $\mathbb{D}^*$. The total sum of the angles of lunes is α. Recall that the resulting peripheral monodromy after grafting is an elliptic rotation of angle α; we can assume that it fixes the point $\infty\in\overline{\mathbb{R}}\subset\mathbb{C}\mathrm{P}^1$. Let F be a fundamental domain of the $\mathbb{Z}$-action on H, bounded by the vertical lines $\{\mathrm{Re}(z)=0\}$ and $\{\mathrm{Re}(z)=1\}$. On F the developing map is a conformal map, that takes these two sides of F to two circular arcs incident at ∞ that differ by an elliptic rotation of angle α fixing ∞. Such a conformal map is the map $z\mapsto e^{-2\pi i\alpha z}$; indeed, this takes the two boundary lines of F to the circular arcs $\{e^{2\pi x}:x\in\mathbb{R}^+\}$ and $\{e^{2\pi i\alpha}\cdot e^{2\pi x}:x\in\mathbb{R}^+\}$. On the punctured disk $\mathbb{D}^*$, the developing map descends to the map $f:w\mapsto w^{-\alpha/2\pi}$, since the universal covering map from H to $\mathbb{D}^*$ is $z\mapsto w:=e^{2\pi iz}$.

Moreover, the Schwarzian derivative of the developing map on F descends to the Schwarzian derivative of f on $\mathbb{D}^*$. Computing this Schwarzian derivative using (7), we obtain

$$S(f)=\frac{4\pi^2-\alpha^2}{8\pi^2w^2}dw^2.$$

Since $\alpha\neq\pm2\pi$, this quadratic differential has a pole of order 2 at the puncture.

(ii) **α is an integer multiple of 2π** Let $\alpha = 2n\pi$, $n \geq 0$. In the case that $c = 0$, note that $n > 0$, and we have, from the same discussion as above, that the developing map descends to a map of the form $w \mapsto w^n$ in a neighborhood of the puncture, and hence the Schwarzian derivative is $\frac{1-n^2}{2w^2}dw^2$, which has a pole of order 2 at the puncture if $n \neq 1$, and a zero at the puncture if $n = 1$.

Henceforth, let us assume that $c \neq 0$, hence the monodromy of the structure around the puncture is a parabolic element. In this case, the developing map descends to the map $f\colon w \mapsto w^{-n} + \log(w)$ on the punctured disk. On the universal cover H, the developing map is $z \mapsto e^{-2\pi i n z} + 2\pi i z$. Indeed, this takes the lines $\{\mathrm{Re}(z) = 0\}$ and $\{\mathrm{Re}(z) = 1\}$ to the circular arcs $\{e^{2\pi n x} - 2\pi x : x \in \mathbb{R}^+\}$ and $\{e^{2\pi n x} - 2\pi x + 2\pi i : x \in \mathbb{R}^+\}$ in $\mathbb{CP}^1$ both incident at ∞, and having the same tangential direction there. The monodromy around the puncture in this case is given by $z \mapsto z + 2\pi i$. As before the developing map "wraps" the infinite strip $\{0 \leq \mathrm{Re}(z) \leq 1\}$ on $\mathbb{CP}^1$ by $2n\pi$.

Computing the Schwarzian derivative of f, we obtain

$$S(f) = \frac{w^{2n} - w^n(2n^3 + 2n) - n^2(n^2 - 1)}{2w^2(w^n - n)^2}dw^2.$$

If $n = 0$, this equals $\frac{1}{2}w^{-2}dw^2$, hence it has a pole of order 2. For $n \geq 1$, we can expand near 0 to get $\bigl(\frac{1}{2}(1-n^2)w^{-2} - 2nw^{n-2} + O(w^{2n-2})\bigr)dw^2$. Clearly, it has a pole of order 2 if $n \geq 2$. If $n = 1$, then it has a pole of order 1. (This final observation results in the next Corollary 3.4.)

- For a geodesic boundary end, we can take the universal cover of its neighborhood to be the set $B = \bigl\{z : \frac{\pi}{2} - \epsilon < \arg(z) < \frac{\pi}{2}\bigr\} \subset \mathbb{H}^2$ with the line $\{\mathrm{Re}(z) = 0\}$ mapping onto the geodesic. Here, we are assuming that the lift of the surface lies to the right of this geodesic line. The end is biholomorphic to $B/\langle z \to \lambda z\rangle$, where $\log(\lambda)$ is the length of the geodesic end. The geodesics entering the end can spiral in one of two directions, clockwise or anticlockwise. Correspondingly, the geodesics either have endpoints at 0 and the positive real axis, or at ∞ and the positive real axis, respectively.

Let us first consider the case of geodesics spiraling anticlockwise into the boundary component. Let the geodesics entering the end of a fundamental domain $U := \{z \in \mathbb{H}^2 : 1 \leq \mathrm{Re}(z) < \lambda\}$ be given by $\{\gamma_i : \mathrm{Re}(z) = a_i\}$, assuming $1 \leq a_1 < \cdots < a_r < \lambda$. Recall the definitions of α_j, ω_j and α; α_j are the weights of the grafting geodesics, $\alpha = \sum \alpha_j$ is the total weight, and $\omega_j = e^{i\alpha_j}$.

This time, the monodromy after grafting can be computed as follows. Defining $E_i(w)$ to be the elliptic elements as before, and T to be the map $z \mapsto \lambda z$, we have from [16, Lemma 5.5] that the monodromy after bending will be conjugate to $E_1(a_1) \circ E_2(a_2) \circ \cdots \circ E_r(a_r) \circ T$, which is computed to be $z \mapsto \lambda\omega_1^{-1}\omega_2^{-1}\ldots\omega_r^{-1}z + c$, for $c = (a_1 - 1) + \sum \omega_1^{-1}\omega_2^{-1}\ldots\omega_i^{-1}(a_{i+1} - a_i)$ (where $a_{r+1} = \lambda$). Since $\lambda > 1$, the monodromy is either a loxodromic or hyperbolic element.

As before, we now compute the Schwarzian derivative of the developing map, and split into two cases depending on the total angle α of the leaves of the lamination spiraling onto the geodesic boundary component:

(i) $\boldsymbol{\alpha = 0}$ (see also Lemma 3.4 of [25]) In this case, recall that the weight on the geodesic boundary is infinite, and hence we perform an "infinite grafting" on any lift in the universal cover. This amounts to attaching a "logarithmic end", or "semi-infinite lune" $\mathcal{L}_\infty$, ie semi-infinite chain of $\mathbb{C}\mathrm{P}^1$-s each slit along an identical arc, to any such lift. (See Section 4.2 of [27].) Take such a lift to be the vertical geodesic $\mathrm{Re}(z) = 0$ in $\mathbb{H}^2$, with the lift of the surface lies to its right. The infinite grafting does not change the monodromy of the end, so it remains $\langle z \to \lambda z \rangle$. Recall that the semi-infinite lune that is grafted in, descends to what is conformally a punctured disk $\mathbb{D}^*$ on the surface (see the proof of Lemma 4.3 in [27]). The developing map restricted to the universal cover of the punctured disk is thus the conformal diffeomorphism $f : H \to \mathcal{L}_\infty$ defined by $z \mapsto i e^{\log(\lambda) z}$. (Recall here that we are taking $H/\langle z \mapsto z+1 \rangle \cong \mathbb{D}^*$ as before,)

This time, the developing map descends to the map $g(w) = i w^{\frac{1}{2\pi i} \log(\lambda)}$ on $\mathbb{D}^*$, which has a Schwarzian derivative given by

$$S(g) = \frac{4\pi^2 + \log(\lambda)^2}{8\pi^2 w^2} dw^2,$$

which clearly has a pole of order 2.

(ii) $\boldsymbol{\alpha \neq 0}$ In this case, we first identify a suitable fundamental domain for the universal cover B of the geodesic boundary end for which it is easier to compute a uniformizing map. As earlier, consider the domain $U := \{z \in \mathbb{H}^2 : 1 \le \mathrm{Re}(z) < \lambda\}$, and let V be its image under the map $z \mapsto \frac{1}{z}$. Then, V is the region bounded by the semicircles joining 0 to $\frac{1}{\lambda}$ and 0 to 1, in the lower half-plane (the shaded region in Figure 2). The grafting geodesics thus become semicircles in the lower half-plane joining 0 to $\frac{1}{a_i}$. We now follow the same argument as before.

First, the monodromy around the puncture is a loxodromic element that fixes 0, given by $z \mapsto \lambda^{-1} e^{i\alpha} z$ (this is using the monodromy formula we found earlier, conjugated by the map $z \mapsto \frac{1}{z}$). Next, grafting introduces lunes of weight α_i along the semicircles in the lower-half plane. This results in a total angle of α between the tangents at 0 to the boundary semicircles after grafting. Using these observations, this time the developing map descends to the map $g(w) = w^{\frac{1}{2\pi}(\alpha + i \log(\lambda))}$ on $\mathbb{D}^*$. Indeed, working in the punctured disk, the two images of a radial slit under g have an angle of α between them, and the monodromy element mapping one to the other is given by multiplication with $(e^{2\pi i})^{\frac{1}{2\pi}(\alpha + i \log(\lambda))} = \lambda^{-1} e^{i\alpha}$ as desired.

Computing the Schwarzian derivative, we get

$$S(g) = \frac{4\pi^2 - (\alpha + i \log(\lambda))^2}{8\pi^2 w^2} dw^2.$$

Since $\alpha + i \log(\lambda) \neq \pm 2\pi$, we again get a pole of order 2.

Now, it only remains to consider the case of geodesics spiraling clockwise into the end, ie the geodesics are semicircles in the upper half-plane. If we take the map $z \mapsto \frac{1}{z}$, B goes to its complex conjugate in $\mathbb{C}$ while the geodesics go to the lines $\mathrm{Re}(z) = \frac{1}{a_i}$ in the lower half-plane. (See Figure 2.) This is clearly the conjugate image of the earlier case of anticlockwise spiraling geodesics, with the grafting geodesics

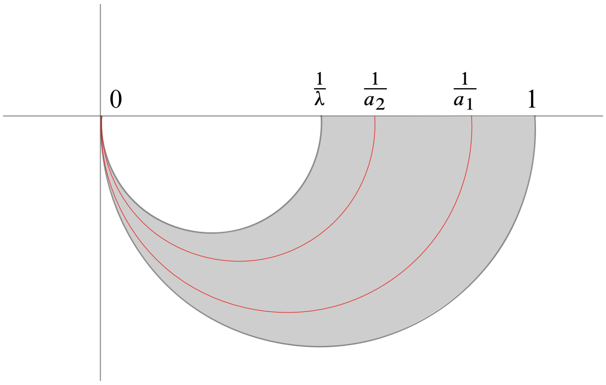


Figure 2: Grafting a geodesic boundary end: a fundamental domain in the universal cover (shown shaded) with the lifts of the spiraling weighted geodesic (shown in red).

being given by $\mathrm{Re}(z) = \frac{\lambda}{a_i}$ in the domain U defined earlier. Thus, it follows that if $z \mapsto f(z)$ is the developing map for the earlier case, the new developing map is its Schwarz reflection, $z \mapsto \overline{f(\bar{z})}$. From this, it follows that the holonomy will be conjugate to $z \mapsto \lambda\omega_1\omega_2 \dots \omega_r z + c$, for

$$c = \left(\frac{\lambda}{a_r} - 1\right) + \sum \omega_r\omega_{r-1} \dots \omega_{r+1-i}\left(\frac{\lambda}{a_{r-i}} - \frac{\lambda}{a_{r+1-i}}\right).$$

This time, developing map descends to a map g on the punctured disk of the form $w \mapsto w^{\frac{1}{2\pi}(\alpha - i\log(\lambda))}$, with Schwarzian derivative

$$S(g) = \frac{4\pi^2 - (\alpha - i\log(\lambda))^2}{8\pi^2 w^2} dw^2$$

having a pole of order 2. □

Remark As a result of our computation in the above proof, we obtain a description of the grafting configuration that gives rise to poles of order 1 — it occurs if and only if we graft a cusp end with $\alpha = 2\pi$ and $c \neq 0$. Thus, we have:

Corollary 3.4 *If there is a pole of order* 1, *it must be obtained by grafting a cusp end along a measured lamination with total weight of leaves going into the cusp equal to* 2π. *Conversely, generically we obtain a pole of order* 1 *by grafting at a cusp when the total weight of the leaves of the lamination going into the cusp is* 2π *— the only exception is when* $c = 0$ (*see* (9)).

Example 1 Consider the projective structure on $\mathbb{CP}^1 \setminus \{0, 1, \infty\}$, ie the thrice punctured sphere, obtained by the trivial chart via the inclusion of this surface in $\mathbb{CP}^1$. Then, if one does the inverse grafting construction via constructing a pleated plane, one would obtain that the pleated plane would be the ideal triangle in the interior of the ball $\mathbb{H}^3$ with vertices at $\{0, 1, \infty\} \in \partial_\infty\mathbb{H}^3$. So, a grafting description of this structure would be obtained by taking a hyperbolic sphere with three cusps and grafting in lunes of

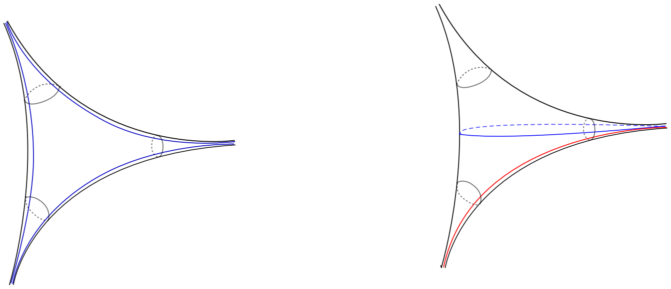


Figure 3: Grafting description of the projective structures in Examples 1 (left) and 2 (right). The blue geodesics have weight π and the red geodesic has weight 2π.

weight π along three geodesic lines running between pairs of cusps (see Figure 3). Since the developing map is just the identity map, its Schwarzian derivative is identically zero. Now, we can compute the constant c (see (9)) in this case by considering the geodesics going into the cusp $\mathbb{H}/\langle z \mapsto z+1\rangle$ given by $a_1 = 0, a_2 = \frac{1}{2}$ (according to the notation used in Lemma 3.3) both having weight π. The computation yields $c = 0 + (-1)\big(\frac{1}{2} - 0\big) + (1)\big(1 - \frac{1}{2}\big) = 0$. This illustrates the final statement of Corollary 3.4 — even though the total weight at each puncture is 2π, they are not poles of order 1.

Example 2 Consider another projective structure on $\mathbb{CP}^1 \setminus \{0, 1, \infty\}$, obtained by the developing map that descends to the map $f(z) = \log(z) + \frac{1}{z}$ (more precisely, this structure descends from a developing map on the intermediate cover $\mathbb{C} \setminus \{2n\pi i : n \in \mathbb{Z}\}$ given by $z \mapsto z + e^{-z}$). As in the previous case, one can perform the inverse grafting operation by constructing a pleated plane, to obtain the following grafting description for this structure: take the thrice punctured sphere and graft along two weighted geodesics: one having both ends going into 1 and of weight π, and the other with ends going into 1 and 0, with a weight of 2π. (See Figure 3.) Since the punctures corresponding to ∞ and 1 have total weight of lamination not equal to 2π, and they are poles of order 2. Computing the constant c at the puncture corresponding to 0, we get $c = 1 \neq 0$. So, we expect 0 to be a pole of order 1. We can compute the Schwarzian derivative of the map f at the punctures to get that indeed 1 and ∞ are poles of order 2, while 0 is a simple pole. This verifies Corollary 3.4 in this case.

We note the following observation culled from the proof of the above lemma (see also Lemma 3.2 of [25]):

Corollary 3.5 *The following information can be inferred from the monodromy around a regular singularity obtained by grafting a cusp or geodesic boundary end*:

(i) *the type of end* (*ie cusp or geodesic boundary*) *and in the case of the latter*, *the length of the geodesic boundary*;

(ii) *the total weight of the leaves of the grafting lamination incident at that end*, *up to positive integer multiples of* 2π.

Proof We simply note the following from the proof of Lemma 3.2:

(i) If the monodromy is not loxodromic, we can infer the type of end to be a cusp. If the monodromy is loxodromic and conjugate to the map $z \mapsto \lambda z$, then the length of the boundary is given by $|\log(\lambda)|$.

(ii) If the monodromy is parabolic or identity, clearly the total weight is 0 modulo 2π. Otherwise, the monodromy is conjugate to $z \mapsto \lambda z$, and the total weight of leaves is given by $\arg(\lambda)$ modulo 2π. □

Since Lemmas 3.2 and 3.3 involve local computations in a neighborhood of each pole, by combining them we obtain:

Proposition 3.6 *The result of grafting a marked hyperbolic surface $X \in \mathcal{T}(\mathbb{S}, \mathbb{M})$ along a measured lamination $(L, \mu) \in \mathcal{ML}(X)$ is a marked projective structure in $\mathcal{P}(\mathbb{S}, \mathbb{M})$.*

Finally, we note a lemma that relates the grafting description at a geodesic or cusp end, to the corresponding exponent of the Schwarzian derivative of the resulting projective structure:

Lemma 3.7 *Suppose we obtain a projective structure by grafting a geodesic boundary end of length l (which can be zero, giving a cusp end) and total weight of the leaves of the grafting lamination spiraling onto that end being α. Let $q(z)dz^2$ denote the Schwarzian derivative of the projective structure at the puncture corresponding to the end. Let the exponent of this quadratic differential be r. Then, if the geodesics spiral anticlockwise, we have $r = \pm(l - i\alpha)$. If the geodesics spiral clockwise, we have $r = \pm(l + i\alpha)$.*

Proof This immediately follows from a simple calculation of the exponent using the Schwarzian derivatives we computed in the proof of Lemma 3.3. Recall that the exponent r is given as $r = \pm 2\pi i\sqrt{1-2a}$, where a is the coefficient of $z^{-2}dz^2$ in the Schwarzian derivative. We have the following cases:

- $l = 0$, ie we have a cusp end:
  - (i) α is not an integer multiple of 2π: the exponent is $\pm i\alpha$.
  - (ii) $\alpha = 2\pi n$ for $n \geq 0$: in both the cases considered, the coefficient of $z^{-2}dz^2$ in the Schwarzian derivative is $\frac{1-n^2}{2}$, so the exponent is $\pm 2\pi i n = \pm i\alpha$.
- We have a geodesic boundary end with $l > 0$:
  - (i) $\alpha = 0$: the exponent is $\pm\log(\lambda) = \pm l$.
  - (ii) $\alpha \neq 0$: If the spiraling is anticlockwise, the exponent is $\pm(i\alpha - \log(\lambda)) = \pm(l - i\alpha)$. If the spiraling is clockwise, it is $\pm(l + i\alpha)$. □

3.2 Defining the unsigned grafting map

In this and the following sections, we omit the pair $(\mathbb{S}, \mathbb{M})$ for the parameter spaces if there is no source of confusion. We define the unsigned grafting map,

$$\mathrm{Gr}' \colon \mathcal{TML} \to \mathcal{P}(\mathbb{S}, \mathbb{M}), \tag{10}$$

where the domain is the space of pairs

$$\mathcal{TML} = \{(X,\lambda) : X \in \mathcal{T},\ \lambda \in \mathcal{ML}(X)\},$$

which is a quotient of the product of signed spaces $\mathcal{T}^{\pm}(\mathbb{S},\mathbb{M}) \times \mathcal{ML}^{\pm}(\mathbb{S},\mathbb{M})$, obtained by identifying the pairs differing only in the signings.

There is a forgetful projection map

$$F\colon \mathcal{T}^{\pm} \times \mathcal{ML}^{\pm} \to \mathcal{TML} \tag{11}$$

that is described as follows: Let us take an element $(X_\sigma, \lambda_\tau) \in \mathcal{T}^{\pm} \times \mathcal{ML}^{\pm}$, where σ, τ denote the respective signings. We need to define the corresponding element on the right-hand side. We take X to be the underlying hyperbolic structure of X_σ. To define the measured lamination λ_τ it remains to prescribe the direction in which the leaves of the lamination spiral on entering the geodesic boundary ends. We define the spiraling direction to be clockwise (with respect to the orientation of X) if the signs σ and τ match at the end, and anticlockwise otherwise.

Remark There is a natural continuous map $\mathcal{TML} \to \mathcal{T}$ that takes the pair (X,λ) to X. However, this map is not a fiber bundle, since the fiber over an element X is $\mathcal{ML}(X)$ which by Proposition 2.9 is homeomorphic to $\mathbb{R}^{\chi-p} \times \mathbb{R}^{p}_{\geq 0}$ (where p is the number of cusp ends of X), and these fibers are not homeomorphic to each other. In particular, $\mathcal{TML}$ cannot be written as a natural product $\mathcal{T} \times \mathcal{ML}$.

Proposition 3.8 *The map* Gr' *is continuous.*

Proof We only provide a sketch of the proof here since this is essentially identical to the proof of continuity of the grafting map for closed surfaces (which in turn follows from the discussion of the continuity of the "quakebend cocycle" in [17, Chapter II.3.11]).

In the closed case, the crucial observation is that if two pairs (X_1,λ_1) and (X_2,λ_2) are close in $\mathcal{T}_g \times \mathcal{ML}$, the corresponding lifted laminations would be close in the universal cover $\mathbb{H}^2$, suitably normalized by fixing three ideal points. We describe this further: consider a $(1+\epsilon)$-quasi-isometry from $\mathbb{H}^2$ to $\mathbb{H}^2$ (fixing, say, $0, 1, \infty \in \partial_\infty \mathbb{H}^2$) that descends to a $(1+\epsilon)$-quasi-isometry between the two surfaces. (Here $\epsilon > 0$ is small if the surfaces X_1 and X_2 are close.) Such an almost-isometry extends to a homeomorphism of the boundary that is close to the identity map. Now recall a measured lamination can be thought of as a Borel measure on $\partial_\infty \mathbb{H}^2 \times \partial_\infty \mathbb{H}^2 \setminus \Delta$ (where Δ is the diagonal subspace). The fact that the ideal boundary correspondence is close to the identity map then implies that the lifts of same lamination $\lambda \in \mathcal{M})$ on the two hyperbolic surfaces will be close (in the Hausdorff metric) on any compact subset K of the universal cover $\mathbb{H}^2$, where the respective universal covers are now identified via the $(1+\epsilon)$-quasi-isometry mentioned above. (In this argument "close" means ϵ'-close where $\epsilon' \to 0$ as $\epsilon \to 0$.) Moreover, the topology on $\mathcal{ML}$ is the weak topology on the space of such measures, and nearby laminations in this topology will also be Hausdorff-close on K (see, for example, Proposition 1.9 of [22] that follows arguments in [42]).

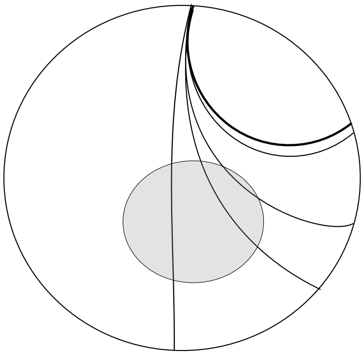


Figure 4: Lifts of a spiraling leaf to the universal cover $\mathbb{H}^2$ accumulates to a lift of the geodesic boundary component. The compact set K (shown shaded) intersects finitely many of such lifts.

As a consequence, if one restricts to such a compact subset $K \subset \mathbb{H}^2$, the resulting developing maps after grafting (X_1, λ_1) and (X_2, λ_2) will be close in $\mathbb{C}\mathrm{P}^1$. (Recall that these developing maps are obtained by grafting in "lunes" along the leaves of the lamination λ_1 and λ_2 respectively.)

The only two new cases to consider in our nonclosed setting are:

(A) Grafting along a hyperbolic surface with geodesic boundary, where the grafting lamination spirals on to the boundary. The leaves spiraling on to the boundary component are isolated geodesics, and in the universal cover, lifts to a sequence of geodesic lines accumulating to the lift of the boundary component. (See Figure 4.) Hence any compact set $K \subset \mathbb{H}^2$ intersects finitely many such lines. As in the closed-surface argument above, as one varies the hyperbolic surface (this might vary the length of the geodesic boundary component), and the lamination (ie varying the weight of the spiraling leaves), the picture of the laminations restricted to K varies continuously, and hence the grafting map is continuous.

(B) Grafting along a hyperbolic surface with a cusp, with finitely many isolated leaves of the grafting lamination going out of the cusp. Once again, in the universal cover any compact set K will intersect finitely many of lifts of such leaves. This time, varying the hyperbolic structure includes the possibility of the cusp "opening out" to a geodesic boundary component (of some small length). It is a consequence of the Collar Lemma that a sequence of hyperbolic surfaces with the length of a geodesic boundary tends to zero converges to a cusped hyperbolic surface in the Gromov–Hausdorff sense (see, for example, Lemma 2.15 of [36]). In particular, even in this case the restriction of the lamination to a fixed compact subset K will vary continuously, and so will developing map after grafting.

Thus, in both cases, the grafting map is still continuous. □

3.3 Bijectivity of the unsigned grafting map

Before we turn to the signed grafting map, we shall first prove that the *unsigned* grafting map defined in Section 3.2 is a bijection.

First, let us note a lemma we will use; this is implicitly used in [27] (see the proof of Proposition 4.5 there). In what follows, we refer to Theorem 3.1 and the brief discussion that follows it.

Lemma 3.9 *Let $f : \widetilde{X} \to \mathbb{CP}^1$ be the developing map of a projective structure on a hyperbolic surface X with a cusp. Let the lift of the cusp to the ideal boundary of $\widetilde{X}$ be denoted by p. Suppose that there is a point $q \in \mathbb{CP}^1$ such that for $x \in \widetilde{X}$ that continuously varies along a nonconstant path in the horocyclic neighborhood of p, there is an embedded disk $U_x \ni x$ in $\overline{\mathbb{H}}$ such that $p \in \partial U_x$, $f(U_x)$ is a round disk in $\mathbb{CP}^1$ tangent to q, and f extends continuously over U_x to p such that $f(p) = q$. Then, if the boundary circles of the disks $f(U_x)$ are not tangent to each other at q, the pleated plane corresponding to f has an isolated pleating line incident at q.*

Proof Each U_x is contained in a maximal disk V_x on $\widetilde{X}$, since its image is a round disk on $\mathbb{CP}^1$. However, we have that the V_x must be distinct, since if $U_x, U_y \subseteq V$, then $f(V)$ is also a round disk in $\mathbb{CP}^1$, containing the round disks $f(U_x)$, $f(U_y)$, but their boundaries must intersect at the common point q, contradicting that the boundaries of $f(U_x)$, $f(U_y)$ are not tangent to each other. Hence, we get a family of maximal disks whose images have q as a common boundary point. Now, consider the interior of the convex hull associated to each maximal disk V_x, ie the convex hull in $\mathbb{H}^3$ of the set $f(\overline{U_x} \cap \partial_\infty \widetilde{X}) \subset \partial_\infty \mathbb{H}^3$. Following Section 4 of [35], since V_x are distinct maximal balls, it follows that $C(f(\overline{U_x} \cap \partial_\infty \widetilde{X}))$ are distinct as well. These define the pleated plane corresponding to the projective structure; in the totally geodesic copy of $\mathbb{H}^2$ obtained by a "straightening" of this pleated plane, each convex hull is either a plaque or a pleat with one boundary point at q. Now, if we have at least two among these family of convex hulls which straighten to a plaque, the plaques must have a nonzero angle between them, which implies that we must have some weight of geodesic lamination incident at q. Otherwise, we would have a continuous family of convex hulls such that all of them straighten to pleating lines incident on q, which again implies that q will have an incident geodesic pleating lamination of nonzero weight. By Lemma 2.6, the geodesic lamination must in fact have a pleating line incident on q. □

Next, we proceed to construct an inverse to the (unsigned) grafting map:

Proposition 3.10 *Given a projective structure $P \in \mathcal{P}(\mathbb{S}, \mathbb{M})$, there is a unique hyperbolic surface $X \in \mathcal{T}(\mathbb{S}, \mathbb{M})$ and a unique measured lamination $(L, \mu) \in \mathcal{ML}(X)$ such that P is obtained by grafting X along (L, μ).*

Proof As a consequence of Theorem 3.1, we know that P can be obtained by grafting a unique hyperbolic surface homeomorphic to $S_{g,k}$ along a unique measured lamination on it. Now, to prove the proposition, it suffices to show that the grafting lamination and the grafting ends are as described in Section 2.1, ie a crown end, a cusp end or a geodesic end, with a finite number of geodesics going into the ends. Hence, this is a completely local argument, and henceforth let us assume that P has no irregular singularities. The case for irregular singularities was established in Proposition 4.5 of [27], and our proof for that case follows from their result.

Let the pair of hyperbolic surface and measured lamination be (X', L). Then, it follows that X' is of genus g and has k ends, which are either cusps or flares. In the case of a cusp, basic hyperbolic geometry implies that any leaf of L incident entering a horocyclic neighborhood of the cusp must be isolated and exiting out of the cusp end (see Lemma 2.6). So, it suffices to show that if X' has a flaring end bounded by a closed geodesic γ, then there is a leaf of $\widetilde{L}$ in the universal cover that is a geodesic line incident to one of the end-points of the lift of γ. This would imply that the leaves of L incident at that end either spirals and accumulates onto γ, or is γ itself (with infinite weight). In other words, we need to only rule out the possibility that there is a leaf of L that exits the flaring end "transverse" to γ.

We shall do this in the remainder of the proof; for simplicity we shall assume that X' has only one flaring end and no cusps.

We shall now use the fact that in a neighborhood of a regular puncture that is conformally $\mathbb{D}^*$, the developing map for the projective structure descends to explicit maps expressed in a local coordinate w on $\mathbb{D}^*$, obtained by solving the Schwarzian equation (1). (See Section 4.1.1 for a discussion.) In what follows, the neighborhood of the puncture is $\mathbb{D}^* = H/\langle z \mapsto z+1\rangle$ where $H = \{z \mid \operatorname{Im}(z) > h_0\}$ is a horodisk centered at ∞.

We have the following cases:

- **$f(w) = w^{-n} + \log(w)$ (on the punctured unit disk) or $f(z) = e^{2\pi i\alpha z}$ (on the universal cover $\mathbb{H}^2$) with α real** In this case, the monodromy around the puncture is identity or elliptic or parabolic. We shall also assume that the asymptotic value the puncture is $0 \in \mathbb{CP}^1$. Now, if $\theta \neq 0$, for each z in the horodisk H, there is a neighborhood of z which maps to a round disk in $\mathbb{CP}^1$, centered at $f(z)$ and tangent to $0 \in \mathbb{CP}^1$. Varying z along a horocycle, the images of the disks have continuously varying tangents at 0. Hence, by Lemma 3.9, we get that there is a pleating line incident at 0, which belongs to a $\rho_{\mathbb{C}}$-equivariant family of geodesics. On straightening the plane, we get a $\rho_{\mathbb{R}}$-equivariant family of geodesics on the disk. (Recall here that $\rho_{\mathbb{C}}$ is the original representation into $\mathrm{PSL}_2(\mathbb{C})$, and $\rho_{\mathbb{R}}$ is the representation into $\mathrm{PSL}_2(\mathbb{R})$ obtained after straightening, see Theorem 3.1.) Since the elliptic or parabolic monodromy around the puncture fixes the point 0, we have that the $\rho_{\mathbb{C}}$-equivariant family of pleating lines is also incident at 0, hence after straightening as well they will be incident at 0. Thus, the $\rho_{\mathbb{R}}$-image of the peripheral loop around the puncture will be a hyperbolic or parabolic element. In case of being a hyperbolic element, the geodesics of the lamination will either approach the geodesic fixed by the hyperbolic element, or it will be that fixed geodesic with infinite weight. In either of these cases, we know from the computations in Lemma 4.2. that the monodromy after grafting will be hyperbolic or loxodromic, hence this case is not possible. Thus, it follows that the cusp monodromy will be parabolic, so we obtain that X' has a cusp end with a geodesic lamination going into it.

- **$f(w) = w^{\alpha}$ (on the punctured disk) with α not real** In this case, instead of varying w along a horocycle in the universal cover, we vary it along a path $\{(t, \theta(t)) : t \in (0, \epsilon)\}$ in polar coordinates on $\mathbb{D}^*$

where $\theta(t) = \frac{\alpha_r}{\alpha_i}\log(t) + k$ (here α_r, α_i are the real and imaginary parts of α respectively). Then, the image of this path under the developing map will be

$$(te^{i\theta})^{\alpha} = e^{(\log t + i\theta)(\alpha_r + i\alpha_i)} = e^{(\log(t)\alpha_r - \theta\alpha_i) + i(\theta\alpha_r + \log(t)\alpha_i)} = e^{k\alpha + i(|\alpha|^2/\alpha_i)\log(t)},$$

and as $t \to 0$, $\log(t) \to -\infty$. Therefore, the image of the path is clearly a circle on $\mathbb{CP}^1$ separating the fixed points of the monodromy. So, it is again easy to construct neighborhoods of z which map to round disks in $\mathbb{CP}^1$ passing through $0 \in \mathbb{C}\mathrm{P}^1$ satisfying the hypotheses of Lemma 3.9. Since the monodromy around the puncture is loxodromic, we know that the $\rho_{\mathbb{R}}$-monodromy is hyperbolic. By Lemma 3.9 the grafting lamination L has a leaf converging to a fixed point of the loxodromic element that is the monodromy around the puncture.

Hence in all the cases, it follows that on a neighborhood of a regular singularity, the projective structure is obtained by grafting along a cusp end with some weight of the geodesic lamination going into the cusp (in the case when the monodromy is parabolic, elliptic, or identity), or by grafting a geodesic boundary component having infinite weight (in the case when the monodromy is hyperbolic) or grafting along weighted geodesics spiraling onto a geodesic boundary component (in the case when the monodromy is loxodromic). □

Remark In order to apply Lemma 3.9 in the first case above, we only require f to extend continuously over a disc U_x and not over the whole neighborhood of the lift of a cusp end. This is an important distinction, as we shall see in Section 4.1.1 when we discuss the asymptotic behavior of the map $z \mapsto z^{-n} + \log(z)$.

3.4 The signed grafting map and proof of Theorem 1.1

We shall now define the signed grafting map

$$\widehat{\mathrm{Gr}}\colon \mathcal{T}^{\pm}(\mathbb{S},\mathbb{M}) \times \mathcal{ML}^{\pm}(\mathbb{S},\mathbb{M}) \to \mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$$

and then prove that it is a homeomorphism (proving Theorem 1.1).

First, given an interior marked point $p \in \mathbb{P}$, we define a complex function c_p on $\mathcal{T}^{\pm} \times \mathcal{ML}^{\pm}$ as

$$c_p(X_\sigma, \mu_\tau) = l + i\alpha, \tag{12}$$

where α and l are the signed weight of geodesic lamination and signed length of geodesic boundary at p.

Now define the map

$$c\colon \mathcal{T}^{\pm} \times \mathcal{ML}^{\pm} \to \mathbb{C}^m \tag{13}$$

as

$$(X_\sigma, \mu_\tau) \mapsto (c_{p_1}, c_{p_2}, \dots, c_{p_m}),$$

where $p_1, p_2, \dots, p_m$ are the interior marked points $\mathbb{P}$ and c_p is the complex parameter defined above.

We already have defined an unsigned grafting map $\mathrm{Gr}'\colon \mathcal{TML} \to \mathcal{P}(\mathbb{S}, \mathbb{M})$. Composing with the projection $F\colon \mathcal{T}^{\pm} \times \mathcal{ML}^{\pm} \to \mathcal{TML}$, we get a map $\mathrm{Gr}' \circ F\colon \mathcal{T}^{\pm} \times \mathcal{ML}^{\pm} \to \mathcal{P}$. This map is not injective: more precisely, suppose $(X, \lambda) \in \mathcal{T}^{\pm} \times \mathcal{ML}^{\pm}$ such that there is a geodesic boundary component of X with a leaf of λ spiraling onto it, then in the signed spaces there is a sign associated with boundary, as well as the leaf, and by our conventions, when both these signs agree, then they spiral in exactly the same way. The map $\mathrm{Gr}' \circ F$ would then take both these signed pairs to the same projective structure. We need to then argue that in such a case, there is nevertheless a map to the signed space of projective structures $\mathcal{P}^{\pm}$ that is injective. This is most efficiently described using the map c defined above, and the notion of the fiber product (see Definition 2.17), as we shall now see.

Recall that the space of signed meromorphic structures $\mathcal{P}^{\pm}$ introduced in Section 2.3 is a fiber product $\mathcal{P}(\mathbb{S}, \mathbb{M}) \times_{\mathbb{C}^m} \mathbb{C}^m$ where $|\mathbb{P}| = m$, with respect to the map $r^2\colon \mathcal{P}(\mathbb{S}, \mathbb{M}) \to \mathbb{C}^m$ and squaring map $\mathrm{sq}\colon \mathbb{C}^m \to \mathbb{C}^m$. Here, recall that r records the exponents (defined up to sign) of the Schwarzian derivative at the punctures in $\mathbb{P}$.

By Lemma 2.18, in order to define a map from $\mathcal{T}^{\pm} \times \mathcal{ML}^{\pm}$ to $\mathcal{P}^{\pm}$ that "lifts" the map $\mathrm{Gr}' \circ F$, it suffices to define a map to $\mathbb{C}^m$ and check that its composition with sq equals the composition $r^2 \circ (\mathrm{Gr}' \circ F)$. This is exactly the map c defined above; indeed, Lemma 3.7 ensures that we have $\mathrm{sq} \circ c = r^2 \circ (\mathrm{Gr}' \circ F)$. By the universal property of fiber products (see Definition 2.17), this defines the signed grafting map $\widehat{\mathrm{Gr}}$ (as in the statement of Theorem 1.1). This also shows that $\widehat{\mathrm{Gr}}$ is continuous.

We can now formally complete the proof of our main grafting theorem:

Proof of Theorem 1.1 Recall that we have defined the grafting map $\widehat{\mathrm{Gr}}$ by first defining the unsigned grafting map Gr' (see (10)) and then using the universal property of fiber products (see the preceding discussion). By Proposition 3.10, the map Gr' is a bijection. We will show that so is $\widehat{\mathrm{Gr}}$, by working locally at the punctures — since signings are local and independent parameters at the punctures, it is sufficient to prove the bijection working locally. At a puncture p, given an unsigned projective structure, there are two signed projective structures with different signings at p that descend to the given structure, except when the exponent at p is zero. Similarly, if we look at the pair $(|l_p|, |\alpha_p|)$ at p, there are exactly two signed pairs with different signings at p that descend to it — either $(\pm l_p, \pm\alpha_p)$ or $(\pm l_p, \mp\alpha_p)$ depending on the direction of spiraling of the incident lamination geodesics — except when $l_p = \alpha_p = 0$, and indeed, by Lemma 3.7 the exponent at a regular singularity is 0 if and only if we graft along a cusp with no leaves of the grafting lamination incident on it. Thus, due to the way the signings are defined, we get that $\widehat{\mathrm{Gr}}$ is bijective at the level of fibers over the signed spaces, hence $\widehat{\mathrm{Gr}}$ is bijective.

Also, the domain of $\widehat{\mathrm{Gr}}$ is a real-dimensional cell of dimension 2χ, by Theorem 2.3 and Lemma 2.10. Thus, by the bijectivity of $\widehat{\mathrm{Gr}}$ and the invariance of domain, it follows that $\widehat{\mathrm{Gr}}$ is a homeomorphism. □

4 The monodromy map

In Section 6 of [3], they define the framed monodromy of a signed projective structure without apparent singularities. The first goal of this section is to extend their definition to the collection of *all* signed projective structures, that is, we want to define a map

$$\widehat{\Phi}\colon \mathcal{P}^{\pm}(\mathbb{S},\mathbb{M}) \to \widehat{\chi}(\mathbb{S},\mathbb{M})$$

such that $\widehat{\Phi}$ agrees with the map defined in [3] on the set of signed projective structures without apparent singularities. Then, we shall characterize the image of $\widehat{\Phi}$, proving Theorem 1.2 and finally show that it is a local homeomorphism (Theorem 1.4).

4.1 Constructing the map $\widehat{\Phi}$

Given a signed and marked meromorphic projective structure $P \in \mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$, we shall construct a framing by considering the asymptotic values of the developing map at the lifts of the marked points $\mathbb{M}$. Recall the following classical notion:

Definition 4.1 (asymptotic value) An asymptotic value of a (meromorphic) function at an ideal point p is a point in $\mathbb{C}\mathrm{P}^1$ that is a limiting value along a path that diverges to p.

At an irregular singularity, there are exactly $n-2$ asymptotic values (see [26, Corollary 4.1]); these can be assigned to the marked points on the corresponding boundary of $\mathbb{S}$. and the corresponding equivariant assignment of lifts at the universal cover defines the framing at the lifts of such marked points. Thus, it remains to define the framing at regular singularities, namely at the marked points $\mathbb{P} \subset \mathbb{M}$. First, we determine the asymptotic values at these points.

4.1.1 Asymptotics of the developing map at regular singularities Let P be defined by a meromorphic quadratic differential q, and consider a regular singularity of P, which recall is an interior puncture in $\mathbb{S}$. We endow the neighborhood of the puncture with the end-extension topology, as defined in Section 3.1. of [10]. Now, we study the asymptotics of these developing maps at the puncture, according to the value of the exponent r of the Schwarzian derivative (see Definition 2.14).

The possible developing maps are obtained by studying the Schwarzian equation (1) — see the discussion in Section 2.3 of [25]. There is also a discussion regarding these asymptotics in Lemma 4.1.1 of [10].

The asymptotics are as follows:

(i) $\boldsymbol{r = 2\pi i\theta}$ **for** $\boldsymbol{\theta \in \mathbb{C}\setminus\mathbb{R}}$ The developing map is of the form $y(z) = z^{\theta}$ — the monodromy around the puncture in this case is given as multiplication with $e^{2i\pi\theta}$, which is either a hyperbolic or loxodromic element. There are two asymptotic values, 0 and ∞. A pair of paths that limit to these asymptotic values are exactly paths spiraling into the cusp in opposite directions. As a consequence, it is not possible to continuously extend the developing map to the puncture.

(ii) $\boldsymbol{r = 2\pi i\theta}$ **for** $\boldsymbol{\theta \in \mathbb{R}\setminus\mathbb{Z}}$ The developing map is $y(z) = z^\theta$ and the monodromy around the puncture is elliptic, given by multiplication with $e^{2i\pi\theta}$. It is possible to continuously extend the developing map to the puncture by setting the value at the puncture to be 0. As a consequence, it has a unique asymptotic value.

(iii) $\boldsymbol{r = 2\pi i n}$ **for** $\boldsymbol{n \in \mathbb{Z}}$ The developing map is either $y(z) = z^n$ — in which case the monodromy is identity, or $y(z) = z^{-|n|} + \log(z)$ — in this case the monodromy is a parabolic element. In the former case, it is possible to extend the map continuously to the puncture by again setting the value at the puncture to zero. In the latter case, it is not possible to extend the developing map continuously to the puncture. However, there is a unique asymptotic value, ∞. To see this, we can write the developing map in polar coordinates as $f(se^{i\theta}) = s^{-n}e^{-in\theta} + i\theta + \log(s) = (s^{-n}\cos(n\theta) + \log(s)) + i(s^{-n}\sin(n\theta) + \theta)$. Now, for s sufficiently small, $|s^{-n}| \gg |\log(s)|$. Hence, taking θ to be zero and $s \to 0$, $f(s) \to \infty$. However, for s sufficiently small, we can also choose θ to make the real part of the above zero, add an arbitrary multiple of $2\pi/n$ to θ so that the magnitude of the imaginary part becomes smaller than $2\pi/n$. Thus, f cannot be extended continuously to ∞. To see that the asymptotic value of ∞ is unique, note that if $(s(t), \theta(t))$ is a curve converging to the puncture, we have $s(t) \to 0$ as $t \to \infty$. Now if $f(s(t), \theta(t))$ approaches a bounded asymptotic value $\alpha + i\beta$, then clearly $\theta(t) \neq 0$ for t sufficiently large, since otherwise, the real part of f, ie $s(t)^{-n} + \log(s(t))$ would become arbitrarily large. So, for t sufficiently large, $2\pi k/n < \theta(t) < 2\pi(k+1)/n$ for an integer k, and moreover $\cos(\theta(t)) \to 0$. However, then the imaginary part of f will be unbounded, a contradiction.

Remark In each case the asymptotic values of the developing map determined above are also fixed points of the monodromy around the puncture. Moreover, in the case that the monodromy is loxodromic or parabolic, the fixed points are precisely the asymptotic values — see cases (i) and (iii) above. However, when the monodromy is elliptic as in case (ii), there would be two fixed points, but the asymptotic value is only one of them. A key advantage with using asymptotic values is that it is also uniquely defined at an apparent singularity, when the peripheral monodromy is the identity element.

4.1.2 Defining the framing

We define the framing at a puncture in $\mathbb{P}$ as follows:

(i) If the monodromy is identity or parabolic, we define the unique asymptotic value of the developing map as the framing. These asymptotic values are computed in Section 4.1.1 for the model developing maps described there, which assume that the puncture is at $0 \in \mathbb{C}\mathrm{P}^1$. More generally, the developing map differs from the model map by postcomposing with an element of $\mathrm{PSL}_2(\mathbb{C})$, and the asymptotic values differ accordingly.

(ii) If the monodromy is elliptic or loxodromic, we define the framing by considering the sign of the projective structure at the puncture, exactly as in [3, Section 6]. (There is such a sign since the exponent around such a puncture is necessarily nonzero; indeed, if $r = 0$ then the peripheral monodromy is either identity or parabolic.) Namely, we assign the framing to be one of the two fixed points of the monodromy around the puncture according to the sign. To be precise, since the fixed points in $\mathbb{C}\mathrm{P}^1$ are determined by the eigenlines corresponding to the two eigenvalues $e^{\pm\lambda}$, we choose the eigenvalue with the same sign in the exponent. By the preceding remark, in the case that the peripheral monodromy is loxodromic, these fixed points are exactly the asymptotic values of the developing map at the puncture. However, in the case where the peripheral monodromy is elliptic, the framing is not necessarily equal to the unique asymptotic value.

Clearly, the association of points in $\mathbb{CP}^1$ with the punctures prescribed gives a well-defined framing β. Together with the usual monodromy representation $\rho\colon \pi_1(\mathbb{S} \setminus \mathbb{P}) \to \mathrm{PSL}_2(\mathbb{C})$, we obtain a framed representation $\hat{\rho} = (\rho, \beta) \in \hat{\chi}(\mathbb{S}, \mathbb{M})$. Recall that we had started with a signed projective structure P at the beginning of the section; the assignment $\hat{\Phi}(P) = \hat{\rho}$ thus defines the framed monodromy map $\hat{\Phi}$. By the preceding remark, our definition agrees with that of Allegretti–Bridgeland in the case where there are no apparent singularities.

4.2 Nondegenerate framed representations and flips

First, we recall the definition of a nondegenerate framed representation (see [3, Section 4.2; 25, Definition 2.4; 27, Definition 2.6]).

Definition 4.2 A framed representation $\hat{\rho} = (\rho, \beta)$ is *degenerate* if one of the following holds:

(i) The image of β is a single point $\{p\}$ and the monodromy around each puncture is parabolic with fixed point p or the identity.

(ii) The image of β has two points $\{p, q\}$ and the monodromy around each puncture fixes both p and q.

(iii) There is a boundary segment I of $\mathbb{S}$ joining two consecutive marked points p_1, p_2, and a lift $\tilde{I}$, such that its endpoints are given by lifts $\widetilde{p_1}, \widetilde{p_2} \in F_\infty$ satisfying $\beta(\widetilde{p_1}) = \beta(\widetilde{p_2})$.

The framed representation is said to be nondegenerate if it is not degenerate.

Remark The above notion is related to, but distinct from, the notion of a *nondegenerate representation*, which was introduced in [25] while dealing with the case when the marked and bordered surface had no boundary components (ie $\mathbb{M} = \mathbb{P}$). In that case, Definition 4.2(iii) holds vacuously, and a framed representation is nondegenerate if the underlying representation is nondegenerate (see [25, Proposition 3.1]). However, the converse is not true in the presence of apparent singularities. As an example, consider the once-punctured torus, and a representation of its fundamental group given by $\alpha \mapsto \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, $\beta \mapsto \left(\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right)$, where α, β are the two generators of the fundamental group. Then, since both $\rho(\alpha)$, $\rho(\beta)$ fix the point ∞, and the monodromy around the puncture $[\rho(\alpha), \rho(\beta)] = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ is the identity element, ρ is a degenerate representation. However, we can frame this representation, by sending the one lift of the puncture to the point $1 \in \mathbb{C}$ and then extending equivariantly. The image of the Farey set under the framing is then the set of integers $\mathbb{Z} \subset \mathbb{C}$, and the framing is nondegenerate by the definition above.

The image under the framing map β of a lift of a puncture is a fixed point for the peripheral monodromy around the puncture. If the monodromy around a puncture is elliptic or loxodromic, we can get a new framing by equivariantly choosing the other fixed point. We define a *flip* of a framing to be a change of framing at a subset of the interior punctures $\mathbb{P}$ of $\mathbb{S}$, which have peripheral monodromy that is either elliptic or loxodromic, by choosing of the other fixed point of that monodromy element. We note the following lemma, which is in essence Remark 4.4(v) of [3] (see also Lemma 9.4 of that paper):

Lemma 4.3 *A framed representation is nondegenerate if and only if all of its flips are nondegenerate.*

We note another useful lemma about framed representations, which is essentially taken from [3, Section 9]:

Lemma 4.4 *Given a framed representation $\widehat{\rho} = (\rho, \beta)$, the following are equivalent*:

(i) *The framed representation is nondegenerate.*

(ii) *We can flip the framing so that there exists an ideal triangulation such that the Fock–Goncharov coordinates associated to the triangulation are well defined.*

(iii) *We can flip the framing so that the image of the framing has at least three points in its image and there is no boundary component with two adjacent marked points having the same framing.*

Proof (i) $\Longrightarrow$ (ii) is a consequence of Theorem 9.1 of [3]. (ii) $\Longrightarrow$ (iii) follows since an ideal triangulation which gives well-defined Fock–Goncharov coordinates must have the property that the set of ideal vertices of a pair of adjacent ideal triangles has at least three distinct points. (iii) $\Longrightarrow$ (i) follows from Definition 4.2 of nondegeneracy. ☐

Remark It follows from the definition of the framing in the previous subsection that changing signs of a signed projective structure at a subset of the punctures in $\mathbb{P}$ precisely results in the flipping the framing of its framed monodromy at those punctures.

4.3 Characterizing the image of $\widehat{\Phi}$

We shall now prove Theorem 1.2 which characterizes the image of the monodromy map $\widehat{\Phi}$. We begin with the following observation:

Lemma 4.5 *Let Z be a hyperbolic structure on a connected marked and bordered surface with a nonempty set of marked points $\mathbb{M}$, such that it has at least one geodesic boundary component in the case that $|\mathbb{M}| = 1$. Then there exists an immersed ideal triangle T in Z with vertices in $\mathbb{M}$ that lifts to an embedded ideal triangle $\widetilde{T}$ in the universal cover of Z. Moreover, if $\mathcal{C}$ is a finite collection of isolated geodesic lines in Z with endpoints in $\mathbb{M}$, then we can choose T such that the interior of T is disjoint from each element of $\mathcal{C}$.*

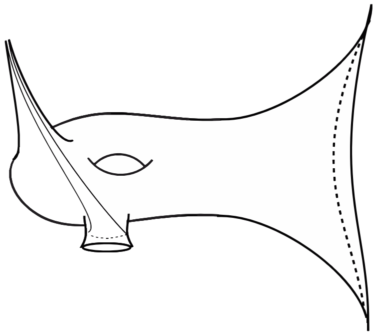


Figure 5: A hyperbolic surface Z as in Lemma 4.5. One can choose three arcs from the marked point p_0 (at the cusp end) to itself that wind around the geodesic boundary component and bound an immersed ideal triangle.

Proof When $|\mathbb{M}| \geq 3$, there exists an ideal triangle T embedded in Z with ideal vertices at three distinct points in $\mathbb{M}$: connect each pair of such points by arcs such that the arcs are pairwise disjoint, and then take their geodesic representatives. For the second statement, we can choose the geodesic sides of T to be either from $\mathcal{C}$, or disjoint from each geodesic line in $\mathcal{C}$; this would imply in particular that the interior of T does not intersect any geodesic line in $\mathcal{C}$.

In the case when $\mathbb{M}$ has exactly *two* points (say p, q), we consider the ideal triangle T where two of the geodesic sides coincide and is exactly the geodesic between p and q, and the third geodesic side is the geodesic arc from one of the points (p or q) to itself that goes around the other point. In the universal cover, such an ideal triangle will lift to an embedded ideal triangle. For the second statement, each geodesic in $\mathcal{C}$ starts and ends in the set $\{p, q\}$, and hence we can ensure that the two geodesics we chose to be sides of T, either coincide with elements of $\mathcal{C}$ or are disjoint from them.

Finally, assume that $\mathbb{M}$ has a single point p_0. Recall that in this case we also assume that the hyperbolic surface Z has a geodesic boundary component c. (See Figure 5.) Identify the universal cover $\widetilde{Z}$ as the geodesically convex subset of $\mathbb{H}^2$. The ideal boundary points of $\widetilde{Z}$ will include infinitely many lifts of p_0; this uses the fact that Z has a nontrivial closed geodesic c — lifts of arcs from p_0 to itself that twist around c a different number of times will lift to arcs between different lifts of p_0. Choose three such lifts of p_0, and connect them with geodesic lines; by convexity of $\widetilde{Z}$ this defines an embedded ideal triangle in the universal cover. Once again, we choose the three geodesic lines above to be either from the set $\mathcal{C}$ or disjoint from its elements. The resulting ideal triangle will have its interior disjoint from $\mathcal{C}$. □

We shall use the above lemma in the proof of Theorem 1.2, which recall, states the image of $\widehat{\Phi}$ is precisely the set of nondegenerate framed representations.

Proof of Theorem 1.2 The proof of one inclusion follows from the constructions in [18; 26], so we refer the reader to those papers for details. Briefly, let (ρ, β) be a given nondegenerate framed representation in $\widehat{\chi}(\mathbb{S}, \mathbb{M})$. Then, by Lemma 4.4, we can flip the framing to (ρ, β') and construct a triangulation of the universal cover of the surface $(\mathbb{S} \setminus \mathbb{M})$, such that the Fock–Goncharov coordinates for this triangulation are well defined. Then, we can construct the pleated plane and projective structure from this collection of Fock–Goncharov coordinates (as described in [18, Section 3; 25, Section 3; 26, Section 3]) to obtain a signed projective structure with framed monodromy (ρ, β'). Finally, we can change signing at punctures in $\mathbb{P}$ to recover the framed monodromy (ρ, β), since as noted in Section 4.2, changing signs is identical to a flip of the framing. So, in what follows we shall focus on proving the other inclusion.

Let (P, μ) be a signed meromorphic projective structure in $\mathcal{P}^{\pm}$, where μ denotes the choice of a signing on the subset of $\mathbb{P}$ having nonzero exponents. We know by Proposition 3.10, that the underlying unsigned projective structure P is obtained by grafting a pair (X, λ), where X is a hyperbolic structure on $(\mathbb{S}, \mathbb{M})$ and λ is a measured geodesic lamination. We shall show that the developing image of P has three distinct asymptotic values at the lifts of its marked points at F_∞. Since such an asymptotic value is a fixed point of the monodromy around that puncture, it follows that for some signing μ', the corresponding framing will have at least three distinct points in its image, and hence the framed representation of the signed projective structure (P, μ') is nondegenerate. Since μ is obtained by changing the sign of μ', the framed monodromy of (P, μ) is nondegenerate as well by Lemma 4.4. The fact that there is no boundary component with adjacent marked points having the same image under the framing follows from the asymptotics of the developing map at irregular singularities, and is discussed in [27], where it is attributed to [45, Chapter 8].

The easiest case is when there is an embedded ideal triangle T in $X \setminus \lambda$ whose boundary edges are bi-infinite leaves of the lamination λ with vertices in $\mathbb{M}$. This lifts to an ideal triangle $\widetilde{T}$ in the universal cover, with ideal vertices in F_∞. Before grafting the developing images of these three points are distinct; indeed, the developing image of $\widetilde{T}$ embeds in $\mathbb{C}\mathrm{P}^1$. Since the interior of $\widetilde{T}$ is disjoint from the lift $\widetilde{\lambda}$ of the bending lamination, its image under the developing map remains embedded after grafting. In particular, the images of the three ideal vertices will be distinct, and are asymptotic values of the developing map of the projective structure P. Thus, we conclude that the image of β has at least three points, and we are done.

The argument works when there is an embedded ideal triangle in the *universal cover* with vertices in F_∞ and that is disjoint from the lift of the bending lamination λ.

An observation that will be useful in what follows is:

Claim *An isolated leaf of λ is either a simple closed geodesic, or a bi-infinite geodesic between two points in $\mathbb{M}$. (The latter possibility includes the case when both ends of the geodesic spirals onto geodesic boundary components.)*

Proof of claim This is essentially Corollary 1.7.4 of [42] and follows from the structure theory of measured geodesic laminations (see, for example, [42, Corollary 1.7.3]): the only other possibility is to have a leaf spiraling at either (or both) ends onto a compactly supported geodesic lamination, but that is not possible since the finite measure of the accumulating leaf will endow that lamination with infinite transverse measure. □

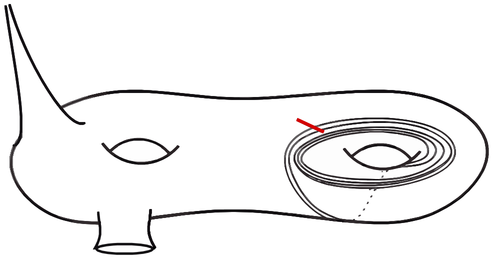


Figure 6: A possible hyperbolic surface X and geodesic lamination λ': the metric completion of $X \setminus \lambda'$ is shown in Figure 5.

Let $\lambda' \subset \lambda$ be the subset of the lamination supported in a compact part of the surface (ie away from the ends including the geodesic boundaries). (See Figure 6.) By the claim above, the leaves of $\lambda \setminus \lambda'$ are isolated bi-infinite geodesics between points of $\mathbb{M}$. Let $Y = \widehat{X \setminus \lambda'}$ be the metric completion of the complement of λ'. (See Figure 5.) Note that Y is a hyperbolic surface that contains ends of X corresponding to the points of $\mathbb{M}$, together with some additional ends that are adjacent to λ' on X, which we shall refer to as the "lamination-ends". The lamination-ends are either geodesic boundary components or crowns.

Let Z be a connected component of Y containing a nonempty set of points from $\mathbb{M}$, and let $\mathcal{C}$ be the collection of isolated leaves of the grafting lamination with endpoints in $\mathbb{M}$ contained in Z. We now apply Lemma 4.5 to obtain an embedded ideal triangle in the universal cover of X, with ideal vertices in F_∞, completely contained in the subset that is the lift of Z (and hence lying in the complement of the lift of λ). As observed above, the existence of such an embedded ideal triangle suffices to complete the proof.

The only case where Lemma 4.5 will not apply is if Z contains a single point of $\mathbb{M}$, and Z has no geodesic boundary components. In this case, there is a lamination-end of Z which is a crown. Recall that a lamination-end is adjacent to the compactly supported lamination λ' on X; the geodesic sides of the crown cannot be isolated leaves of the lamination λ', from the claim above. Hence, there are infinitely many (in fact uncountably many) leaves of λ' that are accumulating onto any such geodesic side. (Indeed, from the structure theory of geodesic laminations an arc transverse to λ' there will intersect it in a Cantor set.) In this case consider a homotopically nontrivial arc γ from p_0 to itself that intersects λ', but such that the transverse measure is small. Such a γ can be described thus: it starts from p_0, crosses the geodesic side of the lamination-end mentioned above, reaches one of the "gaps" in the Cantor-set cross-section that also belongs to Z, and subsequently remains in Z and returns to p_0. In the universal cover, the end-points of its lift $\tilde{\gamma}$ will be points $p_\pm \in F_\infty$ that are two lifts of p_0. (See Figure 7.)

Let $\widetilde{Z}_1$ and $\widetilde{Z}_2$ be the corresponding two lifts of Z which $\tilde{\gamma}$ starts and ends in, respectively. Before grafting, these regions of the universal cover develop into disjoint regions in $\mathbb{CP}^1$; in particular, the developing image of $p_\pm$ are distinct points. This latter fact remains true after grafting, since the transverse measure between $\widetilde{Z}_1$ and $\widetilde{Z}_2$ is small; this implies that for the developing map of P, the grafted region

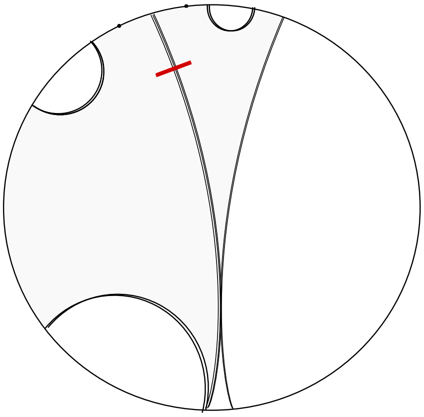


Figure 7: Two complementary regions (plaques) of the lift of the lamination separated by a transverse arc (shown in red) with small measure. If two points in F_∞ lie in the ideal boundary of such plaques, after grafting they will define two distinct asymptotic values of the developing map.

in between their images has small angular width. Thus we obtain two points in F_∞ whose developing images in $\mathbb{C}\mathrm{P}^1$ are distinct.

Repeating the argument for another (homotopically distinct) choice of arc from p_0 to itself with small intersection with λ', that lifts to an arc, say from p_- to another point $p' \in F_\infty$, we can conclude that the image of p_- and p' are distinct under the developing map of P. Then, a third arc from p_+ to p' will also have small transverse measure, since it will be at most the sum of the transverse measures between p_- and p' and between p_- and p_+. Once again, before grafting these images of the three points $\{p', p_-, p_+\}$ are distinct, and since after grafting the relative bending between them (which is determined by the transverse measures) is small, they remain distinct. We can then conclude that the three points have distinct images under the developing map of P. Since the framing β for P is determined by the asymptotic values of the developing map, its image has at least three points, and we are done. □

We shall now give a characterization of the representations underlying the framed representations in the image of $\widehat{\Phi}$. Let

$$\pi : \widehat{\chi}(\mathbb{S}, \mathbb{M}) \to \chi(\mathbb{S} \setminus \mathbb{P})$$

be the forgetful map to the $\mathrm{PSL}_2(\mathbb{C})$-representation variety of the punctured-surface group $\pi_1(\mathbb{S} \setminus \mathbb{P})$, and let $\Phi = \pi \circ \widehat{\Phi}$ be the unframed monodromy map. As a corollary of Theorem 1.2, we can characterize the image Φ as follows (see Theorem A of [18] for the case when $\mathbb{P} = \mathbb{M}$):

Corollary 4.6 *Recall that m denotes the number of punctures of $\mathbb{S}$, k is the number of boundary components, and $\{n_i - 2\}_{1 \le i \le k}$ are the numbers of marked points on the boundary components, so that $N = \sum_{i=1}^{k}(n_i - 2)$ denotes the total number of marked boundary points. If $N \ge 3$, then any representation is in the image of Φ, ie the unframed monodromy map is surjective. For $N \le 2$, a representation ρ lies in the image of Φ if and only if one of the following hold:*

- *ρ is a nondegenerate representation.*
- *$k = 0$ and ρ is a degenerate representation with at least one apparent singularity, excluding the following cases:*
  - *ρ is the trivial representation, for $g > 0$ and $m = 1$ or $m = 2$,*
  - *the image of ρ is a group of order 2 and $g > 0$, $m = 1$.*
- *$N = 1$ and ρ is a degenerate representation, excluding the following cases:*
  - *ρ is the trivial representation and $m = 0$ or $m = 1$,*
  - *the image of ρ is a group of order 2 and $m = 0$.*
- *$N = 2$ and ρ is a degenerate representation, excluding the case where ρ is the trivial representation and $m = 0$.*

Proof By Theorem 1.2, it suffices to show that these are the only representations that can be framed to obtain a nondegenerate framed representation. This is a consequence of the following four cases/observations:

(i) If ρ is nondegenerate, we can arbitrarily assign a framing to obtain a nondegenerate (ρ, β), by the first part of [25, Proposition 4.1]. If $N \geq 3$, then again we can arbitrarily assign a framing to the marked points on boundaries so that the image has at least 3 distinct points, making it nondegenerate.

(ii) In the case $k = 0$, we are reduced to the cases considered in Theorem A of [18]: If ρ is degenerate and has no apparent singularities, any framed representation will also be degenerate by [25, Proposition 4.1], which cannot arise as the framed monodromy of a projective structure by [3, Theorem 6.1]. If ρ is degenerate and has an apparent singularity, we reduce to the following cases:

(a) **ρ is the trivial representation** In this case, is it easy to check that by Definition 4.2 the framed representation will be nondegenerate if and only if the image of the framing consists of at least 3 points. Thus, one can define a nondegenerate framing in this case if and only if $m \geq 3$.

(b) **ρ is not the trivial representation and $m \geq 2$** In this case, suppose that the punctures are $\{p_1, p_2, \dots, p_m\}$ and p_1 is an apparent singularity. We shall use the same labels for their lifts to a fundamental domain in the universal cover of $\mathbb{S}$. Let Q denote the set of points that are fixed by all elements of the image of ρ, as in the definition of a degenerate representation. Now, since the image of p_1 under the framing can be arbitrary, and ρ is nontrivial, we can choose a point to be the image of p_1 and extend equivariantly such that there are at least two distinct images of the lifts of p_1 under the ρ-equivariant framing, and moreover these images are disjoint from Q. Thus, the image of the resulting framing has at least three points, ensuring that it is nondegenerate.

(c) **ρ is not the trivial representation and $m = 1$** Let the only puncture be p_1, which is also an apparent singularity. If the image of ρ is a group of order two, then clearly the image of the framing can have at most two points, and those set of points will be fixed by the image of ρ. Thus, it would be a degenerate framed representation by Definition 4.2. On the other hand if the image of ρ is a group of order more than two, clearly we can choose a point $p \in \mathbb{C}\mathrm{P}^1$ such that its orbit under the action of this group consists of at least three points. Thus, we can define the framing by mapping a lift of the puncture to p and extending equivariantly; the resulting framed representation is nondegenerate because the image of the framing has at least three points. So, in this case, the representation has such a framing if and only if the image of ρ is not a group of order two.

(iii) Suppose the total number N of marked points on boundaries be equal to 1, and ρ is a degenerate representation. If the image of ρ is a group of order at least 3, we can frame the marked boundary points so that its image under the action of ρ also has at least 3 points, hence we would obtain a nondegenerate framing. Now we consider the rest of the cases:

(a) **ρ is the trivial representation** Clearly if $m \leq 1$, the image of the framed representation has exactly 2 points which are fixed points of the monodromy around all loops, hence the representation is degenerate. If $m \geq 2$, then we can arbitrarily assign a framing to the punctures and the marked boundary points to obtain a framing with 3 points, getting a nondegenerate representation.

(b) **The image of ρ is a group of order 2** If $m = 0$, the image of any framing will necessarily consist of at most 2 points. Moreover, since the image of ρ is abelian, the monodromy around the boundary must be identity. Hence, it follows that the framed representation is degenerate. If $m \geq 1$, then we can define a framing that sends the (lift of the) puncture to a point fixed by the monodromy around all loops, and sends a lift of the marked point on the boundary to a point in $\mathbb{C}\mathrm{P}^1$ whose orbit under the image of ρ has order 2. Then again, we obtain a framing with 3 points, making it nondegenerate.

(iv) Finally, if the total number of marked boundary points is 2, and there is at least 1 apparent singularity, we can assign points in $\mathbb{C}\mathrm{P}^1$ to lifts of the marked boundary to a fundamental domain arbitrarily, and extend equivariantly to obtain a framing with at least 3 points, making it a nondegenerate framing. If there are no punctures, then again we can obtain 3 points in the image of the framing if ρ is not the trivial representation. If ρ is trivial, then clearly any framed representation will be degenerate. □

Remark The projective structures in [18] and their resulting *framed* monodromies are consistent with the characterization in Theorem 1.2, ie they are nondegenerate framed representations. For example, in the case that the number of punctures $m \geq 2$, and ρ is a nontrivial affine representation (ie with image in the affine group $\mathrm{Aff}(\mathbb{C}) \subset \mathrm{PSL}_2(\mathbb{C})$), then ρ is a degenerate representation. However, the corresponding affine structure with holonomy ρ constructed in the proof of [18, Theorem C] does define a nondegenerate framing by considering the asymptotic values of the developing map as in Section 4.1.2.

4.4 Showing $\widehat{\Phi}$ is a local homeomorphism

In this section, we prove that the monodromy map is a local homeomorphism (Theorem 1.4). For the case of projective structures with only irregular singularities, this was proved in [27], and we shall use results from there, together with our grafting theorem (Theorem 1.1).

Before starting the proof, we record the following observation that we shall use; this is also [8, Lemma 5.4].

Lemma 4.7 *Let E be either a cuspidal end or a collar neighborhood of a geodesic boundary component of a hyperbolic surface, and let L_1 and L_2 be two distinct collections of pairwise-disjoint weighted geodesics incident at that end, ie either going into the cusp or spiraling onto the boundary geodesic. Suppose that*

- (i) *the total sum of weights for both collections are the same,*
- (ii) *in case of a geodesic boundary end, the leaves spiral in the same direction,*
- (iii) *the resulting monodromy around the puncture is the same in both cases.*

Then, the developing maps on the universal cover of the end E after grafting along L_1 and L_2, respectively, are $\mathbb{Z}$-equivariantly isotopic.

Proof Let $\widetilde{E}$ be a lift of E to the universal cover. We shall focus our attention on a fundamental domain $F \subset \widetilde{E}$ of the $\mathbb{Z}$-action. The geodesic lines in L_1 and L_2 lift to a collection of pairwise-disjoint geodesic lines passing through F, each asymptotic to the same point in the ideal boundary. Here, we choose the fundamental domain F such that there are finitely many such lifts in F. (See Figure 2 for the case of a geodesic boundary end.) Grafting along L_1 (and L_2) inserts lunes along these lines in the universal cover, of angular widths equal to the corresponding weights. The intermediate regions of F can be thought of as lunes of zero angular width (ie regions bounded on two sides by circular arcs incident to a common point in $\mathbb{C}\mathrm{P}^1$ where they share a tangent line). After grafting, the new fundamental domain of the $\mathbb{Z}$-action on the target is a portion of a lune in $\mathbb{C}\mathrm{P}^1$ of angular width equal to the sum of the weights on the grafting lines. Since we are also assuming that the monodromy in both cases is identical, the two circular arcs bounding this fundamental domain can be taken to be exactly the same. The remaining boundary edge of F might result in distinct arcs after grafting (that depends on the locations of the grafting lines). However these arcs will be isotopic to each other, and this isotopy can be extended to a $\mathbb{Z}$-equivariant isotopy on $\widetilde{E}$. □

Remark The monodromy around the puncture can differ even if the hypotheses (i) and (ii) in the statement of Lemma 4.7 are satisfied, as the following example illustrates: Let $E = H/\langle z \mapsto z+1\rangle$ be a cusp, let L_1 comprise two geodesics, both with weight π, which lift to the vertical lines $\mathrm{Re}(z) = 0$ and $\mathrm{Re}(z) = \frac{1}{3}$, and let L_2 comprise two geodesics with weight π, this time lifting to vertical lines $\mathrm{Re}(z) = 0$ and $\mathrm{Re}(z) = \frac{1}{2}$. To compute the monodromy from (8) and (9), we use $a_1 = 0$, $\omega_1 = \omega_2 = e^{i\pi} = -1$, which implies that $c = 1 - 2a_2$ where a_2 is the real coordinate of the second geodesic. Thus, in the first case $a_2 = \frac{1}{3} = c$ and the monodromy is a parabolic element, while in the second case $a_2 = \frac{1}{2}$, $c = 0$ and the monodromy is the identity element.

Proof of Theorem 1.4 We shall use the strategy of the proof in [27]. By the invariance of domain it suffices to show that $\widehat{\Phi}$ is locally injective.

Let $\widehat{\rho}$ be a framed representation in $\widehat{\chi}(\mathbb{S},\mathbb{M})$ that lies in the image of $\widehat{\Phi}$ and let $P \in \mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ such that $\widehat{\Phi}(P)=\widehat{\rho}$.

By Theorem 1.1, we know that there exists $(X,\lambda) \in \mathcal{T}^{\pm} \times \mathcal{ML}^{\pm}$ such that we obtain the projective structure P by grafting X along λ. We shall show that there is a neighborhood U of P in $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$ such that if $P' = \widehat{\mathrm{Gr}}(X',\lambda')$ for some $(X',\lambda') \in U$ has the same framed monodromy $\widehat{\rho}$, then $X = X'$ and $\lambda = \lambda'$.

By Corollary 3.5 we know that the monodromy around any regular puncture determines the type of the corresponding end of X and X' (including, in the case of a geodesic boundary end, its length) and the total weight (modulo 2π) of the leaves of λ that are incident at that end. Since λ and λ' are close to each other, the homotopy classes of the leaves are identical and the corresponding transverse measures are close; in our context this implies that the leaves incident to the cusp or geodesic boundary end of X (and X') will have *identical* total weights. Moreover, in the case of a geodesic boundary end with leaves of λ of *nonzero* weight spiraling into it, one can ensure that the neighborhood U in $\mathcal{T}^{\pm} \times \mathcal{ML}^{\pm}$ is small enough such that the direction of spiraling of the leaves are the same. Since, as already mentioned, the geometry of such an end E is the same for X and X', we conclude from Lemma 4.7 that one can modify P and P' by an isotopy (this does not change the corresponding points in the deformation space) such that the restriction of their developing maps to the universal cover of E are identical.

Proposition 6.3 of [27] asserts that the same holds at irregular singularity: for each such singularity, the framed representation $\widehat{\rho}$ determines the corresponding hyperbolic crowns of X and X', as well as the leaves (and weights) of λ and λ' that intersect the crown end.

Thus, as in [27], we are reduced to an application of the Ehresmann–Thurston principle for manifolds with boundary (see, for example, Theorem I.1.7.1 of [17] or Proposition 1 of [15]). Briefly, let S_0 be the compact surface-with-boundary obtained by removing the ends of the marked-and-bordered surface $\mathbb{S}$, and consider the space $\mathcal{D}$ of developing maps for projective structures on S_0 such that on each boundary component they all restrict to the same map. Then the Ehresmann–Thurston principle asserts that there is a neighborhood of any developing map D_0 in this relative deformation space $\mathcal{D}$ such that any other developing map in that neighborhood with identical holonomy will be equivariantly isotopic to D_0. We apply this to conclude that the projective structures P and P' determine the same point in $\mathcal{P}^{\pm}(\mathbb{S},\mathbb{M})$. □

4.5 Proof of Corollary 1.5

Following the remark at the end of Section 2.3.3, there are nonconstant holomorphic functions $r_p : \mathcal{P}^{\pm} \to \mathbb{C}$ that map signed projective structures to the exponent at each puncture p in $\mathbb{P}$ (which corresponds to a regular singularity). We know from the computations in the proofs of Lemmas 3.3 and 3.7 that if a puncture p is an apparent singularity for the projective structure, then the exponent $r_p = 2\pi i n$ for some $n \in \mathbb{Z}$. Therefore, it follows that the projective structures having apparent singularities are contained in the analytic subset of $\mathcal{P}^{\pm}$ locally cut out by finitely many equations of the form $\{r_p = 2\pi i n_p : p \in \mathbb{P}\}$ for a tuple of integers $(n_p)_{p\in\mathbb{P}} \in \mathbb{Z}^{|\mathbb{P}|}$.

By Theorem 1.1 of [3], we know that $\widehat{\Phi}$ is a holomorphic map on the subset $\mathcal{P}^* \subset \mathcal{P}^{\pm}$ of signed projective structures with no apparent singularities. Therefore, by Riemann's removable singularity theorem (see, for example, Proposition 1.1.7 of [32]), it follows that $\widehat{\Phi}$ is a holomorphic map on all of $\mathcal{P}^{\pm}$. Combining this with Theorem 1.4, we conclude that the map $\widehat{\Phi}$ is a local biholomorphism. □

References

[1] **D G L Allegretti**, *Stability conditions, cluster varieties, and Riemann–Hilbert problems from surfaces*, Adv. Math. 380 (2021) art. id. 107610 MR

[2] **D G L Allegretti**, *Stability conditions and Teichmüller space*, Math. Ann. 390 (2024) 3827–3890 MR

[3] **D G L Allegretti**, **T Bridgeland**, *The monodromy of meromorphic projective structures*, Trans. Amer. Math. Soc. 373 (2020) 6321–6367 MR

[4] **C L Alley**, *On the monodromy of meromorphic cyclic opers on the Riemann sphere*, Int. Math. Res. Not. 2021 (2021) 16693–16725 MR

[5] **S Baba**, *2π-grafting and complex projective structures, I*, Geom. Topol. 19 (2015) 3233–3287 MR

[6] **S Baba**, *2π-grafting and complex projective structures with generic holonomy*, Geom. Funct. Anal. 27 (2017) 1017–1069 MR

[7] **S Baba**, *On Thurston's parameterization of $\mathbb{C}\mathrm{P}^1$-structures*, from "In the tradition of Thurston — geometry and topology" (K Ohshika, A Papadopoulos, editors), Springer (2020) 241–254 MR

[8] **S Baba**, *Neck-pinching of $\mathbb{C}\mathrm{P}^1$-structures in the $\mathrm{PSL}_2\mathbb{C}$-character variety*, J. Topol. 18 (2025) art. id. e70010 MR

[9] **M Bainbridge**, **C Johnson**, **C Judge**, **I Park**, *Haupt's theorem for strata of abelian differentials*, Israel J. Math. 252 (2022) 429–459 MR

[10] **S A Ballas**, **P L Bowers**, **A Casella**, **L Ruffoni**, *Tame and relatively elliptic $\mathbb{CP}^1$-structures on the thrice-punctured sphere*, Algebr. Geom. Topol. 24 (2024) 4589–4650 MR

[11] **F Bonahon**, *Geodesic laminations on surfaces*, from "Laminations and foliations in dynamics, geometry and topology", Contemp. Math. 269, Amer. Math. Soc., Providence, RI (2001) 1–37 MR

[12] **T Bridgeland**, **I Smith**, *Quadratic differentials as stability conditions*, Publ. Math. Inst. Hautes Études Sci. 121 (2015) 155–278 MR

[13] **D Chen**, **G Faraco**, *Period realization of meromorphic differentials with prescribed invariants*, Forum Math. Sigma 12 (2024) art. id. e90 MR

[14] **S Chenakkod**, **G Faraco**, **S Gupta**, *Translation surfaces and periods of meromorphic differentials*, Proc. Lond. Math. Soc. 124 (2022) 478–557 MR

[15] **J Danciger**, *A geometric transition from hyperbolic to anti-de Sitter geometry*, Geom. Topol. 17 (2013) 3077–3134 MR

[16] **D Dumas**, *Complex projective structures*, from "Handbook of Teichmüller theory, II", IRMA Lect. Math. Theor. Phys. 13, Eur. Math. Soc., Zürich (2009) 455–508 MR

[17] **D B A Epstein**, **A Marden**, *Convex hulls in hyperbolic space, a theorem of Sullivan, and measured pleated surfaces*, from "Analytical and geometric aspects of hyperbolic space", London Math. Soc. Lecture Note Ser. 111, Cambridge Univ. Press (1987) 113–253 MR

[18] **G Faraco**, **S Gupta**, *Monodromy of Schwarzian equations with regular singularities*, Geom. Topol. 29 (2025) 549–617 MR

[19] **V Fock**, **A Goncharov**, *Moduli spaces of local systems and higher Teichmüller theory*, Publ. Math. Inst. Hautes Études Sci. 103 (2006) 1–211 MR

[20] **V V Fock**, **A B Goncharov**, *Dual Teichmüller and lamination spaces*, from "Handbook of Teichmüller theory, I", IRMA Lect. Math. Theor. Phys. 11, Eur. Math. Soc., Zürich (2007) 647–684 MR

[21] **E Frenkel**, **D Ben-Zvi**, *Vertex algebras and algebraic curves*, 2nd edition, Mathematical Surveys and Monographs 88, Amer. Math. Soc., Providence, RI (2004) MR

[22] **D Gabai**, *Almost filling laminations and the connectivity of ending lamination space*, Geom. Topol. 13 (2009) 1017–1041 MR

[23] **D Gallo**, **M Kapovich**, **A Marden**, *The monodromy groups of Schwarzian equations on closed Riemann surfaces*, Ann. of Math. 151 (2000) 625–704 MR

[24] **W M Goldman**, *Locally homogeneous geometric manifolds*, from "Proceedings of the International Congress of Mathematicians, II", Hindustan Book Agency, New Delhi (2010) 717–744 MR

[25] **S Gupta**, *Monodromy groups of $\mathbb{C}\mathrm{P}^1$-structures on punctured surfaces*, J. Topol. 14 (2021) 538–559 MR

[26] **S Gupta**, **M Mj**, *Monodromy representations of meromorphic projective structures*, Proc. Amer. Math. Soc. 148 (2020) 2069–2078 MR

[27] **S Gupta**, **M Mj**, *Meromorphic projective structures, grafting and the monodromy map*, Adv. Math. 383 (2021) art. id. 107673 MR

[28] **A E Hatcher**, *Measured lamination spaces for surfaces, from the topological viewpoint*, Topology Appl. 30 (1988) 63–88 MR

[29] **O Haupt**, *Ein Satz über die Abelschen Integrale 1. Gattung*, Math. Z. 6 (1920) 219–237 MR

[30] **D A Hejhal**, *Monodromy groups and linearly polymorphic functions*, Acta Math. 135 (1975) 1–55 MR

[31] **D A Hejhal**, *Monodromy groups for higher-order differential equations*, Bull. Amer. Math. Soc. 81 (1975) 590–592 MR

[32] **D Huybrechts**, *Complex geometry: an introduction*, Springer (2005) MR

[33] **Y Kamishima**, **S P Tan**, *Deformation spaces on geometric structures*, from "Aspects of low-dimensional manifolds", Adv. Stud. Pure Math. 20, Kinokuniya, Tokyo (1992) 263–299 MR

[34] **M Kapovich**, *Periods of abelian differentials and dynamics*, from "Dynamics: topology and numbers" (P Moree, A Pohl, L Snoha, T Ward, editors), Contemp. Math. 744, Amer. Math. Soc., Providence, RI (2020) 297–315 MR

[35] **R S Kulkarni**, **U Pinkall**, *A canonical metric for Möbius structures and its applications*, Math. Z. 216 (1994) 89–129 MR

[36] **L Lang**, *Harmonic tropical morphisms and approximation*, Math. Ann. 377 (2020) 379–419 MR

[37] **S Lang**, *Algebra*, 3rd edition, Graduate Texts in Mathematics 211, Springer, New York (2002) MR

[38] **T Le Fils**, *Periods of abelian differentials with prescribed singularities*, Int. Math. Res. Not. 2022 (2022) 5601–5616 MR

[39] **T Le Fils**, *Holonomy of complex projective structures on surfaces with prescribed branch data*, J. Topol. 16 (2023) 430–487 MR

[40] **F Luo**, *Monodromy groups of projective structures on punctured surfaces*, Invent. Math. 111 (1993) 541–555 MR

[41] **G Nascimento**, *Monodromies of projective structures on surface of finite-type*, Geom. Dedicata 218 (2024) art. id. 1 MR

[42] **R C Penner**, **J L Harer**, *Combinatorics of train tracks*, Annals of Mathematics Studies 125, Princeton University Press (1992)

[43] **H P de Saint-Gervais**, *Uniformization of Riemann surfaces: revisiting a hundred-year-old theorem*, European Mathematical Society, Zürich (2016) MR

[44] **T Sérandour**, *Meromorphic projective structures, opers and monodromy*, Int. Math. Res. Not. 2025 (2025) rnaf199 MR

[45] **Y Sibuya**, *Global theory of a second order linear ordinary differential equation with a polynomial coefficient*, North-Holland Mathematics Studies 18, North-Holland, Amsterdam (1975) MR

[46] **H Tanigawa**, *Grafting, harmonic maps and projective structures on surfaces*, J. Differential Geom. 47 (1997) 399–419 MR

[47] **D Thurston**, *On geometric intersection of curves in surfaces*, unpublished notes

The Mathematical Institute, Oxford University
Oxford, United Kingdom

Department of Mathematics, Indian Institute of Science
Bangalore, India

`spandan.ghosh@maths.ox.ac.uk`, `subhojoy@iisc.ac.in`

Received: 4 December 2023 Revised: 30 November 2024

Algebraic & Geometric Topology **25**:8 (2025) 4827–4838
DOI: 10.2140/agt.2025.25.4827
Published: 20 November 2025

A new twist on modular links from an old perspective

KHANH LE

We show that the complement of arithmetic modular links found by Pinsky, Purcell and Rodríguez-Migueles (Pacific J. Math. 327 (2023) 337–358) is homeomorphic to the complement of augmented chain links. In particular, these link complements arise as n-fold cyclic covers of the Whitehead link complement.

57K10, 57K32

1 Introduction

The *modular surface* Σ_{Mod} is an orbifold obtained as the quotient space of the hyperbolic plane $\mathbb{H}^2$ by the modular group $\mathrm{PSL}_2(\mathbb{Z})$. Since the action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{H}^2$ is by orientation-preserving isometries, Σ_{Mod} is an oriented 2-orbifold equipped with a hyperbolic metric. Any closed oriented geodesic $\bar{\gamma}(t)$ on Σ_{Mod} has a canonical lift $\gamma(t) := (\bar{\gamma}(t), \bar{\gamma}'(t))$ to the unit tangent bundle $\mathrm{UT}(\Sigma_{\mathrm{Mod}})$. Milnor showed that $\mathrm{UT}(\Sigma_{\mathrm{Mod}})$ is homeomorphic to the complement of the trefoil knot $T_{2,3}$ in S^3 [12]. Therefore, every nonempty finite collection of canonical lifts $\Gamma \subset \mathrm{UT}(\Sigma_{\mathrm{Mod}})$ of oriented closed geodesics in Σ_{Mod} together with the trefoil knot determines an $(n+1)$-component link $\Gamma \cup \{T_{2,3}\}$ in S^3 for $n \geq 1$. Following Ghys [10], we refer to the collection Γ, without the trefoil knot, as a *modular link* when $|\Gamma| \geq 2$ and *modular knot* when $|\Gamma| = 1$. Here $|\cdot|$ denotes the number of connected components. The *complement of modular links* refers to $M_\Gamma := \mathrm{UT}(\Sigma_{\mathrm{Mod}}) \setminus \Gamma$. For emphasis, the complement of the modular link Γ is the complement of the $(n+1)$-component link $\Gamma \cup \{T_{2,3}\}$ in S^3 where $|\Gamma| = n$ is the number of components of the modular link.

Modular links have attracted attention of mathematicians due to their connections to dynamics, low-dimensional topology and number theory. For example, in [10], Ghys showed that the isotopy classes of modular knots coincide with the isotopy classes of Lorenz knots which are periodic orbits of a 3-dimensional differential equation [3]. Furthermore, Ghys proved that the linking number in S^3 between the canonical lift γ to $\mathrm{UT}(\Sigma_{\mathrm{Mod}})$ of an oriented closed geodesic $\bar{\gamma}$ in Σ_{Mod} and the trefoil knot $T_{2,3}$ is given by the Rademacher function, a classical arithmetic function coming from number theory [10]. The latter result has been generalized to the setting of arbitrary (p, q, ∞)-triangle groups in [11].

The complement of modular links M_Γ is known to be hyperbolic [9]. More recently, there have been many works relating the hyperbolic volume to the length of the geodesics [2; 5; 6; 14]. Recently, Pinsky, Purcell and Rodríguez-Migueles [13] found an infinite family $\mathcal{F}\Sigma_{\mathrm{Mod}}$ of modular links whose complement M_Γ

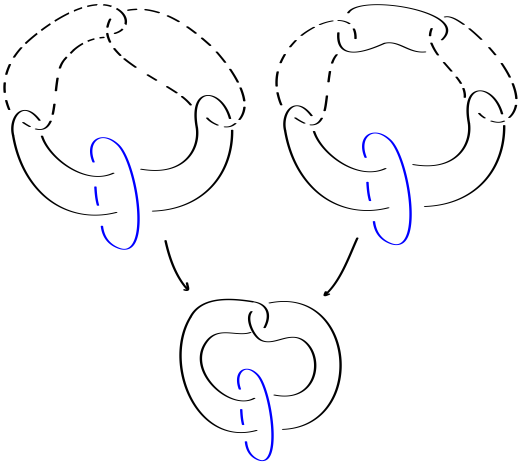


Figure 1: The Whitehead link $S^3 \setminus C_1$, a 3-fold cyclic cover $S^3 \setminus C_3$ (top left) and a 4-fold cyclic cover $S^3 \setminus C_4$ (top right) both branched over the blue component. Forgetting the dotted components in $S^3 \setminus C_3$ and $S^3 \setminus C_4$, we obtain two split links whose complement are handlebodies of genus 2 and 3, respectively. Consequently, we obtain surjective homomorphisms $\pi_1(S^3 \setminus C_3) \to F_2$ and $\pi_1(S^3 \setminus C_4) \to F_3$ in each case.

admits an arithmetic hyperbolic structure [13, Theorem 1.1]. For any $n \geq 3$, the family $\mathcal{F}\Sigma_{\mathrm{Mod}}$ contains at least two modular links with n-component. See Section 2.2 for a precise parametrization of modular links in $\mathcal{F}\Sigma_{\mathrm{Mod}}$ in terms of the Farey graph. The hyperbolic structures of M_Γ for any collection Γ in $\mathcal{F}\Sigma_{\mathrm{Mod}}$ are all commensurable to that of the Bianchi orbifold $\mathbb{H}^3/\operatorname{PSL}_2(\mathbb{Z}[i])$. Furthermore, there is a unique modular knot Γ_0 in the family $\mathcal{F}\Sigma_{\mathrm{Mod}}$. The complement M_{Γ_0} is known to be homeomorphic to that of the Whitehead link [13]. In general, it is an open question that Γ_0 is the only arithmetic modular knot [13].

The main result of this paper is to explicitly identify the complements of modular links in $\mathcal{F}\Sigma_{\mathrm{Mod}}$ as the complements of augmented chain links in S^3. These *augmented chain link complements* $S^3 \setminus C_n$ can be obtained by taking the n-fold cyclic cover branched over the unknotted component of the Whitehead link $S^3 \setminus C_1$; see Figure 1. In particular, C_n is a link in S^3 with $n+1$ components.

Theorem 1.1 *Let $\Gamma \in \mathcal{F}\Sigma_{\mathrm{Mod}}$ be an n-component modular link. The complement M_Γ is homeomorphic to the complement $S^3 \setminus C_n$.*

Using the work of Cooper and Long [4], we obtain the following corollary of Theorem 1.1:

Corollary 1.2 *Let $\Gamma \in \mathcal{F}\Sigma_{\mathrm{Mod}}$. The complement M_Γ fibers. Furthermore, if $|\Gamma| \notin \{1, 2, 3, 5\}$, then the complement M_Γ contains a closed embedded essential surface.*

The fact that the complement of modular links fibers was shown by Dehornoy in [8]. In fact, Dehornoy proved a much more general fact: the complement of every finite collection of periodic orbits of the geodesic flow on the unit tangent bundle of the triangle orbifold (p, q, ∞) fibers [8, Corollary 1.5].

Modular knots, without the trefoil component, considered in [13] are examples of Berge knots, namely the family of knots which lie as simple closed curves on the fiber of the trefoil knot complement. Chain links have also played an important role in the study of the topology and geometry of these Berge knots. In particular, Baker gave a surgery description of Berge knots on the fiber of the trefoil knot using chain links [1, Proposition 3.1]. Using this, he proved that this family of Berge knots contains hyperbolic knots with arbitrary large volume [1, Theorem 4.1]. As a consequence, there is no surgery description for these Berge knots on a single link in S^3 [1].

Acknowledgements

The author would like to thank Alan Reid for helpful conversations, his enthusiasm for the project, and for his comments on an earlier draft of the paper. The author would also like to thank Neil Hoffman and Ken Baker for some helpful conversations regarding [1]. Figure 4 is drawn using SnapPy [7]. The author thanks the anonymous referee for pointing out some mistakes in the previous version of the paper and for making suggestions that have improved the exposition.

2 Preliminaries

We begin by stating some definitions, collecting some standard facts about modular links.

2.1 Definitions and background

The *modular surface*, Σ_{Mod}, is the quotient space of $\mathbb{H}^2$ by the group $\mathrm{PSL}_2(\mathbb{Z})$. This group is generated by two elliptic isometries: U which rotates about i by an angle of π, and V which rotates about $\frac{1}{2}\big(1+i\sqrt{3}\big)$ by an angle of $\frac{2}{3}\pi$. As elements of $\mathrm{PSL}_2(\mathbb{Z})$, U and V have the form

$$U=\pm\begin{pmatrix}0&-1\\1&0\end{pmatrix}\quad\text{and}\quad V=\pm\begin{pmatrix}0&-1\\1&-1\end{pmatrix}.$$

A fundamental domain of the action of $\mathrm{PSL}_2(\mathbb{Z})$ is the triangle with a real vertex at $\frac{1}{2}\big(1+i\sqrt{3}\big)$ and two ideal vertices at 0 and ∞; see Figure 2. The hyperbolic metric on $\mathbb{H}^2$ descends to a hyperbolic metric on Σ_{Mod} with two points of cone angles π and $\frac{2}{3}\pi$ and a single cusp. An oriented simple closed geodesic on Σ_{Mod} corresponds to a conjugacy class of a primitive hyperbolic elements in $\mathrm{PSL}_2(\mathbb{Z})$. Each oriented simple closed geodesic γ on Σ_{Mod} has a representative in the corresponding conjugacy class that admits a factorization into a product of

$$L=\pm\begin{pmatrix}1&1\\0&1\end{pmatrix}\quad\text{and}\quad R=\pm\begin{pmatrix}1&0\\1&1\end{pmatrix}.$$

Therefore, we associate to an oriented simple closed geodesic γ in Σ_{Mod}: a word, w_γ, in the positive powers of L and R that is not a power of any subword. The correspondence between γ and w_γ is well defined up to a cyclic permutation of w_γ.

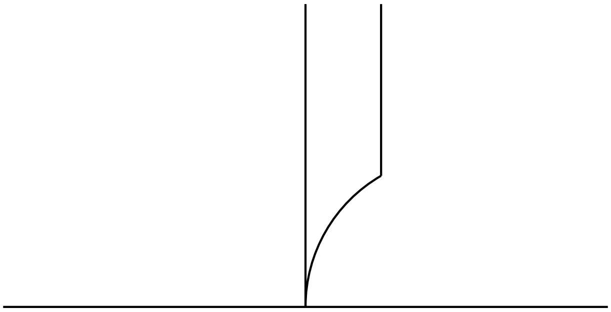


Figure 2: A fundamental domain of Σ_{Mod} in $\mathbb{H}^2$.

Since Σ_{Mod} comes equipped with a hyperbolic metric, there exists a natural flow on the unit tangent bundle $\mathrm{UT}(\Sigma_{\mathrm{Mod}})$ which is the geodesic flow Ψ_t defined as follows. Given a pair of a point and a unit vector based at the point, $(x, v) \in \mathrm{UT}(\Sigma_{\mathrm{Mod}})$, the geodesic flow moves the point (x, v) in unit speed along the geodesic starting at x tangent to v. Each oriented simple geodesic γ on Σ_{Mod} has a canonical lift to $\mathrm{UT}(\Sigma_{\mathrm{Mod}})$. The periodic orbits of Ψ_t on $\mathrm{UT}(\Sigma_{\mathrm{Mod}})$ correspond precisely to the canonical lift to $\mathrm{UT}(\Sigma_{\mathrm{Mod}})$ of oriented simple geodesics on Σ_{Mod}.

As noted in the introduction, $\mathrm{UT}(\Sigma_{\mathrm{Mod}})$ is homeomorphic to the complement of the trefoil knot $S^3 \setminus T_{2,3}$. In [10], Ghys showed that periodic orbits of Ψ_t can be isotoped to lie on a branched surface in $S^3 \setminus T_{2,3}$ which is known as the Lorenz template, $\mathcal{T}$; see Figure 3.

The Lorenz template supports a flow which can parametrized as follows. We identify the branching locus of the surface with the open interval $(0, 1)$. Starting at any point $x < \frac{1}{2}$, the flow line follows the left side of the template and returns to the branching locus at the point $2x \bmod 1$. If $x > \frac{1}{2}$, the flow line follows the right side of the template and comes back to the branching locus at the point $2x \bmod 1$. Any periodic orbit of this flow can be determined by a periodic orbit of the times-2 map on the interval $[0, 1]$. Given a sequence of LR-word w_γ, we can obtain the corresponding point in $(0, 1)$ by converting LR into a

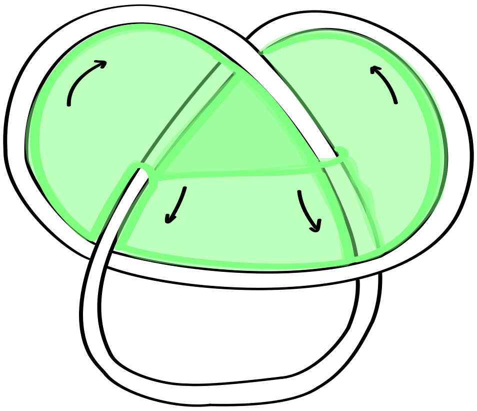


Figure 3: The modular template $\mathcal{T}$ together with a flow.

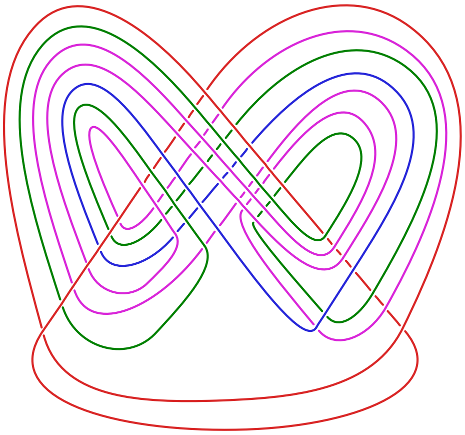


Figure 4: The 3-component modular link $\{LR, L^2R^2, L^2RLR^2\}$ (where these components are drawn in blue, green, and magenta, respectively) and the trefoil knot $T_{2,3}$ (red). The LR-component corresponds to the sequence of periodic orbit $\left\{\frac{1}{3}, \frac{2}{3}\right\}$ on $(0,1)$. The L^2R^2-component corresponds to the sequence of periodic orbit $\left\{\frac{1}{5}, \frac{2}{5}, \frac{4}{5}, \frac{3}{5}\right\}$ on $(0,1)$. Finally, the L^2RLR^2-component corresponds to the sequence of periodic orbit $\left\{\frac{11}{63}, \frac{22}{63}, \frac{44}{63}, \frac{25}{63}, \frac{50}{63}, \frac{37}{63}\right\}$ on $(0,1)$.

binary sequence by the rule $L \mapsto 0$ and $R \mapsto 1$. Let $\bar{w}_\gamma$ be the decimal number that corresponds to the binary sequence and $|w_\gamma|$ be the length of the LR-word. The point in $(0,1)$ that corresponds to w_γ is given by

$$\frac{\bar{w}_\gamma}{2^{|w_\gamma|} - 1}.$$

Therefore, given a collection of LR-words representing a modular link, we can draw the modular link on the Lorenz template $\mathcal{T}$ by computing the corresponding sequences of periodic orbit on $(0,1)$ and connect them by the flow line on $\mathcal{T}$. For an example of a 3-component modular link $\{LR, L^2R^2, L^2RLR^2\}$, see Figure 4.

2.2 A construction of arithmetic modular links

Now we will review the construction of a family of arithmetic modular links $\mathcal{F}\Sigma_{\mathrm{Mod}}$ from [13]. First consider the six-fold cyclic cover of Σ_{Mod} by the once-punctured torus $\Sigma_{1,1}$:

$$\bar{\pi} : \Sigma_{1,1} \to \Sigma_{\mathrm{Mod}}.$$

Viewing $\Sigma_{1,1}$ as the quotient $(\mathbb{R}^2 \setminus \mathbb{Z}^2)/\mathbb{Z}^2$, we see that $\Sigma_{1,1}$ can be identified with the square torus with a point removed; see Figure 5. A geodesic connecting the cone point of order 2 and the cusp of Σ_{Mod} lifts to a collection of three cusp-to-cusp geodesics on $\Sigma_{1,1}$.

A line in $\mathbb{R}^2$ with slope p/q and disjoint from $\mathbb{Z}^2$ projects to an essential simple closed curve in $\Sigma_{1,1}$. Conversely, an essential simple closed curve in $\Sigma_{1,1}$ lifts to a line in $\mathbb{R}^2$ with slope p/q and disjoint

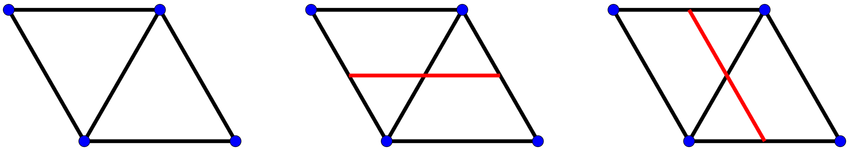


Figure 5: The once-punctured torus $\Sigma_{1,1}$ with a punctured removed (blue). The $\frac{0}{1}$ curve is shown in the middle. The $\frac{1}{0}$ curve is shown on the right.

from $\mathbb{Z}^2$. We see that the isotopy classes of essential simple closed curve in $\Sigma_{1,1}$ correspond to $\mathbb{Q} \cup \{\frac{1}{0}\}$. They are organized by the Farey tessellation of $\mathbb{H}^2$; see Figure 6. In particular, the ideal vertices of the Farey triangulation coincide with $\mathbb{Q} \cup \{\infty\}$. The edges of the Farey triangulation connect p/q and r/s if and only if the corresponding simple close curves have geometric intersection number 1.

We can parametrize isotopy classes of oriented essential simple closed curve in $\Sigma_{1,1}$ by the set of vectors

$$\mathcal{U} = \left\{ \begin{pmatrix} p \\ q \end{pmatrix} \,\middle|\, \gcd(p,q) = 1,\ p = \pm 1 \text{ if } q = 0,\ q = \pm 1 \text{ if } p = 0 \right\},$$

the set of rational direction in $\mathbb{R}^2$. The vector $\binom{1}{0}$ corresponds to the positive y-direction of $\mathbb{R}^2$ while the vector $\binom{0}{1}$ corresponds to the positive x-direction of $\mathbb{R}^2$. By abusing notation, we will use elements of $\mathcal{U}$ to denote isotopy classes of oriented essential simple closed curve in $\Sigma_{1,1}$. Similarly, we will use elements of $\mathbb{Q} \cup \{\frac{1}{0}\}$ to the denote the unoriented counterpart in $\Sigma_{1,1}$.

Since the deck group of $\overline{\pi} : \Sigma_{1,1} \to \Sigma_{\mathrm{Mod}}$ acts by isometries on $\Sigma_{1,1}$, we have an associated 6-fold cyclic covering $\pi : \mathrm{UT}(\Sigma_{1,1}) \to \mathrm{UT}(\Sigma_{\mathrm{Mod}})$. The unit tangent bundle $\mathrm{UT}(\Sigma_{1,1})$ can be trivialized as a product $\Sigma_{1,1} \times S^1$ where $S^1 = \mathbb{R}/2\pi\mathbb{Z}$. In particular, the oriented curve $\binom{p}{q} \in \mathcal{U}$ on $\Sigma_{1,1}$ determines a canonical lift to the oriented curve

$$\begin{pmatrix} p \\ q \end{pmatrix} \times \left\{ \arg \begin{pmatrix} p \\ q \end{pmatrix} \right\},$$

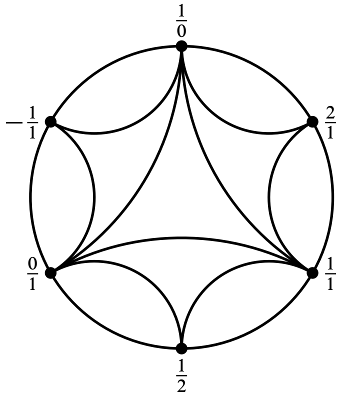


Figure 6: The Farey graph parametrizing essential simple closed curves on $\Sigma_{1,1}$.

where $\arg\colon \mathcal{U} \to [0, 2\pi)$ is the angle from $\binom{0}{1}$ to $\binom{p}{q}$ in the counterclockwise direction. Since an oriented curve in $\mathcal{U} \subset \Sigma_{1,1}$ completely determines its canonical lift to $\mathrm{UT}(\Sigma_{1,1})$, we also use elements in $\mathcal{U}$ to denote this canonical lift.

By [13, Lemma 5.1], the action of a generator ν of the deck group of $\bar{\pi}\colon \Sigma_{1,1} \to \Sigma_{\mathrm{Mod}}$ on the oriented curve is by the order-6 matrix in $\mathrm{SL}_2(\mathbb{Z})$

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

Since an oriented curve in $\mathcal{U} \subset \Sigma_{1,1}$ determines its canonical lift, we can also denote the action of the deck group of $\pi\colon \mathrm{UT}(\Sigma_{1,1}) \to \mathrm{UT}(\Sigma_{\mathrm{Mod}})$ on the set of canonical lifts $\mathcal{U} \subset \mathrm{UT}(\Sigma_{1,1})$ by the same matrix ν.

The following lemma from [13] explains the relationship between canonical lifts of oriented closed geodesic in $\Sigma_{1,1}$ in $\mathcal{U}$ and canonical lifts of oriented closed geodesic in Σ_{Mod}.

Lemma 2.1 [13, Lemma 5.1] *Suppose that $\bar{\gamma}$ is an oriented closed geodesic in Σ_{Mod} obtained by projecting the simple closed curve $p/q \subset \Sigma_{1,1}$ via the covering map $\bar{\pi}\colon \Sigma_{1,1} \to \Sigma_{\mathrm{Mod}}$. Then the canonical lift $\gamma \subset \mathrm{UT}(\Sigma_{\mathrm{Mod}})$ has six lifts. These lifts are*

$$\left\{\pm\begin{pmatrix} p \\ q \end{pmatrix}, \pm\begin{pmatrix} q \\ q-p \end{pmatrix}, \pm\begin{pmatrix} p-q \\ p \end{pmatrix}\right\} \subset \mathrm{UT}(\Sigma_{1,1}).$$

A main result of [13] is the following theorem:

Theorem 2.2 [13, Theorem 4.3, 5.3] *Suppose that $\Delta := \left\{\binom{a_j}{b_j}\right\} \subset \mathrm{UT}(\Sigma_{1,1})$ such that*

1. $|\Delta| < \infty$,
2. *Δ is invariant under the action of $\nu = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 1 \end{smallmatrix}\right)$, and*
3. *for every $\binom{a_j}{b_j}$, there exists $\binom{a_i}{b_i}$ and $\binom{a_k}{b_k}$ such that $\left|\det\left(\begin{smallmatrix} a_i & a_j \\ b_i & b_j \end{smallmatrix}\right)\right| = \left|\det\left(\begin{smallmatrix} a_j & a_k \\ b_j & b_k \end{smallmatrix}\right)\right| = 1$.*

Then the manifolds $\mathrm{UT}(\Sigma_{1,1}) \setminus \Delta$ and $\mathrm{UT}(\Sigma_{\mathrm{Mod}}) \setminus \pi(\Delta)$ are both arithmetic.

Let us denote by $\mathcal{F}\Sigma_{1,1}$ the collection of Δ where Δ is the union of canonical lifts of oriented closed geodesics in $\Sigma_{1,1}$ to $\mathrm{UT}(\Sigma_{1,1})$ satisfying the conditions of Theorem 2.2. The collection of arithmetic modular links that was found in [13] is described as

$$\mathcal{F}\Sigma_{\mathrm{Mod}} := \{\Gamma \subset \mathrm{UT}(\Sigma_{\mathrm{Mod}}) \mid \pi^{-1}(\Gamma) \in \mathcal{F}\Sigma_{1,1}\}.$$

We end with the following observation from [13] underpinning their construction:

Lemma 2.3 [13, Lemma 4.1] *Let $N_{\alpha,\beta}$ be the manifold*

$$N_{\alpha,\beta} := (\Sigma_{1,1} \times [0,1]) \setminus \{\alpha \times \{0\} \cup \beta \times \{1\}\},$$

where α and β are p/q and r/s curves on $\Sigma_{1,1}$ such that $|ps - qr| = 1$. Then $N_{\alpha,\beta}$ is homeomorphic to $N_{0,1}$.

Remark 2.4 The homeomorphism between $N_{\alpha,\beta}$ and $N_{0,1}$ is induced by the linear map that sends α to 0 and β to 1. If we orient all the curves α, β, 0 and 1, then there exists a unique linear transformation that preserves the orientations of the curves and induces the homeomorphism between $N_{\alpha,\beta}$, $N_{0,1}$.

3 Proof of the main results

3.1 Proof of Theorem 1.1

In this section, we give a proof of Theorem 1.1. We begin with the following observation.

Lemma 3.1 *For any $\Delta \in \mathcal{F}\Sigma_{1,1}$, Δ contains*

$$\Delta_0 := \left\{\pm\begin{pmatrix}0\\1\end{pmatrix}, \pm\begin{pmatrix}1\\1\end{pmatrix}, \pm\begin{pmatrix}1\\0\end{pmatrix}\right\}.$$

Consequently, Δ_0 is the smallest collection in $\mathcal{F}\Sigma_{1,1}$ ordered by inclusion.

Proof We project Δ to $\Sigma_{1,1}$ to get a collection of essential simple closed curves $\overline{\Delta} \subset \Sigma_{1,1}$. The fact that Δ is ν-invariant implies that $\overline{\Delta}$ is V-invariant where we view $\overline{\Delta} \subset \mathbb{Q} \cup \{\infty\}$. Since $\overline{\Delta}$ is V-invariant, $|\overline{\Delta}| = 3n$ for some $n \geq 1$. Furthermore, there are exactly n curves represented by vertices of the Farey graph in the intervals from $\frac{0}{1}$ to $\frac{1}{1}$, from $\frac{1}{1}$ to $\frac{1}{0}$ and from $\frac{1}{0}$ to $\frac{0}{1}$ all oriented counterclockwise. The third condition for Δ is satisfied only if $\left\{\frac{0}{1}, \frac{1}{1}, \frac{1}{0}\right\} \subseteq \overline{\Delta}$. Lifting these curves to $\mathrm{UT}(\Sigma_{1,1})$, we get the desired conclusion for Δ. □

Let $\Gamma_0 = \left\{\pi\begin{pmatrix}0\\1\end{pmatrix}\right\}$. Up to a reparametrization, the manifold M_{Γ_0} is

$$M_{\Gamma_0} = \frac{\Sigma_{1,1} \times [0,1] \setminus \left\{\frac{0}{1} \times \{0\} \cup \frac{1}{1} \times \{1\}\right\}}{(x,0) \sim (\nu(x),1)}.$$

Let $\phi\colon M_{\Gamma_0} \to S^1$ be a surjection coming from projecting onto the second factor which induces a surjective homomorphism $\phi_*\colon \pi_1(M_{\Gamma_0}) \to \mathbb{Z}$. The map ϕ_* sends the meridian of the trefoil to 1 (up to taking inverse) and the meridian of the $\frac{0}{1}$ geodesic to 0. Let M_n be the cover of M_{Γ_0} that corresponds to $\phi_*^{-1}(n\mathbb{Z})$ for some positive integer n, then up to a reparametrization of S^1 the manifold M_n is

$$M_n = \frac{\Sigma_{1,1} \times [0,n] \setminus \left\{\nu^i\left(\frac{0}{1}\right) \times \{i\}\right\}_{i=0}^{n}}{(x,0) \sim (\nu^n(x),n)}.$$

See Figure 7 for an example of M_2.

Lemma 3.2 *The manifold M_n is homeomorphic to the complement of the n-component augmented chain link $S^3 \setminus C_n$.*

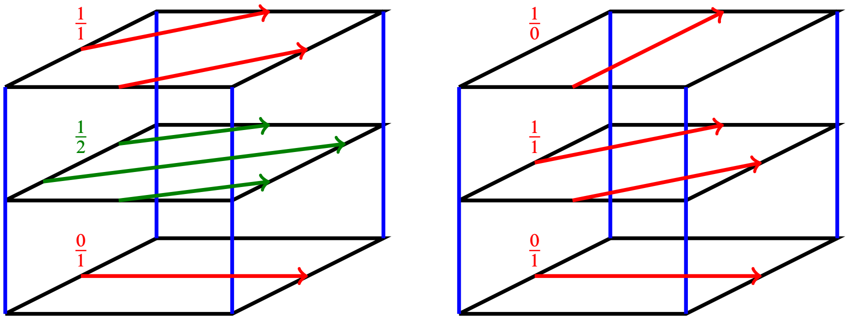


Figure 7: On the left is the manifold M_Γ where $\Gamma = \{\pi\binom{0}{1}, \pi\binom{1}{2}\}$ realized as a once-punctured torus bundle with some curves in red and green drilled out. The gluing map on the left is given by the matrix $\nu = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}$ which glues the bottom to the top. On the right is the manifold M_2, the cover of M_{Γ_0} corresponds to $\phi_*^{-1}(2\mathbb{Z})$. The gluing map on the right is given by the matrix $\nu^2 = \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}$ which glues the bottom to the top.

Proof The manifold M_{Γ_0} is homeomorphic to the complement of the Whitehead link by a homeomorphism $h\colon M_{\Gamma_0} \to S^3 \setminus C_1$ described in [13, Figure 8]. Following [13, Figure 8], the homeomorphism h is obtained by performing a Rolfsen twist about the component Γ_0. The homeomorphism h takes the meridian of the trefoil component in the link $\Gamma_0 \cup \{T_{2,3}\}$ to the meridian of a component of the Whitehead link. Therefore, the homeomorphism h lifts to a homeomorphism between M_n and a n-fold cyclic cover branched over $h(N(T_{2,3}))$ where $N(T_{2,3})$ is a neighborhood of the trefoil knot. Since the two components of the Whitehead link are symmetric, the latter manifold is $S^3 \setminus C_n$. $\square$

Remark 3.3 The manifold M_n can be thought of as the complement of a link in the n -fold cyclic covering of the trefoil complement. In general, cyclic coverings of the trefoil complement do not embed into S^3. It is surprising that after drilling out some link components they do always embed in S^3.

Proof of Theorem 1.1 Let $\Gamma \subset \mathrm{UT}(\Sigma_{\mathrm{Mod}})$ be any modular link in $\mathcal{F}\Sigma_{\mathrm{Mod}}$, $n = |\Gamma|$ and $M_\Gamma = \mathrm{UT}(\Sigma_{\mathrm{Mod}}) \setminus \Gamma$. Given Lemma 3.2, our goal is to show that M_Γ and M_n are homeomorphic. We lift Γ to obtain a collection $\Delta \subset \mathrm{UT}(\Sigma_{1,1})$ that is ν-invariant. By Lemma 3.1, Δ contains $\binom{0}{1}$ and $\binom{1}{1}$. Up to a reparametrization of S^1, M_Γ is

$$M_\Gamma = \frac{(\Sigma_{1,1} \times [0,n]) \setminus \{\gamma_i \times \{i\}\}_{i=0}^n}{(x,0) \sim (\nu(x), n)}$$

for $0 \le i \le n$ where γ_0 and γ_n are $\frac{0}{1}$ and $\frac{1}{1}$ curve on $\Sigma_{1,1}$. Cutting both manifolds M_Γ and M_n along the thrice-punctured sphere $\Sigma_{1,1} \times \{0\} \setminus \{\frac{0}{1} \times \{0\}\}$, we get

$$P_\Gamma = \Sigma_{1,1} \times [0,n] \setminus \{\gamma_i \times \{i\}\}_{i=0}^n \quad \text{and} \quad P_n = \Sigma_{1,1} \times [0,n] \setminus \left\{\nu^i \binom{0}{1} \times \{i\}\right\}_{i=0}^n.$$

By Lemma 2.3, for each $0 \le i \le n-1$ we have a homeomorphism

$$h_i\colon N_{\gamma_i,\gamma_{i+1}} \to N_{\nu^i(0),\nu^{i+1}(0)}.$$

By Remark 2.4, we can choose h_i so that they are induced by linear maps that preserve the orientations of the removed curves. Note that $\Sigma_{1,1} \times \{i+1\} \setminus \{\gamma_{i+1} \times \{i+1\}\}$ is homeomorphic to a thrice-punctured sphere. Our choice of h_i ensures that the composition $h_{i-1}^{-1} \circ h_i$ on $\Sigma_{1,1} \times \{i+1\} \setminus \{\gamma_{i+1} \times \{i+1\}\}$ is a homeomorphism of the thrice-punctured sphere that preserves the punctures. Up to isotopy, we can glue the homeomorphism h_i's together and get a homeomorphism $h\colon P_\Gamma \to P_n$. Note that h is the identity on $\Sigma_{1,1} \times \{0\} \setminus \left\{\frac{0}{1} \times \{0\}\right\}$. Gluing the bottom of P_Γ to the top, we get a homeomorphism

$$h\colon M_\Gamma \to \frac{\Sigma_{1,1} \times [0,n] \setminus \left\{\nu^i\left(\frac{0}{1}\right) \times \{i\}\right\}_{i=0}^n}{(x,0) \sim ((h_{n-1} \circ \nu)(x), n)}.$$

The manifolds M_n and $h(M_\Gamma)$ are obtained from P_n by gluing the bottom to the top via the two maps ν^n and $h_{n-1} \circ \nu$, respectively. The two gluing maps differ on $\Sigma_{1,1} \times \{0\} \setminus \left\{\frac{0}{1} \times \{0\}\right\}$ by $\nu^{-n} \circ h_{n-1} \circ \nu$. We will show that this homeomorphism is isotopic to the identity. Since $\Sigma_{1,1} \times \{0\} \setminus \left\{\frac{0}{1} \times \{0\}\right\}$ is a thrice-punctured sphere, it suffices to show that the map $\nu^{-n} \circ h_{n-1} \circ \nu$ preserves the punctures of $\Sigma_{1,1} \times \{0\} \setminus \left\{\frac{0}{1} \times \{0\}\right\}$. These punctures comprises of the puncture of $\Sigma_{1,1} \times \{0\}$ and the two sides of the removed geodesic $\left\{\frac{0}{1} \times \{0\}\right\}$. The homeomorphism $\nu^{-n} \circ h_{n-1} \circ \nu$ preserves the puncture coming from $\Sigma_{1,1} \times \{0\}$. The homeomorphism ν up to isotopy is an orientation preserving linear map on $\Sigma_{1,1}$. In particular, ν preserves the orientation of any oriented simple closed curve. By Remark 2.4, the homeomorphisms h_i can be chosen to preserves the removed geodesic as an oriented curve on $\Sigma_{1,1}$. Therefore, the map $\nu^{-n} \circ h_{n-1} \circ \nu$ is isotopic to the identity. Therefore, $h(M_\Gamma)$ is homeomorphic to M_n. □

3.2 Proof of Corollary 1.2

The claim about containing a closed embedded essential surface follows from the work of Cooper and Long [4]. For completeness, we give a brief summary of their article focusing on the pertinent details. In this article, the authors studied pure braids from the representation-theoretic and the geometric perspective. On the representation-theoretic side, they introduced and studied the derivative variety associated to an element of the pure braid group [4, Section 2,3].

On the geometric side, they studied the complement $S^3 \setminus \hat{\sigma}$ of the closure of a braid $\sigma \in B_n$ [4, Section 4]. In particular, they showed that $S^3 \setminus \hat{\sigma}$ contains a closed essential surface where σ is a pure 4-braid lying in the kernel of the Grassner representation [4, Theorem 4.8]. To establish this result, they give general criteria for a link complement in S^3 to contain a closed essential surface [4, Theorem 4.1 and Corollary 4.6]. In Theorem 4.1, they showed that if the $\mathrm{SL}_2(\mathbb{C})$-representation variety of an n-component link $L \subset S^3$ contains a component of dimension $> n+3$ and has an irreducible representation, then $S^3 \setminus L$ contains a closed essential surface. They pointed out a sufficient condition for the hypothesis of Theorem 4.1 is that the link group $\pi_1(S^3 \setminus L)$ surjects a nonabelian free group of rank k such that $3k > n+3$. This is the essential point of [4, Corollary 4.4]. The surjection $\pi_1(S^3 \setminus L) \to F_k$ where F_k is the free group of rank k allows one to embed the representation variety of F_k into that of $\pi_1(S^3 \setminus L)$ by pullbacks. The representation variety of the nonabelian free group of rank k contains a component with an irreducible representation and has dimension $3k > n+3$. It follows that the hypothesis of Theorem 4.1 is satisfied if $\pi_1(S^3 \setminus L)$ surjects a nonabelian free group of rank k such that $3k > n+3$.

Remark 3.4 Though Corollary 4.4 of [4] is stated as removing one component of the link, to apply their argument, one just needs the fact that the fundamental group of the link complement surjects a nonabelian free group of sufficiently large rank.

Proof of Corollary 1.2 Theorem 1.1 shows that M_Γ is a $|\Gamma|$-fold cyclic cover of M_{Γ_0}. Since the complement of the Whitehead link fibers, M_Γ also fibers.

The claim about containing a closed embedded essential surface follows from the work of Cooper and Long [4]. Suppose that $n = |\Gamma| > 1$, then we write $n = 2k$ or $n = 2k+1$ where $k \geq 1$ is an integer. The group $\pi_1(S^3 \setminus C_n)$ has a surjection onto the free group of rank $k+1$ coming from deleting k components when $n = 2k$ and $k+1$ components when $n = 2k+1$; see Figure 1. Therefore, $\pi_1(M_\Gamma)$ surjects a free group of rank $k+1$ for an appropriate k. Similar to [4, Corollary 4.4], the surjection shows that there exists a component of characters of irreducible $\mathrm{SL}_2(\mathbb{C})$-representations of dimension $3k$. The number of cusp of $S^3 \setminus C_n$ is $|\Gamma| + 1 = n+1$. When $n = 2k$, $3k$ is strictly greater than $n+1$ if and only if $k > 1$. When $n = 2k+1$, $3k$ is strictly greater than $n+1$ if and only if $k > 2$. That is, if $|\Gamma| \not\in \{1,2,3,5\}$, then $S^3 \setminus C_n$ satisfy the hypothesis of [4, Theorem 4.1]. It follows from [4, Theorem 4.1] that if $|\Gamma| \not\in \{1,2,3,5\}$, then $S^3 \setminus C_n$, and hence M_Γ, contains a closed embedded essential surface. □

References

[1] **K L Baker**, *Surgery descriptions and volumes of Berge knots, I: Large volume Berge knots*, J. Knot Theory Ramifications 17 (2008) 1077–1097 MR

[2] **M Bergeron**, **T Pinsky**, **L Silberman**, *An upper bound for the volumes of complements of periodic geodesics*, Int. Math. Res. Not. 2019 (2019) 4707–4729 MR

[3] **J S Birman**, **R F Williams**, *Knotted periodic orbits in dynamical systems, I: Lorenz's equations*, Topology 22 (1983) 47–82 MR

[4] **D Cooper**, **D D Long**, *Derivative varieties and the pure braid group*, Amer. J. Math. 115 (1993) 137–160 MR

[5] **T Cremaschi**, **J A Rodríguez-Migueles**, *Hyperbolicity of link complements in Seifert-fibered spaces*, Algebr. Geom. Topol. 20 (2020) 3561–3588 MR

[6] **T Cremaschi**, **J A Rodrigurź-Migueles**, **A Yarmola**, *On volumes and filling collections of multicurves*, J. Topol. 15 (2022) 1107–1153 MR

[7] **M Culler**, **N M Dunfield**, **M Goerner**, **J R Weeks**, *SnapPy, a computer program for studying the geometry and topology of 3-manifolds* Available at `http://snappy.computop.org`

[8] **P Dehornoy**, *Geodesic flow, left-handedness and templates*, Algebr. Geom. Topol. 15 (2015) 1525–1597 MR

[9] **P Foulon**, **B Hasselblatt**, *Contact Anosov flows on hyperbolic* 3*-manifolds*, Geom. Topol. 17 (2013) 1225–1252 MR

[10] **E Ghys**, *Knots and dynamics*, from "International Congress of Mathematicians, I", Eur. Math. Soc., Zürich (2007) 247–277 MR

[11] **T Matsusaka**, **J Ueki**, *Modular knots, automorphic forms, and the Rademacher symbols for triangle groups*, Res. Math. Sci. 10 (2023) art. id. 4 MR

[12] **J Milnor**, *Introduction to algebraic K-theory*, Ann. of Math. Stud. 72, Princeton Univ. Press (1971) MR

[13] **T Pinsky**, **J S Purcell**, **J A Rodríguez-Migueles**, *Arithmetic modular links*, Pacific J. Math. 327 (2023) 337–358 MR

[14] **J A Rodríguez-Migueles**, *Periods of continued fractions and volumes of modular knots complements*, J. Knot Theory Ramifications 32 (2023) art. id. 2350063 MR

Department of Mathematics, Rice University
Houston, TX, United States

`khanh.le@rice.edu`

Received: 13 December 2023 Revised: 17 September 2024

Algebraic & Geometric Topology **25**:8 (2025) 4839–4896
DOI: 10.2140/agt.2025.25.4839
Published: 20 November 2025

Flat fully augmented links are determined by their complements

CHRISTIAN MILLICHAP
ROLLAND TRAPP

We show that two flat fully augmented links with homeomorphic complements must be equivalent as links in $\mathbb{S}^3$. This requires a careful analysis of how totally geodesic surfaces and cusps intersect in these link complements and behave under homeomorphism. One consequence of this analysis is a complete classification of flat fully augmented link complements that admit multiple reflection surfaces. In addition, our work classifies those symmetries of flat fully augmented link complements which are not induced by symmetries of the corresponding link.

57K10, 57K32, 57M50

1 Introduction

Two links, L_1 and L_2, in $\mathbb{S}^3$ are *equivalent* if there exists an orientation-preserving homeomorphism of pairs from $(\mathbb{S}^3, L_1)$ to $(\mathbb{S}^3, L_2)$. An equivalence of links induces a homeomorphism between the link complements $\mathbb{S}^3 \setminus L_1$ and $\mathbb{S}^3 \setminus L_2$, which shows that links determine their complements. However, the converse of this statement is generally not true. For instance, if a link contains an unknotted component, then a Dehn twist along this component is a homeomorphism of the complement that will frequently produce a nonequivalent link. Whitehead used this technique in [25] to show that there are infinitely many distinct links with the same complement as the Whitehead link. Some other known constructions for producing distinct links with the same complement were found by Berge [4] and Gordon [10, Section 6]. In contrast to links, the Gordon–Luecke theorem [11] shows that knots are determined by their complements in $\mathbb{S}^3$. Knots in certain closed, oriented 3-manifolds other than $\mathbb{S}^3$ are also determined by their complements; see the work of Rong [24], Matignon [18], Gainullin [9], and Ichihara–Saito [14] for some examples. Moving forward, we will always assume knots and links are embedded in $\mathbb{S}^3$. This contrast between links with multiple components and knots motivated the following question raised by Mangum and Stanford [17]:

Question Is there a set of links $\mathcal{S}$ such that if $L_1, L_2 \in \mathcal{S}$ and $\mathbb{S}^3 \setminus L_1$ is homeomorphic to $\mathbb{S}^3 \setminus L_2$, then L_1 is equivalent to L_2?

In the same paper, Mangum–Stanford show that the set of homologically trivial and Brunnian links (called HTB links in their paper) provide an affirmative answer to this question [17, Theorem 3]. As far as the authors know this is the only example in the literature of an infinite set of links, other than the set of knots, with this property. This motivates the main goal of our paper, which is to show that the family of *flat fully augmented links* (flat FALs) also have this property.

Flat FALs are a family of hyperbolic links which can be obtained from link diagrams, meeting certain diagrammatic conditions, in the following way. Given a link diagram add a trivial component (called a crossing circle) enclosing each twist region, then remove all twists so the strands of the diagram run parallel through the crossing circles (see Figure 2). Knot circles of the resulting flat FAL are components that lie in the projection plane of the resulting diagram. Flat FALs, and more generally, FALs, are a rich family of hyperbolic links, which have received much attention in the last twenty years due to the explicit combinatorial descriptions of their geometric structures and connections to highly twisted knots via Dehn surgery; see [5; 6; 7; 8; 13; 15; 19; 20; 21; 22; 23] for some examples from the literature. We direct the reader to the beginning of Section 2 for more details on essential properties of flat FALs used in this paper. We can now formally state our main theorem, recalling that an isotopy of links in $\mathbb{S}^3$ induces an equivalence of links.

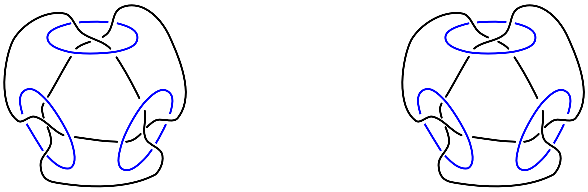
Theorem 1.1 *Let $\mathcal{A}$ and $\mathcal{A}'$ be flat FALs. Then $(\mathbb{S}^3, \mathcal{A})$ is isotopic to $(\mathbb{S}^3, \mathcal{A}')$ if and only if $\mathbb{S}^3 \setminus \mathcal{A}$ is homeomorphic to $\mathbb{S}^3 \setminus \mathcal{A}'$.*

Theorem 1.1 provides a new infinite set of links that are determined by their complements, distinct from knots and HTB links. By definition, every flat FAL contains at least three components, and so, no flat FALs are knots. At the same time, there are infinitely many flat FALs that are not HTB links. Specifically, any flat FAL with at least two knot circles is not an HTB link since it will contain a Hopf sublink. Such a sublink violates the homologically trivial property that the linking number is 0 for any two components.

Another way our work differs from both Gordon–Luecke and Mangum–Stanford is in the techniques used. The proof of Theorem 1.1 greatly leverages the hyperbolic structure of flat FAL complements and relies on analyzing the behavior of totally geodesic surfaces and cusps under isometries induced by homeomorphism. This geometric approach differs from the purely topological approaches used by Gordon–Luecke and Magnum–Stanford, where these authors determine which Dehn surgeries on a knot or HTB link produce $\mathbb{S}^3$.

The *flat* hypothesis of Theorem 1.1 is necessary, making the result as general as possible within the class of FALs. To see this, consider the twisted FALs of Figure 1. They differ by a Dehn twist on the top circle, so their complements are homeomorphic. The links are distinct, however, since all components are unknots in Figure 1(a) while one component is a trefoil in Figure 1(b).

The proof of Theorem 1.1 breaks into two cases, depending on the number of reflection surfaces contained in a flat FAL complement. Intuitively, a reflection surface for a flat FAL complement $\mathbb{S}^3 \setminus \mathcal{A}$ is a totally



(a) all components unknotted (b) one trefoil component

Figure 1: Distinct twisted FALs with homeomorphic complements.

geodesic surface that corresponds with the projection plane for an FAL diagram of $\mathcal{A}$; see Section 2 for more explicit details and properties of reflection surfaces. As an intermediary step in the proof of Theorem 1.1, we classify flat FAL complements with multiple reflection surfaces and show that homeomorphisms between flat FAL complements preserve reflection surfaces. For emphasis, we now state the classification of flat FALs with multiple reflection surfaces (see Figure 6 and the beginning of Section 3.2 for descriptions of the links P_n and O_n).

Theorem 1.2 *Suppose $M = \mathbb{S}^3 \setminus \mathcal{A}$ is a flat FAL complement with multiple distinct reflection surfaces. Then either*

- *$\mathcal{A}$ is equivalent to the Borromean rings and M contains exactly three distinct reflection surfaces*, *or*
- *$\mathcal{A}$ is equivalent to P_n with $n \geq 3$, or O_n with $n \geq 2$, and M contains exactly two distinct reflection surfaces.*

A slightly more general version of this classification result is given in Theorem 3.11, along with several useful corollaries on how reflection surfaces behave under homeomorphism.

The proof of Theorem 1.2 relies on an analysis of how cusps and totally geodesic surfaces behave relative to different reflection surfaces in a flat FAL complement. With Theorem 1.2 in hand, basic properties of the links P_n and O_n show that two flat FALs whose complements admit multiple reflection surfaces are equivalent as links if and only if their respective complements are homeomorphic. This is highlighted in Corollary 3.12.

Theorem 1.2 also implies that a flat FAL with multiple reflection surfaces cannot be homeomorphic to a flat FAL with a single reflection surface; see Corollary 3.13. As a result, we now only need to consider the case where there exists a homeomorphism $h\colon M \to M'$ between two flat FAL complements, each with unique reflection surfaces. This is a far more challenging task, and relies on the topology and geometry of thrice-punctured spheres in a flat FAL complement that are not contained in the reflection surface, which were classified by Morgan–Ransom–Spyropoulos–Trapp–Ziegler in [20]. In particular, we make extensive use of sets of thrice-punctured spheres that separate M, whose homeomorphic images in M' are greatly restricted by the classification in [20]. These restrictions help us describe how the knot and crossing circles which intersect such thrice-punctured spheres must behave under homeomorphism.

This, in turn, allows us to show that any flat FAL complement that admits a unique reflection surface and a nontrivial homeomorphism must have a particular link structure, which we call a signature link; see Definition 5.1. Furthermore, only one type of nontrivial homeomorphism is possible in this case, which we call a full-swap. We carefully describe full-swap homeomorphisms of signature links in terms of compositions of Dehn twists along sets of Hopf sublinks; see Definition 5.4. In addition, the induced action of any such full-swap homeomorphism on the corresponding flat FAL (which must be a signature link) can be made explicit on a diagrammatic level, which makes it easy to construct a specific isotopy between any such pair of flat FALs with homeomorphic complements.

As partially noted in the previous paragraph, our work not only shows that flat FALs are determined by their complements, but classifies the types of homeomorphisms that can exist between flat FAL complements and what types of flat FAL complements can admit certain types of homeomorphisms. These severe geometric restrictions on self-homeomorphisms of flat FAL complements lead to a concise comparison between their symmetry groups and those of their complements. Recall that the symmetry group of a link $L \subset \mathbb{S}^3$ is the group of homeomorphisms of pairs $(\mathbb{S}^3, L)$ up to isotopy, which we denote by $\mathrm{Sym}(\mathbb{S}^3, L)$. Similarly, the symmetry group of the corresponding link complement is the group of self-homeomorphisms of $\mathbb{S}^3 \setminus L$ up to isotopy, denoted by $\mathrm{Sym}(\mathbb{S}^3 \setminus L)$. Any self-homeomorphism of $(\mathbb{S}^3, L)$ induces a self-homeomorphism of $\mathbb{S}^3 \setminus L$, and so, $\mathrm{Sym}(\mathbb{S}^3, L) \subseteq \mathrm{Sym}(\mathbb{S}^3 \setminus L)$. However, this can be a strict containment; see [12] for some examples. The following theorem classifies symmetries of flat FAL complements that are not induced by symmetries of the corresponding link.

Theorem 1.3 *Let $\mathcal{A}$ be a flat FAL. Then either*

- *$\mathcal{A}$ is not a signature link and both $\mathcal{A}$ and its complement $M = \mathbb{S}^3 \setminus \mathcal{A}$ have the same symmetry group, or*
- *$\mathcal{A}$ is a signature link and full-swaps on $\mathcal{A}$ generate symmetries of $M = \mathbb{S}^3 \setminus \mathcal{A}$ which are not restrictions of symmetries of $\mathcal{A}$ to M.*

Part of Theorem 1.3 is proved in Theorem 3.14 at the end of Section 3.1. The rest of the proof of this theorem is completed at the end of Section 7.

We now describe the organization of the rest of this paper. In Section 2 we introduce flat FALs, the necessary terminology related to cusps and totally geodesic surfaces contained in flat FAL complements, and review some essential facts from the literature on the geometry of FAL complements. In Section 3, we prove Theorem 1.2. This allows us to focus the rest of the paper on homeomorphisms between flat FAL complements, each with a unique reflection surface and this transition is discussed in Section 4. Then Section 5 discusses "signature link" complements, a special class of flat FAL complements, along with a "full-swap" homeomorphism that can be performed on any signature link complement. The image of a full-swap homeomorphism is another signature link complement, and it is straightforward to construct an explicit isotopy between their corresponding links. In Section 6, we prove some useful facts about sets of thrice-punctured spheres that separate a flat FAL complement (with a unique reflection surface) and how they behave under homeomorphisms. Finally, in Section 7 we build off the tools from Section 6 to show that any homeomorphism between flat FAL complements (each with a unique reflection surface) must essentially be a "full-swap" homeomorphism between "signature link" complements. This allows us to construct an isotopy between the corresponding links and prove our main theorem.

Acknowledgement We would like to thank Jeffrey Meyer for comments and suggestions on preliminary work done for this paper.

2 Totally geodesic surfaces and cusps in flat FALs

This section provides a brief introduction to flat FALs, compiles some known results about totally geodesic surfaces in their complements, and introduces two concepts used extensively in Section 3: reflection-like surfaces and their induced structures on a flat FAL complement. Aside from reflection-like surfaces and their induced structures, this section is a review of results in the literature.

We first describe how to construct a flat FAL. For this process, start with a link L and a diagram $D(L)$. We can build a diagram $D(\mathcal{F})$ for an FAL $\mathcal{F}$ corresponding to L by augmenting each twist region in $D(L)$ with a circle and undoing all full-twists from each respective twist region. After this procedure, twist regions of $D(L)$ that had contained an odd number of crossings will still contain a single crossing in $D(\mathcal{F})$. If we remove all of these remaining single crossings, then we will have constructed a diagram $D(\mathcal{A})$ for a flat FAL $\mathcal{A}$, as illustrated in Figure 2.

In our work, we will solely be interested in the case where a flat FAL $\mathcal{A}$ is hyperbolic in the sense that the complement $\mathbb{S}^3 \backslash \mathcal{A}$ admits a complete metric of constant negative curvature. As noted in [23, Theorem 2.5], a (flat) FAL $\mathcal{A}$ is hyperbolic if and only if it there exists a corresponding link diagram $D(L)$ that is nonsplittable, prime, twist-reduced, and contains at least two twist regions, where $D(\mathcal{A})$ is obtained

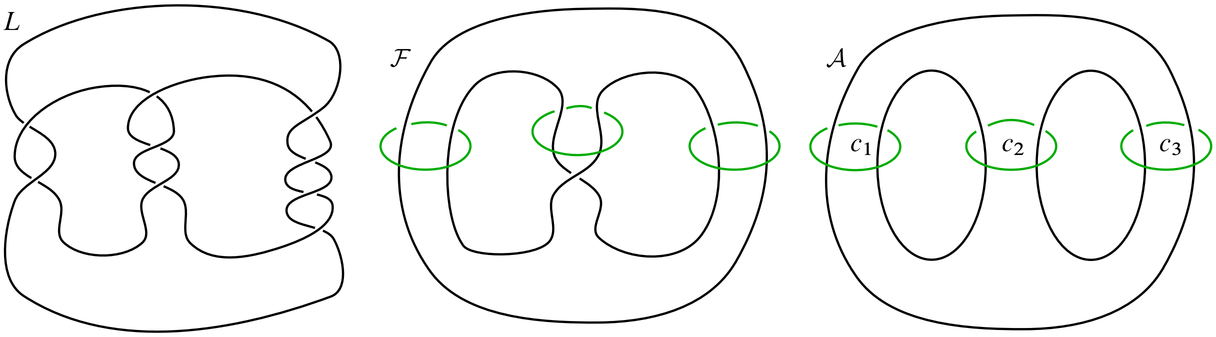


Figure 2: On the left is a diagram of a link L with three twist regions. The middle diagram shows the corresponding FAL $\mathcal{F}$ obtained from fully augmenting L. The right diagram shows the corresponding flat FAL $\mathcal{A}$. Crossing circles of $\mathcal{A}$ are labeled by c_i, for $i = 1, 2, 3$.

from $D(L)$ by the augmentation process described above. We refer the reader to [8, Section 1] for these diagrammatic definitions. Moving forward, we will assume any flat FAL is hyperbolic, and refer to $D(L)$ as the corresponding diagram that was augmented to construct $D(\mathcal{A})$.

This augmentation process partitions the components of $\mathcal{A}$ into *crossing circles*, the trivial components added via augmentation, and *knot circles*, components coming from the original link L. Observe that each crossing circle in a flat FAL bounds a *crossing disk*, a disk in $\mathbb{S}^3$ punctured twice by parallel knot circle arcs that replace an original twist region of $D(L)$. We define a flat FAL diagram to be a link diagram together with this additional structure. More precisely, we have:

Definition 2.1 A link diagram $D(\mathcal{A})$ is a *flat FAL diagram* for a (hyperbolic) link $\mathcal{A}$ if $D(\mathcal{A})$ was constructed by fully augmenting a corresponding diagram $D(L)$ and removing all crossings from each twist region. In addition, we assume $D(\mathcal{A})$ carries with it the partition of components of $\mathcal{A}$ into crossing- and knot-circles, as well as the choice of crossing disks, relative to this augmentation of $D(L)$. A link $\mathcal{A}$ is a *flat FAL* if it admits a flat FAL diagram $D(\mathcal{A})$.

The geometric structures of augmented links were first studied in Adams [2]. A particularly nice geometric decomposition of (flat) FAL complements into pairs of identical right-angled ideal polyhedra was described by Agol–Thurston in the appendix of [16]. This geometric decomposition has proved to be a fruitful tool for analyzing geometric and topological properties of FALs; see [6; 7; 8; 13; 15; 23] for a few examples. Infinite subclasses of FALs have also been examined in the literature. For instance, Meyer–Millichap–Trapp [19] studied the arithmeticity, invariant trace fields, symmetries, and hidden symmetries of FALs obtained by fully augmenting pretzel links, and Purcell examined FALs whose complements admit a decomposition into regular ideal hyperbolic octahedra in [23]. In both cases, each subclass contains an infinite number of flat FALs. Thus the family of flat FALs is a large set of links with many interesting properties.

We now collect some important facts about totally geodesic surfaces and cusps contained inside a flat FAL complement. These properties will be essential for proving both Theorems 1.1 and 1.2. Most of the results stated here are known in the literature and we refer the reader to [8; 20; 23] for more details on the geometric properties of FAL complements discussed here.

First, we describe reflection surfaces, which are an important type of totally geodesic surface contained in every flat FAL complement. Given a flat FAL $\mathcal{A}$ in $\mathbb{S}^3 \cong \mathbb{R}^3 \cup \{\infty\}$, position the crossing circles and disks so that they are orthogonal to the projection plane and embed the knot circles in the projection plane. Such an isotopy is always possible based on the diagrammatic definition of a flat FAL. Then reflection in the projection plane maps every component of $\mathcal{A}$ to itself and fixes the projection plane pointwise. Mostow–Prasad rigidity implies that this projection plane corresponds with a totally geodesic surface $R \subset M = \mathbb{S}^3 \setminus \mathcal{A}$ and there exists an orientation-reversing involution $\iota_R \colon M \to M$ that fixes R pointwise corresponding with reflection in the projection plane. Since $\mathcal{A}$ could admit FAL diagrams where different surfaces play the role of the projection plane, we provide the following definition.

Definition 2.2 Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ be a flat FAL complement and let $R \subset M$ be an embedded totally geodesic surface. If there exists an FAL diagram $D(\mathcal{A})$ in the projection plane P for which $R = P \setminus \mathcal{A}$, then we say that R is a *reflection surface* of M (relative to the diagram $D(\mathcal{A})$).

An FAL diagram $D(\mathcal{A})$ partitions the cusps of $M = \mathbb{S}^3 \setminus \mathcal{A}$ into *crossing circle cusps* and *knot circle cusps*, corresponding with crossing circles and knot circles of $\mathcal{A}$, respectively. We also say that this partition is induced by a reflection surface R, though this ultimately depends on the FAL diagram since reflection surfaces are defined relative to the projection plane for a diagram. As we will see in Section 3, it is possible for a flat FAL complement to admit distinct reflection surfaces that induce different partitions on the components of $\mathcal{A}$. At the same time, it is possible for a flat FAL complement to admit one reflection surface that induces different partitions on the components of $\mathcal{A}$; this phenomenon will be examined in Section 5. Here, we say a cusp of M is an *R-knot circle cusp* (respectively, *R-crossing circle cusp*) if the corresponding link component of $\mathcal{A}$ is a knot circle (respectively, crossing circle) relative to R. In this paper, we frequently use the same notation to refer to a component of $\mathcal{A}$ and the corresponding cusp on M. In addition, each R-crossing circle C of $\mathcal{A}$ bounds an *R-crossing disk D*, which is a disk twice punctured by the two (not necessarily distinct) knot circles going through C. As noted in [23, Lemma 2.1], each such R-crossing disk is an embedded totally geodesic thrice-punctured sphere in M. Furthermore, R intersects D orthogonally in the set of three simple, nonseparating geodesics on D; see Figure 3 for a visual of simple geodesics on a thrice-punctured sphere.

A flat FAL diagram $D(\mathcal{A})$ (or, alternatively, its reflection surface R) determines more structure than the partitioning of cusps of M into knot and crossing circle cusps. It also determines a meridian and longitude on each boundary torus of a cusp neighborhood. In this context, meridians and longitudes will be considered *slopes*, or unoriented isotopy classes of simple closed curves on this boundary torus.

Each component L of a flat FAL is topologically an unknot in $\mathbb{S}^3$, so the torus boundary T_L of a tubular neighborhood V_L of L has a natural choice of meridional and longitudinal slopes. The unknotted torus $T_L \subset \mathbb{S}^3$ bounds a solid torus on each side. A meridian m is the slope of T_L that bounds a disk in V_L, while a longitude ℓ is a slope of T_L that bounds a disk in $\mathbb{S}^3 \setminus V_L$. The unoriented curves m and ℓ are well defined up to ambient isotopy on T_L, so give well-defined slopes on T_L.

From a geometric perspective, the slopes m and ℓ can be described as the intersection of T_L with totally geodesic surfaces. More precisely, let $\mathcal{A}$ be a flat FAL with reflection surface R. If K is a knot circle of $\mathcal{A}$ with torus boundary T_K of a cusp neighborhood of K, then $R \cap T_K$ is a pair of simple closed curves that represent longitudes on T_K. Similarly, for any crossing circle C of $\mathcal{A}$, the set $R \cap T_C$ is a pair of simple closed curves that represent meridians on T_C. Crossing disks relative to the reflection surface R are orthogonal to it, and so intersect T_K in meridians and T_C in longitudes. See [8, Lemma 2.3] for more details.

The longitudes and meridians just described will be referred to as *R-meridians* and *R-longitudes* when we want to emphasize the FAL diagram (and corresponding reflection surface) used to determine them.

Since each torus boundary of a cusp T of M is rectangular (see [8, Lemma 2.3]), curves orthogonal to $R \cap T$ determine the second generator for the fundamental group of each such torus. We refer to the basis for each $\pi_1(T)$ described above as the *peripheral structure* relative to the FAL diagram $D(\mathcal{A})$.

The following proposition summarizes the essential features of a reflection surface in a flat FAL complement that we just discussed.

Proposition 2.3 *Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ be a flat FAL complement with an FAL diagram $D(\mathcal{A})$ which partitions the components of $\mathcal{A}$ into crossing circles and knot circles. Then M contains an embedded totally geodesic R with the following features*:

- *R corresponds with $P \setminus \mathcal{A}$, where P is the projection plane for $D(\mathcal{A})$.*
- *R is fixed pointwise by an orientation reversing involution $i_R\colon M \to M$.*
- *R intersects each R-crossing disk of M orthogonally in its three simple nonseparating geodesics.*
- *R intersects every cusp of M in two parallel curves.*
- *R intersects the boundary torus T_C of each crossing circle cusp C in a pair of meridians and provides a peripheral structure on $\pi_1(T_C)$ relative to $D(\mathcal{A})$.*
- *R intersects the boundary torus T_K of each knot circle cusp K in a pair of longitudes and provides a peripheral structure on $\pi_1(T_K)$ relative to $D(\mathcal{A})$.*

Our main goal is to show that flat FALs are determined by their complements among the set of all flat FALs. For this reason we will need to consider homeomorphic images of reflection surfaces and the structures associated with them that are highlighted in the previous proposition. By Mostow–Prasad rigidity, any homeomorphism between flat FALs induces an isometry between them, and so, we can restrict our analysis to the isometric image of a reflection surface.

Let M, M' be flat FAL complements with reflection surfaces R, R', respectively, and let $\rho\colon M' \to M$ be an isometry between these hyperbolic 3-manifolds. Since the definition of a reflection surface is diagram dependent, we can not immediately assume that $\rho(R')$ is a reflection surface for M. This motivates the following definition.

Definition 2.4 We say that S is a *reflection-like* surface for a flat FAL complement M if there exists an isometry $\rho\colon M' \to M$ between flat FAL complements such that $S = \rho(R')$, where R' is a reflection surface for M'.

Any reflection surface for M is also a reflection-like surface for M via the identity map. In addition, any homeomorphic flat FAL complements have the same number of reflection-like surfaces. Indeed, a homeomorphism $h\colon M' \to M$ between flat FAL complements M' and M induces a unique isometry $\rho_h\colon M' \to M$. By Definition 2.4, the image of each reflection-like surface in M' is a reflection-like surface in M. At the same time, ρ_h^{-1} preserves the property of being reflection-like, so M' and M have the same number of reflection-like surfaces. Similarly, equivalent flat FALs have the same number of reflection-like surfaces since an equivalence of links induces a homeomorphism between their respective complements.

Many important topological and geometric properties of reflection surfaces will be preserved under isometry, which we highlight below. These directly follow from Proposition 2.3.

Proposition 2.5 *Let $\rho\colon M' \to M$ be an isometry between flat FAL complements, let $R' \subset M'$ be a reflection surface, and let $S = \rho(R')$ denote the corresponding reflection-like surface in M. Then:*

- *S is fixed pointwise by an orientation reversing involution $i_S\colon M \to M$.*
- *S intersects the image of each R'-crossing disk of M' orthogonally in its three simple nonseparating geodesics.*
- *S intersects every cusp of M in two parallel curves.*
- *Let T be the boundary torus of a cusp of M. Then S determines a peripheral structure on T, namely, the image of the peripheral structure that R' determines on $\rho^{-1}(T)$.*

Our work in Section 3 will show that every reflection-like surface in an FAL complement is a reflection surface. This is highlighted in Corollary 3.13. After proving this result, we will drop the term reflection-like surface, and instead, only use reflection surface.

A reflection-like surface $S \subset M$ determines a partition of the components of $\mathcal{A}$ into S-crossing circles and S-knot circles, coming from the partition of $\mathcal{A}'$ into R'-crossing circles and R'-knot circles. We use the term *S-structure* on M to refer to this partition of the components of $\mathcal{A}$, the peripheral structures on each cusp determined by S, and a fixed choice of images of R'-crossing disks in M. Suppose M contains distinct reflection-like surfaces R and S. A component L of $\mathcal{A}$ *changes type* if L is an R-knot circle and an S-crossing circle (or vice versa). This terminology will be useful when discussing flat FAL complements that admit multiple reflection surfaces.

In addition to reflection surfaces, (totally geodesic) thrice-punctured spheres contained in flat FAL complements will serve as powerful tools for analyzing homeomorphisms between these manifolds. We now review properties of, and results regarding, embedded, totally geodesic, thrice-punctured spheres in flat FAL complements.

Any thrice-punctured sphere has precisely six simple geodesics, three of which are separating, three of which are nonseparating, and none of which are closed, as depicted in Figure 3. Any intersection of a totally geodesic thrice-punctured sphere D with another totally geodesic surface in a flat FAL complement must be some subset of these six simple geodesics on D that are pairwise disjoint. This puts strong restrictions on the behavior of intersections between D and other totally geodesic surfaces, which we shall exploit.

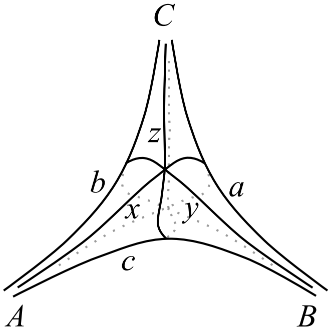


Figure 3: A thrice-punctured sphere with separating geodesics labeled x, y, z and nonseparating geodesics labeled a, b, c.

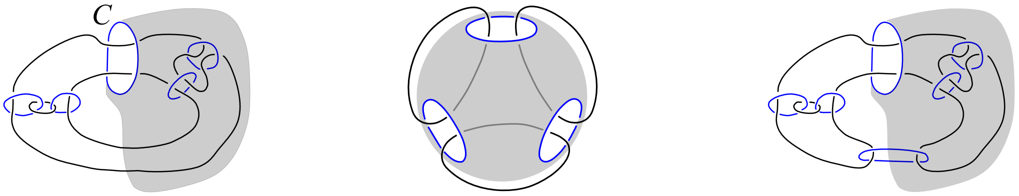
Every FAL complement contains many totally geodesic thrice-punctured spheres that are not contained in a reflection surface, which we will call *nonreflection*, *thrice-punctured spheres* (relative to a designated reflection surface). Crossing disks in M, for example, constitute one category of such spheres. Results in [20] imply there are three types of nonreflection thrice-punctured spheres in flat FAL complements, classified in terms of their intersections with R and their punctures. Here, a *puncture* refers to the intersection of a totally geodesic surface in M with a torus boundary of a cusp of M, which will produce a set of simple closed curves, each of which represents a slope on this torus. We sometimes use the term longitudinal puncture to refer to a puncture that is a representative for a longitude on this torus and we also use the term meridional puncture for a puncture that is a representative for a meridian on this torus (here meridian and longitude refer to the peripheral structure induced by a reflection surface $R \subset M$). We now define the three types of nonreflection, thrice-punctured spheres in flat FAL complements.

Definition 2.6 Let D be a nonreflection thrice-punctured sphere in the flat FAL complement $M = \mathbb{S}^3 \setminus \mathcal{A}$ with reflection surface R.

- D is a *crossing disk* if $D \cap R$ consists of the three nonseparating geodesics of D and the punctures of D are one crossing circle longitude and two knot circle meridians.
- D is a *longitudinal disk* if $D \cap R$ consists of the three nonseparating geodesics of D and the punctures of D are longitudes of three distinct crossing circles.
- D is a *singly separated disk* if $D \cap R$ consists of one separating geodesic of D and the punctures of D are a longitude of one crossing circle C and two meridians of another crossing circle C'.

Collectively, we refer to crossing disks and longitudinal disks as *N-disks* since their intersections with R both consist of a nonseparating geodesics of D.

This definition of crossing disk broadens the typical meaning of the term. The crossing circle C illustrated in Figure 4(a) bounds two distinct crossing disks, the one interior to C as well as the shaded disk. The



(a) a crossing disk (b) a longitudinal disk (c) a singly separated disk

Figure 4: Types of nonreflection, thrice-punctured spheres.

reason for extending the meaning of “crossing disk” is that any such disk D for a crossing circle C could be chosen as a crossing disk in $\mathcal{A}$ by replacing the current crossing disk with D. The two crossing disk structures come from fully augmenting diagrams of the same link obtained by flyping one twist region to a different part of the link diagram (see [20]).

Parts (b) and (c) of Figure 4 illustrate a longitudinal and singly separated disk, respectively. The following theorem from [20] classifies nonreflection thrice-punctured spheres in flat FAL complements.

Theorem 2.7 [20, Theorem 3.11] *Let D be a nonreflection thrice-punctured sphere in the flat FAL complement $M = \mathbb{S}^3 \setminus \mathcal{A}$ with reflection surface R. Then D is orthogonal to R in M and D is either a crossing, longitudinal, or singly separated disk.*

In our work, we will frequently make use of Theorem 2.7 and the qualifications of nonreflection thrice-punctured spheres given in Definition 2.6. The properties of separating and nonseparating are topological, so if a reflection surface intersects a thrice-punctured sphere in its separating geodesic, so will any homeomorphic image of them. Thus, if D is singly separated by the reflection surface R, then $h(D)$ is also separated by $h(R)$ along one separating geodesic. Similarly, the homeomorphic image of an N-disk relative to the reflection surface R must intersect the reflection-like surface $h(R)$ in its nonseparating geodesics. In the case when $h(R)$ is again a reflection surface, we say that a homeomorphism preserves the type of a singly separated disk, and preserves the property of being an N-disk. The only type changes possible when $h(R)$ is a reflection surface, then, are between crossing and longitudinal disks. These can be recognized by analyzing the images of punctures under this homeomorphism.

Finally, we define a *separating pair* to be a pair of disjoint thrice-punctured spheres that separate M. The following theorem from [20] tells us exactly which pairs of thrice-punctured spheres are separating pairs in FAL complements.

Theorem 2.8 [20, Theorem 4.8] *Let $\{S_1, S_2\}$ be a pair of disjoint essential thrice-punctured spheres in the complement $M = \mathbb{S}^3 \setminus \mathcal{A}$ of the FAL $\mathcal{A}$. The pair $\{S_1, S_2\}$ is a separating pair if and only if each is either a crossing disk or a singly separated disk and their longitudinal slopes coincide.*

It is clear that any two crossing and/or singly separated disks that share longitudinal crossing-circle punctures separate a flat FAL complement. Theorem 2.8 states that these are the only separating pairs in FAL complements.

3 Flat FAL complements with multiple reflection surfaces

Proposition 2.3 shows that a reflection surface determines a lot of the geometric structure of a flat FAL complement. In this section, we will determine exactly how many reflection surfaces can exist in a flat FAL complement and exactly which flat FAL complements admit multiple reflection surfaces. This will be a useful first step towards our main goal of showing that flat FALs are determined by their complements. We begin with some preliminary observations which specify how distinct reflection-like surfaces can intersect boundary tori of cusp neighborhoods.

As described in Section 2, reflection-like surfaces induce a peripheral structure on boundary tori of cusp neighborhoods. Recall that the peripheral structure consists of two orthogonal, unoriented isotopy classes of simple closed curves (slopes), one labeled meridian and the other longitude. Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ denote a flat FAL complement with distinct reflection-like surfaces R and S, and let T denote the boundary torus of a cusp of M. Our first lemma shows that the meridian-longitude pairs determined by R and S are the same setwise.

Lemma 3.1 *Let M be a flat FAL complement with two distinct reflection-like surfaces R and S and let T be the boundary torus of a cusp of M. Then an R-meridian of T is either an S-meridian or S-longitude. Likewise, an R-longitude of T is either an S-meridian or S-longitude.*

Proof The result will follow from the fact that these meridian-longitude pairs are orthogonal.

Let T denote the boundary torus of a cusp neighborhood of M for some cusp expansion of M. On this torus, let m, ℓ denote the R-meridian and R-longitude, with geodesic lengths μ, λ, respectively. Likewise, let m', ℓ' denote the S-meridian and S-longitude with lengths μ', λ', respectively. Since meridian-longitude pairs are orthogonal

$$\mu\lambda = \mathrm{Area}(T) = \mu'\lambda'.$$

Given an arbitrary orientation on the slopes $\{m, \ell\}$ and $\{m', \ell'\}$, there are integers p, q, r, s with $m' = pm + q\ell$ and $\ell' = rm + s\ell$. By orthogonality we have

$$\mu' = \sqrt{(p\mu)^2 + (q\lambda)^2} \quad \text{and} \quad \lambda' = \sqrt{(r\mu)^2 + (s\lambda)^2}.$$

At least one of p or q is nonzero, which implies $\mu' \geq \min\{\mu, \lambda\}$, with equality exactly when m' is the shortest curve of the set $\{m, \ell\}$. Moreover, when neither p nor q are zero, $\mu' > \max\{\mu, \lambda\}$. Similar inequalities hold for λ'. If both p and q are nonzero, then

$$\mu'\lambda' > \max\{\mu, \lambda\} \min\{\mu, \lambda\} = \mu\lambda.$$

Since $\mu\lambda = \mu'\lambda'$, however, this implies the oriented and simple m' is one of $\pm m$ or $\pm \ell$. The slope m', then, is one of the slopes m or ℓ. Similarly, ℓ' equals the other slope and the sets $\{m', \ell'\}$ and $\{m, \ell\}$ are equal. □

Thus, if a flat FAL complement admits multiple reflection-like surfaces $\{R_i\}$, each such R_i-structure still induces the same basis on a boundary torus of a cusp, though their induced peripheral structures may differ since meridians and longitudes might switch roles (ie some components of $\mathcal{A}$ could change type).

Remark The proof of Lemma 3.1 would not work if we assumed M was just an FAL complement. If the corresponding FAL has some number of half twists, then certain meridian-longitude pairs might no longer be orthogonal, and so, our arguments used above would no longer apply. We point this out since multiple essential results for this paper rely on Lemma 3.1, and we want to make sure the reader knows why our work doesn't immediately apply more broadly to the class of FALs and not just flat FALs.

The next two propositions clarify how a different reflection-like surface S behaves relative to a given R-structure by considering R-knot circle and R-crossing circle cusps separately.

Proposition 3.2 *Let M be a flat FAL complement with two distinct reflection-like surfaces R and S. If K is an R-knot circle and T_K the boundary torus of a cusp corresponding to K, then $S \cap T_K$ is a pair of R-meridians of K.*

Proof Lemma 3.1 implies that $S \cap T_K$ is either a pair of R-meridians or R-longitudes. For the sake of contradiction, suppose that $S \cap T_K$ is a pair of R-longitudes.

Since K is an R-knot circle, $R \cap T_K$ is a pair of R-longitudes. First suppose that $R \cap T_K = S \cap T_K$. Then the composition $\iota_R \circ \iota_S$ acts as the identity on T_K and preserves the normal direction. Therefore $\iota_R \circ \iota_S$ is the identity on M by [3, Proposition A.2.1], and $R = S$. This is a contradiction, which implies $S \cap T_K \neq R \cap T_K$.

Now consider the case where $S \cap T_K$ are R-longitudes distinct from $R \cap T_K$, and let D be an R-crossing disk punctured by K. Note that D intersects T_K in R-meridians, which are orthogonal to the R-longitudes $S \cap T_K$. Thus S intersects D orthogonally. Moreover, R intersects D in the nonseparating geodesics of D, which implies that $S \cap D$ consists of a single separating geodesic with both "ends" on K. The other two punctures of D must be on distinct cusps because one is an R-knot circle puncture, and the other an R-crossing circle puncture. Since S intersects D orthogonally along a separating geodesic, the reflection ι_S preserves D and interchanges the two punctures on distinct cusps. This contradicts the fact that ι_S preserves all cusps, and $S \cap T_K$ cannot be R-longitudes distinct from $R \cap T_K$.

Since S cannot meet T_K in R-longitudes, the pair of curves $S \cap T_K$ are R-meridians. □

Proposition 3.3 *Let M be a flat FAL complement with two distinct reflection-like surfaces R and S. Let C be an R-crossing circle bounding the R-crossing disk D. Then either*

(i) *D is a component of S, or*

(ii) *S intersects D orthogonally along the separating geodesic of D with punctures on C. In this case the R-crossing circle C is also an S-crossing circle.*

Proof Let C be the R-crossing circle bounding D. By Lemma 3.1 we know that $S \cap T_C$ is either two R-meridians or two R-longitudes of C.

Consider the case where $S \cap T_C$ consists of two R-meridians of C. Since C is an R-crossing circle the intersection $R \cap T_C$ consists of R-meridians as well. If the surfaces R and S intersect T_C in the same curves then, as in the proof of Proposition 3.2, one arrives at the contradiction that $R = S$. Thus the R-meridians $S \cap T_C$ are distinct from those of $R \cap T_C$. Now S and D are an embedded totally geodesic surfaces in M, so their intersection is a union of disjoint, complete, simple geodesics. This follows from the observation that the local picture of $S \cap D$, as seen in the universal cover, consists of two planes intersecting along a geodesic. The only complete, simple geodesics of D that intersect T_C are two nonseparating geodesics, say γ_1, γ_2, and one separating geodesic, label it γ. Since R contains the geodesics γ_1, γ_2, their intersection with T_C lies in $R \cap T_C$. This implies that S cannot contain γ_1, γ_2, since S and R intersect T_C in distinct curves. Thus $S \cap D$ contains the separating geodesic γ and neither γ_1 nor γ_2. All other complete, simple geodesics on D intersect γ, which implies $\gamma = S \cap D$ and we are in case (ii). Orthogonality follows from the fact that the R-meridians $S \cap T_C$ are orthogonal to the R-longitude $D \cap T_C$.

It remains to prove that C is an S-crossing circle as well. Since S is reflection-like, there is a flat FAL M' with reflection surface R' and an isometry $h\colon M' \to M$ with $S = h(R')$. Note that $D' = h^{-1}(D)$ is not contained in R' since D is not contained in S. Moreover, since S intersects D in a separating geodesic, the geodesic $R' \cap D'$ separates D' as well. The classification of nonreflection thrice punctured spheres given in Theorem 2.7 shows that D' must be a singly separated disk, and all punctures of singly separated disks are crossing-circle punctures (see Figure 4). Thus all punctures of D are S-crossing circles, and C must be an S-crossing circle.

Now suppose $S \cap T_C$ consists of two R-longitudes of T_C. We will show $D \subset S$. Since C is an R-crossing circle we know $R \cap T_C$ is an R-meridian, and since D is an R-crossing disk for C it intersects T_C in a single R-longitude of T_C. Thus $D \cap T_C$ is a single curve parallel to, or equal to one of, the curves in $S \cap T_C$.

First consider the case where $D \cap T_C$ equals one of the curves of $S \cap T_C$, call it γ. Then S and D are embedded totally geodesic surfaces whose intersection contains γ which, despite being geodesic in the induced Euclidean metric on T_C, is not a geodesic in M. If S and D met transversally, their intersection could only contain geodesics. As D is connected and intersects S in a nongeodesic curve, it must be a subset of S.

Now consider the case where $D \cap T_C$ is parallel to, and distinct from, the curves of $S \cap T_C$. We will show this case leads to a contradiction. As above, let M' be a flat FAL with reflection surface R' that admits an isometry $h\colon M' \to M$ for which $S = h(R')$. Further, let $D' = h^{-1}(D)$ and $T' = h^{-1}(T_C)$.

Since we are assuming $D \cap T_C$ is disjoint from $S \cap T_C$, so also $D' \cap T'$ and $R' \cap T'$ are disjoint and D' is not a subset of R'. Then D' is a nonreflection, thrice-punctured sphere in M' and must be one of the types described in Theorem 2.7 (see Figure 4). The classification shows that punctures of nonreflection, thrice-punctured spheres are either knot meridians, crossing longitudes or crossing meridians. Of these, only the crossing meridians of singly separated disks are disjoint from the reflection surface (see Figure 4(c)).

Thus D' is a singly separated disk in M' and $D' \cap T'$ consists of two crossing circle meridians. This, however, contradicts the fact that $D \cap T_C$ is a single component.

Therefore, if $S \cap T_C$ consists of two R-longitudes of T_C, then D is a component of S and we are in case (i) of the theorem. □

3.1 Three reflection surfaces

In this subsection we classify flat FAL complements that contain more than two distinct reflection-like surfaces.

Theorem 3.4 *Suppose a flat FAL complement* $M = \mathbb{S}^3 \setminus \mathcal{A}$ *admits at least three distinct reflection-like surfaces. Then* $\mathcal{A}$ *is equivalent to the Borromean rings and* M *contains exactly three reflection-like surfaces, all of which are reflection surfaces.*

Proof First, we show that a flat FAL complement M can admit at most three distinct reflection-like surfaces. This follows from the observation that at most one reflection-like surface can satisfy each case of Proposition 3.3. To see this, let D be a crossing disk with respect to a reflection-like surface R, and suppose S_1 and S_2 are reflection-like surfaces distinct from R in M. If both S_1 and S_2 contain D (so satisfy Proposition 3.3(i)), then $\iota_{S_2} \circ \iota_{S_1}$ is the identity on D and preserves the normal direction. By [3, Proposition A.2.1] we have $S_1 = S_2$. On the other hand, if S_1 and S_2 both satisfy Proposition 3.3(ii), then both intersect D orthogonally along the same separating geodesic. This implies $\iota_{S_2} \circ \iota_{S_1}$ fixes any point on this geodesic as well as the tangent space there, so again $S_1 = S_2$.

Thus a flat FAL complement has at most three distinct reflection-like surfaces, and now suppose M has three distinct reflection-like surfaces: R, S, and P. Moreover, let D be an R-crossing disk bounded by the R-crossing circle C. Then the argument given above shows that (up to relabeling) P contains D while S and D satisfy Proposition 3.3(ii). In particular, R and S intersect T_C in disjoint pairs of R-meridians. Our goal is to show that P consists of exactly two (disjoint) thrice-punctured spheres, which allows us to quickly classify all such flat FAL complements with such a reflection surface.

Since S and D satisfy Proposition 3.3(ii), C is an S-crossing circle. We will let D' be an S-crossing disk for C. Note that $D' \neq D$ since S intersects D in a separating geodesic so D cannot be an S-crossing disk. Now consider the surfaces R and P relative to D'. Since D' intersects S orthogonally, and $R \cap T_C$ is parallel to $S \cap T_C$, we have R intersects D' orthogonally. Proposition 3.3 applied to D' and R implies that they satisfy case (ii). Hence P and D' must satisfy case (i). Thus D' is a component of P, and $D \cup D' \subset P$.

Now consider $D \cup D'$ relative to the reflection-like surface S. The disk D is an S-singly-separated disk which shares a longitudinal crossing-circle puncture with the S-crossing disk D'. Thus D' and D form a separating pair by Theorem 2.8. Since no proper subset of a reflection-like surface separates, $P = D \cup D'$.

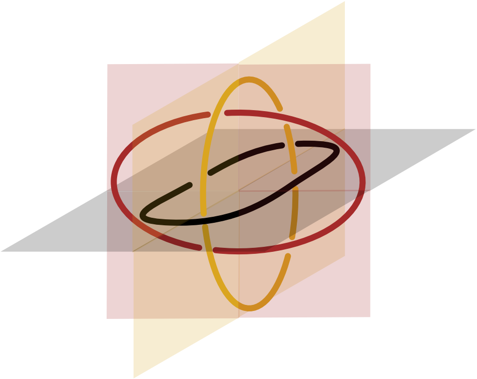


Figure 5: Three reflection surfaces in the Borromean rings complement.

The reflection-like surface $P = D \cup D'$ has a total of six punctures, and must intersect the boundary of each cusp in two curves; so there are three cusps in M. The only three-component flat FAL is the Borromean rings, so $\mathcal{A}$ must be the Borromean rings.

Finally, note that the Borromean rings complement contains three distinct reflection surfaces, as depicted in Figure 5. Each shaded plane is a reflection surface with the link component it contains serving as the one knot circle. □

3.2 Two reflection surfaces

Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ be a flat FAL complement with reflection-like surface R and an additional (distinct) reflection-like surface R'. Our goal in this subsection is to show any such flat FAL must be equivalent to either the Borromean rings, P_n with $n \geq 3$, or O_n with $n \geq 2$. See Figure 6 for FAL diagrams of P_4 and O_4.

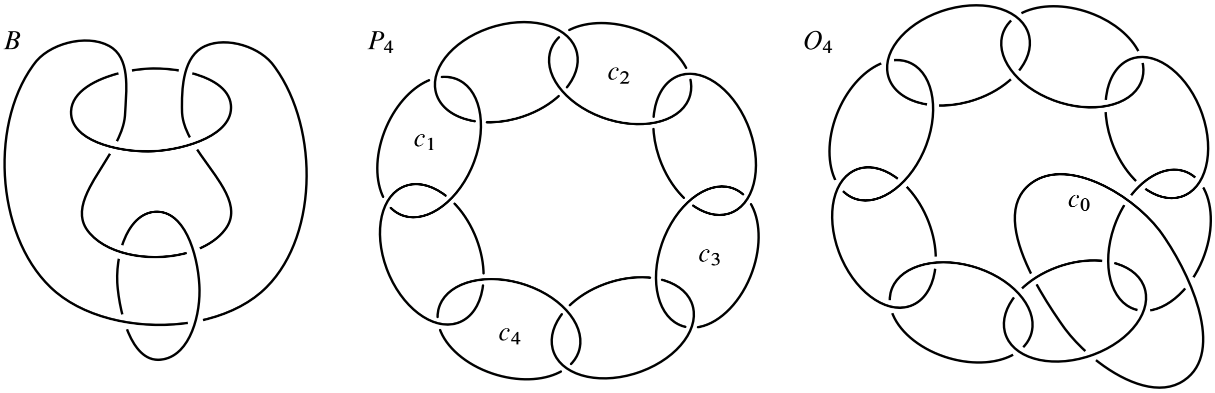


Figure 6: On the left, the Borromean rings, B, is depicted as a flat FAL. In the middle, P_4 is depicted with its four crossing circles labeled. On the right, O_4 is depicted, which can be constructed by adding the crossing circle C_0 to P_4.

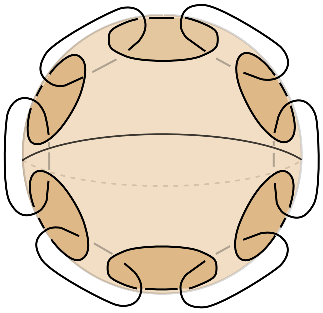


Figure 7: The second reflection-like surface in P_6.

More generally, P_n is the minimally twisted chain with $2n$ components. Thus P_n is a flat FAL that admits an FAL diagram with n crossing circles and n knot circles, linked together in a chain alternating between crossing circles and knot circles. This link can also be described as fully augmenting a pretzel link with n twist regions with an even number of crossings in each twist region, and performing a homeomorphism of the complement to undo all twists from each twist region. The link O_n is a flat FAL that admits an FAL diagram with $n+1$ crossing circles and n knot circles, where this FAL diagram can be obtained by adding a single crossing circle to the FAL diagram for P_n, just as depicted in Figure 6.

Each of P_n and O_n contain (at least) two distinct reflection surfaces: one corresponding with the projection plane in Figure 6 and one corresponding with a 2-sphere containing $C_1, \dots, C_n$ and meeting $K_1, \dots, K_n$ orthogonally in these figures; in the case of O_n, this 2-sphere meets the link component C_0 orthogonally. A 90° rotation of $\mathbb{S}^3$ about the axis of the chain interchanges the two reflection surfaces (see Figure 7 for the second reflection-like surface in P_6).

We now prove a useful fact about thrice-punctured spheres contained in a reflection surface, which we will make use of throughout this subsection.

Lemma 3.5 *Let M be a flat FAL complement with reflection-like surface R, and let D be a thrice-punctured sphere component of R. Then D has an odd number of R-knot circle punctures.*

Proof The properties involved are topological, so without loss of generality we assume that R is the reflection surface of the flat FAL complement $M = \mathbb{S}^3 \setminus \mathcal{A}$. Further let $\mathbb{S}^2$ be the projection two-sphere in M, so that $R = \mathbb{S}^2 \setminus \mathcal{A}$ is the reflection surface. Finally, let D be a thrice-punctured sphere component of R. Since each component of R has at least one knot circle puncture, we must show that D does not have two knot circle punctures.

Suppose, on the contrary, that exactly two knot circles, labeled J and K, puncture D. We will show that one of J or K has at most one crossing circle linking it, contradicting the fact that every knot circle is linked by at least two crossing circles. Some notation will be useful.

Note that $\mathbb{S}^2 \setminus (J \cup K)$ consists of two disks and an annulus. Let D_J and D_K denote the disk components bounded by J and K, respectively, and note that D is the annular component. Then D is a once-punctured annulus and the additional puncture, label it the point p, comes from a crossing circle C. Now C punctures $\mathbb{S}^2$ twice, and only once in D, so (possibly after relabeling) C also punctures D_K. Now D is a thrice-punctured sphere, so there is a unique geodesic γ_{pK} in D joining the punctures p and K. Thus, if D_C is a crossing disk bounded by C then $\gamma_{pK} = D_C \cap D$.

Now let C' be a crossing circle other than C with a crossing disk D' punctured by J. As C is the only crossing circle puncturing D, we know C' is disjoint from D. However, since J punctures D' we know $D' \cap D$ is a nonempty subset of geodesics on D with at least one endpoint on J. The separating geodesic γ_J of D with both endpoints on J intersects γ_{pK} nontrivially, since γ_{pK} is the nonseparating geodesic of D opposite J. Since $\gamma_{pK} = D_C \cap D$ and crossing disks are disjoint, D' cannot intersect D in γ_J. Further, $D' \cap D$ cannot be the nonseparating geodesic of D joining J and p since p is on the crossing circle C and $C' \neq C$. Thus $D' \cap D$ must be the nonseparating geodesic γ_{JK} of D joining J and K.

We have shown that any crossing disk punctured by J intersects D in the geodesic γ_{JK}. Since crossing disks are orthogonal to the reflection surface, at most one crossing disk intersects D along a given geodesic. Thus at most one crossing circle bounds a crossing disk punctured by J, contradicting the fact that each knot circle in an FAL is linked by at least two crossing circles. □

Each reflection-like surface partitions the cusps corresponding to components of $\mathcal{A}$ into a set of knot circle cusps and a set of crossing circle cusps. If some component of $\mathcal{A}$ changes type, then the geometry of M is greatly restricted. It is also possible that no "swapping" occurs, ie, R and R' induce the same partition on the components of $\mathcal{A}$ into knot circle (cusps) and crossing circle (cusps). We now show that the latter case can not happen when M has two distinct reflection surfaces.

Lemma 3.6 *Suppose $M = \mathbb{S}^3 \setminus \mathcal{A}$ is a flat FAL complement with two reflection-like surfaces R and R' that induce the same partitions on the cusps of M into crossing circle cusps and knot circle cusps. Then $R = R'$.*

Proof Proposition 4.6 of [13] shows that there exists some R-crossing circle C of $\mathcal{A}$ such that the corresponding R-crossing disk D is the unique totally geodesic thrice-punctured sphere in M that intersects the torus boundary of the cusp corresponding to C in a longitude and intersects the boundary of two R-knot circle cusps (not necessarily distinct) in a meridian on each of these cusps. By assumption, C is also an R'-crossing circle and the two R-knot circle cusps intersecting D are also R'-knot circle cusps. By Lemma 3.5, D can not be a subset of R', and so, D is a nonreflection thrice-punctured sphere for R'. Since D has two R'-knot circle and one R'-crossing circle punctures, Theorem 2.7 implies D must also be the corresponding R'-crossing disk for C. Let T_K be the boundary torus of a cusp neighborhood for one of the knot circles that intersect D. Then $D \cap T_K$ is orthogonal to both $R \cap T_K$ and $R' \cap T_K$. However, if $R \neq R'$, then Proposition 3.2 implies that $R \cap T_K$ and $R' \cap T_K$ are orthogonal, which provides a contradiction. Thus, $R = R'$ under these conditions. □

In the rest of this subsection, we consider the case where a component of $\mathcal{A}$ changes type. First, we give a useful lemma.

Lemma 3.7 *Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ be a flat FAL complement with two distinct reflection-like surfaces R and R'. Suppose L is an R-knot circle and an R'-crossing circle. If D' is an R'-crossing disk for L, then $D' \subset R$; moreover, D' is unique.*

Proof To see this, note that since L is an R-knot circle, the reflection surfaces R' and R intersect T_L in orthogonal pairs of curves by Proposition 3.2. As an R-knot circle there are at least two R-crossing disks that link L. Then $\partial D'$ is a single curve on T_L orthogonal to $R' \cap T_L$, so it must be an R-longitude parallel to $R \cap T_L$. As such D' intersects R-crossing disks linking L, and that intersection must be in a nonseparating geodesic with one puncture on L. To see this let D be an R-crossing disk punctured by L. If D' intersected D in a separating geodesic of D with endpoints on L then $\partial D'$ would consist of two R-longitudes of L, but $\partial D' \cap T_L$ is a single R-longitude since D' is an R'-crossing disk.

Then R and D' have a common intersection with R-crossing disks that link L, and both are orthogonal to such crossing disks, so $D' \subset R$.

To see that D' is unique, assume there is a second R'-crossing disk D^* for L. By the above argument, $D^* \subset R$ as well. Then $D' \cup D^*$ form a separating pair contained in R. Since a proper subset of R cannot separate, R must equal $D' \cup D^*$. As in the proof of Theorem 3.4, a cusp count shows manifold has three cusps and must be Borromean rings. The Borromean rings, however, do not have a crossing circle that bounds two crossing disks in same R'-structure; therefore, D' is unique. □

Suppose K is a component of $\mathcal{A}$ that corresponds with an R-knot circle cusp and an R'-crossing circle cusp. Let D' be the R'-crossing disk bound by K. Then Lemma 3.7 guarantees that D' is a component of R. Since K is an R-knot circle, there exists some R-crossing circle C that links K, along with an R-crossing disk D which K punctures. We now consider the intersection patterns for $D \cap D'$. The work of Yoshida [26, Proposition 3.1] shows that there are three possible cases to consider:

(i) $D \cap D'$ is a single nonseparating geodesic on both D and D',

(ii) $D \cap D'$ is a pair of nonseparating geodesics on both D and D', and

(iii) $D \cap D'$ is a separating geodesic on one and nonseparating on the other.

Since $D' \subset R$ and $D \cap R$ consists of nonseparating geodesics on D, the curve $D \cap D'$ must be nonseparating on D. Hence, in case (iii), the separating geodesic must be on D'. See Figure 8 for corresponding diagrams. These cases are considered separately in the next three propositions.

Proposition 3.8 *Suppose $M = \mathbb{S}^3 \setminus \mathcal{A}$ is a flat FAL complement with two distinct reflection-like surfaces R and R'. Let K be a component of $\mathcal{A}$ that is an R'-crossing circle and an R-knot circle, and let D' be the R'-crossing disk K bounds. Let C be an R-crossing circle bounding the R-crossing disk D, with K puncturing D. If the intersection $D' \cap D$ is a single nonseparating geodesic on each thrice-punctured sphere (case* (i) *of Figure 8), then $\mathcal{A}$ is either P_n with $n \geq 3$, or O_n with $n \geq 2$.*

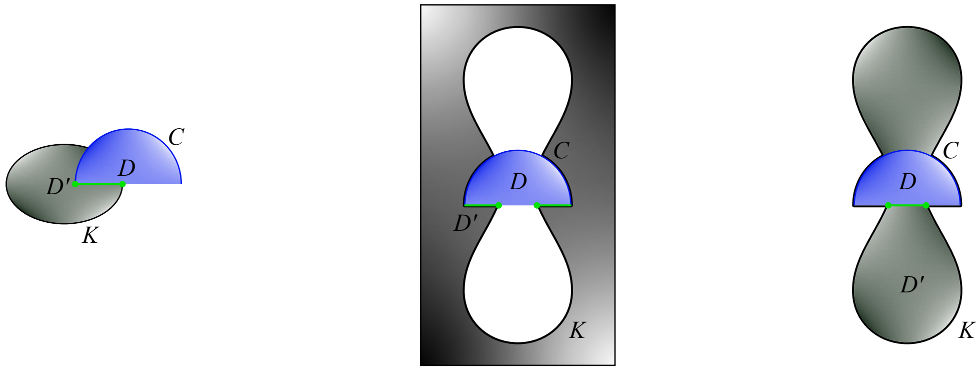


Figure 8: The three cases for intersection patterns for $D' \cap D$ where the projection plane here is determined by R. In each figure, the thrice-punctured spheres D and D' are shaded blue and gray, respectively. Only the top half of the crossing disk D is drawn in each diagram. Intersections between D and D' are highlighted in green.

Proof Let K, D', C and D be as in the statement of the proposition. By Lemma 3.7 the disk D' is a component of R and note that, since $D' \cap D$ is a single nonseparating geodesic on both, D' punctured once each by K and C. Lemma 3.5 implies that the remaining puncture of D' comes from an R-crossing circle, say J.

Since D' has punctures along three distinct cusps, the R-crossing disk D_J that J bounds intersects D' in a single nonseparating geodesic for both disks (case (i) of Figure 8). In the R-structure, then, the cusps J, K, C form a chain of three components.

Now consider J, K and C in the R'-structure. We know that K is an R'-crossing circle and that D' is its corresponding R'-crossing disk. Thus the other two punctures, J and C, must be R'-knot circles. By Lemma 3.7, then, the R-crossing disks D and D_J are components of the reflection surface R'.

Applying Lemma 3.5 to the disks D and D_J in the R'-structure implies the third puncture in each is an R'-crossing circle puncture. Translating the picture to the R-structure, and applying Lemma 3.7 to each end of the chain produces a chain of four or five components. There are four components if the punctures unaccounted for in D and D_J correspond to the same R'-crossing circle.

Since $\mathcal{A}$ has finitely many components, iterating this argument terminates in a sublink of $\mathcal{A}$ isotopic to P_n, and which we denote by $\mathcal{P}$. The R-crossing disks for $\mathcal{P}$ are components of R', and vice versa. For convenience, let $K_1, \dots, K_n$ denote the R-knot circles of $\mathcal{P}$, and let $S_1, \dots, S_n$ be the corresponding thrice-punctured sphere components of R that they bound (so the S_i are R'-crossing disks). Further let $C_1, \dots, C_n$ denote the R-crossing circles of $\mathcal{P}$, and let $D_1, \dots, D_n$ be the respective R-crossing disks that they bound.

Consider the action of $\iota_{R'}$ on the components of the reflection surface R. Each of the S_i reflect to themselves since they are R'-crossing disks. Now let $\widehat{R} = R \setminus \left(\bigcup_{i=1}^{n} S_i\right)$. Our goal is to show that the K_i are the only R-knot circle punctures of $\widehat{R}$. Cutting $\widehat{R}$ along its intersection with the D_i separates $\widehat{R}$ into two subsets $\widehat{R}_0$ and $\widehat{R}_1$, which are R'-reflections of each other. Let N be an R-knot circle distinct from the K_i that punctures $\widehat{R}$. Then N cannot pass through the D_i, which are already punctured twice by the K_i. Thus N is contained entirely in one of $\widehat{R}_0$ or $\widehat{R}_1$, and $\iota_{R'}(N)$ is in the other. This contradicts the fact that $\iota_{R'}$ preserves each cusp; therefore, no such N exists and the K_i are the only R-knot circle punctures of $\widehat{R}$. Since $R = \widehat{R} \cup \bigl(\bigcup_{i=1}^{n} S_i\bigr)$ and the only R-knot circle puncturing S_i is K_i for $i = 1, \dots, n$, we have that the K_i are the only R-knot circles puncturing R, and so, they are also the only R-knot circles of $\mathcal{A}$.

At this stage, any components of $\mathcal{A}$ that are not in $\mathcal{P}$ must be R-crossing circles that puncture $\widehat{R}$ twice. Our goal is to show there is at most one such R-crossing circle. Let $\widetilde{C}$ be an R-crossing circle of $\mathcal{A}$ that is not a component of $\mathcal{P}$, and let $\widetilde{D}$ be an R-crossing disk bounded by $\widetilde{C}$. Since the K_i are the only R-knot circles of $\mathcal{A}$, the disk $\widetilde{D}$ intersects some K_j and the thrice-punctured sphere $S_j \subset R$ that it bounds. The crossing disks of $\mathcal{P}$ intersect S_j in the two nonseparating geodesics on S_j with endpoints on K_j, so $\gamma_j = S_j \cap \widetilde{D}$ must be a separating geodesic on S_j since R-crossing disks must be disjoint.

Suppose there are at least two R-crossing circles in $\mathcal{A} \setminus \mathcal{P}$, and label a second one $\widetilde{C}'$ with R-crossing disk $\widetilde{D}'$. We now describe how to construct an open, embedded, essential annulus $A \subset M$, with boundary (in $\mathbb{S}^3$) curves $\widetilde{C}$ and $\widetilde{C}'$. Consider the standard genus one Heegaard decomposition of $\mathbb{S}^3 = T_1 \cup T_2$, where T_1 and T_2 are solid tori. A sufficiently small closed neighborhood of $\bigcup_{i=1}^{n}(K_i \cup S_i \cup C_i \cup D_i)$ produces a solid torus in $\mathbb{S}^3$ that is isotopic to one of these solid tori, say T_1, and disjoint from $\widetilde{C}$, $\widetilde{C}'$, as well as any additional R-crossing circles not in $\mathcal{P}$. In addition, by taking appropriate neighborhoods, we can assume each of $\widetilde{D}$ and $\widetilde{D}'$ intersects T_1 in a meridional disk of T_1. Note that $\partial T_1 \setminus (\widetilde{D} \cup \widetilde{D}')$ produces two embedded annuli in M, and label one of these A'. Then $A = A' \cup (\widetilde{D} \setminus \operatorname{int}(T_1)) \cup (\widetilde{D}' \setminus \operatorname{int}(T_1))$ provides the desired annulus, contradicting hyperbolicity. Thus, $\mathcal{A} \setminus \mathcal{P}$ is either empty or contains exactly one R-crossing circle that intersects $\widehat{R}$ twice, once in $\widehat{R}_0$ and once in $\widehat{R}_1$, as needed. □

Proposition 3.9 *Suppose $M = \mathbb{S}^3 \setminus \mathcal{A}$ is a flat FAL complement with two distinct reflection-like surfaces R and R'. Let K be a component of $\mathcal{A}$ that is an R'-crossing circle and R-knot circle, and let D' be the R'-crossing circle disk K bounds. Let C be an R-crossing circle bounding the R-crossing disk D, with K puncturing D. If the intersection $D' \cap D$ is a pair of nonseparating geodesics on each thrice-punctured sphere* (case (ii) *in Figure 8*), *then $\mathcal{A}$ is the Borromean rings.*

Proof By the work of Yoshida [26, Lemma 3.7], any hyperbolic 3-manifold with two thrice-punctured spheres intersecting in this manner must be a (possibly empty) Dehn filling of one of three manifolds: a certain double cover of the Whitehead link complement, the Borromean rings complement, or the minimally twisted hyperbolic 4-chain link complement. See Figure 17 in [26] for diagrams of these links. All three of these manifolds have the common hyperbolic volume of $2v_8$, where v_8 denotes the volume of a regular ideal hyperbolic octahedra. At the same time, the work of Purcell [23, Proposition 3.6] shows that the Borromean rings complement is the unique minimal volume flat FAL complement. Since nonempty Dehn filling strictly decreases volume, the only flat FAL complement with thrice-punctured spheres intersecting in this manner is the Borromean rings complement, as needed. □

Proposition 3.10 *Suppose* $M = \mathbb{S}^3 \setminus \mathcal{A}$ *is a flat FAL complement with two distinct reflection-like surfaces* R *and* R'*. Let* K *be a component of* $\mathcal{A}$ *that is an* R'*-crossing circle and* R*-knot circle, and let* D' *be the* R'*-crossing circle disk* K *bounds. Let* C *be an* R*-crossing circle bounding the* R*-crossing disk* D*, with* K *puncturing* D*. If the intersection* $D' \cap D$ *is a separating geodesic on* D' *and a nonseparating geodesic on* D (*case* (iii) *in Figure 8*)*, then* $\mathcal{A}$ *is either the Borromean rings or* O_n *with* $n \geq 2$.

Proof Let K, D', C and D be as in the statement of the proposition. By Lemma 3.7 the disk D' is a component of R, and note that it is punctured once by K. Since D' is a thrice-punctured sphere, we know that it must have two more punctures. We break down this proof into cases depending on whether those punctures come from R-crossing circles or R-knot circles. Since $D \cap D'$ is a separating geodesic γ_K on D', the thrice-punctured sphere D partitions D' into two regions separated by γ_K.

Case I Suppose the other two punctures of D' come from R-crossing circles C_1 and C_2. Further, suppose $C_1 = C_2$. Then C_1 and C_2 must puncture different regions of $D' \setminus D$ since D' must have a puncture on each side of the separating geodesic γ_k. In this case, $\mathcal{A}$ contains a Borromean rings sublink, $L_B = K \cup C \cup C_1$. Let D_1 designate the R-crossing disk corresponding to C_1. As an R-crossing disk, D_1 intersects R in the three nonseparating geodesics on D_1. Two of these geodesics must be contained in $D' \subset R$ since C_1 punctures D' in two different regions and D_1 can not intersect γ_K. Since $D' \cap D_1$ is a pair of nonseparating geodesics on D_1, it follows that this intersection is also a pair of nonseparating geodesics on D' by the work of Yoshida [26, Proposition 3.1]. Following the proof of Proposition 3.9 with D_1 replacing D implies that $A = L_B$, ie, A is the Borromean rings.

Now suppose that C_1 and C_2 are distinct R-crossing circles, each of which puncture D'. Then C_1 has an R-crossing disk D_1, which is punctured by K and some distinct R-knot circle K_1. This implies that the intersection $D' \cap D_1$ is a single nonseparating geodesic on each of these thrice-punctured spheres. Then Proposition 3.8 with D_1 replacing D shows that L is either P_n or O_n. However, P_n does not contain a crossing circle C whose crossing disk D separates a region of R, which implies that $\mathcal{A}$ must be O_n here.

Case II Suppose that at least one of the other two punctures of D' comes from an R-knot circle, K_1. We will show this case leads to a contradiction. By Lemma 3.5, D' must have an odd number of R-knot circle punctures, and so, the third puncture of D' comes from an R-knot circle distinct from K and K_1, which we label K_2. Furthermore, since γ_K is a separating geodesic on D', K_1 and K_2 must puncture different regions of $D' \setminus D$. Any R-crossing disk punctured by K_1, call it D_1, is disjoint from D and so must intersect D' in geodesic(s) disjoint from $\gamma_K = D \cap D'$. The only geodesics of D' disjoint from $D \cap D'$ are the nonseparating geodesics $\{\gamma_1, \gamma_2\}$ joining K to the other punctures K_1 and K_2, respectively. Since D_1 is punctured by K_1, it's one other R-knot circle puncture must come from K, and so, $\gamma_1 = D_1 \cap D'$. Now there is at most one connected, embedded, totally geodesic surface orthogonal to R and containing γ_1; therefore, at most one crossing disk intersects D' along γ_1. This implies at most one crossing circle links K_1, contradicting the fact that every knot circle must be linked by at least two crossing circles in an FAL. □

We can now give the following classification of flat FAL complements that admit multiple reflection surfaces. A slightly less general version of this theorem was originally stated in Theorem 1.2 in Section 1.

Theorem 3.11 *Suppose* $M = \mathbb{S}^3 \setminus \mathcal{A}$ *is a flat FAL complement with multiple distinct reflection-like surfaces. Then either*

- $\mathcal{A}$ *is equivalent to the Borromean rings, and* M *contains exactly three distinct reflection-like surfaces, all of which are reflection surfaces, or*
- $\mathcal{A}$ *is equivalent to* P_n *with* $n \geq 3$, *or* O_n *with* $n \geq 2$, *and* M *contains exactly two distinct reflection-like surfaces, both of which are reflection surfaces.*

Proof Suppose M is a flat FAL complement with at least two distinct reflection-like surfaces. Then Lemma 3.6 shows that some R'-crossing circle K of $\mathcal{A}$ must switch to become an R-knot circle (or vice versa). Let D' be the R'-crossing disk corresponding to K. By Lemma 3.7, $D' \subset R$. At the same time, since K is an R-knot circle, it punctures at least two R-crossing disks, one of which we label D. Then D and D' are both totally geodesic thrice-punctured spheres in M that intersect nontrivially since D must intersect R on each side of K (thinking of K as a simple closed curve in the projection plane), and one of these components is D'. The work of Yoshida [26, Proposition 3.1] shows that there are exactly three possibilities for such intersection patterns, which are covered in Propositions 3.8, 3.9, and 3.10. Combined, these propositions tell us that $\mathcal{A}$ is equivalent to either the Borromean rings, P_n with $n \geq 3$, or O_n with $n \geq 2$. Theorem 3.4 distinguishes the Borromean rings as the only flat FAL whose complement admits at least three distinct reflection-like surfaces. In this case, this FAL complement has exactly three reflection-like surfaces, all of which are reflection surfaces, as noted in Theorem 3.4; see Figure 5 for a visualization of the three reflection surfaces. If an FAL complement has exactly two reflection-like surfaces, then the corresponding link is either P_n with $n \geq 3$ or O_n with $n \geq 2$. The discussion at the beginning of Section 3.2 shows that the corresponding FAL complements for these links each admit two distinct reflection surfaces, and so, every reflection-like surface is also a reflection surface in all of these cases. □

This classification theorem implies that within the family of flat FALs, those whose complements admit multiple reflection surfaces are determined by their complements. We highlight this result in the following corollary.

Corollary 3.12 *Let* $\mathcal{A}$ *and* $\mathcal{A}'$ *be flat FALs, and suppose the complement of* $\mathcal{A}$ *admits multiple distinct reflection surfaces. Then* $\mathbb{S}^3 \setminus \mathcal{A}$ *is homeomorphic to* $\mathbb{S}^3 \setminus \mathcal{A}'$ *if and only if* $\mathcal{A}$ *and* $\mathcal{A}'$ *are equivalent links.*

Proof If $\mathcal{A}$ and $\mathcal{A}'$ are equivalent links, the orientation-preserving homeomorphism between the pairs $(\mathbb{S}^3, \mathcal{A})$ and $(\mathbb{S}^3, \mathcal{A}')$ induces one between their complements. So, suppose $\mathbb{S}^3 \setminus \mathcal{A}$ and $\mathbb{S}^3 \setminus \mathcal{A}'$ are homeomorphic flat FAL complements where $\mathbb{S}^3 \setminus \mathcal{A}$ admits multiple reflection surfaces. Then these flat FAL complements are isometric and $\mathbb{S}^3 \setminus A'$ also contains multiple reflection-like surfaces. By Theorem 3.11, A' is equivalent to either the Borromean rings B, P_n with $n \geq 3$, or O_n with $n \geq 2$. Thus, we just need to consider the cases where $\mathbb{S}^3 \setminus \mathcal{A}$ is homeomorphic to the complement of one of these links. Note that B has three components, each P_n has $2n$ components with $n \geq 3$, and O_n has $2n+1$ components with $n \geq 2$. Thus the number of components distinguishes the links B, $\{P_n\}_{n=3}^{\infty}$ and $\{O_n\}_{n=2}^{\infty}$. Since there is a one-to-one correspondence between components of a link and cusps of the corresponding link complement, this shows that the number of cusps of one of these link complements determines the corresponding link, completing the proof. □

The work from this section places an important restriction on the behavior of homeomorphisms between flat FAL complements.

Corollary 3.13 *Let M and M' be flat FAL complements and suppose there exists a homeomorphism $h\colon M \to M'$, which induces an isometry ρ_h. Then R is a reflection-like surface for M if and only if R is a reflection surface for M. In particular, ρ_h provides a one-to-one correspondence between reflection surfaces.*

Proof By the comments immediately following Definition 2.4, the isometry ρ_h produces a one-to-one correspondence between reflection-like surfaces. If we show that every reflection-like surface is actually a reflection surface, we will be done.

Theorem 3.11 covers the multiple reflection-like surface case. On the other hand, the reflection surface in a flat FAL is one reflection-like surface. Therefore, if a flat FAL has a unique reflection-like surface it must be the reflection surface, completing the proof. □

Before moving on, we make two useful observation that follow from Corollary 3.13. First off, we will no longer use the term reflection-like surface since a reflection-like surface is a reflection surface. In addition, if $h\colon M \to M'$ is a homeomorphism, M has a unique reflection surface R, and M' contains a reflection surface R', then M' also has a unique reflection surface and $\rho_h(R) = R'$.

Our final result from this section shows that in most cases, every symmetry of a flat FAL complement with multiple reflection surfaces is induced by a symmetry of that link. In particular, the following theorem proves the first statement from Theorem 1.3 in the introduction.

Theorem 3.14 *Let $\mathcal{A}$ be a flat FAL, other than P_3, whose complement admits multiple reflection surfaces. Then both $\mathcal{A}$ and its complement $M = \mathbb{S}^3 \setminus \mathcal{A}$ have the same symmetry group.*

Proof Since every symmetry of $(\mathbb{S}^3, \mathcal{A})$ induces one of its complement, it's enough to show that every homeomorphism $h\colon M \to M$ extends to an isotopy of $\mathbb{S}^3$.

Since $\mathcal{A}$ is not P_3, Theorem 3.11 implies it is either P_n with $n \geq 4$, O_n with $n \geq 2$, or the Borromean rings, and we consider the cases separately.

Let $\mathcal{A} = O_n$, with $n \geq 2$, and let R be a reflection surface in $M = \mathbb{S}^3 \setminus O_n$. Note that R consists of n thrice-punctured spheres and one $(n+2)$-punctured sphere S^2_{n+2} whose punctures are n longitudes along the R-knot circles of O_n and 2 punctures by the same R-crossing circle C_0. Since $n \geq 2$, S^2_{n+2} is the only component of R with more than three punctures.

Now let $h\colon M \to M$ be a homeomorphism with induced isometry $\rho_h\colon M \to M$. Corollary 3.13, applied to the case where $M' = M$, implies that $R' = \rho_h(R)$ is a reflection surface for O_n. Since R' is a reflection surface for O_n, it has a unique $(n+2)$-punctured sphere component ${S^2_{n+2}}'$. The unique cusp punctured twice by ${S^2_{n+2}}'$ corresponds to an R'-crossing circle while the remaining n cusps punctured by ${S^2_{n+2}}'$ correspond to the R'-knot circles of O_n. Since ρ_h preserves the topology of components of R we have $\rho_h(S^2_{n+2}) = {S^2_{n+2}}'$, and R-knot circles must map to R'-knot circles. Moreover, $C_0' = \rho_h(C_0)$ must be the R'-crossing circle of O_n puncturing ${S^2_{n+2}}'$ twice. The remaining components of O_n are R-crossing circles and, since all R'-knot circles are accounted for, ρ_h must map them to R'-crossing circles.

Thus ρ_h preserves the type of each component of O_n-crossing circles map to crossing circles, and similarly with knot circles. Let L be a component of O_n with image $L' = \rho_h(L)$, and let $\{m, \ell\}$ and $\{m', \ell'\}$ denote the respective R- and R'- meridian-longitude pairs. Lemma 3.1 implies ρ_h maps the set $\{m, \ell\}$ to the $\{m', \ell'\}$, and we must show it takes meridians to meridians.

If K is an R-knot circle of O_n, then $T_K \cap R$ is a pair of simple closed curves, each representing an R-longitude, which we denoted by ℓ earlier. As above, let $\{m', \ell'\}$ denote the R' meridian and longitude for $T_{K'} = \rho_h(T_K)$. Then $\rho_h(T_K \cap R) = T_{K'} \cap R'$, and so maps longitudinal slopes of K to those of K'. As above, Lemma 3.1 implies that ρ_h preserves meridians as well. Thus ρ_h preserves peripheral structures on all R-knot circles.

To see that h preserves peripheral structures on R-crossing circles, let C be an R-crossing circle with R'-crossing circle image $C' = \rho_h(C)$. Using notation similar to the above, we have the calculation $\rho_h(T_C \cap R) = T_{C'} \cap R'$ so ρ_h preserves meridional slopes, and Lemma 3.1 implies it preserves peripheral structures on crossing circles of O_n as well.

Thus ρ_h, and therefore our original $h\colon M \to M$, preserves peripheral structures on all components and extends to an isotopy of $\mathbb{S}^3$. This implies, of course, that a symmetry of M is the restriction of a symmetry of $(\mathbb{S}^3, O_n)$ to its complement.

The proof for P_n, with $n \geq 4$, follows similarly, but is a little more direct. In this case the reflection surface R consists of n thrice punctured spheres and one n-punctured sphere S^2_n. Since $n \geq 4$, the sphere S^2_n is the unique component of R with more than three punctures. The proof follows as above, with the simplification that punctures of S^2_n correspond to the distinct R-knot circles of P_n.

We have proven the theorem for flat FALs with two reflection surfaces and the only case remaining is the flat FAL with three reflection surfaces – the Borromean rings. In this case, a quick check using SnapPy confirms the symmetry group of the Borromean rings and its complement coincide, with common symmetry group $\mathbb{Z}_2 \times G$, where G represents the group of symmetries of the octahedron. $\square$

The proof of Theorem 3.14 does not extend to P_3. For P_3 all components of the reflection surface R are thrice-punctured spheres, with one punctured by three knot circles. A simple puncture count, then, does not guarantee that the isometry ρ_h preserves the component of R punctured by knot circles. Section 5 will show that this is more than just a shortcoming of the above proof. In fact P_3 is a signature link, and it will be seen that signature link complements have more symmetries than the links themselves.

4 Transition to the unique reflection surface case

Corollary 3.13 tells us that homeomorphic flat FAL complements must have the same number of reflection surfaces. This allows us to break down the proof of our main result, Theorem 1.1, into two cases:

(1) There exists a homeomorphism between flat FAL complements, each with multiple reflection surfaces. This case is already covered by Corollary 3.12.

(2) There exists a homeomorphism between flat FAL complements, each with a unique reflection surface.

We can actually put a more narrow focus on the homeomorphisms we need to analyze in case (2). Let M and M' be the complements of the flat FALs $\mathcal{A}$ and $\mathcal{A}'$, respectively, each containing unique reflection surfaces denoted by R and R'. If there exists a homeomorphism $h\colon M \to M'$, which induces isometry ρ_h, then we must have $R' = \rho_h(R)$. Recall that the reflection surfaces determine peripheral structures on each component of $\mathcal{A}$ and $\mathcal{A}'$, respectively. Thus if h preserves both knot circles and crossing circles, then it preserves peripheral structures and extends to an isotopy of $\mathbb{S}^3$, making the links $\mathcal{A}$ and $\mathcal{A}'$ equivalent. For this reason we will be mainly interested in homeomorphisms that change a knot circle K into a crossing circle, or vice versa. In this case we will say that h *changes the type* of a component K of $\mathcal{A}$, and will call h a *type-changing homeomorphism*.

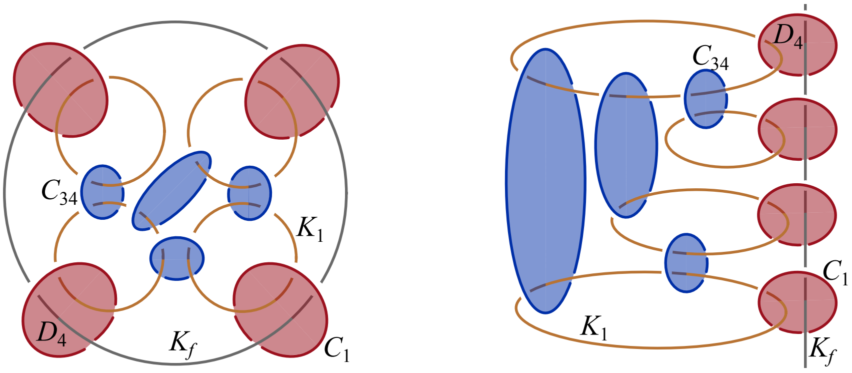
The rest of this paper is dedicated to analyzing type-changing homeomorphisms of flat FAL complements containing unique reflection surfaces. Section 5 will introduce a particular class of type-changing homeomorphisms where we can easily find an isotopy in $\mathbb{S}^3$ between the corresponding FALs. Section 6 examines how separating sets (partially introduced in Section 2) behave under type-changing homeomorphisms to help restrict the behavior of such homeomorphisms. Combining the work of these two sections, we then prove our main result in Section 7.

5 Signature links and full-swap homeomorphisms

In this section we define signature links, which are a special class of FALs, and show that they admit a particular type-changing homeomorphisms, which we call a full-swap, on their complements. Furthermore, we will show that signature links whose complements are homeomorphic via a full-swap exhibit an explicit isotopy in $\mathbb{S}^3$.

Definition 5.1 A *signature link* is a flat FAL $\mathcal{L}$ whose components can be partitioned into four nonempty sets

$$\mathcal{L} = \{K_f\} \cup \mathcal{K} \cup \mathcal{C} \cup \mathcal{C}_{\mathcal{K}},$$



(a) aesthetic diagram (b) pragmatic diagram

Figure 9: Two diagrams of a signature link $\mathcal{L}$.

with the following properties. The first set consists solely of a knot circle K_f which cuts the projection plane into two disks, one of which contains the remaining knot circles $\mathcal{K} = \{K_1, \dots, K_n\}$, which we call the inside of K_f. Each $K_i \in \mathcal{K}$ is linked with K_f by a unique crossing circle C_i, and $\mathcal{C} = \{C_1, \dots, C_n\}$. The set of remaining crossing circles is denoted by $\mathcal{C}_{\mathcal{K}}$, and components in $\mathcal{C}_{\mathcal{K}}$ link two distinct knot circles in $\mathcal{K}$. A crossing circle of $\mathcal{C}_{\mathcal{K}}$ that links $K_i, K_j \in \mathcal{K}$ will be denoted by C_{ij}.

Throughout this section, we let D_i designate a crossing disk for $C_i \in \mathcal{C}$ and we let D_{ij} designate a crossing disk for $C_{ij} \in \mathcal{C}_{\mathcal{K}}$.

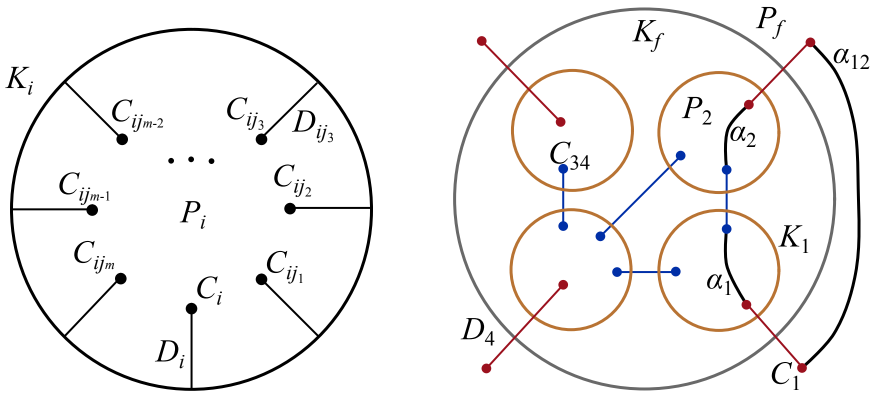
Figure 9(a) depicts a signature link. The knot circles in $\mathcal{K}$ are all inside K_f, and are numbered counterclockwise around K_f (only K_1 is labeled in Figure 9(a)). The diagram of Figure 9(b) will be helpful in visualizing a full-swap, and is obtained by isotoping K_f until it is vertical. We remark that there can be more than one way to decompose $\mathcal{L}$ as a signature link. The knot circle K_4 (unlabeled in Figure 9) could have been designated K_f instead since it is linked to every other knot circle.

According to Definition 5.1 the link P_3 is a signature link, but a quick check verifies that this is the only link whose complement contains multiple reflection surfaces which is a signature link. Thus all other signature link complements have a unique reflection surface.

The fact that $\mathcal{L}$ is hyperbolic places restrictions on the set $\mathcal{C}_{\mathcal{K}}$. The set $\mathcal{C}_{\mathcal{K}}$, for example, can not be empty. More can be said, of course, about properties of $\mathcal{C}_{\mathcal{K}}$ resulting from the hyperbolicity of $\mathcal{L}$, but we content ourselves with a result about longitudinal disks which requires the following technical lemma.

Lemma 5.2 *Two N-disks in an FAL complement are either identical or disjoint.*

Proof To see this, we show that if two N-disks intersect, then they are identical. Let D_1, D_2 be N-disks in an FAL complement M, and so, they each intersect the reflection surface R in their nonseparating geodesics. Lemma 3.4 of [26] states that thrice-punctured spheres in orientable three-manifolds cannot



(a) crossing disks in $\mathcal{P}_i$

(b) α curves for D_ℓ

Figure 10: Crossing disks and a longitudinal disk intersecting the reflection surface.

intersect along a geodesic that is separating in both. Thus, if $\gamma \in D_1 \cap D_2$, then γ is nonseparating in at least one disk, say D_1. The disk D_1 is an N-disk so γ is contained in the reflection surface R and D_2 intersects D_1 along a geodesic in R. By Theorem 2.7 both D_1 and D_2 are orthogonal to R along γ. Therefore, since D_1 and D_2 are both embedded totally geodesic surfaces that intersect in a common geodesic and are orthogonal to R, we can conclude that they must be equal, completing the proof. □

Lemma 5.3 *Let $\mathcal{L}$ be a signature link. Then every crossing circle $C_{ij} \in \mathcal{C}_{\mathcal{K}}$ bounds a totally geodesic, longitudinal disk with crossing circles C_i and C_j.*

Proof Let $\mathcal{L}$ be a signature link with complement $M = \mathbb{S}^3 \setminus \mathcal{L}$. Given a crossing circle $C_{ij} \in \mathcal{C}_{\mathcal{K}}$, we will construct a longitudinal disk with punctures C_i, C_j, C_{ij} by gluing two disks, which are essentially topological descriptions of the geodesic disks described in [20]. Afterwards, we will show this longitudinal disk is totally geodesic.

A knot circle $K_i \in \mathcal{K}$ of a signature link $\mathcal{L}$ bounds two disks in the projection plane, and we define the *inside* of K_i to be the disk $\mathcal{P}_i$ that does not contain K_f. Thus $\mathcal{P}_i$ is punctured once by C_i and once by each crossing circle of $\mathcal{C}_{\mathcal{K}}$ that links K_i.

Now consider how crossing disks intersect $\mathcal{P}_i$. Since all crossing circles in $\mathcal{L}$ link distinct knot circles, and since K_i is the only knot circle puncture of $\mathcal{P}_i$, no crossing disk intersects $\mathcal{P}_i$ in a geodesic arc with both endpoints on K_i. Crossing disks that intersect $\mathcal{P}_i$, then, do so in a geodesic joining a crossing circle puncture to the boundary curve K_i. Further, since crossing disks are disjoint they intersect $\mathcal{P}_i$ in disjoint arcs (see Figure 10(a)). The complement of crossing disks in $\mathcal{P}_i$ is then connected and there is an embedded arc, disjoint from crossing disks, between any two crossing circle punctures of $\mathcal{P}_i$.

Similarly, the knot circle K_f divides the projection plane into two topological disks, one punctured by the knot circles of $\mathcal{K}$ and the other by the crossing circles of $\mathcal{C}$. The *outside* of K_f, denoted by $\mathcal{P}_f$, refers to the disk punctured by the crossing circles of $\mathcal{C}$. The argument of the previous paragraph shows that there is an embedded arc in $\mathcal{P}_f$ joining any two crossing circle punctures that is disjoint from crossing disks of $\mathcal{L}$.

Given a crossing circle $C_{ij} \in \mathcal{C}_{\mathcal{K}}$ we construct a longitudinal disk D_ℓ, with punctures C_i, C_j, C_{ij}, by gluing a disk D_+ above the reflection surface to its reflection D_-. We begin by describing the boundary of D_+. Let $\alpha_i \subset \mathcal{P}_i$ be an embedded arc disjoint from crossing disks that joins C_i and C_{ij} punctures, and define α_j similarly. Also let α_{ij} denote an embedded arc in $\mathcal{P}_f$ which is disjoint from crossing disks and joins the C_i and C_j punctures. Figure 10(b), for example, illustrates the α arcs corresponding to the crossing circle C_{12} of the signature link in Figure 9.

The arcs α_i, α_j, α_{ij}, together with the top halves of the crossing circles C_i, C_j, C_{ij}, form a simple closed curve γ in $\mathbb{S}^3$. Let M^+ be the region of M above the reflection surface and we wish to show that γ bounds a disk in M^+. First, M^+ is a handlebody, since it is a three-ball with arcs removed for each crossing circle. Further, the top half of each crossing disk is a meridional disk for each handle. Thus removing an open neighborhood of each crossing disk from M^+ results in a three-ball with γ in its boundary. The curve γ, then, bounds a disk D_+ in M^+. The disk D_+ is disjoint from crossing disks and intersects the reflection surface along the α arcs in its boundary. Let D_- be the reflection of D_+, and let $D_\ell = D_+ \cup D_-$. Then the crossing circles C_i, C_j, C_{ij} form the boundary of D_ℓ in $\mathbb{S}^3$, and the interior of D_ℓ is an embedded thrice-punctured sphere in M with longitudinal punctures along the crossing circles C_i, C_j, C_{ij}, as desired.

To see that D_ℓ is totally geodesic it is enough to show that it is incompressible and boundary incompressible, by Theorem 3.1 of [1]. The proof given here is essentially that of [23, Lemma 2.1], but slightly simpler because the punctures of D_ℓ are distinct. Suppose α is a curve in D_ℓ that bounds a compressing disk $D \subset M \setminus D_\ell$. Then α separates D_ℓ into two pieces, one of which contains a single puncture C of D_ℓ. Thus $\alpha \cup C$ bound an annulus in D_ℓ whose union with D is a boundary compressing disk for the crossing circle C, contradicting the fact that M is hyperbolic.

When discussing ∂-incompressibility it will be convenient to think of M as the interior of a orientable, closed three-manifold $\overline{M}$ with torus boundary components. In this case, D_ℓ is properly embedded in $\overline{M}$ with longitudinal boundary curves along T_i, T_j, T_{ij}, the boundary tori of $\overline{M}$ corresponding to cusps C_i, C_j, C_{ij} of M. For convenience, and by an abuse of notation, we let C_i, C_j, C_{ij} denote the boundary curves of D_ℓ as well.

Now suppose that D is a ∂-compressing disk for D_ℓ. Then $\partial D = \alpha \cup \beta$ where $\alpha = D \cap D_\ell$ is an arc in D_ℓ with both endpoints on the same boundary curve, say C_i, and β is an arc in T_i. The arc α is isotopic to a separating geodesic of D_ℓ, which decomposes D_ℓ into two annuli, and we let A denote the one containing C_j. The other boundary component of the annulus A consists of the arc α together with a subarc of C_i, call it δ. Gluing the ∂-compressing disk D to A along their intersection α yields another annulus $A \cup D$ with C_j as one boundary component. The other boundary component of $A \cup D$ is the simple closed curve $\beta \cup \delta$ on the boundary torus T_i.

If $\beta \cup \delta$ bounds a disk on T_i, a copy of it in M caps off one boundary component of $A \cup D$, creating a ∂-compressing disk for C_j, which is impossible. On the other hand, suppose $\beta \cup \delta$ is nontrivial on T_i. Then $A \cup D$ is an incompressible annulus which is not boundary parallel since its boundary curves are on separate boundary components of $\overline{M}$. Again, this contradicts the fact that M is hyperbolic.

Thus D_ℓ is incompressible and ∂-incompressible and it has a totally geodesic representative by [1, Theorem 3.1]. □

The remainder of this section is devoted to constructing a "full-swap", which is a type-changing homeomorphism of the complement of a signature link $\mathcal{L}$. Full-swaps, despite changing the types of some components of $\mathcal{L}$, will be shown to produce a link equivalent to $\mathcal{L}$.

To begin, we define an ml-swap homeomorphism on a flat FAL complement $M = \mathbb{S}^3 \setminus \mathcal{A}$ in terms of Dehn twists on a Hopf sublink $\mathcal{H} \subset \mathcal{A}$. In particular, the Hopf sublink will consist of a knot and crossing circle pair that are linked. Zevenbergen, in [27], first constructed a product that exchanged meridional and longitudinal slopes on each component of $\mathcal{H}$ and was able to analyze the effect on the other components of $\mathcal{A}$. We review his construction here, then provide an alternative cut-and-paste construction which highlights how ml-swaps effect reflection surfaces and crossing disks.

Figure 11 illustrates an ml-swap on the Hopf sublink determined by the knot- and crossing-circle pair labeled $\{K, C\}$. Before describing the Dehn twists we observe some features of $\mathcal{A}$ relative to $\mathcal{H}$. Since $\mathcal{H} = K \cup C$ is a Hopf link, the crossing circle C must link distinct knot circles and we label the other one J. Let D be a crossing disk for C and let N be an open regular neighborhood of the cell complex $K \cup C \cup D$ in $\mathbb{S}^3$. Then N is an unknotted open solid torus, so $W = \mathbb{S}^3 \setminus N$ is a closed unknotted solid torus in $\mathbb{S}^3$. The neighborhood N can be chosen so that $N \cap \mathcal{A}$ contains the components K and C, and an arc of J that intersects D (see Figure 11(a)). Then W contains all components of $\mathcal{A} \setminus (K \cup C \cup J)$ together with one arc of J.

We now describe Zevenbergen's Dehn twists that make up an ml-swap and, abusing notation, we refer to components by their original labels throughout the Dehn twist process. First perform a Dehn twist along K, which adds a full twist to W and links C around W as in Figure 11(b). For simplicity, isotope C so that it is flat, twisting J and K vertical in the process, to get Figure 11(c). Now perform a Dehn twist along C that untwists W so that it is returned to its original form. This twist unlinks K and W while linking the arc of J with both W and K, as in Figure 11(d). Finally, perform a Dehn twist on K that unlinks C and J, obtaining the link depicted in Figure 11(e). This composition of Dehn twists is an ml-swap. The result will not be another flat FAL in general, as observed in [27], but we will see that performing multiple ml-swaps on signature links can produce a flat FAL.

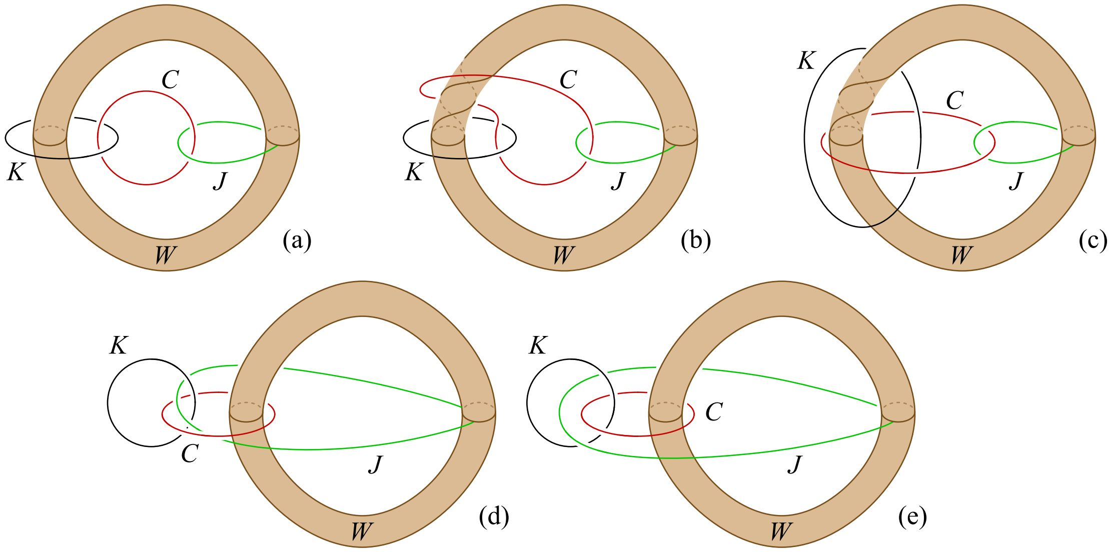


Figure 11: An ml-swap as a product of Dehn twists.

This product of Dehn twists is ultimately a local operation in the sense that changes to the link occur within a 3-ball containing the Hopf sublink. Moreover, if there are multiple Hopf sublinks that are contained in disjoint three-balls, then the result of performing Dehn twists on each Hopf sublink is independent of the order in which they are done. With this background in place we make the following definition.

Definition 5.4 An *ml-swap* on a Hopf sublink $\mathcal{H}$ of a flat FAL $\mathcal{A}$ is the homeomorphism resulting from performing the Dehn twists just described above on the components of $\mathcal{H}$. A *full-swap* on a signature link $\mathcal{L}$ is the composition of all ml-swaps on Hopf sublinks $K_i \cup C_i$, for $i = 1, \dots, n$.

Dehn twists provide a convenient description of an ml-swap, and we now consider an alternative description that highlights the effect of an ml-swap on the reflection surface and crossing disks involved. Figure 12(a) highlights a crossing circle C and knot circle K whose meridians and longitudes will be swapped. The other knot circle linked by C is included to emphasize how the reflection surface moves, but the rest of the link is not pictured. The homeomorphism maps C to the knot circle C' and K to the crossing circle K' depicted in Figure 12(g). An ml-swap preserves the reflection surface, while moving the location of the component R_0 to that of R_0'.

Let us walk through the homeomorphism one step at a time. In Figure 12(b), torus neighborhoods of C and K are pictured to emphasize what happens to meridians and longitudes. Now slice the manifold along the reflection surface and consider the top half pictured in Figure 12(c), which is a handlebody H_+ (the homeomorphism on the bottom half is the reflection of that pictured). In Figure 12(c), the half-tori around K and C, as well as the copy of R_0, form three annuli. The middle row of Figure 12 depicts an isotopy sliding the three annuli along a handle of H_+. In the process meridians and longitudes of K and

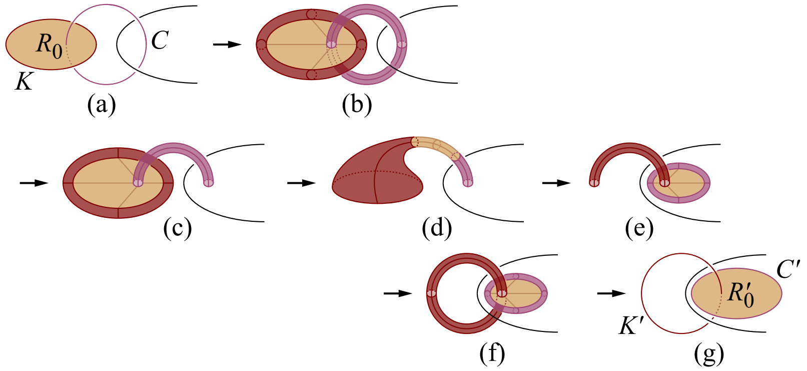


Figure 12: An ml-swap.

C are swapped. The final row depicts regluing the isotoped $H_\pm$ along R, and finally removing the torus neighborhoods of K' and C'. In terms of peripheral structures on cusps of M, an ml-swap swaps meridians and longitudes on cusps corresponding to C and K, and leaves the remaining structures the same.

It is instructive to consider the image under an ml-swap of crossing disks punctured by the knot circle involved. The image of a crossing disk D bounded by C is the natural first choice to consider. Figure 13 illustrates that D is sliced in half, each half rotated by “a third”, then reglued along its nonseparating geodesics. A similar rotation is done on the bottom half, so swapping the types of C and K has the effect of “rotating” D.

Let C^* be a crossing circle other than C linking K. Let D^* a crossing disk bounded by C^*, and consider the image of D^* under the ml-swap. Since the K-puncture of D^* becomes a crossing circle puncture, its image $D^{*\prime}$ has two crossing circle punctures. This is illustrated in Figure 14. Since an N-disk in an FAL complement has either one or three crossing circle punctures, we know $D^{*\prime}$ is not a crossing disk, assuming the image of $M = \mathbb{S}^3 \setminus \mathcal{A}$ under this ml-swap is a flat FAL complement.

In fact, an ml-swap on a flat FAL can result in a link that is not an FAL (see [27]). Performing ml-swaps on all possible (C_i, K_i) pairs in a signature link (ie a full-swap), however, does produce another flat FAL. Consider, for example, the simplest signature link: the chain P_3 in Figure 15(a). A full-swap homeomorphism h_f is realized by performing successive ml-swaps on the Hopf sublinks $C_1 \cup K_1$ and $C_2 \cup K_2$, which yields the sequence of Figure 15.

The image of P_3 is isotopic to P_3. Proposition 5.5 will show that this holds more generally — that the image of a signature link under a full-swap is always isotopic to the original link. While full-swaps

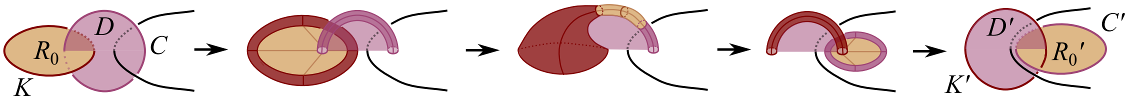


Figure 13: Rotating a crossing disk in an ml-swap.

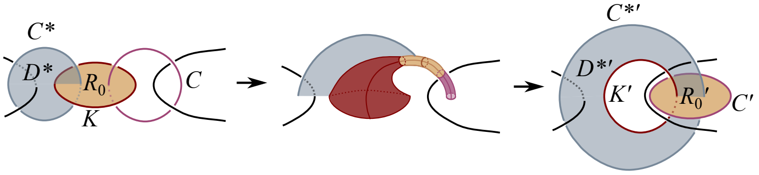


Figure 14: Sliding a crossing disk in an ml-swap.

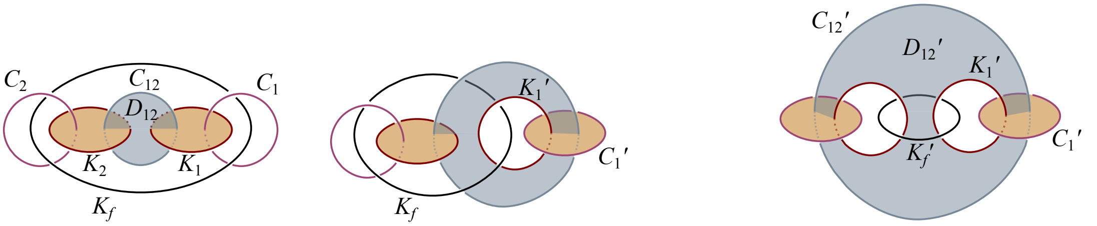
produce equivalent links, they do interchange some crossing- and longitudinal-disks. For example, in Figure 15(c), the image of the crossing disk D_{12} is the longitudinal disk D'_{12}. In addition, the longitudinal disk with punctures $\{C_1, C_2, C_{12}\}$ of Figure 15(a) becomes the crossing disk that C'_{12} bounds (neither are pictured, but reading Figure 15 backwards illustrates the change from longitudinal to crossing disk). The proof of Proposition 5.5 will show that a full-swap on an arbitrary signature link interchanges crossing- and longitudinal-disks for every crossing circle in $\mathcal{C}_{\mathcal{K}}$.

The pragmatic diagram of Figure 9(b) will be more convenient for the proof of Proposition 5.5, so we assume K_f is vertical and crossing circles of $\mathcal{C}$ are ordered from bottom to top. Figure 16 illustrates the effect of a full-swap on such a diagram of a signature link. Do note that h_f *does not* map the crossing disk for C_{24} to that of C'_{24} but to the longitudinal disk with punctures K'_2, K'_4, and C'_{24}. This happens on a more general scale and will be justified in the following proof.

Proposition 5.5 *Let $\mathcal{L}$ be a signature link and let the homeomorphism h_f designate the full-swap on all Hopf sublinks $C_i \cup K_i$. Then $\mathcal{L}$ and $h_f(\mathcal{L})$ are isotopic links. In particular, $h_f(\mathcal{L})$ is a signature link.*

Proof Let $\mathcal{L} = \{K_f\} \cup \mathcal{K} \cup \mathcal{C} \cup \mathcal{C}_{\mathcal{K}}$ be a signature link with $\mathcal{K}$ and $\mathcal{C}$ each containing n components. Since each ml-swap maps a link in $\mathbb{S}^3$ to a link in $\mathbb{S}^3$, we have that $h_f(\mathcal{L})$ is a link in $\mathbb{S}^3$ by construction. We will describe how to build $h_f(\mathcal{L})$ by performing a full-swap on the pragmatic diagram of $\mathcal{L}$, where K_f corresponds with the z-axis and so that each C_j is contained in a vertical translate of the xy-plane at height $z = j$, as depicted on the left side of Figure 16. Here, the yz-plane corresponds with the projection plane for an FAL diagram of $\mathcal{L}$ with each $C_{ij} \in \mathcal{C}_{\mathcal{K}}$ linking only K_i and K_j and meeting the yz-plane orthogonally.

Now, consider the sublink $\mathcal{L}_s = \{K_f\} \cup \mathcal{K} \cup \mathcal{C}$ and its image $h_f(\mathcal{L}_s)$. Recall that h_f is the composition of n ml-swaps, one performed on each Hopf sublink $(C_i, K_i) \in \mathcal{C} \times \mathcal{K}$ for $i = 1, \dots, n$; see Figure 12 for the



(a) original FAL (b) non-FAL after one ml-swap (c) FAL isotopic to original after two ml-swaps

Figure 15: Two ml-swaps.

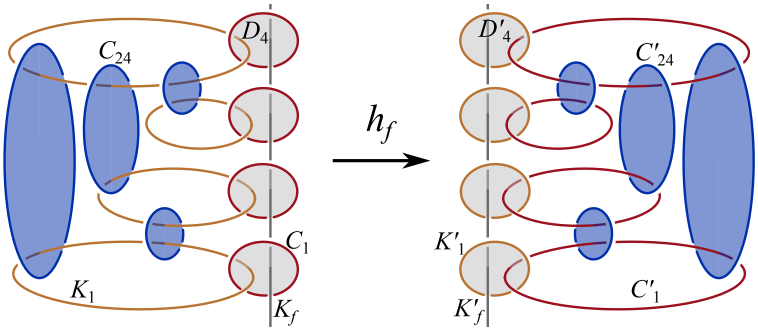


Figure 16: The type-changing homeomorphism h_f applied to a signature link.

local picture of a single ml-swap. Then $h_f(\mathcal{L}_s)$ is constructed from $\mathcal{L}_s$ by keeping K_f fixed as the z-axis, while each Hopf sublink $h_f(K_j \cup C_j)$ links $K'_f = h_f(K_f)$ via $K'_j = h_f(K_j)$. In addition, each K'_j is contained in the vertical translate of the xy-plane at height $z = j$, as depicted on the right in Figure 16.

We will now show that the images of the crossing circles in $\mathcal{C}_{\mathcal{K}}$ under h_f are unlinked unknots, which will help us determine how $h_f(\mathcal{C}_{\mathcal{K}})$ behaves. Every component of $\mathcal{L}$ is unknotted, and links those components used in the Dehn twists of an ml-swap at most once. In this situation, the Dehn twists never knot an unknotted component so the image of each component in $\mathcal{L}$ is an unknot as well. Further, an ml-swap does not link two crossing circles of $\mathcal{C}_{\mathcal{K}}$ because the linking introduced by the first Dehn twist along K in Figure 11 is undone by the following twist along C. In addition, since the ml-swaps that comprise a full-swap occur in disjoint 3-balls, the same applies to a full-swap.

Now unknots in $\mathbb{S}^3$ have a canonical peripheral structure in which a meridian links the component once and a longitude bounds an embedded disk in its complement. An ml-swap preserves this $\mathbb{S}^3$-peripheral structure on all components of $\mathcal{L}$ except K and C, for which it swaps meridians and longitudes. This observation allows us to discuss the *topology* of the images of thrice-punctured spheres under a full-swap by analyzing images of their punctures.

A crossing disk D_{ij} has a longitudinal puncture along C_{ij} and meridional punctures along K_i and K_j. Since meridians of K_i, K_j map to $\mathbb{S}^3$-longitudes of K'_i, K'_j, the image D'_{ij} of D_{ij} under a full-swap is a thrice-punctured sphere with longitudinal punctures along each of C'_{ij}, K'_i and K'_j. Similarly, the longitudinal disk D^{ℓ}_{ij} (guaranteed by Lemma 5.3) has longitudinal punctures along C_{ij}, C_i and C_j. A full-swap maps longitudes of C_i and C_j to meridians in the $\mathbb{S}^3$-peripheral structure of C'_i and C'_j, so $D^{\ell\,\prime}_{ij}$ has a $\mathbb{S}^3$-longitudinal puncture along C'_{ij}, and $\mathbb{S}^3$-meridional punctures along C'_i and C'_j. The component C'_{ij}, then, bounds an embedded disk in $\mathbb{S}^3$ punctured once by each of C'_i and C'_j. This implies C'_{ij} links only C'_i and C'_j.

Using these representatives for $\mathcal{L}$ and $h_f(\mathcal{L})$, we see that a rotation along the z-axis by 180° provides the necessary isotopy between $\mathcal{L}$ and $h_f(\mathcal{L})$ in $\mathbb{R}^3 \cup \{\infty\} \cong \mathbb{S}^3$. Furthermore, $h_f(\mathcal{L})$ is a signature link $\{K'_f\} \cup \mathcal{K}' \cup \mathcal{C}' \cup \mathcal{C}'_{\mathcal{K}}$, where $K'_f = h_f(K_f)$, $\mathcal{K}' = h_f(\mathcal{C})$, $\mathcal{C}' = h_f(\mathcal{K})$, and $\mathcal{C}'_{\mathcal{K}} = h_f(\mathcal{C}_{\mathcal{K}})$. □

6 Separating sets

In this section, we will utilize two important subsets of thrice-punctured spheres whose removal separates an FAL complement: separating pairs and separating quadruples. The behavior of type-changing homeomorphisms between flat FAL complements is significantly restricted by the existence of separating sets.

Separating pairs were introduced in Section 2 as a pair of disjoint thrice-punctured spheres whose union separates an FAL complement. Theorem 4.8 from [20] was also introduced in that section, which states that a pair of thrice-punctured spheres $\{S_1, S_2\}$ is a separating pair if and only if each is either a crossing disk or a singly separated disk and their longitudinal slopes coincide.

We now introduce a particular type of separating set consisting of four thrice-punctured spheres. Let D be a longitudinal disk with longitudinal punctures along the crossing circles C_1, C_2, C_3, and let D_i be a crossing disk with crossing circle puncture C_i. Then the set $Q = \{D, D_1, D_2, D_3\}$ is a separating set of four thrice-punctured spheres. The sets we are concerned with have one additional property.

Definition 6.1 Let D be a longitudinal disk in a flat FAL complement M with crossing circle punctures C_1, C_2, and C_3 that bound crossing disks D_1, D_2 and D_3. Then $Q = \{D, D_1, D_2, D_3\}$ is a *separating quadruple* if each crossing disk D_i is punctured by distinct knot circles.

The signature links of Section 5 contain separating quadruples. To see this, note that in a signature link every crossing circle links distinct knot circles. Moreover, Lemma 5.3 guarantees the existence of a longitudinal disk D^{ℓ}_{ij} for every $C_{ij} \in \mathcal{C}_{\mathcal{K}}$. Thus, each triple (C_i, C_j, C_{ij}) in a signature link generates a separating quadruple Q_{ij}. The remark below will show that Q_{ij} is unique up to a choice of crossing disks.

We introduce some terminology. A general separating quadruple Q is illustrated in Figure 17, which depicts only those components which puncture Q. In a flat FAL knot circles cannot cross each other, so knot circles puncture adjacent crossing disks. Let K_i denote the knot circle puncturing Q that is *opposite* the crossing circle C_i in the sense that they are not linked. By an abuse of terminology, a *component of* (*or in*) Q will refer to components that puncture disks in Q.

Remark We also point out that two separating quadruples with the same crossing circle punctures have the same knot circle punctures and longitudinal disk. Suppose Q and Q' are separating quadruples with the same crossing circle punctures. Lemma 4.5 of [20] shows that there is at most one longitudinal disk containing any two given crossing circle punctures, let alone three, so Q and Q' have the same longitudinal disk. Now recall that crossing circles in a separating quadruple link distinct knot circles, and only those components of $\mathcal{A}$. This implies that every disk they bound is punctured by the same two knot circles. Hence two separating quadruples sharing the same crossing circle punctures can differ only in their crossing disks.

We begin with the following lemma which immediately leads to a special case, Corollary 6.3, of our main result. This lemma will also be essential for establishing some technical results on how separating quadruples can behave under a type-changing homeomorphism.

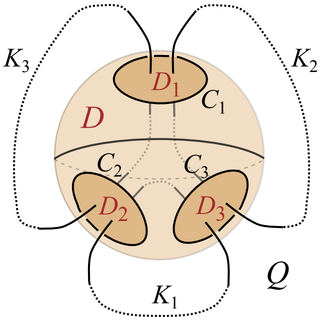


Figure 17: A separating quadruple.

Lemma 6.2 *Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ be a flat FAL complement with a unique reflection surface, and let C be a crossing circle in $\mathcal{A}$ that links the same knot circle K twice. If $h\colon M \to M'$ is a homeomorphism of flat FAL complements, then h preserves the types of both C and K.*

Proof Let C be a crossing circle in a flat FAL complement M containing a unique reflection surface R. Further, let C bound a crossing disk D which is punctured twice by the same knot circle K, and suppose $h\colon M \to M'$ is a homeomorphism of flat FAL complements. The assumption that the reflection surface $R \subset M$ is unique implies that M' has a unique reflection surface R', and that $R' = h(R)$.

We first prove that $K' = h(K)$ must be a knot circle in M'. Suppose, on the contrary, that K' is a crossing circle in M'. Then meridians of K map to longitudes of K' because both are perpendicular to reflection surfaces and $R' = h(R)$. The fact that $R' = h(R)$ further implies that $D' = h(D)$ is a nonreflection thrice-punctured sphere in M', since D is in M. The two meridional K-punctures of D map to two longitudinal K'-punctures of D'. However, by Theorem 2.7, nonreflection thrice-punctured spheres in a flat FAL complement do not have two longitudinal punctures along the same crossing circle. So, K' must be a knot circle in M'.

Since K' is a knot circle puncturing the disk D' twice in M', the remaining puncture of D' must be an M' crossing circle by the characterization of Theorem 2.7. The remaining puncture of D' is the image $C' = h(C)$ of C, and h preserves the types of both C and K. □

If a flat FAL (whose complement has a unique reflection surface) has a single knot circle, as is the case for FALs of two-bridge links with an even number of twist regions (other than the Borromean rings, which has three reflection surfaces), then every crossing circle links the same knot circle twice. Thus any homeomorphism preserves the type of all components and their peripheral structures and can be realized by an isotopy of $\mathbb{S}^3$. This observation leads to the following immediate corollary of Lemma 6.2:

Corollary 6.3 *A flat FAL with a single knot circle is determined by its complement among all flat FALs. In particular, flat FALs corresponding to 2-bridge links with an even number of twists are determined by their complements.*

Proof The only flat FAL with one knot circle and multiple reflection surfaces is the Borromean rings, which we've already seen to be determined by its complement. The unique reflection surface case follows from Lemma 6.2. □

Remark We would like to emphasize the necessity of the unique reflection surface hypothesis in Lemma 6.2. For instance, as noted in Theorem 3.4, the Borromean rings complement admits three distinct reflection surfaces and the flat FAL diagram for this link has two crossing circles and a single knot circle that links each crossing circle twice. However, there exists homeomorphisms of the Borromean rings complement where both a crossing circle and the knot circle switch types.

We now prove a technical lemma considering homeomorphisms that change a crossing disk to a longitudinal disk, or vice versa. It turns out that such a homeomorphism h implies the existence of a separating quadruple Q, and the action of h on Q can be made quite precise.

Lemma 6.4 *Let M and M' be homeomorphic flat FAL complements with unique reflection surfaces, and $h\colon M \to M'$ a homeomorphism that changes the type of an N-disk $D \subset M$. Then*:

- (i) *The disk D is part of a separating quadruple Q in M whose image $Q' = h(Q)$ in M' is also a separating quadruple.*
- (ii) *The homeomorphism h fixes the types of exactly one opposite knot- and crossing-circle pair K_f, C_f in Q.*
- (iii) *The longitudinal disk and exactly one crossing disk in Q change type under h. These disks share the crossing circle puncture C_f.*

Proof We are given that M and M' are homeomorphic flat FAL complements with unique reflection surfaces, and that $h\colon M \to M'$ is a homeomorphism changing the type of an N-disk D. Since M and M' each have unique reflection surfaces R and R', we have $R' = h(R)$. Then the image of a nonreflection thrice-punctured sphere in M is nonreflection in M'. Further, N-disks and singly separated disks are distinguished by the topological property of separating, which implies that h maps N-disks to N-disks.

Consider first the case where D is a longitudinal disk whose image $D' = h(D)$ is a crossing disk. Let C_1, C_2, C_3 denote the crossing circle punctures of D, and let D_i be a choice of crossing disk bounded by C_i. To show D is part of a separating quadruple we must show that the D_i are punctured by distinct knot circle components.

Two of the crossing circle punctures of D, say C_2, C_3, change type because D' is a crossing disk. Thus Lemma 6.2 implies D_2, D_3 are punctured by distinct knot circle punctures. Now consider D_1, whose image D_1' must be a crossing disk or longitudinal disk since N-disks map to N-disks. If D_1' stays a crossing disk, then Theorem 2.8 implies that $\{D_1', D'\}$ is a separating pair since both are crossing disks sharing the crossing circle puncture C_1'. This cannot happen because the pair $\{D_1, D\}$ does not separate in M (again by Theorem 2.8). Therefore, D_1 changes type and D_1' must be a longitudinal disk. Then, by Lemma 6.2, D_1 has distinct knot circle punctures and $Q = \{D, D_1, D_2, D_3\}$ is a separating quadruple in M.

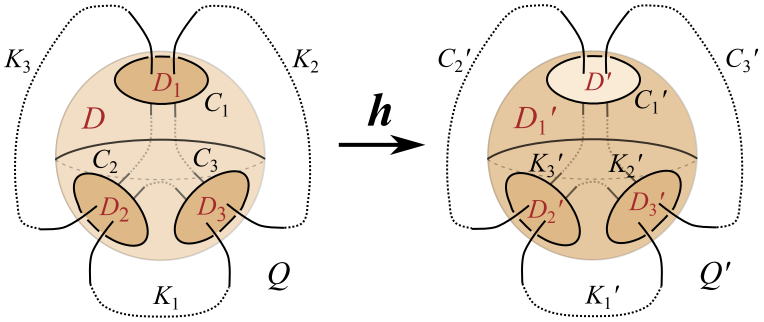


Figure 18: Homeomorphic image of Q when longitudinal disk changes type.

For convenience, label the knot circle punctures K_1, K_2, K_3, where K_i is the knot circle opposite C_i in that it does not puncture D_i.

To see that $Q' = \{D', D_1', D_2', D_3'\}$ is also a separating quadruple, we must show that it consists of three crossing disks, with distinct knot circle punctures, and one longitudinal disk. The disk D' is assumed to be a crossing disk and, in this case, the disk D_1' was shown to be longitudinal. Moreover, since C_2', C_3' are knot circles, the disks D_2', D_3' must be crossing disks in M' because these are the only nonreflection disks with knot circle punctures (Theorem 2.7). Further, each of the disks in Q are punctured by three distinct components, so their images are as well, and each crossing disk in Q' is punctured by distinct knot circle components. Thus Q' is a separating quadruple in M'.

This analysis proves the third conclusion as well, since D, D_1 change type while the crossing disks D_2, D_3 do not.

To see statement (ii) note that C_1', K_2', K_3' are the crossing circle punctures of Q' since they are the punctures of the longitudinal disk D_1'. The remaining punctures of Q' must be knot circles in M', so K_1' is a knot circle. Thus h preserves the type of the opposite knot- and crossing-circle pair C_1, K_1, while changing the type of all other components. Observe that C_1 is the crossing circle puncture shared by the disks that change type, namely D and D_1.

Now suppose D is a crossing disk in M that changes type to a longitudinal disk D' in M'. Apply the previous argument to h^{-1} and D', then note that if h^{-1} and D' satisfy the conclusions of the lemma, then so does h and D. □

Lemma 6.4 can be applied to the full-swap homeomorphisms discussed in Section 5, revealing some of the geometric structure inherent in such homeomorphisms. Before proceeding we remark that any homeomorphism between flat FAL complements with unique reflection surfaces that does not preserve peripheral structures must change the type of a knot circle. Indeed, if a crossing circle changes type, then one of the knot circles it links changes type as well since there are no (nonreflection) thrice-punctured spheres with three knot circle punctures. Thus if h changes the type of a component, there must be a knot circle that changes type.

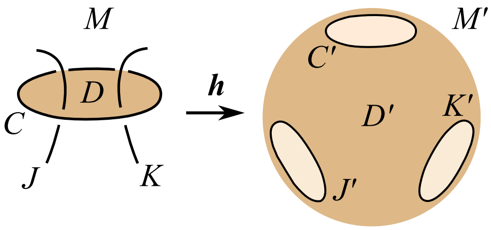


Figure 19: Crossing disk D maps to longitudinal disk D'.

Lemma 6.5 *Let $h\colon M \to M'$ be a homeomorphism between flat FAL complements with unique reflection surfaces, and suppose h changes the type of the knot circle K in M. Then h changes the type of exactly one crossing circle C_1 linked by K and, of all crossing disks punctured by K, h fixes the type of exactly those bounded by C_1.*

Proof First we show h changes the type of at most one crossing circle linking K. Suppose, on the contrary, that C_1, C_2 are distinct crossing circles linking K whose images C_1', C_2' are both knot circles in M'. Let D_1, D_2 be a choice of crossing disks they bound and note that, since C_1, C_2 are distinct, the disks D_1 and D_2 do not form a separating pair by Theorem 2.8. To determine if the image of an N-disk is a crossing disk, it is enough to show that it has a knot circle puncture because homeomorphisms between flat FAL complements with unique reflection surfaces map N-disks to N-disks. Since $D_i' = h(D_i)$ is punctured by the knot circle C_i', the disk D_i' must be a crossing disk in M'. The disks D_1', D_2' also share the crossing circle puncture K' and so form a separating pair in M', again by Theorem 2.8. The homeomorphic image of a nonseparating set, however, cannot be separating, and h changes the type of at most one crossing circle linking K.

Now we argue that the image of at least one crossing circle linking K is a knot circle in M'. Since K is linked by at least two crossing circles, at most one of which can change type, there is a crossing circle C linking K whose image C' is a crossing circle in M'. Let D be a crossing disk bounded by C and J be the knot circle other than K which punctures D. By Lemma 6.2, $J \neq K$. Since $D' = h(D)$ has two crossing circle punctures in K' and C', we find J' must also be a crossing circle. Hence D' is a longitudinal disk in M' (see Figure 19).

Thus D is a crossing disk in M that changes type, and Lemma 6.4 shows that D is part of a separating quadruple $Q = \{D, D_1, D_2, D_3\}$. Assume we've labeled disks so that D_3 is the longitudinal disk, and J punctures D_2 while K punctures D_1. Finally, let C_1 be crossing circle puncture of D_1 and K_f the final knot circle in Q. Figure 20 depicts a "schematic" diagram of this labeling in the sense that components of the FAL which are not in Q are not pictured.

Now h changes the crossing disk D to a longitudinal disk, so Lemma 6.4(iii) implies the images of D_1, D_2 are again crossing disks. Since D_1' is a crossing disk with crossing circle puncture K', the crossing circle C_1 changes type under h. Thus at least one crossing circle linking K maps to a knot circle.

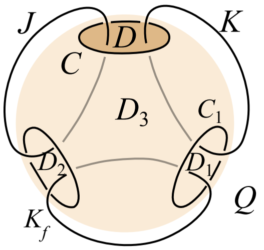


Figure 20: Schematic of Q.

Combining that with the first part of the proof shows that h changes the type of exactly one crossing circle linking K.

To see the last statement, let C_1 be the unique crossing circle linking K that changes type and suppose D_1 is any crossing disk punctured by K which is bounded by C_1. Since C_1 changes type, the image $D_1' = h(D_1)$ has a knot circle puncture in M' and must be a crossing disk.

Conversely, suppose D is a crossing disk punctured by K and bounded by the crossing circle $C \neq C_1$. Then two punctures of D', both C' and K', are crossing circles in M'. Theorem 2.7 implies the third puncture of D' is also a crossing circle, and D' is a longitudinal disk. □

Lemmas 6.4 and 6.5 allow us to show, in the following lemma, that separating quadruples exist in the presence of type-changing homeomorphisms.

Lemma 6.6 *Let $h\colon M \to M'$ be a homeomorphism between flat FAL complements with unique reflection surfaces, and suppose K is a knot circle that changes type under h. Further, let C_1 be the unique crossing circle linking K that changes type under h. Then each crossing circle $C \neq C_1$ that links K is part of a separating quadruple that includes the punctures K, C_1 and C.*

Proof Since $C \neq C_1$, Lemma 6.5 implies that h fixes the type of C so $C' = h(C)$ is a crossing circle in M'. The image D' of a crossing disk D bounded by C, then, is punctured by the crossing circles K' and C'; therefore, D' must be a longitudinal disk. Thus D is an N-disk that changes type, and Lemma 6.4 implies it is part of a separating quadruple Q.

It remains to show that C_1 must be a crossing circle puncture in Q. First, since K punctures D it is a knot circle component of Q and must puncture one other crossing disk, say D_1, in Q. Moreover, since exactly one crossing disk of Q changes type, by Lemma 6.4(iii), h fixes the type of D_1. Lemma 6.5 then implies that D_1 is bounded by C_1, finishing the proof. □

Lemmas 6.5 and 6.6 demonstrate that flat FALs which admit a type-changing homeomorphism contain features similar to those of signature links. This motivates considering a certain sublink, the signature sublink, which we define in the next section.

7 Complements determine flat FALs

In this section we show that flat FALs are determined by their complements. We start by assuming that a flat FAL complement $M=\mathbb{S}^3\setminus\mathcal{A}$ has a unique reflection surface and admits a type-changing homeomorphism to another flat FAL complement, since all other cases are either trivial or covered by the work in Section 3. With these assumptions, we first show in Section 7.1 that any such $\mathcal{A}$ contains a sublink, $\mathcal{L}_h$, that features many of the properties of a signature link. Then in Section 7.2 we prove some technical tools involving separating quadruples, which are used to introduce an embedded two-sphere, S^2_α, in $\mathbb{S}^3$ that intersects any such $\mathcal{A}$ in only two points on $\mathcal{K}_f$ and separates the other components of $\mathcal{A}$ in a useful manner. From here, our next goal is to show that $\mathcal{L}^c=\mathcal{A}\setminus\mathcal{L}_h$ must be empty and in fact, $\mathcal{A}$ must be a signature link. This is all done in Section 7.3. Essentially, if $\mathcal{L}^c$ is nonempty or $\mathcal{A}$ fails to have any of the features necessary to be a signature link, then we can use our two-sphere S^2_α to show that $\mathcal{A}$ is either a connect-sum or a split link, contradicting hyperbolicity. Once we have proven that $\mathcal{A}$ must be a signature link, then we can quickly show that this type-changing homeomorphism is a full-swap, possibly pre- or post-composed with homeomorphisms that extend to isotopies of $\mathbb{S}^3$. At this point, we can use Proposition 5.5 to obtain the desired result.

7.1 Signature sublinks

The forthcoming lemma highlights the necessary properties for us to define a signature sublink.

Lemma 7.1 *Let $M=\mathbb{S}^3\setminus\mathcal{A}$ be a flat FAL complement with a unique reflection surface, and suppose there is a type-changing homeomorphism $h\colon M\to M'$ to another flat FAL complement M'. Then $\mathcal{A}$ contains a sublink*

$$\mathcal{L}_h=\{K_f\}\cup\mathcal{K}\cup\mathcal{C}\cup\mathcal{C}_{\mathcal{K}},$$

whose components satisfy the following properties:

(i) *The sets $\mathcal{K}$ and $\mathcal{C}$ contain $n\geq 2$ knot- and crossing-circles, respectively.*

(ii) *Each crossing circle of $\mathcal{C}=\{C_1,\dots,C_n\}$ links the corresponding knot circle of $\mathcal{K}=\{K_1,\dots,K_n\}$ and the knot circle K_f. Further, the homeomorphism h changes the types of the components in $\mathcal{C}\cup\mathcal{K}$ and fixes the type of K_f.*

(iii) *Let $\mathcal{C}_{\mathcal{K}}$ denote the set of all crossing circles of $\mathcal{A}$ that link distinct knot circles in $\mathcal{K}$. Then h fixes the type of each component in $\mathcal{C}_{\mathcal{K}}$. Moreover, each knot circle in $\mathcal{K}$ is linked with at least one crossing circle in $\mathcal{C}_{\mathcal{K}}$, and at most one crossing circle in $\mathcal{C}_{\mathcal{K}}$ links two given knot circles in $\mathcal{K}$.*

Proof We begin by showing that h changes the type of at least one knot circle in $\mathcal{A}$. Since M and M' are homeomorphic and M has a unique reflection surface R, Corollary 3.13 say that $R'=h(R)$, where R' is the unique reflection surface for M'. Suppose h only changes the type of crossing circles. Let C be a crossing circle that changes type and D a crossing disk bounded by C. Then $D' = h(D)$ would be a thrice-punctured sphere with three knot circle punctures, and Theorem 2.7 implies that D' must be part of the reflection surface R'. This cannot occur since D is a nonreflection thrice-punctured sphere in M and $R' = h(R)$; therefore, h changes the type of a knot circle, call it K_1.

Since h changes K_1 to a crossing circle, Lemma 6.5 implies there is a unique crossing circle C_1 that links K_1 and changes type. Lemma 6.2 implies that C_1 links distinct knot circles, K_1 and another which we denote by K_f. The type of K_f is fixed by h; otherwise, a crossing disk bounded by C_1 would map to a nonreflection thrice punctured sphere with exactly one knot circle puncture, contradicting Theorem 2.7.

The type-changing homeomorphism h, then, guarantees the existence of a knot circle K_f whose type is fixed, together with a Hopf sublink $\{K_1, C_1\}$ whose types change and for which K_f is linked by C_1.

Define $\mathcal{C}$ to be all crossing circles of $\mathcal{A}$ which link K_f *and* change type. Note that $\mathcal{C}$ contains at least two crossing circles. Indeed, since K_1 and C_1 satisfy the hypotheses of Lemma 6.6 we conclude that every crossing circle $C \neq C_1$ that links K_1 is part of a separating quadruple Q_C that includes the components K_1, C_1, and C. Moreover, since C_1 links K_f, the knot circle K_f is a puncture of Q_C as well. Lemma 6.4(ii) shows that both crossing circles of Q_C that link K_f change type, so that both are in $\mathcal{C}$. Thus $\mathcal{C}$ contains at least two crossing circles of $\mathcal{A}$.

Now each $C_i \in \mathcal{C}$ changes type so links distinct knot circles (Lemma 6.2), K_f and a second knot circle $K_i \in \mathcal{A}$. Moreover, since C_i changes type and the type of K_f is fixed, the argument above for the existence of K_1 shows that K_i must change type. Finally, since each knot circle that changes type is linked by a unique crossing circle that changes type (Lemma 6.5), the K_i are distinct. Let $\mathcal{K} = \{K_1, \dots, K_n\}$ denote the knot circles (other than K_f) linked by crossing circles in $\mathcal{C}$, and note that h changes the type of each knot circle in $\mathcal{K}$.

At this stage we have proven parts (i) and (ii) of the lemma.

Now define $\mathcal{C}_{\mathcal{K}}$ to be all crossing circles of $\mathcal{A}$ that link two knot circles of $\mathcal{K}$, and let C_{ij} denote a crossing circle linking $K_i, K_j \in \mathcal{K}$. To see that h fixes the type of C_{ij}, note that a crossing disk D_{ij} bounded by C_{ij} is punctured by both K_i and K_j. Since h changes K_i and K_j to crossing circles, the disk $h(D_{ij})$ must be a thrice-punctured sphere with at least two crossing circle punctures. Theorem 2.7 implies $h(D_{ij})$ must be a longitudinal disk, so h fixes the type of C_{ij}.

We now address the existence statements of part (iii) of the lemma. First fix a knot circle $K_i \in \mathcal{K}$. We must show there is at least one crossing circle in $\mathcal{C}_{\mathcal{K}}$ linking K_i. As in the above proof that $\mathcal{C}$ contains at least two crossing circles, Lemma 6.6 applies to the Hopf sublink $\{K_i, C_i\}$. Thus each crossing circle $C \neq C_i$ that links K_i is part of a separating quadruple Q_C that includes K_f as a puncture. Again as above, Lemma 6.4(ii) implies that C links two knot circles of $\mathcal{K}$, so that for each K_i there is at least one $C_{ij} \in \mathcal{C}_{\mathcal{K}}$. In this case let Q_{ij} denote the separating quadruple Q_C, and note that the crossing circles of Q_{ij} are C_i, C_j, and C_{ij}.

Now fix a pair of knot circles $K_i, K_j \in \mathcal{K}$, and suppose $C_{ij} \in \mathcal{C}_{\mathcal{K}}$ exists. We begin by showing C_{ij} is a puncture in a separating quadruple Q whose crossing circle punctures are C_i, C_j, and C_{ij}. Since $C_{ij} \neq C_i$ links K_i, Lemma 6.6 implies it is part of a separating quadruple Q containing the punctures K_i, C_i, and C_{ij}. The knot circles K_i, K_j, and K_f are the knot circle punctures of Q, since they are linked by C_i and C_{ij}. The final crossing circle of Q must link K_f and K_j, and must change type since C_{ij} is the only crossing circle of Q whose type is fixed by h. Therefore the final crossing circle is C_j, and Q has crossing circle punctures C_i, C_j, and C_{ij}.

Hence each crossing circle linking K_i and K_j forms a longitudinal disk with C_i and C_j. Now Lemma 4.2 of [20] shows there is at most one longitudinal disk with two given crossing circle punctures, so there is at most one C_{ij}. □

The sublink $\mathcal{L}_h$ of Lemma 7.1 satisfies many of the properties of a signature link. In particular, $\mathcal{L}_h$ consists of knot and crossing circles partitioned into nonempty sets that satisfy all linking requirements of Definition 5.1. The only properties of a signature link not yet verified are that all knot circles in $\mathcal{K}$ lie on the same side of K_f and that $\mathcal{L}_h$ is a flat FAL itself. Proofs of these properties appear in Theorems 7.10 and 7.11, respectively, but several preliminary results are required. We remark that Lemma 7.1 highlights a further similarity: the homeomorphism h changes types on components of $\mathcal{L}_h$ in precisely the same way that a full-swap does on a signature link. Finally, note that the sublink depends on the choice of a type-changing homeomorphism h together with a knot circle K_1 that changes type. The subscript of $\mathcal{L}_h$ is intended to emphasize this dependency.

Lemma 7.1 motivates the following definition.

Definition 7.2 Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ be a flat FAL complement with a unique reflection surface, and suppose M admits a type-changing homeomorphism h that changes the type of the knot circle K_1 of $\mathcal{A}$. Let K_f, $\mathcal{C}$, $\mathcal{K}$, and $\mathcal{C}_{\mathcal{K}}$ be as in Lemma 7.1.

The *signature sublink* $\mathcal{L}_h$ of $\mathcal{A}$ is the union of these components, so that

$$\mathcal{L}_h = \{K_f\} \cup \mathcal{K} \cup \mathcal{C} \cup \mathcal{C}_{\mathcal{K}},$$

endowed with a fixed choice of crossing disk D_i for each $C_i \in \mathcal{C}$. Let $\mathcal{D}$ denote the set of chosen crossing disks $\{D_1, \dots, D_n\}$.

Given a signature sublink there may be several choices for crossing disks bounded by crossing circles in $\mathcal{C}$. For example, the crossing circle C_4 of Figure 9 bounds the crossing disk D_4 pictured as well as a crossing disk passing between K_2 and K_3. Fixing an orientation on K_f, we make the convention that the ordering on $\mathcal{C}$ and $\mathcal{D}$ is determined by traversing K_f from D_1 in the chosen direction. Different choices for $\mathcal{D} = \{D_1, \dots, D_n\}$ can lead to different orderings, so for convenience we assume the choice is fixed throughout.

The next lemma shows that every crossing circle of $\mathcal{A} \setminus \mathcal{C}$ that bounds a crossing disk punctured by a knot circle of $\mathcal{K}$ is an element of $\mathcal{C}_{\mathcal{K}}$. The proof will show both that the crossing circle links distinct knot circles (rather than the same knot circle twice), and that both knot circles are from $\mathcal{K}$. The lemma also associates a unique separating quadruple Q_{ij} to each $C_{ij} \in \mathcal{C}$, and characterizes possible intersections of these separating quadruples. The separating quadruples guaranteed by Lemma 7.3 will be the main tool used in what follows, so we let $\mathcal{Q} = \bigcup_{\mathcal{C}_{\mathcal{K}}} Q_{ij}$ denote their union.

Lemma 7.3 *Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ be a flat FAL complement with a unique reflection surface, suppose there is a type-changing homeomorphism $h\colon M \to M'$ to another flat FAL complement M', and let $\mathcal{L}_h$ be the corresponding signature sublink of $\mathcal{A}$. If C is a crossing circle of $\mathcal{A} \setminus \mathcal{C}$, with a crossing disk punctured by some $K_i \in \mathcal{K}$, then*:

(i) *There is an index j with $C = C_{ij} \in \mathcal{C}_{\mathcal{K}}$.*

(ii) *The crossing circle C_{ij} bounds a unique crossing disk D_{ij} in M.*

(iii) *There is a separating quadruple Q_{ij} uniquely determined by C_i, C_j, C_{ij} and the chosen crossing disks in $\mathcal{D}$.*

(iv) *Distinct separating quadruples $Q_{ij}, Q_{kl} \in \mathcal{Q}$, are either disjoint or share a single crossing disk in $\mathcal{D}$.*

Proof Let C be a crossing circle in $\mathcal{A} \setminus \mathcal{C}$ that bounds a crossing disk D punctured by $K_i \in \mathcal{K}$. Since K_i changes type, C links distinct knot circles by Lemma 6.2, and we let K^* denote the other knot circle puncturing D. Further, again since K_i changes type, C_i is the unique crossing circle linking it that changes type by Lemma 6.5. Now C is not equal to C_i, so there is a separating quadruple Q with components K_i, C_i, and C by Lemma 6.6. The crossing circles C_i and C link all three knot circle components of Q, so K^* and K_f are the other knot circles of Q and must be linked by the final crossing circle, say C^*, of Q. Note that C and K_f are the only components of Q whose type is fixed by the homeomorphism h (by Lemma 6.4(ii)), so h changes the types of C^* and K^*. Thus C^* is a crossing circle linking K_f which changes type under h, implying there is an index j with $C^* = C_j \in \mathcal{C}$, and $K^* = K_j$ as well. The original crossing circle C is then $C_{ij} \in \mathcal{C}_{\mathcal{K}}$, verifying part (i) of the lemma.

Now consider statement (ii) of the lemma. Suppose C_{ij} bounded two distinct crossing disks, the original disk D and another D_1. Then D and D_1 form a separating pair by Theorem 2.8, and every knot circle of $\mathcal{A}$ intersects $D \cup D_1$ an even number of times (possibly zero). This implies both are punctured by $K_i, K_j \in \mathcal{K}$, both of which change to crossing circles under h. Since C_{ij} doesn't change type, the homeomorphism h maps D and D_1 to distinct longitudinal disks with the same punctures. This is impossible, however, because Lemma 4.5 of [20] shows that there is at most one longitudinal disk with two given punctures, let alone three. Hence C_{ij} bounds a unique crossing disk.

The existence portion of statement (iii) is guaranteed by Lemma 6.6, as noted above. To see uniqueness, note that two separating quadruples with the same crossing circle components have the same longitudinal disks by Lemma 4.5 of [20]. Thus two separating quadruples with the same crossing circle components can differ only in their crossing disks. Since C_{ij} bounds a unique crossing disk and, by convention, there is a fixed choice of disks $\mathcal{D}$ for crossing circles of $\mathcal{C}$, the Q_{ij} are unique.

Finally, we prove statement (iv) of the lemma. If Q_{ij} and Q_{kl} are disjoint we are done, so suppose their intersection is nonempty. All N-disks are identical or disjoint (Lemma 5.2), so if Q_{ij} intersects Q_{kl} nontrivially, they share some subset of disks. We will show that if the intersection is other than a single crossing disk in $\mathcal{D}$, the separating quadruples are equal.

An initial observation is that if $\{D_i, D_j\} \subset Q_{ij} \cap Q_{kl}$ then $Q_{ij} = Q_{kl}$. In this case both Q_{ij} and Q_{kl} contain crossing circles C_i and C_j. Lemma 7.1(iii) shows they also contain the unique crossing circle $C_{ij} \in \mathcal{C}_{\mathcal{K}}$, and the proof of statement (iii) above shows that the separating quadruples are equal.

Now if $Q_{ij} \cap Q_{kl}$ contains a longitudinal disk, they are equal by the above argument since they would share the punctures C_i and C_j. Similarly, Q_{ij} and Q_{kl} are equal if they share D_{ij} since this implies they share C_{ij}, and the definition of Q_{ij} implies they share C_i and C_j as well.

Thus the intersection $Q_{ij} \cap Q_{kl}$ of distinct separating quadruples from $\mathcal{Q}$ is either empty or a single disk from $\mathcal{D}$. □

One way to phrase Lemma 7.3(i) is to say that the signature sublink $\mathcal{L}_h$ contains all crossing circles of the FAL $\mathcal{A}$ that link a knot circle of $\mathcal{K}$. Since each knot circle in $\mathcal{K}$ changes type, Lemma 6.2 implies that if K_i punctures a crossing disk bounded by $C \in \mathcal{A}$ then it does so once, and C links distinct knot circles. For convenience let $\mathcal{L}^c = \mathcal{A} \setminus \mathcal{L}_h$ denote the components of $\mathcal{A}$ not in $\mathcal{L}_h$. If C is a crossing circle of $\mathcal{L}^c$, then, it either links the fixed component K_f or only knot circles of $\mathcal{L}^c$.

7.2 The standard ball

Our next objective is to associate a two-sphere S_α^+ with every arc α of $K_f \setminus \mathcal{D}$, with the property that $S_\alpha^+ \cap \mathcal{A}$ consists of two points on K_f. This is formally stated and proved at the end of this subsection in Proposition 7.9. To build this two-sphere, we first need to introduce a number of properties and terminology related to separating quadruples associated with a signature sublink $\mathcal{L}_h$ of $\mathcal{A}$. As noted in the introduction to Section 7, this two-sphere will play an essential role in classifying the flat FALs under consideration in this section.

Choose an orientation on K_f, say counterclockwise, and note that the crossing disks of $\mathcal{D}$ partition K_f into n oriented arcs. Let α be an *open arc* of $K_f \setminus \mathcal{D}$ and use the orientation on K_f to order the disks of $\mathcal{D}$ (and associated components of $\mathcal{L}_h$) so that α goes from D_n to D_1. Index the components of $\mathcal{K}$ and $\mathcal{C}$ so that $C_i \in \mathcal{C}$ bounds D_i and links $K_i \in \mathcal{K}$, and use this ordering to index crossing circles $\mathcal{C}_{\mathcal{K}}$ and separating quadruples $\mathcal{Q}$. Also number the arcs of $K_f \setminus \mathcal{D}$ so that α_i goes from D_{i-1} to D_i, for $2 \le i \le n$ (α could be considered α_1 in this ordering, but we continue to refer to it as α because of the special role it plays). Thus α induces an ordering on components and disks of the signature sublink $\mathcal{L}_h$, other than K_f, which we call the *α-ordering* of $\mathcal{L}_h$.

Let S^2_{ij} denote the separating quadruple $Q_{ij} \in \mathcal{Q}$ thought of as a two-sphere embedded in $\mathbb{S}^3$. The crossing circles $\{C_i, C_j, C_{ij}\}$ and disks $\{D_i, D_j, D_{ij}, D^{\ell}_{ij}\}$ of Q_{ij} are subsets of S^2_{ij}; whereas, the knot circle components $\{K_i, K_j, K_f\}$ each puncture S^2_{ij} twice. Here, D^{ℓ}_{ij} is a longitudinal disk, with longitudinal punctures along the crossing circles C_i, C_j, and C_{ij}, as discussed at the beginning of Section 6. Moreover, $\mathbb{S}^3 \setminus S^2_{ij}$ is two open three-balls, one of which contains α. Define the *inside* $\mathbb{I}_{ij}$ of Q_{ij} (or of S^2_{ij}) to be the open three-ball component of $\mathbb{S}^3 \setminus S^2_{ij}$ that does not contain α (thinking of α as *outside* each Q_{ij}). Observe that S^2_{ij} is the boundary of $\mathbb{I}_{ij}$, so the closure is $\overline{\mathbb{I}_{ij}} = \mathbb{I}_{ij} \cup S^2_{ij}$.

We highlight the consequence of Lemma 7.3(iv) that insides are either disjoint or nested.

Lemma 7.4 *Let $\mathcal{A}$ be a flat FAL whose complement admits a unique reflection surface and a type-changing homeomorphism h to another flat FAL complement. Let $\mathcal{L}_h$ be a signature sublink of $\mathcal{A}$ and $\mathcal{Q}$ be the set of all separating quadruples determined by $\mathcal{C}_{\mathcal{K}}$. Finally, let α be an open arc of $K_f \setminus \mathcal{D}$ inducing insides for each separating quadruple of $\mathcal{Q}$.*

If $Q_{ij}, Q_{kl} \in \mathcal{Q}$ are distinct separating quadruples, then their insides $\mathbb{I}_{ij}$ and $\mathbb{I}_{kl}$ are either disjoint or properly nested (so $\mathbb{I}_{ij}$ is a proper subset of $\mathbb{I}_{kl}$, or vice versa).

Proof We let α_{ij} denote the arc of K_f that is inside Q_{ij}, so that α_{ij} runs from D_i to D_j and is disjoint from α. If α_{ij} and α_{kl} overlap, but are not nested, then the disks D_i, D_j alternate with D_k, D_l around K_f. This implies the separating quadruples Q_{ij} and Q_{kl} intersect but not along a crossing disk in $\mathcal{D}$, contradicting Lemma 7.3(iv). Thus if the arcs α_{ij} and α_{kl} overlap then they are nested, and we conclude the insides $\mathbb{I}_{ij}$ and $\mathbb{I}_{kl}$ are either nested or disjoint. □

We now use the insides of separating quadruples and set inclusion to define a partial ordering on the set $\mathcal{Q}$.

Definition 7.5 Let α be an open arc of $K_f \setminus \mathcal{D}$ inducing insides on elements of $\mathcal{Q}$. The separating quadruple Q_{ij} is *inside* Q_{kl}, denoted by $Q_{ij} \prec_{\alpha} Q_{ij}$, if $\mathbb{I}_{ij}$ is a proper subset of $\mathbb{I}_{kl}$.

Observe that the inside relation, being defined using set inclusion, is indeed a strict partial ordering on $\mathcal{Q}$. First, the subset relation is transitive so the inside relation is as well. Second, the *proper* subset restriction implies the inside relation is neither reflexive nor symmetric, so it is a strict partial ordering.

The inside relation is most easily seen by choosing the midpoint of α to be infinity in $\mathbb{S}^3 = \mathbb{R}^3 \cup \{\infty\}$ and viewing the link from infinity as in Figure 21. The separating quadruple Q_{12} is not related to any other separating quadruple of $\mathcal{Q}$, while both relations $Q_{35} \prec_{\alpha} Q_{25}$ and $Q_{56} \prec_{\alpha} Q_{57}$ hold.

We now introduce more terminology that will be helpful in the ensuing discussion.

The term *inside* will frequently be used in the context of link components or thrice-punctured spheres to imply containment within a separating quadruple. For example, the disk D_3 of Figure 21 is inside the separating quadruple Q_{25} since $D_3 \subset \mathbb{I}_{25}$, as are the open arcs $\alpha_3, \alpha_4, \alpha_5$ of K_f. More generally, recall that two separating quadruples Q_{ij} and Q_{kl} intersect in at most one disk of $\mathcal{D}$ (together with its punctures). This implies that if $Q_{ij} \prec_{\alpha} Q_{kl}$, then the components C_{ij}, D_{ij}, and D^{ℓ}_{ij} of Q_{ij} are inside Q_{kl} as well.

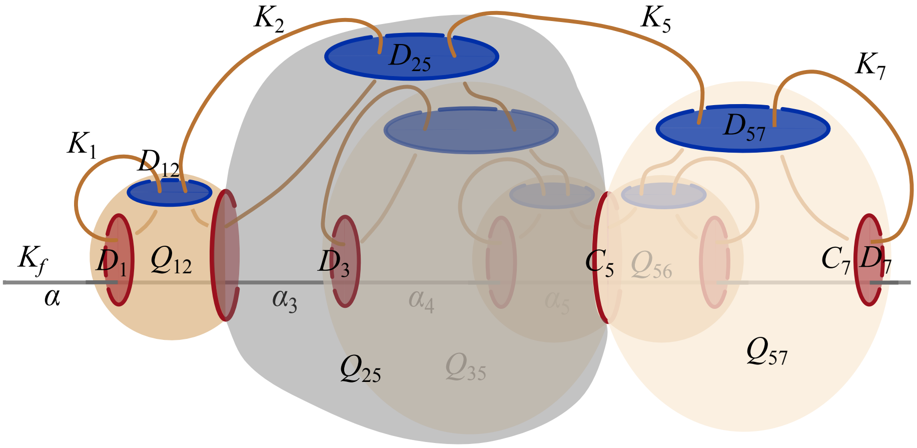


Figure 21: A signature sublink with three outermost quadruples.

A component, disk, or arc is *outside* the separating quadruple Q_{ij} if it is in the open three-ball of $\mathbb{S}^3 \setminus Q_{ij}$ containing α. Thus the disks D_1 and D_{12} of Figure 21 are outside Q_{25}. More generally, if disjoint separating quadruples are not comparable in the inside partial ordering, then they are outside each other.

We will also have occasion to describe separating quadruples as being on the same or opposite sides of a common disk in $\mathcal{D}$. Suppose two separating quadruples $Q, Q' \in \mathcal{Q}$ share a common disk $D \in \mathcal{D}$. Then the closed 3-balls $\overline{\mathbb{I}}$ and $\overline{\mathbb{I}'}$ share a common boundary disk D and have interiors that are either nested or disjoint. Define Q and Q' to be on the *same side* of the crossing disk $D \in \mathcal{D}$ if their insides are nested, and on *opposite sides* if they are disjoint. For example, the separating quadruples Q_{56} and Q_{57} in Figure 21 are on the same side of D_5 while Q_{25} and Q_{57} are on opposite sides.

Recall that an element $Q \in \mathcal{Q}$ is *maximal* if $Q_{ij} \prec_\alpha Q$ whenever Q_{ij} and Q are comparable. Thus the separating quadruples Q_{12}, Q_{25}, and Q_{57} of Figure 21 are maximal elements. The term *outermost quadruple* will be used for maximal elements in the inside partial ordering on $\mathcal{Q}$, as it is more intuitive.

We now highlight some important properties of this partial ordering.

Lemma 7.6 *Let $\mathcal{A}$ be a flat FAL whose complement admits a unique reflection surface and a type-changing homeomorphism h to another flat FAL complement. Let $\mathcal{L}_h$ be a signature sublink, α an open arc of $K_f \setminus \mathcal{D}$, and let $\prec_\alpha$ denote the induced inside relation on $\mathcal{Q}$. The inside relation is a strict partial ordering on $\mathcal{Q}$ with the following properties*:

(i) *Distinct separating quadruples of $\mathcal{Q}$ are not comparable if and only if their insides are disjoint.*

(ii) *Two separating quadruples that share a crossing disk in $D \in \mathcal{D}$ are comparable if and only if they are on the same side of D.*

(iii) *Each $Q_{ij} \in \mathcal{Q}$ is either outermost or contained in a unique outermost element of $\mathcal{Q}$.*

(iv) *The disk $D_1 \in \mathcal{D}$ is part of a unique outermost quadruple, and is not inside any element of $\mathcal{Q}$. The same result is true of the disk $D_n \in \mathcal{D}$.*

Proof Let Q and Q' be distinct separating quadruples in $\mathcal{Q}$. When their insides are properly nested, Q and Q' are comparable; however, if Q and Q' have disjoint insides they are not comparable. Lemma 7.4 shows these are the only two cases, proving statement (i). To see that statement (ii) holds, suppose $Q, Q' \in \mathcal{Q}$ are distinct separating quadruples that share the crossing disk $D \in \mathcal{D}$. By statement (i) they are not comparable if and only if their insides are disjoint which, by definition, is equivalent to saying they are on opposite sides of D.

Now, we consider statement (iii). Suppose $Q \in \mathcal{Q}$ is not outermost, so that there is a $Q' \in \mathcal{Q}$ with $Q \prec_\alpha Q'$. If Q'' is another separating quadruple with $Q \prec_\alpha Q''$, then the insides of Q' and Q'' intersect nontrivially, and statement (i) implies they are comparable. Thus the set of all separating quadruples larger than Q is a finite linearly ordered subset, and so contains a unique maximal element.

The argument for statement (iv) follows from the facts that every disk in $\mathcal{D}$ is either inside or on an outermost element of $\mathcal{Q}$, and that disks adjacent to α are not inside any element of $\mathcal{Q}$. To see the first fact, note that the knot circle K_i punctures D_i, and Lemma 7.1(iii) proves that K_i is linked by some crossing circle $C_{il} \in \mathcal{C}_{\mathcal{K}}$. The crossing circle C_{il} generates a separating quadruple $Q_{il} \in \mathcal{Q}$ which contains the disk D_i, by Lemma 7.3. We just verified that Q_{il} is either outermost or inside an outermost $Q \in \mathcal{Q}$. If Q_{il} is outermost then D_i is part of an outermost quadruple; otherwise, D_i is either on or inside Q. Thus every $D_i \in \mathcal{D}$ is either inside or on an outermost element of $\mathcal{Q}$. Now suppose $D_i \subset \mathbb{I}_{jk}$ for some disk $D_i \in \mathcal{D}$ and where $\mathbb{I}_{jk}$ is the inside of an outermost quadruple $Q_{jk} \in \mathcal{Q}$. Then both arcs of K_f adjacent to D_i are inside Q_{jk} as well. By definition of inside, however, the arc $\alpha \subset K_f$ is outside every element of $\mathcal{Q}$ so disks adjacent to α are not inside any element of $\mathcal{Q}$. Since α is adjacent to D_1 and D_n, they cannot be inside any separating quadruple and must be part of an outermost separating quadruple. To see uniqueness, note that all separating quadruples containing D_1 must be on the side of D_1 opposite α. Hence every pair of separating quadruples containing D_1 are comparable, making all such separating quadruples a linearly ordered subset, which must have a unique maximal element. The same observations hold for the disk D_n. □

Consider the set of all outermost separating quadruples, which we denote by $\{Q_1, Q_2, \dots, Q_l\}$. No two outermost separating quadruples are comparable, so their insides $\{\mathbb{I}_i\}$ are disjoint by Lemma 7.6(i). The open arcs $\beta_i = \mathbb{I}_i \cap K_f$, therefore, are disjoint as well. By definition of inside, the arc α is disjoint from all $\{\mathbb{I}_i\}$, so the arcs $\{\beta_i\}$ can be ordered as they are encountered starting at α and traversing K_f according to its orientation. We assume the sequence $\{Q_1, Q_2, \dots, Q_l\}$ is listed using this order on the $\{\beta_i\}$.

As an example, for a given α, there is a unique outermost quadruple $\{Q_1\}$ if and only if $Q_1 = Q_{1n}$. An alternative characterization is that there is a crossing circle $C_{1n} \in \mathcal{C}_{\mathcal{K}}$ linking the first and last knot circle in the α-ordering of $\mathcal{K}$.

If Q_i, Q_j are not consecutive in the ordered sequence $\{Q_1, Q_2, \dots, Q_l\}$, there is some Q_k between them along K_f and they cannot share a disk of $\mathcal{D}$. Hence, nonconsecutive separating quadruples in the sequence $\{Q_1, Q_2, \dots, Q_l\}$ are disjoint. Consecutive outermost quadruples Q_i, Q_{i+1}, on the other hand, can either share a disk or be disjoint. They are disjoint if there is an arc of K_f between β_i and β_{i+1}. We say that consecutive quadruples Q_i, Q_{i+1} are *adjacent* if $Q_i \cap Q_{i+1} = D_{j_i}$ for some $D_{j_i} \in \mathcal{D}$. Since the insides of Q_i and Q_{i+1} are disjoint, if they are adjacent they are on opposite sides of D_{j_i} (Lemma 7.6(ii)). The subsequence $\{Q_i, \dots, Q_j\}$ of $\{Q_1, Q_2, \dots, Q_l\}$, is a *maximally adjacent subsequence* if consecutive quadruples in the subsequence are adjacent while the pairs $\{Q_{i-1}, Q_i\}$ and $\{Q_j, Q_{j+1}\}$ are disjoint. The ordered sequence of all outermost quadruples $\{Q_1, Q_2, \dots, Q_l\}$ partitions into maximally adjacent subsequences. For example, the maximally adjacent subsequence of Figure 21 is $\{Q_{12}, Q_{25}, Q_{57}\}$.

Let $\{Q_1, \dots, Q_m\}$ be the initial maximally adjacent subsequence, so that consecutive quadruples of the subsequence are adjacent, while Q_m and Q_{m+1} are not. We will use the subsequence $\{Q_1, \dots, Q_m\}$ to define the standard ball associated with α.

Some elementary observations are in order before we define the standard ball. Let Q_j be a outermost separating quadruple with inside $\mathbb{I}_j$, and note that its closure $\overline{\mathbb{I}_j}$ is a closed three-ball with boundary sphere Q_j (we abuse notation and use Q_j to refer to the two-sphere embedded in $\mathbb{S}^3$ corresponding to this separating quadruple). Now suppose Q_{j-1} and Q_j are adjacent, outermost separating quadruples which share the crossing disk $D_{l_j} \in \mathcal{D}$. Since Q_{j-1} and Q_j are outermost, Lemma 7.6(i) implies their insides are disjoint. The union $\overline{\mathbb{I}_{j-1}} \cup \overline{\mathbb{I}_j}$ is a closed 3-ball, since it is two closed 3-balls (with disjoint interiors) glued along a common disk in their boundary spheres. The open disk $D^\circ_{l_j}$ is interior to $\overline{\mathbb{I}_{j-1}} \cup \overline{\mathbb{I}_j}$, so the boundary is $\partial(\overline{\mathbb{I}_{j-1}} \cup \overline{\mathbb{I}_j}) = (Q_{j-1} \cup Q_j) \setminus D^\circ_{l_j}$. These observations extend to subsequences $\{Q_i, \dots, Q_k\}$ of adjacent, outermost separating quadruples. For each outermost quadruple Q_j in the subsequence, form the closed three-ball $\overline{\mathbb{I}_j} = \mathbb{I}_j \cup Q_j$. Then $\bigcup_{j=i}^k \overline{\mathbb{I}_j}$ is a closed three-ball, because it is a sequence of closed three-balls with disjoint interiors in which only consecutive balls are glued together disk on their boundary spheres. Moreover, the boundary sphere of $\bigcup_{j=i}^k \overline{\mathbb{I}_j}$ is given explicitly by $\bigl(\bigcup_{j=i}^k Q_j\bigr) \setminus \bigl(\bigcup_{j=i+1}^k D^\circ_{l_j}\bigr)$, where $D_{l_j} = Q_{j-1} \cap Q_j$. Applying this to the initial maximally adjacent subsequence $\{Q_1, \dots, Q_m\}$ yields the standard ball associated with α.

Definition 7.7 Let α be an arc of $K_f \setminus \mathcal{D}$ with initial maximally adjacent subsequence $\{Q_1, \dots, Q_m\}$ of outermost separating quadruples. Let $\mathbb{I}_j$ be the inside of Q_j, and $\overline{\mathbb{I}_j}$ its closure. The *standard ball* of α, denoted by $\mathbb{B}^3_\alpha$, is the union

$$\mathbb{B}^3_\alpha = \bigcup_{j=1}^m \overline{\mathbb{I}_j},$$

and let $S^2_\alpha = \bigl(\bigcup_{j=1}^m Q_j\bigr) \setminus \bigl(\bigcup_{j=2}^m D^\circ_{l_j}\bigr)$ denote the boundary sphere of $\mathbb{B}^3_\alpha$, where $D_{l_j} = Q_{j-1} \cap Q_j$.

In what follows, components of $\mathcal{A}$ that intersect S^2_α will be important so, for convenience, we introduce some notation. Each $Q_j \in \{Q_1, \dots, Q_m\}$ is some $Q_{l_j l_{j+1}} \in \mathcal{Q}$, and it is with this notation we defined $D_{l_j} = Q_{j-1} \cap Q_j$, for $2 \le j \le m$. Rather than using double-subscripts, we adopt lowercase letters to represent components of the Q_j. For example, denote the knot circles $K_{l_j}, K_{l_{j+1}}$ of $Q_j = Q_{l_j l_{j+1}}$ by k_j and k_{j+1}, respectively, let $c_{j,j+1} = C_{l_j l_{j+1}}$, while $d^\ell_{j,j+1}$ denotes the longitudinal disk $D^\ell_{l_j l_{j+1}}$. We continue to use K_f for the fixed knot circle puncturing Q_j. Please refer to Figure 22 for an illustration of this notation.

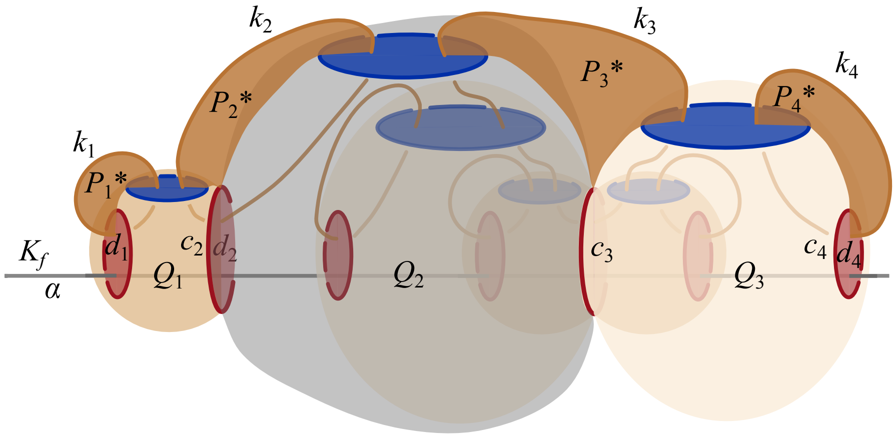


Figure 22: The standard ball with disks $\mathcal{P}_i^*$.

Now consider the boundary S^2_α of the standard ball with this notation. The crossing disk shared by the outermost quadruples Q_{j-1} and Q_j is $d_j = D_{l_j}$, and k_j is the knot circle puncturing them. Further, by Lemma 7.6(iv), the disk $D_1 = d_1$ lies on the separating quadruple Q_1, so that $k_1 = K_1$. This implies $S^2_\alpha = \left(\bigcup_{j=1}^m Q_j\right) \setminus \left(\bigcup_{j=2}^m d_j^\circ\right)$, and S^2_α is punctured twice by each of the knot circles $K_f, k_1, k_2, \dots, k_m, k_{m+1}$.

The boundary S^2_α of the standard ball, then, is a two-sphere embedded in $\mathbb{S}^3$ that is intersected by the link $\mathcal{A}$ in many components of $\mathcal{L}_h$. Our immediate goal is to extend S^2_α to an embedded two-sphere S^+_α that intersects $\mathcal{A}$ in exactly two points of K_f. Since $\mathcal{A}$ is hyperbolic, the desired S^+_α cannot define a connect-sum decomposition, and one component of $\mathbb{S}^3 \setminus S^+_\alpha$ must be a standard ball-arc pair. This significantly restricts the link $\mathcal{A}$ and allows us to prove that $\mathcal{A} = \mathcal{L}_h$ (Theorem 7.11), which ultimately leads to our main result.

Some preliminary definitions, and a technical lemma, are necessary before constructing the two-sphere S^+_α. Lemma 7.3 shows that the structure of a signature link outlined in Lemma 5.3 (a longitudinal disk and separating quadruple) persists in signature sublinks, even in the (potential) presence of additional components. In the proof of Lemma 5.3 the inside $\mathcal{P}_i$ of the knot circle K_i can be described as the component of the reflection surface bounded by K_i and not containing K_f.

Now let k_i be a knot circle in $\mathcal{K}$ puncturing the standard ball. Analogously define the inside $\mathcal{P}_i$ of k_i to be that component of the reflection surface not containing K_f, including the boundary curve k_i. Each k_i punctures the boundary S^2_α twice, so half of the closed disk $\mathcal{P}_i$ is inside and half outside of $\mathbb{B}^3_\alpha$. We let $\mathcal{P}_i^*$ denote the portion of $\mathcal{P}_i$ outside of $\mathbb{B}^3_\alpha$. Precisely we have $\mathcal{P}_i^* = \mathcal{P}_i \setminus \mathbb{B}^3_\alpha$. Note $\mathcal{P}_i^*$ is a half-open disk with the arc of k_i outside $\mathbb{B}^3_\alpha$ part of its boundary. For $2 \le i \le m$, the interior of $\mathcal{P}_i^*$, then, is an open disk whose boundary consists of an arc of k_i, geodesics on each of Q_{i-1} and Q_i, and single point of the crossing circle c_i on $Q_{i-1} \cap Q_i$ (see Figure 22). Note that, for $i = 1, m+1$, $\mathcal{P}_i^*$ is also an open disk, but whose boundary only consists of an arc of k_i and a geodesic on Q_i.

Before constructing S_α^+, we prove a technical lemma showing that no components of $\mathcal{A}$ intersect $\mathcal{P}_i^*$ (other than the boundary arc of k_i). Components of $\mathcal{L}_h$ and its complement $\mathcal{L}^c = \mathcal{A} \setminus \mathcal{L}_h$ will be considered separately.

Lemma 7.8 *Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ be a flat FAL complement with a unique reflection surface that admits a type-changing homeomorphism h to another flat FAL complement. Let $\mathcal{L}_h$ denote the signature sublink of $\mathcal{A}$. Let α be an arc of $K_f \setminus \mathcal{D}$ with standard ball $\mathbb{B}_\alpha^3$ generated by the initial maximally adjacent subsequence $\{Q_1, \dots, Q_m\}$ of outermost quadruples.*

If $\mathcal{P}_i^*$ is the portion of the reflection surface inside of $k_i \in \mathcal{K}$ but outside $\mathbb{B}_\alpha^3$, then the components of $\mathcal{A}$ are disjoint from the interior of $\mathcal{P}_i^*$.

Proof Initially focus on components of $\mathcal{L}_h$ and consider $\mathcal{P}_i \cap \mathcal{L}_h$. First observe that k_i is the only knot circle of $\mathcal{L}_h$ that is a boundary curve of $\mathcal{P}_i$. To see this, suppose that K is a knot circle of $\mathcal{A}$ interior to $\mathcal{P}_i$. Then K and K_f are on opposite sides of k_i, and cannot be linked by a crossing circle. Since each $K_j \in \mathcal{K}$ is linked to K_f by C_j, we see $K \notin \mathcal{K}$ so K is not in $\mathcal{L}_h$. Thus $\mathcal{P}_i^* \cap \mathcal{L}_h$ consists only of punctures by crossing circles of $\mathcal{L}_h$.

We turn our attention to crossing circles of $\mathcal{L}_h$ which intersect $\mathcal{P}_i^*$. The crossing circle C_i is the only one of $\mathcal{C}$ that intersects $\mathcal{P}_i$, and it does so in one point on the boundary of, not interior to, $\mathcal{P}_i^*$. Now suppose $C \in \mathcal{C}_\mathcal{K}$ is a crossing circle that links k_i, and let Q be the separating quadruple it generates (guaranteed by Lemma 7.3). In this case, Q contains the crossing disk D_i and is comparable to exactly those separating quadruples of $\mathcal{Q}$ on the same side of D_i as Q (Lemma 7.6((ii))). Consider the cases $i = 1$, $2 \le i \le m$, and $i = m + 1$ separately.

In the case $i = 1$, the disk d_1 is $D_1 \in \mathcal{D}$ by Lemma 7.6(iv), and every separating quadruple containing D_1 is opposite to α. Thus Q is comparable to Q_1 and, by maximality of Q_1, we have $Q \prec_\alpha Q_1$. This implies the crossing circle C is inside Q_1 and disjoint from $\mathcal{P}_i^*$.

In the case $2 \le i \le m$, the separating quadruple Q shares the disk d_i with both Q_{i-1} and Q_i. The outermost quadruples Q_{i-1} and Q_i are not comparable so must be on opposite sides of d_i. Thus Q is comparable to one of either Q_{i-1} or Q_i, making C interior to that outermost quadruple and disjoint from $\mathcal{P}_i^*$.

Finally consider $\mathcal{P}_{m+1}^*$ which is bounded by an arc of the knot circle k_{m+1}. In this case Q shares the disk d_{m+1} with Q_m (eg disk $\mathcal{P}_4^*$ of Figure 22). If Q and Q_m are on the opposite sides of d_{m+1}, then Q is contained in a unique outermost quadruple Q_{m+1} on the opposite side of d_{m+1} from Q_m (Lemma 7.6). But then the sequence $\{Q_1, Q_2, \dots, Q_m\}$ can be extended by Q_{m+1} to a longer sequence of adjacent, outermost separating quadruples. This contradicts the definition of $\{Q_1, Q_2, \dots, Q_m\}$, so Q and Q_m are on the same side of d_{m+1}. Maximality of Q_m implies that C is inside Q_m and disjoint from $\mathcal{P}_i^*$.

In all cases, then, the interior of $\mathcal{P}_i^*$ is disjoint from crossing circles of $\mathcal{L}_h$. The preceding argument showed that the same is true of knot circles in $\mathcal{L}_h$, so the interior of $\mathcal{P}_i^*$ is disjoint from the signature sublink $\mathcal{L}_h$.

It remains to show that components of $\mathcal{L}^c = \mathcal{A} \setminus \mathcal{L}_h$ do not intersect the interior of $\mathcal{P}_i^*$. The proof amounts to showing that if the set of components of $\mathcal{L}^c$ intersecting $\mathcal{P}_i^*$ is nontrivial, then $\mathcal{A}$ is a split link, hence $\mathcal{L}^c$ must be disjoint from $\mathcal{P}_i^*$.

First recall that any crossing disk punctured by k_i is bounded by a crossing circle in $\mathcal{L}_h$ by Lemma 7.3, and consider a crossing circle $C \in \mathcal{L}^c$ that punctures the interior of $\mathcal{P}_i^*$. Then, if D a crossing disk bounded by C it is disjoint from k_i as well as from separating quadruples in $\mathcal{Q}$ — in other words, D is disjoint from the boundary of $\mathcal{P}_i^*$. Now C punctures $\mathcal{P}_i^*$, so D intersects the reflection surface entirely within $\mathcal{P}_i^*$. In particular, C punctures the interior of $\mathcal{P}_i^*$ twice and any knot circle(s) linked by C are interior to $\mathcal{P}_i^*$.

Now let K be a knot circle of $\mathcal{L}^c$ and recall that every crossing circle in $\mathcal{L}_h$ links only knot circles in $\mathcal{L}_h$. Thus all crossing circles of $\mathcal{A}$ that link K are contained in $\mathcal{L}^c$. Now suppose K intersects $\mathcal{P}_i^*$. Knot circles of $\mathcal{A}$ are disjoint in the reflection surface and K, being in $\mathcal{L}^c$, is disjoint from $\mathcal{Q}$, so K is contained in the interior of $\mathcal{P}_i^*$. Then any crossing circle C that links K punctures $\mathcal{P}_i^*$, and the above argument shows that C links only knot circles interior to $\mathcal{P}_i^*$.

Now suppose that $\mathcal{L}^c$ intersects $\mathcal{P}_i^*$ nontrivially, and let $\mathcal{L}_i^c$ denote the sublink of components of $\mathcal{A}$ that intersect $\mathcal{P}_i^*$. The above argument shows that crossing circles of $\mathcal{L}_i^c$ link only knot circles of $\mathcal{L}_i^c$, and vice versa. Thus the components of flat FAL $\mathcal{A}$ partition into two nonempty subsets, $\mathcal{L}_i^c$ and its complement, in which crossing circles only link knot circles within their respective subset. In an FAL this results in a split link, contradicting the fact that $\mathcal{A}$ is hyperbolic.

Thus $\mathcal{P}_i^*$ is disjoint from $\mathcal{L}^c$, completing the proof that the components of $\mathcal{A}$ are disjoint from the interior of $\mathcal{P}_i^*$. □

Proposition 7.9 *Let $\mathcal{A}$ be a flat FAL whose complement admits a type-changing homeomorphism h, with associated signature sublink $\mathcal{L}_h$. Let K_f be the knot circle component of $\mathcal{L}_h$ whose type is fixed by h, and α be an arc of $K_f \setminus \mathcal{D}$. The standard ball $\mathbb{B}^3_\alpha$ has a neighborhood $N(\mathbb{B}^3_\alpha)$ in $\mathbb{S}^3$ whose boundary is a two-sphere S^+_α such that $S^+_\alpha \cap \mathcal{A}$ is precisely two distinct points of K_f. Moreover, S^+_α can be chosen so that*

(i) *every component, other than K_f, of $\mathcal{A}$ that intersects $\mathbb{B}^3_\alpha$ is inside S^+_α, and*

(ii) *every component of $\mathcal{A}$ that is outside $\mathbb{B}^3_\alpha$ is also outside S^+_α.*

Proof Let K_f be oriented with α an arc of K_f between two consecutive disks of $\mathcal{D}$ and endow components of $\mathcal{L}_h$ with the α-ordering. Further, let $\{Q_1, \dots, Q_m\}$ be the initial maximally adjacent subsequence and let $\mathbb{B}^3_\alpha$ be the standard ball of α with boundary sphere S^2_α. The proof consists of constructing a cell complex X consisting of $\mathbb{B}^3_\alpha$ together with portions of the projection plane. The desired sphere S^+_α will be the boundary of an appropriately chosen regular neighborhood $N(X)$ of this cell complex.

Let $\mathcal{R}$ denote the unique reflection surface of M. As in Lemma 7.8, let $\mathcal{P}_i$ denote the disk of $\mathcal{R} \setminus k_i$ that does not contain K_f, together with its boundary curve k_i. Now define X to be the cell complex

$$X = \mathbb{B}^3_\alpha \cup \Bigl(\bigcup_{1 \le i \le m+1} \mathcal{P}_i \Bigr).$$

Again following Lemma 7.8, the standard ball intersects each disk $\mathcal{P}_i$ and we let $\mathcal{P}_i^*$ denote the portion of $\mathcal{P}_i$ outside $\mathbb{B}^3_\alpha$ (ie $\mathcal{P}_i^* = \mathcal{P}_i \setminus \mathbb{B}^3_\alpha$ — see Figure 22).

The first statement of the proposition involves components of $\mathcal{A}$ that intersect $\mathbb{B}^3_\alpha$, and we show that each of these is contained in X. Since the components of the quadruples $\{Q_1, \dots, Q_m\}$ are the only components of $\mathcal{A}$ that intersect the boundary S^2_α of the standard ball, we address those first. All crossing circles in these quadruples lie on the boundary of the standard ball and are, therefore, subsets of X. Recall that the knot circles of $\mathcal{A}$ that puncture S^2_α have been labeled $K_f, k_1, k_2, \dots, k_m, k_{m+1}$, which are not subsets of $\mathbb{B}^3_\alpha$. Since X includes the closed disks $\mathcal{P}_i$, however, each k_i is a subset of X. Thus all components of $\mathcal{A}$ that intersect S^2_α, other than K_f, are subsets of X. The remaining components of $\mathcal{A}$ that intersect $\mathbb{B}^3_\alpha$ are interior to $\mathbb{B}^3_\alpha$. Thus every component of $\mathcal{A}$ that intersects $\mathbb{B}^3_\alpha$ is contained in X and, therefore, interior to every neighborhood of X in $\mathbb{S}^3$.

The goal is to show that an appropriate open neighborhood $N(X)$ of X is an open three-ball whose boundary sphere satisfies the requirements for S^+_α. First observe that each $\mathcal{P}_i^*$ can be retracted into $\mathbb{B}^3_\alpha$ along the disk $\mathcal{P}_i$. Since small enough neighborhoods of $\mathbb{B}^3_\alpha$ are three-balls, the same is true for X. Moreover, since the knot circle K_f punctures S^2_α twice, the same will be true of small enough neighborhoods of X. From now on we assume $N(X)$ is an open three-ball neighborhood of X with boundary sphere S^2_X that is punctured twice by K_f. The previous paragraph demonstrates that S^2_X contains all other components of $\mathcal{A}$ that intersect $\mathbb{B}^3_\alpha$, so S^2_X satisfies the first statement of the proposition as well.

We have shown that components of $\mathcal{A}$ that intersect $\mathbb{B}^3_\alpha$ behave as desired relative to S^2_X. Furthermore, the neighborhood $N(X)$ can be chosen small enough so that its boundary sphere doesn't intersect any components of $\mathcal{A}$ that are disjoint from X. Thus any component of $\mathcal{A}$ that is disjoint from (or outside) X is also outside S^2_X.

To finish the proof of statement (ii), then, we must show that every component of $\mathcal{A}$ that is outside $\mathbb{B}^3_\alpha$ is also outside X. Or, contrapositively, show that every component of $\mathcal{A}$ that intersects $X \setminus \mathbb{B}^3_\alpha = \bigl(\bigcup_{1 \le i \le m+1} \mathcal{P}_i^*\bigr)$ also intersects $\mathbb{B}^3_\alpha$. Lemma 7.8, however, demonstrates that the only components of $\mathcal{A}$ intersecting the $\mathcal{P}_i^*$ are components of the outermost quadruples $\{Q_1, \dots, Q_m\}$, which also intersect $\mathbb{B}^3_\alpha$. Thus every component of $\mathcal{A}$ that is outside $\mathbb{B}^3_\alpha$ is also outside X, and S^2_X satisfies the second statement of our proposition, making it our desired S^2_+. □

As an example, the sphere S^+_α of Figure 22 contains the entire signature sublink $\mathcal{L}_h$ except the arc α. In fact, the twice-punctured sphere S^+_α of Proposition 7.9 will be used to provide a connect-sum decomposition of $\mathcal{A}$ if $\mathcal{A} \ne \mathcal{L}_h$, contradicting hyperbolicity. Hence it is the key to finishing the proof that $\mathcal{A} = \mathcal{L}_h$.

7.3 Hyperbolicity conditions

We can now use the properties of S_α^+ from Proposition 7.9 along with hyperbolicity conditions, specifically a hyperbolic link can not be a split link and can not be a connect-sum, to prove our main results. Before continuing, let us compare the *signature sublinks* of Definition 7.2 to the *signature links* of Definition 5.1. One glaring difference is that a signature link is assumed to be a flat FAL while a signature sublink, being a sublink of a flat FAL, is not necessarily a flat FAL by itself. The other difference is that all knot components of $\mathcal{K}$ in a signature link are assumed to be on the same side of K_f, and this has yet to be proven for signature sublinks. By *same side of* K_f we mean the same component of its complement in the projection plane. Theorem 7.10 shows that all knot circles in $\mathcal{K}$ for a signature sublink $\mathcal{L}_h$ are indeed on the same side of K_f. Theorem 7.11 proves that $\mathcal{L}_h$ is a flat FAL by proving (the stronger result) that it equals the flat FAL $\mathcal{A}$. With these results in hand it is not hard to finish, in Theorem 7.12, the proof that flat FALs are determined by their complements.

Theorem 7.10 *Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ and $M' = \mathbb{S}^3 \setminus \mathcal{A}'$ be flat FAL complements, each with unique reflection surfaces. Suppose $h\colon M \to M'$ is a type-changing homeomorphism, and let $\mathcal{L}_h$ be the signature sublink of $\mathcal{A}$. Then all knot circles in $\mathcal{K}$ are on the same side of K_f.*

Proof The result will follow once we show that if knot circles of $\mathcal{K}$ are on opposite sides of K_f, then there is an arc α in $K_f \setminus \mathcal{D}$ such that the two-sphere S_α^+ of Proposition 7.9 provides a connect-sum decomposition of $\mathcal{A}$.

A crossing circle in $\mathcal{C}_\mathcal{K}$ links two knot circles of $\mathcal{K}$ that are on the same side of K_f, since a crossing circle of $\mathcal{A}$ that punctures opposite sides of K_f necessarily links K_f. Therefore, if $C_{ij}, C_{kl} \in \mathcal{C}_\mathcal{K}$ link knot circles on opposite sides of K_f, then the sets of crossing circles $\{C_i, C_j, C_{ij}\}$ and $\{C_k, C_l, C_{kl}\}$ are disjoint. Since separating quadruples can intersect only along common crossing disks (Lemma 7.3(iv)), this implies Q_{ij} and Q_{kl} are disjoint. In particular, outermost separating quadruples that link knot circles of $\mathcal{K}$ on opposite sides of K_f are not adjacent.

Now suppose there are knot circles in $\mathcal{K}$ that are on opposite sides of K_f, and label knot circles of $\mathcal{K}$ so that K_n and K_1 are on opposite sides of K_f. Let α denote the arc of K_f between crossing disks D_n and D_1, and let $\mathbb{B}_\alpha^3$ be the standard ball of α. Since K_f intersects D_1 and D_n at the endpoints of α, they are contained in outermost separating quadruples Q_{1i} and Q_{jn} by Lemma 7.6(iv). The maximally adjacent subsequence $\{Q_1, \dots, Q_m\}$ contains only outermost quadruples on the same side of K_f as Q_1, and create the standard ball $\mathbb{B}_\alpha^3$. Since the outermost quadruple Q_{jn} is on the opposite side from those in $\{Q_1, \dots, Q_m\}$, its components are outside $\mathbb{B}_\alpha^3$. There are components of $\mathcal{A}$, then, on both sides of the two-sphere S_α^+ constructed in Proposition 7.9. But then S_α^+ provides a nontrivial connect-sum decomposition of $\mathcal{A}$, contradicting the hyperbolicity of $\mathcal{A}$. □

Theorem 7.11 *Let $M = \mathbb{S}^3 \setminus \mathcal{A}$ and $M' = \mathbb{S}^3 \setminus \mathcal{A}'$ be flat FAL complements each with unique reflection surfaces. Suppose $h\colon M \to M'$ is a type-changing homeomorphism. Then $\mathcal{A}$ is a signature link.*

Proof To prove this result we show that $\mathcal{A}$ equals its signature sublink $\mathcal{L}_h$. Once we've shown $\mathcal{L}_h = \mathcal{A}$, the signature sublink is a flat FAL and satisfies the final property of Definition 5.1. We assume that $\mathcal{L}^c = \mathcal{A} \setminus \mathcal{L}_h$ is nonempty, and will arrive at a contradiction.

We begin by showing that if C is a crossing circle of $\mathcal{L}^c$, then any crossing disk D bounded by C is punctured only by knot circles in $\mathcal{L}^c$. First, since $C \in \mathcal{A} \setminus \mathcal{L}_h$, Lemma 7.3 implies that D is not punctured by any $K_i \in \mathcal{K}$.

Now suppose that D is punctured by K_f. The crossing disks $\mathcal{D}$ cut K_f into n arcs. Let α be an arc of K_f that punctures the disk D, then construct the standard ball $\mathbb{B}^3_\alpha$ of α and the associated twice-punctured sphere S^+_α of Proposition 7.9. Now the crossing disk D must be outside $\mathbb{B}^3_\alpha$ since it is punctured by α and is disjoint from $S^2_\alpha = \partial\mathbb{B}^3_\alpha$. Then $C = \partial D$ is also outside $\mathbb{B}^3_\alpha$, and C must be outside S^+_α by Proposition 7.9. There are also components of $\mathcal{A}$ inside S^+_α, by Proposition 7.9, since S^+_α contains the components of the maximally adjacent subsequence $\{Q_1, \dots, Q_m\}$, which form a nontrivial subset of $\mathcal{A}$. Thus S^+_α provides a nontrivial connect-sum decomposition of $\mathcal{A}$. Since $\mathcal{A}$ is hyperbolic this is a contradiction, and every crossing circle $C \in \mathcal{L}^c$ links knot circles in $\mathcal{L}^c$.

Further recall that crossing circles of $\mathcal{L}_h$ link only knot circles in $\mathcal{L}_h$. At this stage the components of $\mathcal{A}$ have been partitioned into two nonempty subsets $\mathcal{L}_h$ and $\mathcal{L}^c$ with the property that crossing circles from one subset bound crossing disks punctured only by knot circle(s) from the same set. This is a contradiction, since such an FAL admits a disconnected diagram, indicating it came from a splittable link. Thus the link that generated $\mathcal{A}$ was split, contradicting the hyperbolicity of $\mathcal{A}$.

We conclude that $\mathcal{L}^c$ is empty, and $\mathcal{L}_h = \mathcal{A}$. □

The following theorem highlights our main result: flat FALs are determined by their complements.

Theorem 7.12 *Let $\mathcal{A}$, $\mathcal{A}'$ be flat FALs with homeomorphic complements. Then $\mathcal{A}$ is isotopic to $\mathcal{A}'$.*

Proof First, if $h\colon M \to M'$ is not type changing, then it preserves peripheral systems. Under this assumption, h extends to an isotopy of $\mathbb{S}^3$, making $\mathcal{A}$ and $\mathcal{A}'$ isotopic links. Moving forward, we will assume that h is type changing. Our proof breaks down into just two cases since two flat FAL complements each with a different number of reflection surfaces can not be homeomorphic by Corollary 3.12.

Case I If M contains multiple distinct reflection surfaces, then Corollary 3.12 tells us that $\mathcal{A}$ and $\mathcal{A}'$ are isotopic links.

Case II Suppose M and M' each contain a unique reflection surface. Then by Theorem 7.11 we know that $\mathcal{A}$ has the structure of a signature link $\mathcal{L}$. Thus, there is a unique knot circle K_f whose type is fixed by h, and h changes the type of every crossing circle C_i linking K_f. Each C_i links a knot circle K_i that changes type and h preserves the type of all other crossing circles, each of which link two of the K_i. The fact that $\mathcal{A}$ has all these properties is justified at the beginning of this section in Lemma 7.1.

Since $\mathcal{A}$ is a signature link, we can consider the full-swap homeomorphism $h_f\colon M \to M''$, discussed in Section 5, where $M'' = \mathbb{S}^3 \setminus \mathcal{A}''$ and $\mathcal{A}''$ is also a signature link. This homeomorphism has the same effect on peripheral structures as h and Proposition 5.5 tells us that $\mathcal{A}$ and $\mathcal{A}''$ are isotopic links. Now h_f and h act identically on peripheral systems, so $h_f \circ h^{-1}\colon M' \to M''$ preserves peripheral systems and extends to an isotopy of $\mathbb{S}^3$. This implies that the links $\mathcal{A}'$ and $\mathcal{A}''$ are isotopic. Thus, $\mathcal{A}$ and $\mathcal{A}'$ are isotopic, finishing this case. □

By combining Theorem 3.14 with the work from this section, we can now provide a complete proof of Theorem 1.3, which we first restate.

Theorem 1.3 *Let $\mathcal{A}$ be a flat FAL. Then either*

- *$\mathcal{A}$ is not a signature link and both $\mathcal{A}$ and its complement $M = \mathbb{S}^3 \setminus \mathcal{A}$ have the same symmetry group, or*
- *$\mathcal{A}$ is a signature link and full-swaps on $\mathcal{A}$ generate symmetries of $M = \mathbb{S}^3 \setminus \mathcal{A}$ which are not restrictions of symmetries of $\mathcal{A}$ to M.*

Proof Let $\mathcal{A}$ be a flat FAL with $M = \mathbb{S}^3 \setminus \mathcal{A}$. Note that P_3 is a signature link (see Figure 15) whose complement contains multiple reflection surfaces.

First, suppose A is not a signature link. If M contains multiple reflection surfaces, then Theorem 3.14 tells us that $\operatorname{Sym}(\mathbb{S}^3, \mathcal{A}) = \operatorname{Sym}(\mathbb{S}^3 \setminus \mathcal{A})$. If M contains a unique reflection surface and A is not a signature link, then Theorem 7.11 implies every self-homeomorphism of M is not type changing. Thus, in this case, every self-homeomorphism of M preserves peripheral structures, and so, extends to an isotopy of $\mathbb{S}^3$, as needed.

Now, suppose A is a signature link. Then M admits a full-swap homeomorphism h_f. As discussed in Section 5, full-swaps are compositions of ml-swaps, where these ml-swaps exchange meridional and longitudinal slopes on (distinct) Hopf sublinks of $\mathcal{A}$. Such homeomorphisms do not extend to isotopies of $\mathbb{S}^3$, and so, h_f is a representative for an element of $\operatorname{Sym}(M)$ that does not restrict to an element of $\operatorname{Sym}(\mathbb{S}^3, \mathcal{A})$. □

References

[1] **C C Adams**, *Thrice-punctured spheres in hyperbolic* 3*-manifolds*, Trans. Amer. Math. Soc. 287 (1985) 645–656 MR

[2] **C C Adams**, *Augmented alternating link complements are hyperbolic*, from "Low-dimensional topology and Kleinian groups", Lond. Math. Soc. Lect. Note Ser. 112, Cambridge Univ. Press (1986) 115–130 MR

[3] **R Benedetti**, **C Petronio**, *Lectures on hyperbolic geometry*, Springer (1992) MR

[4] **J Berge**, *Embedding the exteriors of one-tunnel knots and links in the* 3*-sphere*, unpublished lecture notes (1993)

[5] **R Blair**, **D Futer**, **M Tomova**, *Essential surfaces in highly twisted link complements*, Algebr. Geom. Topol. 15 (2015) 1501–1523 MR

[6] **E Chesebro**, **J DeBlois**, **H Wilton**, *Some virtually special hyperbolic* 3*-manifold groups*, Comment. Math. Helv. 87 (2012) 727–787 MR

[7] **R Flint**, *Intercusp geodesics and cusp shapes of fully augmented links*, PhD thesis, City University of New York (2017) Available at `https://www.proquest.com/docview/1898760342`

[8] **D Futer**, **J S Purcell**, *Links with no exceptional surgeries*, Comment. Math. Helv. 82 (2007) 629–664 MR

[9] **F Gainullin**, *Heegaard Floer homology and knots determined by their complements*, Algebr. Geom. Topol. 18 (2018) 69–109 MR

[10] **C M Gordon**, *Links and their complements*, from "Topology and geometry: commemorating SISTAG", Contemp. Math. 314, Amer. Math. Soc., Providence, RI (2002) 71–82 MR

[11] **C M Gordon**, **J Luecke**, *Knots are determined by their complements*, J. Amer. Math. Soc. 2 (1989) 371–415 MR

[12] **S R Henry**, **J R Weeks**, *Symmetry groups of hyperbolic knots and links*, J. Knot Theory Ramifications 1 (1992) 185–201 MR

[13] **N R Hoffman**, **C Millichap**, **W Worden**, *Symmetries and hidden symmetries of* (ϵ, d_L)*-twisted knot complements*, Algebr. Geom. Topol. 22 (2022) 601–656 MR

[14] **K Ichihara**, **T Saito**, *Knots in homology lens spaces determined by their complements*, Bull. Korean Math. Soc. 59 (2022) 869–877 MR

[15] **S Knavel**, **R Trapp**, *Embedded totally geodesic surfaces in fully augmented links*, Comm. Anal. Geom. 31 (2023) 563–593 MR

[16] **M Lackenby**, *The volume of hyperbolic alternating link complements*, Proc. Lond. Math. Soc. 88 (2004) 204–224 MR

[17] **B Mangum**, **T Stanford**, *Brunnian links are determined by their complements*, Algebr. Geom. Topol. 1 (2001) 143–152 MR

[18] **D Matignon**, *On the knot complement problem for non-hyperbolic knots*, Topology Appl. 157 (2010) 1900–1925 MR

[19] **J S Meyer**, **C Millichap**, **R Trapp**, *Arithmeticity and hidden symmetries of fully augmented pretzel link complements*, New York J. Math. 26 (2020) 149–183 MR

[20] **P Morgan**, **B Ransom**, **D Spyropoulos**, **R Trapp**, **C Ziegler**, *Belted sum decompositions of fully augmented links*, New York J. Math. 31 (2025) 1–42 MR

[21] **J S Purcell**, *Volumes of highly twisted knots and links*, Algebr. Geom. Topol. 7 (2007) 93–108 MR

[22] **J S Purcell**, *Cusp shapes under cone deformation*, J. Differential Geom. 80 (2008) 453–500 MR

[23] **J S Purcell**, *An introduction to fully augmented links*, from "Interactions between hyperbolic geometry, quantum topology and number theory", Contemp. Math. 541, Amer. Math. Soc., Providence, RI (2011) 205–220 MR

[24] **Y W Rong**, *Some knots not determined by their complements*, from "Quantum topology", Ser. Knots Everything 3, World Sci., River Edge, NJ (1993) 339–353 MR

[25] **J H C Whitehead**, *On doubled knots*, J. Lond. Math. Soc. 12 (1937) 63–71

[26] **K Yoshida**, *Unions of* 3*-punctured spheres in hyperbolic* 3*-manifolds*, Comm. Anal. Geom. 29 (2021) 1643–1689 MR

[27] **M Zevenbergen**, *Crushtaceans and complements of fully augmented and nested links*, honors thesis, University of Rochester (2021) Available at `https://www.sas.rochester.edu/mth/undergraduate/honorspaperspdfs/zevenbergen2021.pdf`

Department of Mathematics, Furman University
Greenville, SC, United States

Department of Mathematics, California State University, San Bernardino
San Bernardino, CA, United States

`christian.millichap@furman.edu`, `rtrapp@csusb.edu`

Received: 18 December 2023 Revised: 31 July 2024

Algebraic & Geometric Topology **25**:8 (2025) 4897–4919
DOI: 10.2140/agt.2025.25.4897
Published: 20 November 2025

BNSR-invariants of surface Houghton groups

NOAH TORGERSON
JEREMY WEST

The surface Houghton groups $\mathcal{H}_n$ are a family of groups generalizing Houghton groups H_n, which are constructed as asymptotically rigid mapping class groups. We give a complete computation of the BNSR-invariants $\Sigma^m(P\mathcal{H}_n)$ of their intersection with the pure mapping class group. To do so, we prove that the associated Stein–Farley cube complex is CAT(0), and we adapt Zaremsky's method for computing the BNSR-invariants of the Houghton groups. As a consequence, we give a criterion for when subgroups of H_n and $P\mathcal{H}_n$ having the same finiteness length as their parent group are finite index. We also discuss the failure of some of these groups to be co-Hopfian.

20F65, 57K20, 57M07

1 Introduction

To any group G of type F_k, one can assign a sequence of invariants $\Sigma^m(G)$, for $m \leq k$. These invariants determine which subgroups of G (containing the commutator subgroup) share which finiteness properties of G. Historically, they are difficult to compute. They were defined across several papers (see [6; 7; 18]), primarily by Bieri, Neumann, Strebel, and Renz, hence the names "BNS-invariants" and "BNSR-invariants". For the remainder of this paper, these will be referred to as Σ-invariants for the sake of brevity. One recent collection of groups for which these have been computed is the family of Houghton groups [19; 20].

Houghton defined his groups as permutation groups of infinite sets in [15]. More specifically, H_n is the group of "eventual translations" of the set $\{1, \dots, n\} \times \mathbb{N}$, ie permutations which in each ray are translations outside some finite set. Brown proved, via a suitable simplicial complex, that Houghton's group H_n is of type F_{n-1}, but not of type F_n (in the language of "finiteness length" this is $\mathrm{fl}(H_n) = n-1$). (Brown proved type FP_{n-1} but not type FP_n, and finitely presented, which together imply F_n but not F_{n-1}.) In [2], the authors define a variant of Houghton groups, called surface Houghton groups, as asymptotically rigid mapping class groups of surfaces with infinite genus. We denote the surface Houghton groups by $\mathcal{H}_n$, and the pure (ie end-fixing) subgroups as $P\mathcal{H}_n$. In the same paper, they also show that $\mathrm{fl}(\mathcal{H}_n) = n - 1$, via a cube complex analogous to the Stein–Farley complexes for Thompson groups.

This is not the only asymptotically rigid mapping class group variant of Houghton groups. While the surface Houghton groups replace the $\mathbb{N}$-rays with ends accumulated by genus (without boundary), the braided Houghton groups (defined by Degenhardt in his PhD thesis [11]) can be realized as asymptotically rigid mapping class groups as well (see [12] for details). This construction replaces the $\mathbb{N}$-rays with planar ends, accumulated by punctures; this surface has noncompact boundary. The braided Houghton groups $\mathrm{Br}\, H_n$ also have $\mathrm{fl}(\mathrm{Br}\, H_n) = n-1$ (as proven in [14]). The braided Houghton groups shall play a small role in the final section of this paper.

In [19; 20], Zaremsky computed the Σ-invariants of the Houghton groups, using the cube complex defined implicitly in [8], and explicitly in [16]. We carry out a parallel of Zaremsky's arguments, showing that the equivalent statement holds for pure surface Houghton groups. Namely, we prove in Section 4 the following, where $m(\chi)$ is the number of nonzero coefficients of χ in ascending standard form (details in Section 4):

Theorem 4.1 *Let χ be a nonzero character of $P\mathcal{H}_n$. Then $[\chi] \in \Sigma^{m(\chi)-1}(P\mathcal{H}_n) \setminus \Sigma^{m(\chi)}(P\mathcal{H}_n)$.*

As an application of this, we provide a partial converse (for Houghton groups and surface Houghton groups) of the well-known fact that if $H \leq G$ is a finite-index subgroup, then $\mathrm{fl}(H) = \mathrm{fl}(G)$. Specifically, we show that whenever a subgroup of H_n or $P\mathcal{H}_n$ has finiteness length $n-1$ and intersects the corresponding commutator with finite index, it is finite index in the full group. Further, in order to be finite index, the intersection condition must hold, for general group-theoretic reasons. (These are also true of $\mathcal{H}_n$, but for trivial reasons: the commutator subgroup of $\mathcal{H}_n$ is all of $\mathcal{H}_n$.) This is carried out in Section 5, along with some discussion of co-Hopfianness.

In order to accomplish this, we demonstrate in Section 3 that the cube complex of [2] is CAT(0).

Theorem 3.9 *Let X_n denote the Stein–Farley complex for $\mathcal{H}_n$. Then X_n is* $\mathrm{CAT}(0)$.

To do this, we use a refinement of a proposition from [14]; see Proposition 3.2. In Section 2, we lay out the definitions of the surfaces and groups we are concerned with, of the corresponding cube complex, and of the version of discrete Morse theory we shall employ. Alongside these definitions are various lemmas which we shall need. We also obtain various nice representatives for the vertices (Section 2) and edges (Section 3) of the complex.

In Section 5, we finish with some applications of the Σ-invariants to when subgroups having maximal finiteness length have finite index. This discussion applies to H_n, $\mathcal{H}_n$, and $P\mathcal{H}_n$. To handle the infinite-index case, there is some discussion of the failure of these groups to be co-Hopfian. In particular, we have the following theorem.

Theorem 5.6 *Let H denote either the Houghton group, or the pure surface Houghton group, and suppose $G < H$ has $\mathrm{fl}(G) = \mathrm{fl}(H)$. Then G is finite index in H if and only if $G \cap H'$ is finite index in H', where H' denotes the commutator subgroup of H. Furthermore, there exist subgroups G with $\mathrm{fl}(G) = \mathrm{fl}(H)$ of both finite and infinite index.*

Recently, Marie Abadie [1] analyzed the CAT(0) property for the Stein–Farley cube complexes of a different family of asymptotically rigid mapping class groups, including the braided Houghton groups. Similar methods are used, in particular a version of the same proposition as we use. Thus, one could attempt to apply Zaremsky's methods to compute the Σ-invariants of $\mathrm{Br}\, H_n$ as well. It seems reasonable to guess that they should work out the same.

The recent paper [3] concerns a generalization of these surface Houghton groups, obtained by varying the rigid structure on the same surface. As they show that the groups they consider are finite-index subgroups of our $\mathcal{H}_n$, the Σ-invariants are effectively the same. Additionally, our results in Section 5 apply to these more general surface Houghton groups.

Acknowledgements

We would like to thank Jing Tao for advice and guidance on this project. We would also like to thank Noel Brady and Justin Malestein for valuable conversations, and Peter Patzt and Matt Zaremsky for helpful comments. We also thank the anonymous referee for helpful comments. The authors were partially supported by the OU Bridge Funding Investment Program.

2 Definitions

2.1 The (pure) surface Houghton group

Here we lay out the definitions necessary for the group $P\mathcal{H}_n$, the pure version of the surface Houghton group defined in [2]. Let $\mathcal{O} = \mathcal{O}_n$ be a sphere with n boundary components, and let T be a torus with two boundary components, ∂^- and ∂^+. Fix an orientation-reversing homeomorphism $\lambda\colon \partial^- \to \partial^+$, and for each i an orientation-reversing homeomorphism μ_i from ∂^- to the i^{th} boundary component of $\mathcal{O}$. We construct Σ_n as follows: begin with $M^1 = \mathcal{O}$, then glue a copy of T to each boundary component of M^1 via the μ_i to obtain M^2. For each $j \geq 2$, glue a copy of T to each boundary component of M^j via λ to obtain M^{j+1}. Then the surface Σ_n is the union of all the M^j's; Σ_n is the surface with n ends, all accumulated by genus. We call $\mathcal{O}$ the *center* of Σ_n, each of the closures of the components of $M^j \setminus M^{j-1}$ is called a *piece*, and each piece B has a canonical homeomorphism $\iota_B\colon B \to T$. We will occasionally write B_k^j to denote the j^{th} piece in the k^{th} end.

Call a subsurface of Σ_n *suited* if it is connected and the union of $\mathcal{O}$ and finitely many pieces. Let $\varphi\colon \Sigma_n \to \Sigma_n$ be a homeomorphism. Call φ *asymptotically rigid* if there exists a suited subsurface $Z \subset \Sigma_n$ (called a *defining surface* for φ) such that

- $\varphi(Z)$ is also suited, and
- φ is *rigid away from* Z, that is, for every piece $B \subset \overline{\Sigma_n \setminus Z}$, we have that $\varphi(B)$ is a piece, and $\varphi|_B = \iota_{\varphi(B)}^{-1} \circ \iota_B$.

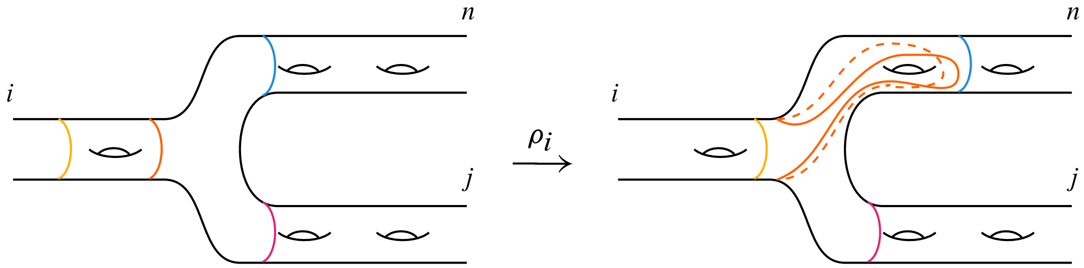


Figure 1: The handle shift ρ_i (which pushes from end i to end n) being applied to various curves in ends i, n, and j.

We may sometimes say that φ is rigid away from a piece B adjacent to $\mathcal{O}$, which we take to mean that φ is rigid away from $\mathcal{O} \cup B$. A special family of suited subsurfaces are those with pieces in only one end. We denote by L_g the suited subsurface consisting of $\mathcal{O}$ and the first g-many pieces in the n^{th} end.

Definition 2.1 (surface Houghton group) The surface Houghton group $\mathcal{H}_n$ is the subgroup of the mapping class group $\operatorname{Map}(\Sigma_n)$ whose elements have an asymptotically rigid representative. The pure surface Houghton group $P\mathcal{H}_n$ is the intersection of $\mathcal{H}_n$ with the pure mapping class group $\operatorname{PMap}(\Sigma_n)$.

We define some special elements of $\mathcal{H}_n$. For σ a permutation of $\{1, \dots, n\}$, choose some homeomorphism of $\mathcal{O}$ permuting the boundary components in the same fashion. For the sake of definiteness, consider σ first as an element of $\operatorname{Map}(S_{0,n})$, the sphere with n punctures; then σ yields a homeomorphism of $\mathcal{O}$, which we also denote by σ. Up to some isotopy, we can assume that the restriction of σ to the i^{th} boundary component of $\mathcal{O}$ is $\mu_{\sigma(i)} \circ \mu_i^{-1}$. Thus, we can obtain a homeomorphism of Σ_n by extending this map to be rigid outside of $\mathcal{O}$; call this extension again by the same name, σ. Such a map is not unique: there are many choices of mapping class in the sphere with n punctures which induce the same permutation.

Next, we define the handle shifts ρ_i, à la [17]. For $i \in \{1, \dots, n-1\}$, we want ρ_i to be a homeomorphism of Σ_n which shifts the i^{th} end towards the n^{th} end by a single genus. Specifically, for $j \geq 1$ map each piece B_n^j to B_n^{j+1}, and map each piece B_i^{j+1} to B_i^j; choose these to be the rigid maps. The ends other than i and n are unchanged, so all that remains is to define how ρ_i acts on $B_i^1 \cup \mathcal{O}$. This is demonstrated by Figure 1. Note that ρ_i is asymptotically rigid outside of $\mathcal{O} \cup B_i^1$.

Lemma 2.2 *Any map $\varphi \in \mathcal{H}_n$ can be written as $\alpha \rho_{i_1} \cdots \rho_{i_j} \sigma$, where α is a compactly supported mapping class, and σ and the ρ_{i_k}'s are as above.*

Proof Consider an arbitrary element $\varphi \in \mathcal{H}_n$: φ clearly induces some permutation σ on the ends of Σ_n; obtain from this $\sigma \in \mathcal{H}_n$. Now we have that $\varphi \circ \sigma^{-1} \in P\mathcal{H}_n$, so we wish to analyze how pure mapping classes act within ends. A rigorous examination of this is contained in Section 2.3. In the meantime, consider how φ acts in the complement of a defining surface: as it does not permute the ends, and takes pieces to pieces, it must shift each end in or out by some integer amount. From this, we obtain a sequence of handle shifts $\rho_{i_1}, \dots, \rho_{i_k}$. Finally, we have that $\varphi \circ \sigma^{-1} \circ (\rho_{i_1} \cdots \rho_{i_k})^{-1}$ is a compactly supported mapping class, α. □

Remark 2.3 The above expression obtained for φ is far from unique: choices were made both in the selection of σ, as well as in the selection of the handle shifts. However, any different choices made for these will simply change α, and the properties of α are rarely relevant for this paper.

2.2 The contractible cube complex

The group $\mathcal{H}_n$ acts on a contractible cube complex X_n, called the Stein–Farley complex. Consider ordered pairs (Z, φ), where Z is a suited subsurface, and $\varphi \in \mathcal{H}_n$. Declare two such pairs (Z_1, φ_1) and (Z_2, φ_2) to be equivalent if the *transition map* $\varphi_2^{-1} \circ \varphi_1$ takes Z_1 to Z_2 homeomorphically, and is rigid elsewhere (ie in the complement of Z_1).

Denote the equivalence class of (Z, φ) by $[Z, \varphi]$, and the set of equivalence classes by $\mathcal{S}$. The group $\mathcal{H}_n$ acts on $\mathcal{S}$ by

$$\psi \cdot [Z, \varphi] = [Z, \psi \circ \varphi].$$

Define the *complexity* of a pair (Z, φ) to be the genus of Z. This extends to be an $\mathcal{H}_n$-invariant height function $f : \mathcal{S} \to \mathbb{Z} \subset \mathbb{R}$. Given vertices $x_1, x_2 \in \mathcal{S}$, we say that $x_1 \prec x_2$ if there are representatives (Z_i, φ_i) of x_i so that $\varphi_1 = \varphi_2$, $Z_1 \subset Z_2$, and $\overline{Z_2 \setminus Z_1}$ is a (nonempty) *disjoint* union of pieces (ie has at most one piece from each end).

We now construct X_n as a cube complex with vertex set $\mathcal{S}$. We declare that whenever $x_1 \prec x_2$, with $d = f(x_2) - f(x_1)$, there is a d-cube with vertex set given by $\{x \mid x_1 \preceq x \preceq x_2\}$. As Σ_n has n ends, the complex X_n is n-dimensional. The height function f extends affinely to an $\mathcal{H}_n$-invariant height function on X_n. Also, let $X_n^{f \le k}$ denote the subcomplex spanned by vertices with height at most k. We have the following properties (see [2, Theorem 4.1]):

Theorem 2.4 *The cube complex X_n is contractible, and the action of $\mathcal{H}_n$ satisfies:*

- *The $\mathcal{H}_n$-stabilizers of cubes are finite extensions of mapping class groups of compact surfaces.*
- *For $k \ge 1$, the subcomplex $X_n^{f \le k}$ is $\mathcal{H}_n$-cocompact.*

We are actually more concerned with the action of $P\mathcal{H}_n$ on X_n. Note that $P\mathcal{H}_n$ is a finite-index subgroup of $\mathcal{H}_n$, and that the finite extension in the above theorem is by permutations of the ends. As mapping class groups of compact surfaces are of type F_∞, we have the following.

Corollary 2.5 *The action of $P\mathcal{H}_n$ on X_n satisfies:*

- *The $P\mathcal{H}_n$-stabilizers of cubes are type F_∞.*
- *For $k \ge 1$, the subcomplex $X_n^{f \le k}$ is $P\mathcal{H}_n$-cocompact.*

We end this section by demonstrating that vertices and edges can be given a nice form. First, we show that every vertex (indeed, every edge) can be expressed with only elements of $P\mathcal{H}_n$, then we provide a canonical form for the suited subsurface component of the vertex. Specifically, we show that any edge can be represented using only elements in the pure group:

Lemma 2.6 *Any edge in X_n can be expressed as $[Z,\varphi]$—$[Z\cup B,\varphi]$, where $\varphi\in P\mathcal{H}_n$.*

Proof Consider a vertex $[Z,\varphi]$, and let B be a piece adjacent to Z, so that $Z\cup B$ is suited. Write $\varphi=\alpha\rho_{i_1}\cdots\rho_{i_j}\sigma$ as in Lemma 2.2. Then we wish to show that $[Z,\varphi]=[\sigma(Z),\varphi\circ\sigma^{-1}]$, and that $[Z\cup B,\varphi]=[\sigma(Z)\cup\sigma(B),\varphi\circ\sigma^{-1}]$. Consider the diagram

$$\begin{tikzcd}
{[Z,\varphi]} \arrow[r, dash] & {[Z\cup B,\varphi]} & {[\sigma(Z),\varphi\circ\sigma^{-1}]} \arrow[r, dash] & {[\sigma(Z\cup B),\varphi\circ\sigma^{-1}]} \arrow[d, equal] \\
 & & & {[\sigma(Z)\cup\sigma(B),\varphi\circ\sigma^{-1}]}
\end{tikzcd}$$

As σ is rigid away from the center $\mathcal{O}$, we have that $\sigma(B)$ is a piece, and of course it must be adjacent to $\sigma(Z)$, so all that remains to be checked is that $[Z,\varphi]=[\sigma(Z),\varphi\circ\sigma^{-1}]$. Note that $(\varphi\circ\sigma^{-1})^{-1}\circ\varphi=\sigma$, which is rigid away from $\mathcal{O}$, and thus takes Z to $\sigma(Z)$ (a suited subsurface), and is certainly rigid outside of Z. □

We assume from here on that any map φ is in $P\mathcal{H}_n$. Vertices have two components, a map and a surface. We now have some control over the map, but what of the surface? For many arguments, it will be convenient to have the following canonical form. Recall that L_g is the surface built by taking the center, and adding in the first g-many pieces in the n^{th} end. Then any edge can be represented with both surfaces an appropriate L_g:

Lemma 2.7 *Any edge in X_n can be expressed as $[L_g,\varphi]$—$[L_{g+1},\psi]$.*

Proof Suppose we are given a vertex $[Z,\varphi']$. Recall that the handle shift ρ_i is rigid away from B_i^1, the piece that it pushes across the center. As $(\varphi)^{-1}\circ\varphi'$ can be chosen to be a composition of handle shifts taking Z to L_g, the only places where $(\varphi)^{-1}\circ\varphi'$ can be nonrigid are pieces in Z. This shows that vertices can be represented as $[L_g,\varphi]$. To extend this to the edge $[Z,\varphi']$ —$[Z\cup B,\varphi']$, consider the diagram

$$\begin{tikzcd}
 & {[L_g\cup B_i^1,\varphi]} \arrow[d, equal] \\
{[L_g,\varphi]} \arrow[ur, dash] \arrow[r, dash] & {[L_{g+1},\varphi\circ\rho_i^{-1}]}
\end{tikzcd}$$

Observe that to obtain an L_{g+1} form for $[Z\cup B,\varphi']$, we can first push Z to L_g, which sends B to a piece B_i^1, and then append one more ρ_i. □

2.3 Characters and Σ-invariants

Recall that a group is of *type F_m* if it admits a proper cocompact action on an $(m-1)$-connected CW-complex. Assume Y is a CW complex and $h\colon Y \to \mathbb{R}$ is continuous. We call the corresponding filtration $(Y^{t\le h})_{t\in\mathbb{R}}$ on Y *essentially $(m-1)$-connected* if for any $t \in \mathbb{R}$, there exists $s \le t$ such that inclusion $Y^{t\le h} \hookrightarrow Y^{s\le h}$ induces the trivial map in π_k for $k \le m-1$.

For any group G, we define a *character* of G to be a homomorphism from G to the additive group $\mathbb{R}$. As characters factor through G^{ab} (the abelianization of G), the space of characters is a vector space with the same dimension as the rank of G^{ab}. Excluding the trivial character and modding out by positive scaling, we obtain the *character sphere* $\Sigma(G)$. The Σ-invariants (also called BNS or BNSR invariants, for Bieri–Neumann–Strebel(–Renz)) are a filtration of the character sphere into subspaces

$$\Sigma^0(G) \supseteq \Sigma^1(G) \supseteq \Sigma^2(G) \supseteq \cdots,$$

defined as follows (we use a slight correction of the definition in [19], as suggested in [20]).

Definition 2.8 Let G be a group of type F_m, and let Y be an $(m-1)$-connected CW complex on which G acts cocompactly. Suppose that the stabilizer of any k-cell is of type F_{m-k} and is contained in the kernel of every character of G.[1] For each nontrivial $\chi \in \operatorname{Hom}(G,\mathbb{R})$, there is a character height function $h_\chi\colon Y \to \mathbb{R}$, a continuous map satisfying $h_\chi(gy) = \chi(g) + h_\chi(y)$ for all $y \in Y$ and $g \in G$. Then $\Sigma^m(G)$ is the set of those $[\chi]$ such that the filtration $(Y^{t\le h_\chi})_{t\in\mathbb{R}}$ is essentially $(m-1)$-connected.

Of what use are these invariants, one might ask? A standard result is the following classification of when finiteness properties of G are preserved to subgroups containing the commutator subgroup.

Proposition 2.9 *Let G be a group of type F_m, and let K be a subgroup so that $[G,G] \le K \le G$. Then K is of type F_m if and only if for every character $\chi \in \operatorname{Hom}(G,\mathbb{R})$ such that $\chi(K)=0$, we have $[\chi] \in \Sigma^m(G)$.*

Definition 2.10 A *(Zaremsky–)Morse function* on an affine cell complex Y is a map $h=(\chi,f)\colon Y \to \mathbb{R}\times\mathbb{R}$ such that both χ and f are affine on cells. The codomain is ordered lexicographically, and we require that f take only finitely many values on $Y^{(0)}$, and that there is some $\varepsilon > 0$ such that adjacent vertices v and w satisfy either $|\chi(v)-\chi(w)| \ge \varepsilon$, or $\chi(v)=\chi(w)$ and $f(v) \ne f(w)$.

We record the following general form of the Morse lemma, as well as a more specific version of it. Both will be used in Section 4.

Lemma 2.11 *Let $-\infty \le p \le q \le r \le +\infty$. If for every vertex $v \in Y^{q<\chi\le r}$ the descending link $\operatorname{lk}^{h\downarrow}_{Y^{p\le\chi}}(v)$ is $(k-1)$-connected, then the pair $(Y^{p\le\chi\le r}, Y^{p\le\chi\le q})$ is k-connected. If for every vertex $v \in Y^{p\le\chi<q}$ the ascending link $\operatorname{lk}^{h\uparrow}_{Y^{\chi\le r}}(v)$ is $(k-1)$-connected, then the pair $(Y^{p\le\chi\le r}, Y^{q\le\chi\le r})$ is k-connected.*

---

[1] That the cell-stabilizers are contained in the kernels is the modification.

Corollary 2.12 *Let $h = (\chi, f)\colon Y \to \mathbb{R} \times \mathbb{R}$ be a Morse function. If Y is $(m-1)$-connected and for every vertex $v \in Y^{\chi<q}$ the ascending link $\mathrm{lk}_Y^{h\uparrow}(v)$ is $(m-1)$-connected, then $Y^{q\le\chi}$ is $(m-1)$-connected.*

Informally, we can think of the characters on $P\mathcal{H}_n$ as being generated by counting the handle shifts $\rho_1, \dots, \rho_{n-1}$, where ρ_i shifts the i^{th} end into the n^{th} end by one piece. Specifically, for $1 \le i \le n-1$, take $\chi_i(\varphi)$ to be the negative of the sum of the powers of ρ_i appearing in φ; for $i = n$, take χ_n to be the sum of the powers of ρ_1 through ρ_{n-1}. We stress that this definition, while convenient to use, is not obviously well-defined. An equivalent definition, which is better suited to demonstrating well-definedness, is as follows (details can be found in Section 3 of [5]).

Let γ be an oriented curve that separates one end E of Σ_n from the rest, oriented so that the end E is on the right-hand side of γ. Then γ defines a nonzero element of $H_1^{\mathrm{sep}}(\Sigma_n, \mathbb{Z})$. To every $\varphi \in \mathrm{PMap}(\Sigma_n)$ and $\gamma \in H_1^{\mathrm{sep}}(\Sigma_n, \mathbb{Z})$, associate an integer $\theta_{[\gamma]}(\varphi)$, as a "signed genus" between γ and $\varphi(\gamma)$. Then the map $\theta_{[\gamma]}\colon \mathrm{PMap}(\Sigma_n) \to \mathbb{Z}$ is a well-defined nontrivial homomorphism, depending only on the homology class of γ. By identifying H^1_{sep} with $\mathrm{Hom}(H_1^{\mathrm{sep}}, \mathbb{Z})$ via the universal coefficients theorem, we obtain a map $\Theta\colon \mathrm{PMap}(\Sigma_n) \to H^1_{\mathrm{sep}}(\Sigma_n, \mathbb{Z})$, by the rule $\Theta(\varphi)[\gamma] = \theta_{[\gamma]}(\varphi)$. Restricting to $P\mathcal{H}_n$, we obtain characters, and it is not difficult to see that they agree with the informal definition above. We emphasize: χ_i measures how much φ pushes the i^{th} end *out*.

Given a vertex, we can define the character height function $h_{\chi_i}([Z, \varphi])$ by adding the number of pieces of Z in the i^{th} end to $\chi_i(\varphi)$. This is well defined: consider two representatives $[Z, \varphi] = [W, \psi]$. We have that $\varphi^{-1}\psi$ takes W to Z, and is rigid elsewhere. This composition can be further decomposed as a mapping class of some suited subsurface composed with a sequence of handle shifts. This compactly supported mapping class affects neither the distribution of the pieces of Z nor the value of any character on φ, so we need consider only handle shifts. Consider $[Z, \varphi] = [\rho_i(Z), \varphi \circ \rho_i^{-1}]$: we have moved one piece from the i^{th} end to the n^{th} end, while increasing the amount of pushing into the i^{th} end and out of the n^{th} end by one. These cancel out in h_{χ_i} and h_{χ_n}. All other basis characters are unchanged under these operations, and thus we have well-defined character height functions on X_n. For an arbitrary character χ, we shall henceforth abuse notation by writing χ when we mean h_χ. The domain will typically be clear.

3 The Stein–Farley complex X_n is CAT(0)

We have a contractible cube complex on which $P\mathcal{H}_n$ acts nicely. In this section, we show that X_n is in fact CAT(0). To begin, we require the following definition.

Definition 3.1 A cube complex X is *cube-complete* if whenever $X^{(1)}$ contains an embedded copy of the 1-skeleton of a d-cube, X contains the entire d-cube.

Once we have shown that X_n is cube-complete, we seek to apply the following proposition, which is a version of [14, Proposition 4.6], where it is taken out of the proof of [9, Theorem 6.1].

Proposition 3.2 *Let X be a cube-complete cube complex, whose 1-skeleton $X^{(1)}$ is a graph with no loops or bigons. Suppose that*

(a) *X is simply connected,*

(b) *$X^{(1)}$ satisfies the 3-square condition.*

Then X is $\mathrm{CAT}(0)$.

The 3-square condition says that whenever the cube complex has 3 squares intersecting in a vertex, and pairwise intersecting in edges (call such an arrangement a 3-wheel), then they are part of a full 3-cube.

Remark 3.3 In [14], there is a third condition: that the 1-skeleton not contain a copy of the complete bipartite graph $K_{2,3}$. This third condition is redundant: A careful examination of the proof of Theorem 6.1 in [9], especially the paragraph beginning "To prove the converse", yields that the other hypotheses imply the nonexistence of $K_{2,3}$'s: A $K_{2,3}$ is a 3-wheel where all the vertices opposite the common vertex are identified, which cannot be the case in an embedded 3-cube. Another, smaller modification: there was a condition that the 1-skeleton be a connected graph, which is automatically true in a simply connected cube complex.

As we have a height function for X_n, and as adjacent vertices always differ by exactly 1, there are two ways in which X_n can fail to be cube-complete.

Definition 3.4 By a *collapsed* cube, we mean the 1-skeleton of a d-cube, such that the difference between its highest and lowest heights is less than d. By an *empty* cube, we mean the 1-skeleton of a d-cube that is not filled by a d-cube.

Given the definition of X_n, we need to check two things: that $X_n^{(1)}$ has no collapsed cubes, and that $X_n^{(1)}$ has no empty cubes.

Choosing vertex representatives to have their homeomorphism components contained in the pure group allows us a consistent definition of "direction" for adding and removing pieces, ie we may consistently label the ends from 1 to n without worrying about any potential reindexing.

In order to show that X_n is cube-complete, we shall proceed as follows: first, we show that given a vertex $[Z,\varphi]$, there is only one ascending edge per direction (see Lemma 3.5). Secondly, we show that there are no collapsed squares. This easily gives that there are no collapsed cubes of any dimension. Finally, we show that there are no empty squares, and from this induct to show that there are no empty cubes.

Lemma 3.5 (uniqueness of ascending edges) *Given a vertex $v=[Z,\varphi]$, there exists exactly one ascending edge in each direction, ie for $i\in\{1,\dots,n\}$, the collection $\{[Z\cup B_i,\varphi]\}$, where B_i is the piece in the i^{th} end adjacent to Z, exhausts the vertices of the ascending link of v.*

Proof What we wish to show is that, given two representatives $[Z,\varphi]=[Z',\varphi']$, whenever we ascend from $[Z',\varphi']$ by adding a piece in the i^{th} end, we obtain the same vertex as by adding a piece in the i^{th} end to $[Z,\varphi]$. Consider the diagram

$$\begin{tikzcd}
{[Z,\varphi]} \arrow[r, dash] \arrow[d, equal] & {[Z\cup B,\varphi]} \arrow[d, dash, dotted, "?"] \\
{[Z',\varphi']} \arrow[r, dash] & {[Z'\cup B',\varphi']}
\end{tikzcd}$$

Our goal is to show that, assuming B' and B are both in the i^{th} end, we have $[Z'\cup B',\varphi']=[Z\cup B,\varphi]$. As $(\varphi')^{-1}\circ\varphi$ is rigid outside Z, it takes B to some piece adjacent to Z'. As $(\varphi')^{-1}\circ\varphi\in P\mathcal{H}_n$, this implies that $(\varphi')^{-1}\circ\varphi(B)$ must still be in the i^{th} end (elements of $P\mathcal{H}_n$ cannot rigidly move pieces from one end into another), and hence is B'. We then have that $(\varphi')^{-1}\circ\varphi$ takes $Z\cup B$ to $Z'\cup B'$, and is rigid elsewhere, ie that $[Z\cup B,\varphi]=[Z'\cup B',\varphi']$. Thus, there can be only one ascending edge per direction. □

Combining this with the L_g representatives of Lemma 2.7, we see that the ascending edges over $[L_g,\varphi]$ are of the form $[L_{g+1},\varphi\rho_i^{-1}]$, for $i\in\{1,\dots,n\}$. (Recall that ρ_n is just the identity.) Next, we prove that there are no collapsed squares.

Lemma 3.6 *Given any square in $X_n^{(1)}$, the difference between the maximal and minimal heights of vertices is* 2.

Proof For the sake of contradiction, suppose there is some collapsed square, which has only two values for the heights of its vertices. Writing the lower height vertices as $[L_g,\varphi]$ and $[L_g,\psi]$, we see that a collapsed square must occur as in Figure 2. That the same representative can be used for both edges from one of the side vertices is justified by Lemma 3.5. To show that the collapsed square in Figure 2 cannot occur, we assume that the middle pair of vertices are distinct, and prove that the leftmost and rightmost vertices cannot be distinct. The equality of the different forms of the middle vertices says that the transition map $\rho_{i_k}\varphi^{-1}\psi\rho_{j_k}^{-1}$ takes L_{g+1} to itself and is rigid elsewhere, for $k=1,2$. This implies that $\rho_{i_k}\varphi^{-1}\psi\rho_{j_k}^{-1}$ has net-zero shifting in all ends, and hence is compactly supported (recall that for asymptotically rigid maps, these are the same). As the middle vertices are distinct, we know that $i_1\neq i_2$ and $j_1\neq j_2$. Thus, we see that $i_k=j_k$ for $k=1,2$.

We will now show that $[L_g,\varphi]=[L_g,\psi]$ by considering the transition map $\varphi^{-1}\psi$ on pieces outside of L_g. The rigidity of the $\rho_{i_k}\varphi^{-1}\psi\rho_{j_k}^{-1}$ maps outside L_{g+1} immediately handles all but three pieces: $B_{i_1}^1$, $B_{i_2}^1$, and B_n^{g+1}. The last of these is again easy: the L_{g+1} transition maps are rigid on the $(g+2)^{\text{nd}}$ piece, which they first push to the $(g+1)^{\text{st}}$ piece, then apply $\varphi^{-1}\psi$, then push back out by one. As this must be sent rigidly to the $(g+2)^{\text{nd}}$ piece, we see that $\varphi^{-1}\psi$ must be rigid on the $(g+1)^{\text{st}}$ piece. For $B_{i_1}^1$, consider that $\varphi^{-1}\psi$ and $\rho_{i_2}\varphi^{-1}\psi\rho_{i_2}^{-1}$ act the same on it. As the latter is rigid there, so is the former. Symmetrically, we have rigidity on $B_{i_2}^1$, and are done. Specifically, we have shown that $\varphi^{-1}\psi$ takes L_g to itself, and is rigid elsewhere, implying that the supposed collapsed square was degenerate to begin with. □

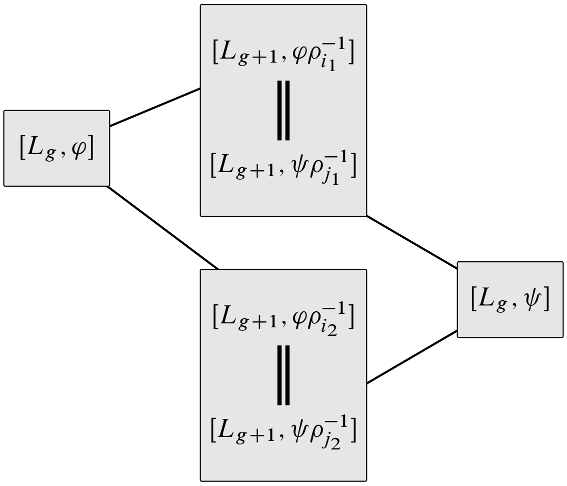

$[L_g, \varphi]$
$[L_{g+1}, \varphi\rho_{i_1}^{-1}]$
$[L_{g+1}, \psi\rho_{j_1}^{-1}]$
$[L_{g+1}, \varphi\rho_{i_2}^{-1}]$
$[L_{g+1}, \psi\rho_{j_2}^{-1}]$
$[L_g, \psi]$

Figure 2: A collapsed square.

Proposition 3.7 *The Stein–Farley complex X_n is cube complete.*

Proof We begin by observing that there can be no collapsed cubes: any such cube must contain a collapsed square. Now we show that there are no empty squares. As there are no collapsed squares, we know that the 1-skeleton of a square must have a lowest vertex, say at height g, two vertices at height $g+1$, and one vertex at height $g+2$. By the uniqueness (per end) of ascending edges (Lemma 3.5), we see that given a representative $[Z, \varphi]$ for the height g vertex, the other vertices must be of the form $[-, \varphi]$, with the blank filled by $Z \cup B_1$, $Z \cup B_2$, and $Z \cup B_1 \cup B_2$, depending on height. The only concern about the $g+2$ vertex would be that the pieces lie in the same end, but we see that this cannot occur. Thus, there are no empty squares.

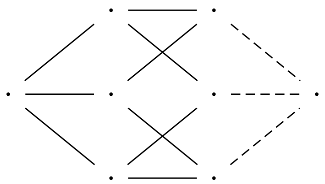
Finally, we show that there are no empty cubes of any dimension. We do so inductively: suppose we already have that there are no empty $(d-1)$-cubes. Consider the 1-skeleton of a d-cube C. Let the bottom vertex of C be $[Z, \varphi]$, at height g. Each of the vertices of C adjacent to $[Z, \varphi]$ is of the form $[Z \cup B_i, \varphi]$, where B_i is in the i^{th} end; each i appears at most once by Lemma 3.5. Also, by the inductive hypothesis, we have that $[Z, \varphi]$ connects to each of the vertices at height $g+d-1$ via a $(d-1)$-cube, so that these vertices can be written as $[Z \cup B_{i_1} \cup \cdots B_{i_{j-1}} \cup \widehat{B}_{i_j} \cup B_{i_{j+1}} \cup \cdots \cup B_{i_d}, \varphi]$, where the hat indicates omission. Choose two such vertices, and consider their common lower vertex (the one missing the two pieces missing in either of the chosen vertices). Applying the doctrine of no empty squares to these 3 vertices and the apex yields that the apex can be written as $[Z \cup B_{i_1} \cup \cdots \cup B_{i_d}, \varphi]$, and thus the cube is filled. □

Remark 3.8 The argument above that ascending edges are unique per direction emphatically does not hold for descending edges. In fact, for any vertex of height at least 1, there are infinitely many descending edges per direction in which it can have a descending edge (given $[Z, \varphi]$, choose $\psi \in \mathcal{PH}_n$ such that $\varphi^{-1}\psi|_Z$ is a homeomorphism, with some nontrivial behavior in the piece to be removed, and such that

$\varphi^{-1}\psi$ is rigid outside Z). It is helpful to keep in mind the following: when ascending, one can always choose coherent representatives from a minimal vertex; when descending, nothing is guaranteed.

Theorem 3.9 *Let X_n denote the Stein–Farley complex for $\mathcal{H}_n$. Then X_n is* CAT(0).

Proof By Proposition 3.7, X_n is cube-complete. By construction, $X_n^{(1)}$ has no loops or bigons. As X_n is contractible, it is simply connected, so item (a) is covered. For item (b), we need to show that 3-wheels can always be completed. There are four possibilities: the common vertex can be lowest height, highest height, or either of the intermediate heights. In the first case, we can have all three edges out of the common vertex use the same representative, each adding a piece/shift in a different end, so that the cube is completed by adding in all three at once. In the intermediate height cases, something similar occurs: we always have the lowest height vertex of the desired cube from which we can build upwards. It is useful to keep in mind the following arrangement of the 1-skeleton of a 3-cube; the solid lines indicate the 3-wheel, and the dashed lines indicate the rest of the 3-cube:



The difficulty is in the remaining case, where the common vertex is maximal in the cube. We begin with the diagram in Figure 3. Our goal is to obtain a representative $[L_{g+2}, \varphi \circ \rho_k]$ for the height $g+2$ vertex currently represented with both φ' and φ''. This will imply the existence of a vertex $[L_g, \varphi \circ \rho_i \rho_j \rho_k]$, which realizes this 3-wheel in a 3-cube. To begin, we have the following lemma that justifies part of the diagram.

Lemma 3.10 *Whenever $[L_g, \varphi \circ \rho_i] = [L_g, \varphi' \circ \rho_j]$ and $[L_{g+1}, \varphi] = [L_{g+1}, \varphi']$, we have that $i = j$.*

Proof First, observe that $[L_g, \varphi \circ \rho_i] = [L_g, \varphi' \circ \rho_j]$ means that $h = \rho_i^{-1}\varphi^{-1}\varphi'\rho_j$ takes L_g to itself, and is rigid elsewhere; also, $[L_{g+1}, \varphi] = [L_{g+1}, \varphi']$ means that $\varphi^{-1}\varphi'(L_{g+1}) = L_{g+1}$. Assume $j \neq n$, and consider the piece B_j^1; let α be the boundary curve it shares with $\mathcal{O}$. What does h do to α? First, ρ_j deforms it into the center and stretches it into the n^{th} end, making it an essential curve in L_g. Then $\varphi^{-1}\varphi'$ moves it around to some essential curve in L_{g+1}. Finally, ρ_i^{-1} pushes and stretches it towards the i^{th} end. As B_j^1 is not in L_g, it must be sent to a piece, and $h(\alpha)$ must be the intersection of this piece with L_g. As all representatives here do not permute ends, it must be that $h(\alpha) = \alpha$. If $i \neq j$, then ρ_i fixes α, so $\rho_i(\alpha) \neq \varphi^{-1}\varphi'\rho_j(\alpha)$, and $h(\alpha) \neq \alpha$. Hence we must have that $i = j$. In the case where $j = n$, equality is obvious. □

Now, we seek the following equality: $[L_{g+2}, \varphi \circ \rho_k] = [L_{g+2}, \varphi' \circ \rho_k]$. That is, we wish to show that $\rho_k^{-1}\varphi^{-1}\varphi'\rho_k$ takes L_{g+2} to itself, and is rigid elsewhere. We already know that this holds for $\rho_i^{-1}\varphi^{-1}\varphi'\rho_i$. The rigidity here occurs precisely when $\varphi^{-1}\varphi'$ does nothing untoward to whatever single piece gets

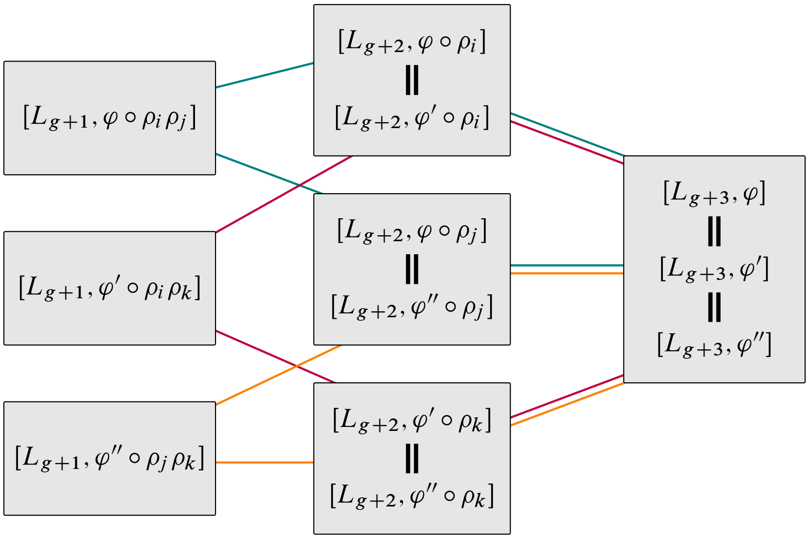

$[L_{g+1}, \varphi \circ \rho_i \rho_j]$
$[L_{g+1}, \varphi' \circ \rho_i \rho_k]$
$[L_{g+1}, \varphi'' \circ \rho_j \rho_k]$
$[L_{g+2}, \varphi \circ \rho_i] \Vert [L_{g+2}, \varphi' \circ \rho_i]$
$[L_{g+2}, \varphi \circ \rho_j] \Vert [L_{g+2}, \varphi'' \circ \rho_j]$
$[L_{g+2}, \varphi' \circ \rho_k] \Vert [L_{g+2}, \varphi'' \circ \rho_k]$
$[L_{g+3}, \varphi] \Vert [L_{g+3}, \varphi'] \Vert [L_{g+3}, \varphi'']$

Figure 3: A 3-wheel with common vertex maximal.

pushed into the n^{th} end, so it doesn't matter which index we choose! (If we have chosen the index n, there is no pushing. To resolve this, simply choose a different height $g+2$ vertex to work with. They can't share ends, as they are in squares with lower vertices, so we can apply Lemma 3.5.)

One small detail remains: we have a bottom vertex connecting to one of the height $g+1$ vertices, but does it connect to the other height $g+1$ vertices? Suppose not: then the cube between $[L_{g+3}, \varphi]$ and $[L_g, \varphi \circ \rho_i \rho_j \rho_k]$ contains a square differing from one of the original 3 squares only at its height $g+1$ vertex (that is, at its lowest vertex). This however means that we would have two squares agreeing on 3 vertices, disagreeing on the 4th, which would contain a collapsed square. As this cannot occur (Lemma 3.6), our height g vertex connects to each of the height $g+1$ vertices we started with. □

4 Σ-invariants of $P\mathcal{H}_n$

The methods of this section are taken from [19; 20]. They need only minimal modification to work in this setting, and are presented partly for the sake of having the entire argument in one place. We begin by noting that the abelianization of $P\mathcal{H}_n$ is $\mathbb{Z}^{n-1}$, and the abelianization map is $(\chi_1, \dots, \chi_{n-1})$ (see Section 6 of [2]). As $\chi_1 + \cdots + \chi_n = 0$, any (nontrivial) character χ can be written (up to renumbering the ends of Σ_n) in ascending standard form, ie $\chi = a_1\chi_1 + \cdots + a_n\chi_n$, with $a_1 \le \cdots \le a_{m(\chi)} < a_{m(\chi)+1} = \cdots = a_n = 0$. We shall henceforth assume all characters are written in such form. Observe that $m(\chi)$ is defined to be the maximal index so that $a_{m(\chi)} < a_n = 0$. Then our goal in this section is to prove the following.

Theorem 4.1 *Let χ be a nonzero character of $P\mathcal{H}_n$. Then $[\chi] \in \Sigma^{m(\chi)-1}(P\mathcal{H}_n) \setminus \Sigma^{m(\chi)}(P\mathcal{H}_n)$.*

4.1 Inclusion

We begin by showing the inclusion into the $(m(\chi)-1)$-layer of $\Sigma(P\mathcal{H}_n)$.

Theorem 4.2 *For any nonzero character* $\chi \in \operatorname{Hom}(P\mathcal{H}_n, \mathbb{R})$, *we have* $[\chi] \in \Sigma^{m(\chi)-1}(P\mathcal{H}_n)$.

We recall here the form of the Morse lemma we shall use in this section.

Corollary 4.3 *Let* $h = (\chi, f)\colon Y \to \mathbb{R} \times \mathbb{R}$ *be a Morse function. If* Y *is* $(m-1)$*-connected and for every vertex* $v \in Y^{\chi<q}$ *the ascending link* $\operatorname{lk}_Y^{h\uparrow}(v)$ *is* $(m-1)$*-connected, then* $Y^{q\le\chi}$ *is* $(m-1)$*-connected.*

From now on, we write X for X_n. The role of Y shall be played by sublevel subcomplexes $X^{f\le q}$ for sufficiently large q. Let $\chi = a_1\chi_1 + \cdots + a_n\chi_n$ be a nontrivial character of $P\mathcal{H}_n$, written in ascending standard form. The function $h = (\chi, f)\colon Y \to \mathbb{R}^2$ is a Morse function. Observe that between any two adjacent vertices, each basis character differs by 1, 0, or -1. We examine h-ascending links in Y.

Lemma 4.4 *Let* $v = [L_g, \varphi]$ *be a vertex in* Y*. An adjacent vertex* $w = [L_{g\pm1}, \psi]$ *is in the* h*-ascending link of* v *if and only if one of the two following conditions holds*:

- $w = [L_{g-1}, \psi]$, *with* $v = [L_g, \psi\rho_i^{-1}]$ *where* $i \le m(\chi)$.
- $w = [L_{g+1}, \varphi\rho_i^{-1}]$, *where* $i \ge m(\chi) + 1$.

Proof If w is f-ascending from v, then $w = [L_{g+1}, \varphi\rho_i^{-1}]$ for some i. If $i \ge m(\chi)+1$, then $\chi(w) = \chi(v)$, so w is h-ascending from v. If $i \le m(\chi)$, then w is χ-descending from v, hence h-descending from v. If w is f-descending, then $v = [L_g, \psi\rho_i^{-1}]$ for some i. If $i \ge m(\chi) + 1$, then we again have $\chi(w) = \chi(v)$, so that w is h-descending from v. If $i \le m(\chi)$, then w is χ-ascending from v, hence h-ascending. □

Remark 4.5 It is worth observing a difference here from the situation in [19]. For the regular Houghton groups, there is one way to go "up" and finitely many ways to go "down" per end (with respect to f); here, while there is only one way to go up per end, there are infinitely many ways to go down. However, the h-ascending link is still a join of its intersection with the f-ascending and f-descending links, as these elements cannot share ends.

We shall need the following fact (see [2, Section 5.2]): Let $v \in X_n$ be a vertex. If $f(v) \ge 2n$, then the f-descending link of v is $(n-2)$-connected. The version of this statement for the regular Houghton groups uses $f(v) \ge 2n - 1$ (see [16, Theorem 3.52]); this discrepancy will result in most of the numbers in the following proposition being either one above or one below the corresponding numbers in [19].

Now set $q = 3n - 2$, so $Y = X^{f\le 3n-2}$.

Proposition 4.6 *Let* v *be a vertex in* Y*. Then* $\operatorname{lk}_Y^{h\uparrow}(v)$ *is* $(m(\chi)-2)$*-connected.*

Proof We have that $f(v)$ is between 0 and $3n-2$. Suppose that $f(v) \leq 2n+m(\chi)-2$. Writing $v=[L_g,\varphi]$, we have that there are $n-m(\chi)$ indices i for which $v'=[L_{g+1},\varphi \circ \rho_i^{-1}]$ is h-ascending. As $(2n+m(\chi)-2)+(n-m(\chi))=3n-2$, the entire f-ascending link of v in X is contained in Y. Thus, the f-ascending part of the h-ascending link is an $(n-m(\chi)-1)$-simplex, which is contractible, so that $\mathrm{lk}_Y^{h\uparrow}(v)$ is contractible.

Suppose instead that $2n+m(\chi)-1 \leq f(v) \leq 3n-2$, and thus that Y does not contain the entire f-ascending part of the h-link of v. We still have its $(3n-f(v)-3)$-skeleton, which, being a skeleton of an $(n-m(\chi)-1)$-simplex, is $(3n-f(v)-4)$-connected. As $f(v) \geq 2n$, we have that the entire f-descending part of the h-ascending link in is Y. This is isomorphic to the f-descending link of a vertex of the same f-height as v in $X_{m(\chi)}$, which we know is $(m(\chi)-2)$-connected so long as $f(v) \geq 2m(\chi)$. As $f(v) \geq 2n+m(\chi)-1$, we need that $2n-1 \geq m(\chi)$, which is certainly true. The join is now $((3n-f(v)-3)+(m(\chi)-1))$-connected. As $f(v) \leq 3n-2$, we have that $(3n-f(v)+m(\chi)-4) \geq m(\chi)-2$, and thus that $\mathrm{lk}_Y^{h\uparrow}(v)$ is $(m(\chi)-2)$-connected. □

4.2 Exclusion

To demonstrate that the result of the last section is sharp, we shall need to employ different techniques. The main result of this section is:

Theorem 4.7 *For any nonzero character* $\chi \in \mathrm{Hom}(P\mathcal{H}_n,\mathbb{R})$, *we have* $[\chi] \notin \Sigma^{m(\chi)}(\mathcal{H}_n)$.

We gather here the propositions and definitions from [20] that we will need, starting with the strong nerve lemma:

Proposition 4.8 *Let* X *be a CW-complex covered by subcomplexes* $(X_i)_{i\in I}$ *and let* L *be the nerve of the cover. Let* $n \geq 1$. *Suppose that any nonempty intersection* $X_{i_1} \cap \cdots \cap X_{i_r}$ *is* $(n-r)$*-connected. Then* $H_k(X) \cong H_k(L)$ *for all* $k \leq n-1$, *and* $H_n(X)$ *surjects onto* $H_n(L)$.

Definition 4.9 For a set of indices $K \subseteq [n]$, consider the subcomplex $\bigcap_{i\in K} X^{\chi_i \leq 0}$ of X. Call any connected component of such a subcomplex a *K-blanket*. By a *blanket* we mean a K-blanket for some unspecified K.

Recall that a subcomplex Z of a CAT(0) cube complex Y is *locally combinatorially convex* if every link in Z of a vertex $z \in Z^{(0)}$ is a full subcomplex of the link of z in Y, and *combinatorially convex* if it is connected and locally combinatorially convex. It is known that combinatorially convex implies CAT(0), hence contractible. In particular, this applies to connected components of locally combinatorially convex subcomplexes.

Lemma 4.10 *For any* K, $\bigcap_{i\in K} X^{\chi_i \leq 0}$ *is locally combinatorially convex. Thus, blankets are combinatorially convex, and hence* CAT(0) *and contractible.*

Proof It suffices to show that each $X^{\chi_i \le 0}$ is locally combinatorially convex. Given a pair of adjacent vertices, we can write them as $v = [L_g, \varphi]$ and $w = [L_{g+1}, \varphi \circ \rho_j^{-1}]$ for some j. Then $\chi_i(w) - \chi_i(v) = \delta_{i,j}$. Thus, if C is a cube containing v, and $w_1, \dots, w_k$ are the vertices of C adjacent to v, then the maximum and minimum values of χ_i on C lie in $\{\chi_i(v), \chi_i(w_1), \dots, \chi_i(w_k)\}$. Thus, whenever $v \in X^{\chi_i \le 0}$ and all these w_j's lie in the link of v in $X^{\chi_i \le 0}$, then the cube C lies in $X^{\chi_i \le 0}$. This implies that the link of v in $X^{\chi_i \le 0}$ is a full subcomplex of the link of v in X. □

As in [20], we have the immediate corollary that intersections of blankets are blankets for the union of the sets of indices. What follows is essentially identical to Zaremsky's approach for the Houghton groups, reproduced here with minor changes for the sake of completeness. Recall now the more general statement of the Morse lemma, Lemma 2.11. For notational convenience, we write $X_{f \le k}$ for $X^{f \le k}$, and $X^{t \le \chi}_{f \le k}$ for the intersection $X_{f \le k} \cap X^{t \le \chi}$.

Lemma 4.11 *If $X^{0 \le \chi}_{f \le 3n-2}$ is not $(m(\chi)-1)$-connected, then $[\chi] \in \Sigma^{m(\chi)}(P\mathcal{H}_n)^c$.*

Proof If $[\chi] \in \Sigma^{m(\chi)}(P\mathcal{H}_n)$, then the filtration $(X^{t \le \chi}_{f \le 3n-2})_{t \in \mathbb{R}}$ is essentially $(m(\chi)-1)$-connected. Every h-ascending link of a vertex in $X_{f \le 3n-2}$ is $(m(\chi)-2)$-connected, so for any $s \le t$ the inclusion $X^{t \le \chi}_{f \le 3n-2} \hookrightarrow X^{s \le \chi}_{f \le 3n-2}$ induces an isomorphism in π_k for $k \le m(\chi) - 2$, and a surjection in $\pi_{m(\chi)-1}$. By assumption, for any t there is some $s \le t$ such that this inclusion induces a trivial map in π_k for $k \le m(\chi) - 1$, implying that $X^{s \le \chi}_{f \le 3n-2}$ is $(m(\chi)-1)$-connected. Rescaling if necessary, we can assume that $s \in \chi(P\mathcal{H}_n)$, and thus we can translate to obtain $X^{s \le \chi}_{f \le 3n-2} \cong X^{0 \le \chi}_{f \le 3n-2}$, so that $X^{0 \le \chi}_{f \le 3n-2}$ is itself $(m(\chi)-1)$-connected. □

In order to show that $X^{0 \le \chi}_{f \le 3n-2}$ is not $(m(\chi)-1)$-connected, we will apply the strong nerve lemma to a covering we now define. For $1 \le i \le n$, let $\{Z_i^\alpha\}$ be the collection of $\{i\}$-blankets in X. The α's are indices in some set, and this index set is not itself important. Set also

$$Y_i^\alpha = Z_i^\alpha \cap X^{0 \le \chi}_{f \le 3n-2}.$$

Lemma 4.12 *The Y_i^α with $1 \le i \le m(\chi)$ cover $X^{0 \le \chi}_{f \le 3n-2}$.*

Proof As the Z_i^α are the connected components of the $X^{\chi_i \le 0}$, it suffices to show that

$$X^{0 \le \chi} \subseteq \bigcup_{i=1}^{m(\chi)} X^{\chi_i \le 0}.$$

As $\chi = a_1 \chi_1 + \cdots + a_{m(\chi)} \chi_{m(\chi)}$, with all coefficients negative, any vertex $v \in X$ with $\chi(v) \ge 0$ must satisfy $\chi_i(v) \le 0$ for some i. This implies the inclusion on vertices. Given a cube in $X^{0 \le \chi}$, let v be its maximal vertex with respect to f. Then all the vertices w of the cube satisfy $\chi_i(w) \le \chi_i(v)$, and hence the entire cube lies in whichever $X^{\chi_i \le 0}$ contains v. □

Lemma 4.13 *Any nonempty intersection of subcomplexes of the form Y_i^α with $1 \le i \le m(\chi)$ is $(m(\chi)-2)$-connected.*

Proof To be nonempty, such an intersection can include at most one term Y_i^α for each i, and can thus be written $Y = Y_{i_1}^{\alpha_1} \cap \cdots \cap Y_{i_r}^{\alpha_r}$, with the i_j all pairwise distinct. Let $Z = Z_{i_1}^{\alpha_1} \cap \cdots \cap Z_{i_r}^{\alpha_r}$, so that $Y = Z \cap X^{0 \le \chi}_{f \le 3n-2}$. We apply Morse-theoretic techniques to Z, this time using Lemma 2.11. As Z is an intersection of blankets, it is a blanket, and thus contractible. Given adjacent vertices $w = [L_g, \varphi]$ and $v = [L_{g+1}, \varphi \circ \rho_i^{-1}]$ with $v \in Z$, we have that $w \in Z$. Thus, for any vertex of Z, the entire f-descending link is in Z. As this is $(n-2)$-connected for $f(v) \ge 2n$, we see that $Z_{f \le 3n-2}$ is $(n-2)$-connected, and so is certainly $(m(\chi)-2)$-connected. Now consider Y as $Z^{0 \le \chi}_{f \le 3n-2}$. As before, the h-ascending link of a vertex v is a join between its f-ascending and f-descending parts. The latter is in Z for the same reasons as above; the former is in Z because it consists of directions i where $m(\chi) + 1 \le i \le n$, on which the considered χ_{i_j} are constant. Thus, the h-ascending link of v is in $Z_{f \le 3n-2}$. As $Z_{f \le 3n-2}$ is $(m(\chi)-2)$-connected, the Morse lemma tells us that Y is $(m(\chi)-2)$-connected. □

Let L be the nerve of the covering of $X^{0 \le \chi}_{f \le 3n-2}$ by the Y_i^α. Since $[\chi] \in \Sigma^{m(\chi)-1}(P\mathcal{H}_n)$, we know that $X^{0 \le \chi}_{f \le 3n-2}$ is $(m(\chi)-2)$-connected, so by the strong nerve lemma L is $(m(\chi)-2)$-connected. The final missing piece is to prove that L is not $(m(\chi)-1)$-acyclic.

Lemma 4.14 *The nerve L is not $(m(\chi)-1)$-acyclic.*

Proof Consider vertices corresponding to Y_i^α and Y_j^β. These vertices can only be adjacent if $i \ne j$, so L is $(m(\chi)-1)$-dimensional. Thus, it suffices to exhibit a nontrivial $(m(\chi)-1)$-cycle. This will come from a collection of $2m(\chi)$ vertices, 2 for each $i \in \{1, \dots, m(\chi)\}$, labeled as $Y_i^{\epsilon_i}$, with $\epsilon \in \{1, 2\}$, with the property $Y_1^{\epsilon_i} \cap \cdots \cap Y_{m(\chi)}^{\epsilon_{m(\chi)}} \ne \varnothing$. This will yield an embedded $(m(\chi)-1)$-sphere in L, which is homologically nontrivial for dimensional reasons.

Recall that $\mathcal{O}$ is the centerpiece of our surface. For each i, take Y_i^1 to be the Y_i^α containing $v_0 = [\mathcal{O}, \mathrm{id}]$, and take Y_i^2 to be the Y_i^α containing $v_i = [\mathcal{O}, \varphi_i]$, where φ_i is some nontrivial mapping class in the i^{th} end; for the sake of definiteness, choose some essential simple closed curve in the standard piece, and let φ_i be the Dehn twist about its image in B_i^1. Any intersection $Y_1^{\epsilon_1} \cap \cdots \cap Y_{m(\chi)}^{\epsilon_{m(\chi)}}$ includes $w = \bigl[\mathcal{O}, \prod \varphi_i\bigr]$, where the product (in an arbitrary order) is taken over those i with $\epsilon_i = 2$, and is therefore nonempty.

It remains to show that $Y_i^1 \ne Y_i^2$ for each i. It suffices to show that $Z_i^1 \ne Z_i^2$. If these are equal (call them Z_i), then we can connect v_0 to v_i via a path in Z_i. In X, one could assume that such a path is one on which f first strictly increases, then strictly decreases; since Z_i is combinatorially convex, this property holds for Z_i as well. Since the path lies in Z_i, χ_i is nonpositive on the whole path. By the uniqueness of ascending edges, this is a path of the form

$$\begin{tikzcd}
& {[Z, \mathrm{id}]} \arrow[r, equal] \arrow[dl, dotted, dash] & {[Z', \varphi_i]} \arrow[dr, dotted, dash] & \\
{[\mathcal{O}, \mathrm{id}]} & & & {[\mathcal{O}, \varphi_i]}
\end{tikzcd}$$

where each of the ascending dotted lines indicates a sequence of edges which only add pieces. Since $\chi_i(v_0) = 0 = \chi_i(v_i)$, none of the edges of the path can be obtained by adding a piece in the i^{th} end. One of Z and Z' must have at least one piece in the i^{th} end, as the transition map is φ_i, which has nonrigid behavior in the i^{th} end. Thus, we have a contradiction. □

5 Subgroups of maximal finiteness length

We use the notation $\mathrm{fl}(G) = n$ to mean that G is a group of type F_n but not of type F_{n+1}. We know that if $G < H$ is finite index, then $\mathrm{fl}(G) = \mathrm{fl}(H)$. But what of the converse, that is, when does $\mathrm{fl}(G) = \mathrm{fl}(H)$ imply that G is finite index in H? For Houghton groups, and for the (pure) surface Houghton groups, the answer is positive for sufficiently large subgroups, in particular for coabelian subgroups. We denote by G' the commutator subgroup of a group G. Recall that the commutator subgroups of the Houghton group H_n and the pure surface Houghton group $P\mathcal{H}_n$ are the finitely supported and compactly supported elements, respectively. (Really, we could say compactly supported in both cases, using the discrete topology for the $\mathbb{N}$-rays on which the Houghton group acts.) For Theorem 5.1 and Proposition 5.2, let H be either H_n or $P\mathcal{H}_n$. For H_n, this is a mild extension of [20, Corollary 2.6].

Theorem 5.1 *Let $G < H$ be a subgroup intersecting the commutator subgroup H' in a finite-index subgroup. If $\mathrm{fl}(G) = n-1$, then G is finite index in H.*

It suffices to show this for groups containing the commutator subgroup. We begin by showing that when G contains the commutator subgroup, the image of G in the abelianization is a maximal rank sublattice, which will then imply finite index.

Proposition 5.2 *Let $G < H$ be as above, ie $\mathrm{fl}(G) = n-1$, and G contains the commutator H'. Write F for the abelianization map to $\mathbb{Z}^{n-1}$. Then $F(G)$ is finite index in $\mathbb{Z}^{n-1}$.*

Proof We start by observing that $F = (\chi_1, \dots, \chi_{n-1})$. As $\Sigma^{n-1}(H)$ is empty, the only way for G to be of type F_{n-1} is that it cannot be killed by any nonzero character of H. We shall construct a sequence of maps $F_i = (F_i^1, \dots, F_i^{n-1}) : H \to \mathbb{Z}^{n-1}$ with $F = F_1$, and a sequence of elements $g_i \in G$ such that $F_n^j(g_i)$ is positive when $i = j$, and zero when $i < j$. Then $F_n(G)$ will be a maximal rank sublattice of $\mathbb{Z}^{n-1}$, obtainable from $F_1(G)$ by integer matrix transformations.

As $\chi_1(G) \neq 0$, we have some smallest positive integer $c_1 \in \chi_1(G)$. Let $g_1 \in \chi_1^{-1}(c_1) \cap G$, and write $F_1(g_1) = (a_1^1, \dots, a_{n-1}^1)$ (note that $a_1^1 = c_1$). Set $F_2 = (\chi_1, a_1^1\chi_2 - a_2^1\chi_1, \dots, a_1^1\chi_{n-1} - a_{n-1}^1\chi_1)$. Then $F_2(g_1) = (c_1, 0, \dots, 0)$. By the same argument, we can choose $c_2 > 0$ minimal in the image of the character $a_1^1\chi_2 - a_2^1\chi_1$, and then $g_2 \in G$ in its preimage. Carrying on, we obtain F_i from F_{i-1} by modifying only the components from i onwards, and build a collection $g_1, \dots, g_{n-1} \in G$ such that the matrix obtained by applying F_n to this collection is lower triangular with integer entries, with positive values on the main diagonal. □

We now prove Theorem 5.1, using the characterization of the commutator subgroup as the elements of compact support. For H_n, the ends correspond to the rays in the obvious way.

Proof Suppose first that G contains the commutator subgroup. Consider the map of coset spaces $\Psi\colon H/G \to \mathbb{Z}^{n-1}/(\boldsymbol{c}\mathbb{Z}^{n-1})$ given by $hG \mapsto F(h)\boldsymbol{c}\mathbb{Z}^{n-1}$. To see that this is well-defined, suppose $h_1G = h_2G$: then $h_1 = h_2g$ for some $g \in G$, and $F(h_1) = F(h_2) + F(g)$. As $F(g) \in \boldsymbol{c}\mathbb{Z}^{n-1}$, we see that $\Psi(h_1G) = \Psi(h_2G)$. We now wish to show that Ψ is injective. Suppose that $\Psi(h_1G) = \Psi(h_2G)$. Then $F(h_1h_2^{-1}) = F(h_1) - F(h_2) = F(g)$ for some $g \in G$. So our question becomes: does $F(h_1h_2^{-1}) \in F(G)$ imply $h_1h_2^{-1} \in G$? The element $h_1h_2^{-1}g^{-1}$ will have no net translation in any end, and hence is compactly supported, and therefore is in G. □

As there are clearly subgroups of finite index in both cases (take the preimage of a finite-index subgroup of $\mathbb{Z}^{n-1}$), there is one loose end remaining: the infinite-index case. Specifically, do there exist subgroups $G < H_n$ (or $G < P\mathcal{H}_n$) whose intersection with the commutator subgroup have infinite index, and such that $\operatorname{fl}(G) = n-1$. First, we see that G is necessarily infinite index in H_n (or $P\mathcal{H}_n$), so all that remains is to check whether this case actually occurs. We thank Noel Brady for the proofs of the following lemmas.

Lemma 5.3 *Suppose $G, K \le H$ are subgroups. If G is finite index in H, then $G \cap K$ is finite index in K.*

Proof Suppose $[H : G] = m < \infty$. Write $H = Gh_1 \cup \cdots \cup Gh_m$, as a disjoint union. For each index i such that $K \cap Gh_i \neq \varnothing$, there is some $g_i \in G$ and $k_i \in K$ such that $k_i = h_ig_i$. Then we can write

$$Gh_i = Gg_i^{-1}k_i = Gk_i.$$

Thus,

$$K \cap Gh_i = K \cap Gk_i = Kk_i \cap Gk_i = (K \cap G)k_i.$$

Intersecting this with our original disjoint decomposition for H, we have

$$K = (K \cap G)k_1 \cup \cdots \cup (K \cap G)k_s,$$

where $s \le m$ is the number of indices for which $K \cap Gh_i \neq \varnothing$. □

Lemma 5.4 *Suppose that a group H has an exhaustion by subgroups, ie there exists a nested chain of subgroups $K_1 \le K_2 \le \cdots \le H$ such that H is the union of all the K_i's. Let $G \le H$. Then $[H : G]$ is finite if and only if the sequence $([K_i : G \cap K_i])$ is eventually constant, in which case it is the limiting value.*

Proof By the proof of Lemma 5.3, we see that the sequence $([K_i : G \cap K_i])$ is bounded by $[H : G] = m$. It is not hard to see that it is nondecreasing, so we need only show that it can't stabilize below m. Choosing a decomposition of H into G-cosets, say $H = Gh_1 \cup \cdots \cup Gh_m$, we see that there must be some index j such that K_j contains all of $\{h_1, \dots, h_m\}$. Therefore, the sequence of indices stabilizes at m.

For the converse, consider a decomposition $H = Gh_1 \cup Gh_2 \cup \cdots$. Intersecting this with K_i, we see as before that the index $[K_i : G \cap K_i]$ is the number of indices j such that $Gh_j \cap K_i$ is nonempty. As every h_j must be in some K_i, we see that if $[H : G]$ is infinite, the sequence $([K_i : G \cap K_i])$ goes to infinity. □

We return to our question on the existence of infinite-index subgroups of Houghton groups with finiteness length $n-1$, We must now split off the ordinary Houghton group from the surface Houghton group. For H_n, the answer to our existence question is easily yes: there are even copies of H_n itself as infinite-index subgroups of H_n! Consider the stabilizer of a single point in $[n]\times\mathbb{N}$: ignoring the fixed point, and sliding its ray back by one to fill in, we have a natural bijection between $[n]\times\mathbb{N}$, on which H_n acts, and $([n]\times\mathbb{N})\setminus\{(i,j)\}$, on which $\mathrm{Stab}(i,j)$ acts, and we see that these actions are the same. (This fact was known to Houghton; see [15].)

By the same argument, the subgroup fixing pointwise any finite set will be a copy of H_n, and the subgroup fixing (as a set, not pointwise) any finite set will be a finite extension of H_n.

In [10], Yves Cornulier defined a stronger failure of co-Hopfianness, which he called "dis-cohopfian". Call a group G *dis-co-Hopfian* if there is some injective homomorphism $\eta\colon G\to G$ such that the intersection of all iterated images of this homomorphism is trivial, ie

$$\bigcap_{n=1}^{\infty}\eta^n(G)=\{e\}.$$

By taking the finite set being stabilized to be the first point in each $\mathbb{N}$-ray, we obtain the following:

Proposition 5.5 *The Houghton groups H_n are not co-Hopfian, and are in fact dis-co-Hopfian.*

To extend the result on the existence of infinite-index subgroups of finiteness length $n-1$ to surface Houghton groups, we shall work out more carefully the embedding $\mathrm{Br}\,H_n\hookrightarrow\mathcal{H}_n$ suggested by [2]. For the braided Houghton group, we take the asymptotically rigid mapping class group of the following surface, which we shall call $\dot{\Sigma}$ (see [12] for details). Begin with a $2n$-gon as the center piece, and take a punctured square for the attached pieces (see Figure 4). In [13], it was shown that the braided Houghton group $\mathrm{Br}\,H_n$ is type F_{n-1} but not type FP_n. This makes it a good candidate for an infinite-index subgroup of $\mathcal{H}_n$ with $\mathrm{fl}(\mathrm{Br}\,H_n)=\mathrm{fl}(\mathcal{H}_n)$.

To obtain a homomorphism $\mathrm{Br}\,H_n\to\mathcal{H}_n$, we replace each puncture with a boundary component, and then pass to the double. Write Σ' for the surface obtained by replacing punctures with boundary components. As for why this yields an injective homomorphism, consider the following diagram, where the upper row is exact:

$$\begin{tikzcd}
1 \arrow[r] & K \arrow[r] & \mathrm{Map}(\Sigma') \arrow[r, "a"] \arrow[d, "b"] & \mathrm{Map}(\dot{\Sigma}) \arrow[r] & 1 \\
 & & \mathrm{Map}(\Sigma) & &
\end{tikzcd}$$

The failure of b to be an injective homomorphism is precisely the subgroup generated by boundary parallel twists. The kernel of a is generated by twists parallel to the compact boundary components. Therefore, we obtain an injective homomorphism $\mathrm{Map}(\dot{\Sigma})\to\mathrm{Map}(\Sigma)$. Restricting to asymptotically rigid subgroups

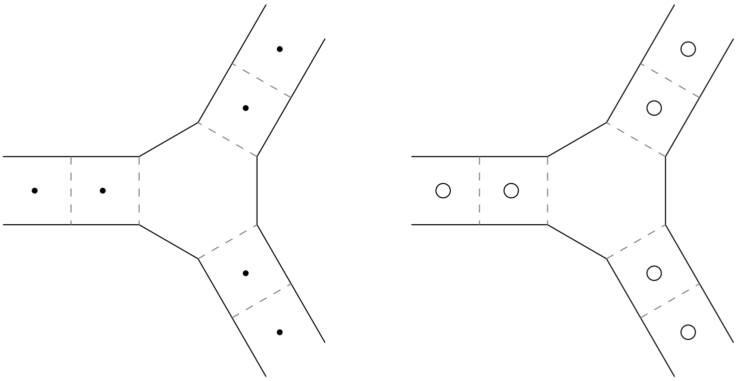


Figure 4: Left: the defining surface for Br H_n. Right: the surface Σ'.

yields an injective homomorphism Br $H_n \to \mathcal{H}_n$. As Br H_n acts trivially on the maximal (ie nonpuncture) ends, this image lies in $P\mathcal{H}_n$.

The image of this homomorphism is a subgroup which fixes (up to isotopy) the multicurve defined by the images of the curve β in Figure 5 in each piece, via the canonical maps ι_B of Section 2.1.

All that remains is to see that this subgroup has infinite index in $P\mathcal{H}_n$. Consider its intersection with the mapping class group of any suited subsurface: here, it must fix the multicurve consisting of the above curves, and is therefore of infinite index. As infinite index in a subgroup implies infinite index in the full group, we have our result.

Between Theorem 5.1, Proposition 5.5, and the above discussion, we have proven the following:

Theorem 5.6 *Let H denote either the Houghton group, or the pure surface Houghton group, and suppose $G < H$ has $\mathrm{fl}(G) = \mathrm{fl}(H)$. Then G is finite index in H if and only if $G \cap H'$ is finite index in H', where H' denotes the commutator subgroup of H. Furthermore, there exist subgroups G with $\mathrm{fl}(G) = \mathrm{fl}(H)$ of both finite and infinite index.*

We finish with further consideration of co-Hopfianness. It is easy to see that Br H_n is not co-Hopfian: there is an inclusion map from the defining surface of Br H_n to itself, sliding the first puncture in some end out, and with all elements moving things only on one side of the skipped puncture (see Figure 6).

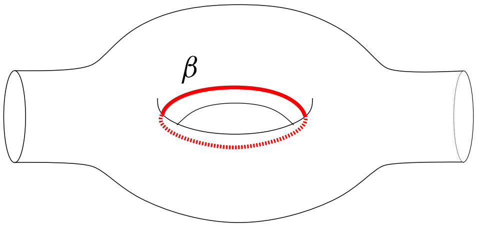


Figure 5: The central curve β in a piece which is fixed by the image of Br H_n in $P\mathcal{H}_n$.

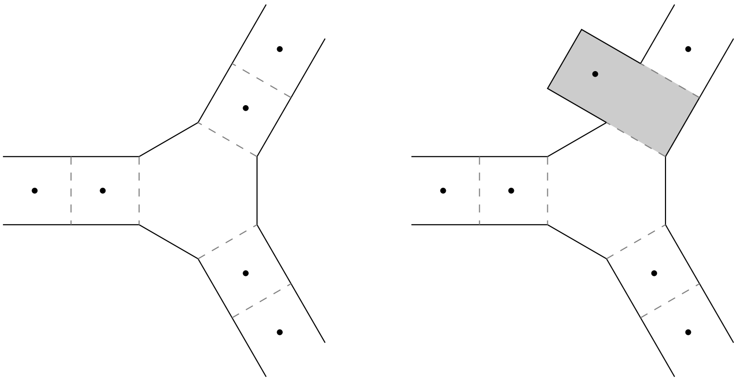


Figure 6: Pushing one puncture “off to the side”.

As with the Houghton groups, doing such a move in all ends simultaneously yields a homomorphism whose iterated images act trivially on arbitrarily large compact subsurfaces. This yields:

Theorem 5.7 *The braided Houghton group* Br H_n *is not co-Hopfian, and is in fact dis-co-Hopfian.*

These approaches are not immediately available for the surface Houghton group, as there is no inclusion of surfaces which skips over a single genus. In fact, in light of the results of [4], if the pure surface Houghton group were to fail to be co-Hopfian, then it must fail either by a non-twist-preserving homomorphism, or by a homomorphism which is not the restriction of a homomorphism on the level of pure mapping class groups. This leaves us with the following question:

Question 5.8 *Is the pure surface Houghton group $P\mathcal{H}_n$ co-Hopfian? If not, is it dis-co-Hopfian?*

References

[1] **M Abadie**, CAT(0) *cube complexes and asymptotically rigid mapping class groups*, master’s thesis, École Polytechnique Fédérale de Lausanne (2024)

[2] **J Aramayona**, **K-U Bux**, **H Kim**, **C J Leininger**, *Surface Houghton groups*, Math. Ann. 389 (2024) 4301–4318 MR

[3] **J Aramayona**, **G Domat**, **C J Leininger**, *Isomorphisms and commensurability of surface Houghton groups*, J. Group Theory 27 (2024) 1129–1141 MR

[4] **J Aramayona**, **C J Leininger**, **A McLeay**, *Big mapping class groups and the co-Hopfian property*, Michigan Math. J. 74 (2024) 253–281 MR

[5] **J Aramayona**, **P Patel**, **N G Vlamis**, *The first integral cohomology of pure mapping class groups*, Int. Math. Res. Not. 2020 (2020) 8973–8996 MR

[6] **R Bieri**, **W D Neumann**, **R Strebel**, *A geometric invariant of discrete groups*, Invent. Math. 90 (1987) 451–477 MR

[7] **R Bieri**, **B Renz**, *Valuations on free resolutions and higher geometric invariants of groups*, Comment. Math. Helv. 63 (1988) 464–497 MR

[8] **K S Brown**, *Finiteness properties of groups*, J. Pure Appl. Algebra 44 (1987) 45–75 MR

[9] **V Chepoi**, *Graphs of some* CAT(0) *complexes*, Adv. Appl. Math. 24 (2000) 125–179 MR

[10] **Y Cornulier**, *Gradings on Lie algebras, systolic growth, and cohopfian properties of nilpotent groups*, Bull. Soc. Math. France 144 (2016) 693–744 MR

[11] **F Degenhardt**, *Endlichkeitseigenschaften gewisser Gruppen von Zöpfen unendlicher Ordnung*, PhD thesis, Goethe-Universität zu Frankfurt am Main (2000)

[12] **L Funar**, *Braided Houghton groups as mapping class groups*, An. Ştiinţ. Univ. Al. I. Cuza Iaşi. Mat. 53 (2007) 229–240 MR

[13] **A Genevois**, **A Lonjou**, **C Urech**, *Asymptotically rigid mapping class groups*, *II*: *Strand diagrams and nonpositive curvature* (2021) arXiv 2110.06721 To appear in Trans. Amer. Math. Soc.

[14] **A Genevois**, **A Lonjou**, **C Urech**, *Asymptotically rigid mapping class groups, I*: *Finiteness properties of braided Thompson's and Houghton's groups*, Geom. Topol. 26 (2022) 1385–1434 MR

[15] **C H Houghton**, *The first cohomology of a group with permutation module coefficients*, Arch. Math. (Basel) 31 (1978) 254–258 MR

[16] **S R Lee**, *Geometry of Houghton's groups*, PhD thesis, University of Oklahoma (2012) Available at `https://www.proquest.com/docview/1044378650`

[17] **P Patel**, **N G Vlamis**, *Algebraic and topological properties of big mapping class groups*, Algebr. Geom. Topol. 18 (2018) 4109–4142 MR

[18] **B Renz**, *Geometrische Invarianten und Endlichkeitseigenschaften von Gruppen*, PhD thesis, Goethe-Universität zu Frankfurt am Main (1988) Available at `https://esb-dev.github.io/mat/diss.pdf`

[19] **M C B Zaremsky**, *On the* Σ*-invariants of generalized Thompson groups and Houghton groups*, Int. Math. Res. Not. 2017 (2017) 5861–5896 MR

[20] **M C B Zaremsky**, *The BNSR-invariants of the Houghton groups, concluded*, Proc. Edinb. Math. Soc. 63 (2020) 1–11 MR

Department of Mathematics, University of Oklahoma
Norman, OK, United States

Department of Mathematics, University of Oklahoma
Norman, OK, United States

`nmtorger@ou.edu`, `jeremy.west-1@ou.edu`

Received: 20 March 2024 Revised: 7 June 2024

Geometry & Topology Publications, an imprint of **mathematical sciences publishers** msp

Algebraic & Geometric Topology **25**:8 (2025) 4921–4947
DOI: 10.2140/agt.2025.25.4921
Published: 20 November 2025

Topological symmetry groups of the generalized Petersen graphs

Angelynn Álvarez
Erica Flapan
Mark Hunnell
John Hutchens
Emille Lawrence
Paul Lewis
Candice Price
Ruth Vanderpool

The topological symmetry group $\mathrm{TSG}(\Gamma)$ of an embedded graph Γ in S^3 is the subgroup of the automorphism group of the graph which is induced by homeomorphisms of (S^3, Γ). If we restrict to orientation-preserving homeomorphisms then we obtain the orientation-preserving topological symmetry group $\mathrm{TSG}_+(\Gamma)$. In this paper, we determine all groups that can be $\mathrm{TSG}(\Gamma)$ or $\mathrm{TSG}_+(\Gamma)$ for some embedding Γ of a generalized Petersen graph other than the exceptional graphs $P(12, 5)$ and $P(24, 5)$.

05C10, 57M15; 92E10

1 Introduction

The field of spatial graph theory was developed in part to classify the symmetries of flexible molecules [13] and in part as an extension of the study of the symmetries of knots and links [4]. However, in contrast with the symmetries of a knot or link, the symmetries of a spatial graph can be understood in terms of the automorphism group of its underlying graph. In particular, we have the following definitions.

Definition 1.1 Let γ be an abstract graph and Γ be an embedding of γ in S^3. The *topological symmetry group of* Γ, denoted by $\mathrm{TSG}(\Gamma)$, is the subgroup of the automorphism group $\mathrm{Aut}(\gamma)$ induced by homeomorphisms of the pair (S^3, Γ). If we only allow orientation-preserving homeomorphisms, we obtain the *orientation-preserving topological symmetry group*, $\mathrm{TSG}_+(\Gamma)$.

Definition 1.2 Let γ be an abstract graph and let G be a subgroup of $\mathrm{Aut}(\gamma)$ such that for some embedding Γ of γ in S^3, we have $G = \mathrm{TSG}(\Gamma)$ or $G = \mathrm{TSG}_+(\Gamma)$. Then we say that the group G is *realizable* or *positively realizable*, respectively, for γ.

Many previous results have been obtained about realizable and positively realizable groups for specific graphs (see, for example, [2; 6; 7; 8]). In this paper, we classify the groups that are realizable and positively

---

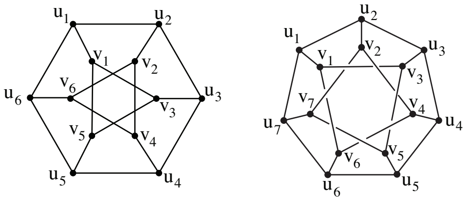


Figure 1: The graph $P(6,2)$ and an embedding of $P(7,2)$.

realizable for the family of *generalized Petersen graphs*. In particular, for $2k < n$, the generalized Petersen graph $P(n,k)$ is obtained from an n-gon with consecutive vertices $u_1, u_2, \dots, u_n$ and a star (possibly with more than one loop) with vertices $v_1, v_2, \dots, v_n$ and edges $\overline{v_i v_{i+k}}$, by adding an edge $\overline{u_i v_i}$ for each $i \le n$ (see the graph $P(6,2)$ and an embedding of $P(7,2)$ in Figure 1). We call the edges $\overline{u_i u_{i+1}}$ *outer edges*, the edges $\overline{v_i v_{i+k}}$ *inner edges*, and the edges $\overline{u_i v_i}$ *spokes*. If $k^2 \equiv \pm 1 \pmod n$, then k and n are relatively prime and hence the star formed by the inner edges of $P(n,k)$ is a single cycle.

Let $B(n,k)$ denote the subgroup of $\operatorname{Aut}(P(n,k))$ which leaves the set of spokes $\{\overline{u_i v_i} \mid i \le n\}$ invariant (either preserving or interchanging the inner and outer edges). Frucht, Graver, and Watkins [11] proved that all but the seven exceptional pairs $(4,1)$, $(5,2)$, $(8,3)$, $(10,2)$, $(10,3)$, $(12,5)$, $(24,5)$ have $\operatorname{Aut}(P(n,k)) = B(n,k)$. Furthermore they showed that

$$B(n,k) = \begin{cases} D_n, & k^2 \not\equiv \pm 1 \pmod n, \\ D_n \rtimes \mathbb{Z}_2, & k^2 \equiv 1 \pmod n, \\ \mathbb{Z}_n \rtimes \mathbb{Z}_4, & k^2 \equiv -1 \pmod n. \end{cases}$$

We classify the realizable and positively realizable subgroups of $\operatorname{Aut}(P(n,k))$ for all nonexceptional pairs (n,k) and for the exceptional pairs $(4,1)$, $(8,3)$, $(10,2)$, $(10,3)$. For the exceptional pair $(5,2)$, the graph $P(5,2)$ is the *Petersen graph* whose topological symmetry groups were determined by Chambers, Flapan, Heath, Lawrence, Thatcher, and Vanderpool [3]. Realizability and positive realizability for $P(12,5)$ and $P(24,5)$ will be considered in a subsequent paper. The exceptional graphs $P(n,k)$ for $n \le 10$ are illustrated in Figure 2.

Our arguments make use of the above theorem of Frucht et al, as well as the theorems listed below. However, for a large number of groups, we prove realizability or positive realizability by constructing intricate embeddings which we show have the required group of symmetries and not a larger group (the

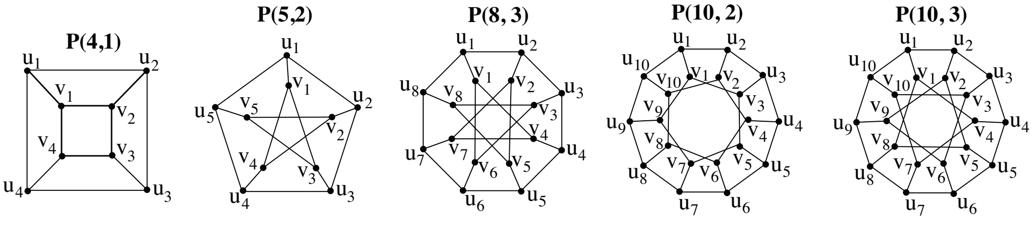


Figure 2: The exceptional graphs $P(4,1)$, $P(5,2)$, $P(8,3)$, $P(10,2)$, and $P(10,3)$.

latter is often the difficult part). These embeddings are described in the text, and augmented by detailed illustrations. Some readers will find these constructions to be the most interesting part of the paper, since they can likely be generalized to other graphs.

If $G = \mathrm{TSG}_+(\Gamma)$ for some embedding Γ of $P(n,k)$, then we can add the same chiral invertible knot to every edge of Γ to get an embedding Γ' with $G = \mathrm{TSG}(\Gamma')$. This means that if we can find an embedding of $P(n,k)$ such that $G = \mathrm{TSG}_+(\Gamma)$, then G is both positively realizable and realizable for $P(n,k)$. Note however, that G may be realizable and not be positively realizable. Also, by putting a distinct knot on each edge of Γ we obtain an embedding Γ'' such that $\mathrm{TSG}(\Gamma'') = \mathrm{TSG}_+(\Gamma'') = \langle e \rangle$. Thus the trivial group is realizable and positively realizable for every $P(n,k)$. Hence we focus on nontrivial groups.

Since $P(n,k)$ is 3-connected, we can use the following theorems.

Theorem 1.3 (Automorphism Rigidity Theorem [5]) *Let G be a 3-connected graph. Suppose that an automorphism σ of G is realizable by a homeomorphism h of some embedding of G in S^3. Then σ is realizable by a homeomorphism f of finite order which is orientation reversing if and only if h is orientation reversing.*

The above theorem allows us to assume realizable automorphisms are induced by finite-order homeomorphisms of S^3, and then use the following simplified version of a theorem of P A Smith [14]. We let $\mathrm{fix}(h)$ denote the fixed-point set of h.

Theorem 1.4 (Smith theory) *Let h be a finite-order homeomorphism of S^3. If h is orientation preserving then $\mathrm{fix}(h)$ is either the empty set or S^1, and if h is orientation reversing then $\mathrm{fix}(h)$ is either two points or S^2.*

We now list several other results that will be used. Note that SO(4) is the group of orientation-preserving isometries of S^3 and SO(3) is the group of orientation-preserving isometries of S^2. The Group Rigidity Theorem below follows from Proposition 3 of [9] together with the geometrization theorem [12].

Theorem 1.5 (Group Rigidity Theorem) *Let γ be a 3-connected graph and $G \leq \mathrm{TSG}_+(\Lambda)$ for some embedding Λ of γ in S^3. Then there is an embedding Γ of γ in S^3 such that $G \leq \mathrm{TSG}_+(\Gamma)$ and $\mathrm{TSG}_+(\Gamma)$ is induced by an isomorphic finite subgroup of* SO(4).

Theorem 1.6 (Involution Theorem [10]) *Let $G \leq \mathrm{SO}(4)$ such that for every involution $g \in G$, we have $\mathrm{fix}(g) \cong S^1$ and no $h \in G$ with $g \neq h$ has $\mathrm{fix}(h) = \mathrm{fix}(g)$. Then G is a subgroup of $D_m \times D_m$ for some odd m or is a finite subgroup of* SO(3).

Theorem 1.7 (Subgroup Theorem [7]) *Let Γ be a 3-connected graph embedded in S^3. Suppose Γ contains an edge whose vertices are not both fixed by a nontrivial element of $\mathrm{TSG}_+(\Gamma)$. Then for every $H \leq \mathrm{TSG}_+(\Gamma)$ there is an embedding Γ' of Γ in S^3 with $H = \mathrm{TSG}_+(\Gamma')$.*

Lemma 1.8 (Edge Embedding Lemma [7]) *Let G be a finite group of diffeomorphisms of S^3 and let γ be a graph whose vertices are embedded in S^3 as a set W such that G induces a faithful action on W (ie, no nontrivial element of G induces the identity on W). Let Y denote the union of the fixed-point sets of all of the nontrivial elements of G. Suppose that the vertices in W satisfy the following*:

(1) *If a pair of adjacent vertices is pointwise fixed by nontrivial elements $h, g \in G$, then $\mathrm{fix}(h) = \mathrm{fix}(g)$.*

(2) *No pair of adjacent vertices is interchanged by an element of G.*

(3) *Any pair of adjacent vertices that is pointwise fixed by a nontrivial $g \in G$ bounds an arc in $\mathrm{fix}(g)$ whose interior is disjoint from $W \cup (Y - \mathrm{fix}(g))$.*

(4) *Every pair of adjacent vertices which are not both fixed by a nontrivial element of G is contained in a single component of $S^3 - Y$.*

Then there is an embedding Γ of the graph γ in S^3 such that Γ is setwise invariant under G.

We also use the website https://people.maths.bris.ac.uk/~matyd/GroupNames/ and the software Sage to identify the isomorphism classes of automorphism groups and their subgroups.

2 Realizability of D_n and its subgroups for all $P(n,k)$

Theorem 2.1 *For every pair (n,k), the dihedral group D_n and all of its subgroups are positively realizable and hence realizable for $P(n,k)$.*

Proof We construct an embedding of $P(n,k)$ as follows. Let $0 < r < R$. We begin with planar circles C and c, centered at the origin, of radius R and r, respectively. Let u_1 and v_1 be the points of intersection between a planar ray from the origin and C and c, respectively. Next let e_1 denote the straight line segment between u_1 and v_1. Then e_1 lies on the ray. Let f denote a rotation of S^3 by $\frac{2\pi}{n}$ about an axis through the origin which is perpendicular to the plane of the projection. Then for each i let $u_i = f^i(u_1)$, $v_i = f^i(v_1)$, $e_i = f^i(e_1)$. Thus we have embedded the spokes $\overline{u_i v_i}$ as the e_i. Next we embed the outer edges $\overline{u_i u_{i+1}}$ as straight line segments. We will embed the inner edges below.

Let g denote a rotation by π about a planar axis that contains the spoke $\overline{u_n v_n}$ if n is odd and contains the two opposite spokes $\overline{u_n v_n}$ and $\overline{u_{\frac{n}{2}} v_{\frac{n}{2}}}$ if n is even. Then the isometry group $D_n = \langle f, g \rangle$ leaves the n-gon of outer edges and the set of spokes invariant, inducing an isomorphic group on the embedded vertices of $P(n,k)$. Now we embed the inner edges $\overline{v_i v_{i+k}}$ so that the collection of edges $\{\overline{v_i v_{i+k}} \mid i \leq n\}$ is pairwise disjoint and setwise invariant under D_n. We can do this, for example, by making the crossings along each inner edge alternate, as illustrated for the embedding of $P(7,2)$ on the right of Figure 1. Let Γ' denote this embedding of $P(n,k)$. Thus Γ' is invariant under $\langle f, g \rangle$, and hence $D_n \leq \mathrm{TSG}_+(\Gamma')$.

We obtain the embedding Γ from Γ' by adding the invertible knot 4_1 to each outer edge. Then Γ is invariant under $\langle f, g \rangle$, and hence $D_n \leq \mathrm{TSG}_+(\Gamma)$. Let L denote the n-gon of outer edges of Γ'. Then L is the only

n-gon containing n copies of the knot 4_1, and hence any element of $\mathrm{TSG}_+(\Gamma)$ must take L to itself. Since no nontrivial automorphism of $P(n,k)$ fixes every u_i, this implies that $D_n \le \mathrm{TSG}_+(\Gamma) \le \mathrm{TSG}_+(L)$. On the other hand, because the automorphism group of an n-gon is D_n, we have $\mathrm{TSG}_+(L) \le D_n$, and hence $D_n = \mathrm{TSG}_+(\Gamma)$.

No outer edge can be pointwise fixed by a nontrivial element of $\mathrm{TSG}_+(\Gamma)$ since that would pointwise fix L and hence all of Γ. Thus by Theorem 1.7, every subgroup of D_n is positively realizable for $P(n,k)$. $\square$

By [11] we know that for $k^2 \not\equiv \pm 1 \pmod n$ where (n,k) is nonexceptional, $\mathrm{Aut}(P(n,k)) = B(n,k) = D_n$. Thus we have the following.

Corollary 2.2 *Let $k^2 \not\equiv \pm 1 \pmod n$ where (n,k) is nonexceptional. Then* $\mathrm{Aut}(P(n,k))$ *and all of its subgroups are positively realizable and hence realizable for $P(n,k)$.*

3 The case $k^2 \equiv 1 \pmod n$

We use the following embedding Λ of $P(n,k)$ in this and the next section. Let U and V denote the cores of complementary isometric solid tori in S^3. Hence U and V are geodesic circles, and for every point on one of these cores, its antipodal point is on the same core. Now, for every $u \in U$ and $v \in V$, there is a unique shortest geodesic e joining u and v, and the length of e is less than π. Since U and V are geodesic circles, it follows that the interior of any such e must be disjoint from $U \cup V$.

Suppose that $k^2 \equiv \pm 1 \pmod n$. Then n and k are relatively prime, and hence the inner edges of $P(n,k)$ form a single loop. The embedding Λ is obtained as follows. We let e_1 be a geodesic of minimal length between U and V, and let u_1 and v_1 be its endpoints on U and V, respectively. Let f be a glide rotation which rotates U by $\frac{2\pi}{n}$ while rotating V by $k\left(\frac{2\pi}{n}\right)$. Then for each i, define $e_{i+1} = f^i(e_1)$, and define the endpoint of e_{i+1} on U to be u_{i+1} and the endpoint on V to be v_{i+1}. Thus $f(e_i) = e_{i+1}$, $f(u_i) = u_{i+1}$, and $f(v_i) = v_{i+1}$.

Since f rotates U by $\frac{2\pi}{n}$, the points u_i and u_{i+1} are consecutive on U. Also, since f rotates V by $k\left(\frac{2\pi}{n}\right)$, we know that f^k rotates V by $k^2\left(\frac{2\pi}{n}\right) \equiv \frac{2\pi}{n} \pmod n$. Since $f^k(v_i) = v_{i+k}$, this implies that v_i and v_{i+k} are consecutive on V. Thus we define the edges $\overline{u_i u_{i+1}}$ and $\overline{v_i v_{i+k}}$ to be minimal arcs on U and V, respectively. It follows that the interiors of the e_i are disjoint from U and V. Furthermore, if we consider the open book decomposition of S^3 whose spine is U, then f rotates the pages of this decomposition around U and each e_i is on a distinct page. Hence the e_i are pairwise disjoint. This gives us an embedding Λ of $P(n,k)$.

Now we orient U and V so that $\overrightarrow{u_i u_{i+1}}$ and $\overrightarrow{v_i v_{i+k}}$ are oriented positively.

Theorem 3.1 *Let $k^2 \equiv 1 \pmod n$. Then $B(n,k) = D_n \rtimes \mathbb{Z}_2$ and all of its subgroups are positively realizable and hence realizable for $P(n,k)$.*

Proof We start with the embedding Λ of $P(n,k)$ described above and we add a 4_1 knot to each edge of U and V to get an embedding Λ' of $P(n,k)$. Recall that f is a glide rotation which rotates U by $\frac{2\pi}{n}$ while rotating V by $k\bigl(\frac{2\pi}{n}\bigr)$ both in the positive direction. Let g be a rotation of S^3 by π around an axis that contains $e_1 = \overline{u_1 v_1}$. Then g takes U and V to themselves, reversing their orientations. Finally, let h be an order-2 rotation of S^3 interchanging the positively oriented U with the positively oriented V, taking u_n to v_n.

In order to show that f, g, and h take Λ' to itself, we need to know that they take edges to edges. First, by definition we know that $f(e_i) = e_{i+1}$. Also, for every i, we have $g(v_{1-ki}) = v_{1+ki}$, and hence $g(v_{1-i}) = g(v_{1-k^2 i}) = v_{1+k^2 i} = v_{1+i}$. Since we also have $g(u_{1-i}) = u_{1+i}$ and we know that the e_i are minimal length geodesics, it follows that $g(e_{1-i}) = e_{i+1}$. Finally, since h interchanges u_n and v_n preserving orientation, we have $h(u_i) = v_{ki}$, and hence $h(u_{ki}) = v_{k^2 i} = v_i$. Thus for every i, we have $h(e_i) = e_{ki}$. It follows that f, g, and h take Λ' to itself.

Now f induces a rotation of order n on U and V, g turns U and V over, and h is an involution interchanging U and V. Thus $\langle f, g, h\rangle$ induces $D_n \rtimes \mathbb{Z}_2$ on Λ', and hence $D_n \rtimes \mathbb{Z}_2 \le \mathrm{TSG}_+(\Lambda')$. However, since every element of $\mathrm{TSG}_+(\Lambda)$ takes the set of edges e_i to themselves, it follows that $\mathrm{TSG}_+(\Lambda') \le B(n,k)$. Now by [11], $B(n,k) = D_n \rtimes \mathbb{Z}_2$ and hence $\mathrm{TSG}_+(\Lambda') = D_n \rtimes \mathbb{Z}_2$.

Finally, no edge of U or V is pointwise fixed by any nontrivial element of $\mathrm{TSG}_+(\Lambda')$. Thus Λ' satisfies the hypothesis of Theorem 1.7, and hence all subgroups of $D_n \rtimes \mathbb{Z}_2$ are positively realizable for $P(n,k)$. □

If (n,k) is nonexceptional then $\mathrm{Aut}(P(n,k)) = B(n,k)$, implying the following.

Corollary 3.2 *Let $k^2 \equiv 1 \pmod n$ where (n,k) is nonexceptional. Then $\mathrm{Aut}(P(n,k))$ and all of its subgroups are positively realizable and hence realizable for $P(n,k)$.*

4 The case $k^2 \equiv -1 \pmod n$

Since $k^2 \equiv -1 \pmod n$, we know that k and n are relatively prime, and hence the inner edges of $P(n,k)$ form a single cycle. Furthermore, if n were divisible by 4, then k would be odd, and hence $k^2 = (2m+1)^2 = 4m^2 + 4m + 1$ for some m. But this implies that $4m^2 + 4m + 1 \equiv -1 \pmod n$, and thus $4m^2 + 4m + 2 = nr$ for some r, which is impossible since n is divisible by 4. Therefore, n cannot be divisible by 4.

According to [11], $B(n,k) = \mathbb{Z}_n \rtimes \mathbb{Z}_4 = \langle \rho, \alpha \mid \rho^n = \alpha^4 = \mathrm{id}, \alpha\rho\alpha^{-1} = \rho^k\rangle$, where $\alpha(u_i) = v_{ki}$, $\alpha(v_i) = u_{ki}$, $\rho(u_i) = u_{i+1}$, and $\rho(v_i) = v_{i+1}$. Observe that every element of $B(n,k)$ can be expressed as $\rho^m \alpha^r$ for some $0 \le m < n$ and $0 \le r < 4$.

Lemma 4.1 *Let $k^2 \equiv -1 \pmod n$. Then an element of $B(n,k)$ has order 4 if and only if it can be expressed as $\rho^m \alpha^{\pm 1}$. Furthermore, if n is odd, then no order-4 element of $B(n,k)$ is positively realizable.*

Proof Since α interchanges the inner and outer cycles of $P(n,k)$ while ρ induces an order-n rotation of both cycles, for any $m < n$, ρ^m rotates both cycles while $\rho^m\alpha^{\pm 1}$ interchanges them. Furthermore, $\alpha^2(u_i) = \alpha(v_{ki}) = u_{k^2 i} = u_{-i}$ and $\alpha^2(v_i) = v_{-i}$. Thus α^2 is an involution which turns over both the inner and outer cycles, and hence so is any $\rho^m\alpha^2$. It follows that every element of the form $\rho^m\alpha^{\pm 1}$ has order 4. Also, since n is not divisible by 4, $D_n = \langle \rho, \alpha^2 \rangle$ has no elements of order 4. Thus an element of $B(n,k)$ has order 4 if and only if it can be expressed as $\rho^m\alpha^{\pm 1}$.

Suppose that an order-4 element β of $B(n,k)$ is induced by an orientation-preserving homeomorphism of some embedding of $P(n,k)$ in S^3. By the above paragraph, β has the form $\rho^m\alpha^{\pm 1}$. Now by Theorem 1.3, there is an embedding Γ' of $P(n,k)$ and a finite-order orientation-preserving homeomorphism h of (S^3, Γ') which also induces β. Now h^2 turns over the inner and outer cycles of Γ, and hence fixes two points on each of these cycles. Thus by Smith theory, $\mathrm{fix}(h^2) = S^1$. Since h is orientation preserving, if h is not fixed-point free, then we also have $\mathrm{fix}(h) = S^1$. Since $\mathrm{fix}(h) \subseteq \mathrm{fix}(h^2)$, we have $\mathrm{fix}(h) = \mathrm{fix}(h^2)$. This implies that h fixes two points on each of the cycles, which is impossible since $\rho^m\alpha^{\pm 1}$ interchanges the two cycles. Thus h must be fixed-point free.

Suppose that n is odd. Since $\beta \in B(n,k)$, h must leave at least one spoke e_i setwise invariant. But this implies that h fixes the midpoint of e_i, which contradicts the above paragraph. Thus no order-4 element of $B(n,k)$ is positively realizable. □

Using the above lemma we see as follows that D_4 is not a subgroup of $B(n,k)$. In particular, if D_4 were a subgroup, then by the lemma it would be generated by an element $\rho^m\alpha$ of order 4 and an element $\rho^r\alpha^2$ of order 2. But $\rho^m\alpha\rho^r\alpha^2 = \rho^{m+nk}\alpha^3$ has order 4, whereas in D_4 the product of a pair of generators has order 2.

Theorem 4.2 *Let $k^2 \equiv -1 \pmod{n}$ and $H \le B(n,k) = \mathbb{Z}_n \rtimes \mathbb{Z}_4$. If n is odd, then H is positively realizable for $P(n,k)$ if and only if $H \le D_n$. If n is even, then H is positively realizable for $P(n,k)$ if and only if either $H \le D_n$ or $H = \mathbb{Z}_4$.*

Proof By Theorem 2.1, D_n and all of its subgroups are positively realizable for $P(n,k)$. If H is not contained in D_n, then H must contain an element of order 4. When n is odd, no order-4 element of $B(n,k)$ is positively realizable by Lemma 4.1, and hence H is not positively realizable.

Thus we assume that n is even. Since n is not divisible by 4, we have $n \equiv 2 \pmod{4}$. In order to construct an embedding of $P(n,k)$ which positively realizes $\mathbb{Z}_4 \le B(n,k)$, we consider the glide rotation h which rotates a standardly embedded solid torus meridionally by $\frac{\pi}{2}$ while rotating it longitudinally by π.

Let U be a meridian of the solid torus with $n-2$ evenly spaced vertices $u_1, \dots, u_{\frac{n}{2}-1}, u_{\frac{n}{2}+1}, \dots, u_{n-1}$. Let $V = h(U)$ with vertices $v_k, v_{2k}, \dots, v_{\frac{n}{2}-k}, v_{\frac{n}{2}+k}, \dots, v_{n-k}$ such that each v_j is $h(u_i)$ for some i. Let D_U and D_V denote disjoint meridional disks bounded by U and V, respectively, such that $h(D_U) = D_V$. Let $C = \mathrm{fix}(h^2)$, and let x denote the midpoint of an arc A of $C - (D_U \cup D_V)$. Let u_n be a vertex on the

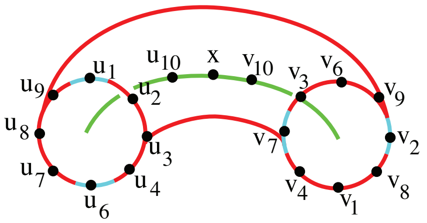


Figure 3: The embedded vertices of $P(10,3)$, except for u_5 and v_5 which are on the (green) core in the front of the solid torus.

arc of $A-\{x\}$ with one endpoint on D_U, and let v_n be a vertex on the arc of $A-\{x\}$ with one endpoint on D_V. Finally, let $v_{\frac{n}{2}} = h(u_n)$ and $u_{\frac{n}{2}} = h(v_n)$. This gives us an embedded set of vertices W.

For example, Figure 3 illustrates the embedded vertices of $P(10,3)$ except for v_5 and u_5, which are in the front of the solid torus. The core is illustrated in green, and h takes the pair of blue arcs on U to the pair of blue arcs on V.

Let $G = \langle h \rangle = \mathbb{Z}_4$. Then G induces a faithful action $H \le B(n,k)$ of W such that no pair of adjacent vertices are fixed by a nontrivial element of H. The pairs of adjacent vertices u_n and v_n and $u_{\frac{n}{2}}$ and $v_{\frac{n}{2}}$ are the only ones which are fixed by h^2. But they are not fixed by any other nontrivial element of G. Also these pairs each bound an arc in C which is disjoint from the other vertices. Hence we can apply Lemma 1.8 to embed the edges of $P(n,k)$ such that the resulting embedded graph is setwise invariant under G.

Next assume that a positively realizable group $H \le B(n,k)$ contains $\mathbb{Z}_4$ as a proper subgroup. Then by Theorem 1.5, for some embedding Γ of $P(n,k)$ in S^3, the group H is induced on Γ by an isomorphic group $H' \le \mathrm{SO}(4)$. Every involution in H has the form $\rho^m\alpha^2$ and hence turns over both the inner and the outer cycle of Γ. Thus every involution in H' fixes two points on each of these cycles. Furthermore, every nontrivial orientation-preserving isometry of (S^3,Γ) which fixes two points on each cycle must be an involution and if two such isometries fix the same points on these cycles then the isometries are equal.

Thus we can apply Theorem 1.6 to conclude that H' is either a subgroup of $D_m \times D_m$ for some odd m or a finite subgroup of $\mathrm{SO}(3)$. However, since $H \le B(n,k)$ and contains $\mathbb{Z}_4$ as a proper subgroup, H has the form $\mathbb{Z}_r \rtimes \mathbb{Z}_4$. But this is impossible since $H' \cong H$. It follows that no subgroup of $B(n,k)$ containing $\mathbb{Z}_4$ as a proper subgroup is positively realizable for $P(n,k)$. □

Theorem 4.3 *Let* $k^2 \equiv -1 \pmod{n}$. *Then* $B(n,k) = \mathbb{Z}_n \rtimes \mathbb{Z}_4$ *and all of its subgroups are realizable for* $P(n,k)$.

Proof Let Λ denote the embedding of the graph $P(n,k)$ from the beginning of Section 3. Let h be an order-2 rotation of S^3 interchanging the positively oriented U with the positively oriented V such that it interchanges u_i and v_{ki}. Now $h(u_{ki}) = v_{k^2i} = v_{-i}$, and hence $h(v_i) = u_{-ki}$. Also, let R be a reflection through a sphere containing the circle V which leaves U setwise invariant, fixing u_n and its antipodal point on U (which is a vertex if n is even). Thus for every i, we have $R(u_i) = u_{-i}$ and $R(v_i) = v_i$. While both h and R take edges contained in U and V to other such edges, neither h nor R takes edges of the form $\overline{u_i v_i}$ to other edges. Thus we are instead interested in the orientation reversing isometry Rh. Observe that $Rh(u_i) = R(v_{ki}) = v_{ki}$ and $Rh(v_i) = R(u_{-ki}) = u_{ki}$. Thus for each spoke e_i, we have $Rh(e_i) = e_{ki}$. It follows that Rh takes Λ to itself interchanging U and V and $(Rh)^2$ turns over U and V, fixing u_n and v_n. In particular, Rh induces α on Λ.

Next, let f be a glide rotation which rotates U by $\frac{2\pi}{n}$ in the positive direction while rotating V by $\frac{2k\pi}{n}$ in the negative direction, so that $f(u_i) = u_{i+1}$ and $f(v_i) = v_{i-k^2} = v_{i+1}$. Thus $f(e_i) = e_{i+1}$. It follows that f takes Λ to itself inducing ρ. Hence $B(n,k) = \mathbb{Z}_n \rtimes \mathbb{Z}_4 = \langle \alpha, \rho \rangle \leq \mathrm{TSG}(\Lambda)$. Now we add the knot 4_1 to each of the spokes e_i of Λ to get an embedding Λ' such that an automorphism is in $B(n,k)$ if and only if it is in $\mathrm{TSG}(\Lambda')$. Thus $\mathbb{Z}_n \rtimes \mathbb{Z}_4$ is realizable for $P(n,k)$.

We show as follows that all proper subgroups of $\mathbb{Z}_n \rtimes \mathbb{Z}_4$ are also realizable. First observe that every subgroup of $B(n,k)$ with no element of order 4 is a subgroup of D_n. Thus by Theorem 2.1 it is positively realizable, and hence it is also realizable.

Let G denote a proper subgroup of $B(n,k)$ which contains an element of order 4. Since n is not divisible by 4, G must be isomorphic to $\mathbb{Z}_r \rtimes \mathbb{Z}_4$, for some $r, m > 1$ with $rm = n$. Now starting with Λ', we add the achiral invertible knot 6_3 to the edge $\overline{u_n u_1}$ and all of the edges in its orbit under $\langle \rho^m, \alpha \rangle$ to obtain an embedding Λ''. Since α interchanges U and V, and α^2 flips U and V over fixing u_n and v_n, Λ'' contains the 6_3 knot on r pairs of adjacent edges on U and r pairs of adjacent edges on V, but not on any other edges. Thus $\mathbb{Z}_r \rtimes \mathbb{Z}_4 = \langle \rho^m, \alpha \rangle \leq \mathrm{TSG}(\Lambda'') \leq \mathbb{Z}_r \rtimes \mathbb{Z}_4$. Hence every subgroup of $\mathbb{Z}_n \rtimes \mathbb{Z}_4$ is realizable. □

If (n,k) is nonexceptional then $\mathrm{Aut}(P(n,k)) = B(n,k)$, implying the following.

Corollary 4.4 *Let $k^2 \equiv -1 \pmod n$ and suppose that (n,k) is nonexceptional. Then $\mathrm{Aut}(P(n,k))$ and all of its subgroups are realizable for $P(n,k)$. If n is odd then the only groups which are positively realizable for $P(n,k)$ are D_n and its subgroups, and if n is even, then the only groups which are positively realizable for $P(n,k)$ are $\mathbb{Z}_4$ and D_n and its subgroups.*

5 The exceptional case $P(4,1)$

We know from [15] that $\mathrm{Aut}((P(4,1)) = S_4 \times \mathbb{Z}_2$.

Proposition 5.1 *Let Γ be the embedding of $P(4,1)$ in S^3 as the 1-skeleton of a cube. Then*

$$\mathrm{TSG}(\Gamma) = \mathrm{TSG}_+(\Gamma) = S_4 \times \mathbb{Z}_2.$$

Proof The group $\mathrm{Aut}((P(4,1)) = S_4 \times \mathbb{Z}_2$ is induced on Γ by the rotations and reflections of a cube. To see that $S_4 \times \mathbb{Z}_2$ is also positively realized by Γ, we flatten out the cube so it is a small square inside of a big square with edges between them. Then the automorphisms induced by reflections of the cube can also be induced by turning the flattened cube over. It follows that $\mathrm{TSG}(\Gamma) = \mathrm{TSG}_+(\Gamma) = S_4 \times \mathbb{Z}_2$. □

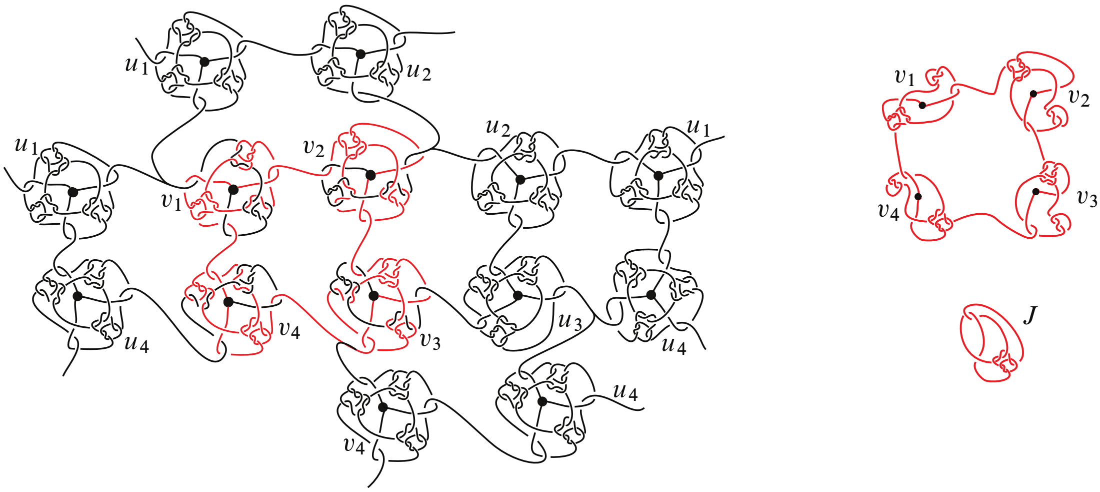


Figure 4: The embedding Γ of $P(4,1)$ with $\mathrm{TSG}_+(\Gamma) = S_4$ is obtained from the left image by identifying vertices with the same labels and pairs of adjacent branched edges. On the right, we see that $\overline{v_1v_2v_3v_4}$ is the connected sum of four copies of J and four trefoils.

The nontrivial automorphisms of $P(4,1)$ which pointwise fix an edge interchange two pairs of nonadjacent vertices and pointwise fix two nonadjacent edges.

Proposition 5.2 *S_4 and all of its subgroups are positively realizable and hence realizable for $P(4,1)$.*

Proof Let Γ be an embedding of $P(4,1)$ as the skeleton of a cube with tangling around the vertices as indicated in the unfolded projection in Figure 4. On the right, we illustrate the knot $\overline{v_1v_2v_3v_4}$.

Observe that the rotations of a solid cube leave Γ invariant, inducing the automorphisms

$$\rho = (u_1u_2u_3u_4)(v_1v_2v_3v_4) \quad \text{and} \quad \sigma = (u_1v_2v_4)(u_2v_3u_4),$$

which generate S_4. Thus

$$S_4 = \langle \rho, \sigma \rangle \leq \mathrm{TSG}_+(\Gamma).$$

Let $\delta = (u_1u_3)(v_1v_3)$ and note that $\mathrm{Aut}((P(4,1)) = S_4 \times \mathbb{Z}_2 = \langle \rho, \sigma, \delta \rangle$ has $S_4 = \langle \rho, \sigma \rangle$ as a maximal subgroup. We show below that $\delta \notin \mathrm{TSG}_+(\Gamma)$.

Suppose that δ is induced on Γ by a homeomorphism h of S^3. Because Γ contains a 3_1 knot but not its mirror image, h must be orientation preserving. Now h induces the automorphism (v_1v_3) of the knot $\overline{v_1v_2v_3v_4}$, which is a connected sum of four copies of J and four 3_1 knots. It follows that h interchanges two copies of J while flipping over the other two copies of J. But by applying the machinery of Bonahon and Siebenmann for algebraic knots [1] to J, we see that J is noninvertible. Thus h cannot exist. Hence $\delta \notin \mathrm{TSG}_+(\Gamma)$, and therefore $\mathrm{TSG}_+(\Gamma) = S_4$.

Since none of the nontrivial elements of $\langle \rho, \sigma \rangle = \mathrm{TSG}_+(\Gamma) = S_4$ pointwise fix an edge, we can use Theorem 1.7 to positively realize all of the subgroups of S_4. □

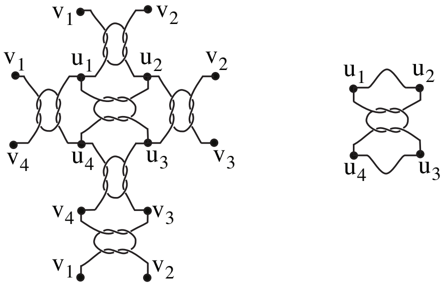


Figure 5: The embedding Γ of $P(4,1)$ with $\mathrm{TSG}_+(\Gamma) = A_4 \times \mathbb{Z}_2$ is obtained from the projection on the left by identifying vertices with the same labels. On the right, we have the projection on a single face.

Proposition 5.3 *$A_4 \times \mathbb{Z}_2$ is positively realizable and hence realizable for $P(4,1)$.*

Proof Let Γ be the embedding of $P(4,1)$ whose unfolded projection on the faces of a cube is illustrated on the left in Figure 5, with the projection on a single face illustrated on the right.

Let $\alpha = (u_2 u_4 v_1)(u_3 v_4 v_2)$, $\beta = (u_1 u_2 u_3 u_4)(v_1 v_2 v_3 v_4)$, and $\delta = (u_1 v_1)(u_2 v_2)(u_3 v_3)(u_4 v_4)$. Then $\langle \alpha, \beta, \delta \rangle = S_4 \times \mathbb{Z}_2$, and has $\langle \alpha, \beta^2, \delta \rangle = A_4 \times \mathbb{Z}_2$ as a maximal subgroup. Observe that α is induced on Γ by a rotation by $\frac{2\pi}{3}$ around an axis through vertices u_1 and v_3, and δ is induced on Γ by a rotation around an equator of the cube that is parallel to $\overline{u_1 u_2 u_3 u_4}$ and $\overline{v_1 v_2 v_3 v_4}$. Also, β^2 is induced by a rotation by π around an axis through center of the faces containing $\overline{u_1 u_2 u_3 u_4}$ and $\overline{v_1 v_2 v_3 v_4}$. However, the knot $\overline{u_1 u_2 u_3 u_4}$ is the connected sum of two trefoils and hence does not have an order-4 symmetry (see the right side of Figure 5). Thus β is not induced by a homeomorphism taking Γ to itself. It follows that $\mathrm{TSG}(\Gamma) = \mathrm{TSG}_+(\Gamma) = A_4 \times \mathbb{Z}_2$. □

Proposition 5.4 *$D_4 \times \mathbb{Z}_2$, $D_3 \times \mathbb{Z}_2$, and all of their subgroups are positively realizable and hence realizable for $P(4,1)$.*

Proof In the proof of Theorem 3.1, we saw that $D_n \rtimes \mathbb{Z}_2 = \langle \rho, \delta, \alpha \rangle$, where ρ, δ, and α are defined there. Furthermore, $\langle \rho, \delta \rangle = D_n$, $\langle \alpha \rangle = \mathbb{Z}_2$, and α commutes with δ. When $k = 1$, α also commutes with ρ. Thus for $P(4,1)$, we have $\langle \rho, \delta, \alpha \rangle = D_4 \times \mathbb{Z}_2$. It follows that $D_4 \times \mathbb{Z}_2$ and all its subgroups are positively realizable for $P(4,1)$.

To show that $D_3 \times \mathbb{Z}_2$ is positively realizable, we let Γ be the skeleton of a cube with identical trefoil knots on the six edges containing u_1 or v_3. A rotation by $\frac{2\pi}{3}$ around an axis through u_1 and v_3 induces the automorphism $\alpha = (u_2 v_1 u_4)(v_2 v_4 u_3)$, a rotation by π around an axis containing $\overline{u_1 v_1}$ and $\overline{u_3 v_3}$ induces $\beta = (u_4 u_2)(v_4 v_2)$, and a rotation by π around an axis through the middle of the edges $\overline{u_4 v_4}$ and $\overline{u_2 v_2}$ induces the automorphism $\delta = (u_4 v_4)(u_2 v_2)(u_1 v_3)(v_1 u_3)$. Now $\langle \alpha, \beta, \delta \rangle = D_3 \times \mathbb{Z}_2$, which is a maximal subgroup of $S_4 \times \mathbb{Z}_2$.

No homeomorphism can induce $(u_1u_2u_3u_4)(v_1v_2v_3v_4)$, since such a homeomorphism would send knotted edges to unknotted edges. Thus we have $\mathrm{TSG}_+(\Gamma) = D_3 \times \mathbb{Z}_2$. Furthermore, the edge $\overline{u_4v_4}$ is not pointwise fixed by any nontrivial element of $\mathrm{TSG}_+(\Gamma)$, and hence by Theorem 1.7 every subgroup of $D_3 \times \mathbb{Z}_2$ is positively realizable. ☐

In summary, we have proved the following.

Theorem 5.5 $\mathrm{Aut}(P(4,1)) = S_4 \times \mathbb{Z}_2$ *and all of its subgroups are positively realizable and hence realizable for* $P(4,1)$.

6 The exceptional case $P(8,3)$

It follows from [15] that $\mathrm{Aut}(P(8,3)) = \mathrm{GL}(2,3) \rtimes \mathbb{Z}_2 = \langle \mu, \beta, \gamma \rangle$, where

$$\begin{aligned}\mu &= (u_1u_7v_8)(u_2v_7v_5)(u_3v_4u_5)(u_6v_3v_1),\\ \beta &= (u_1u_7)(u_2u_6)(u_3u_5)(v_1v_7)(v_2v_6)(v_3v_5),\\ \gamma &= (u_1u_2u_3u_4u_5u_6u_7u_8)(v_1v_2v_3v_4v_5v_6v_7v_8).\end{aligned}$$

Throughout this section, we will use the embeddings of $P(8,3)$ illustrated in Figure 6. The bottom embedding is the same as the one on the right except that the edge $\overline{u_8v_8}$ passes through the point at ∞. The steps in Figure 6 show that the embeddings are all isotopic. Thus we abuse notation and refer to all of them as Γ.

Observe that μ is induced on the leftmost embedding by a rotation by $2\pi/3$ around the axis containing u_8, v_6, u_4, v_2. We see that γ is induced on the right embedding by rotating the u-cycle by $\pi/4$ while rotating the v-cycle by $3\pi/4$. Finally, β is induced on the bottom embedding by a rotation by π around

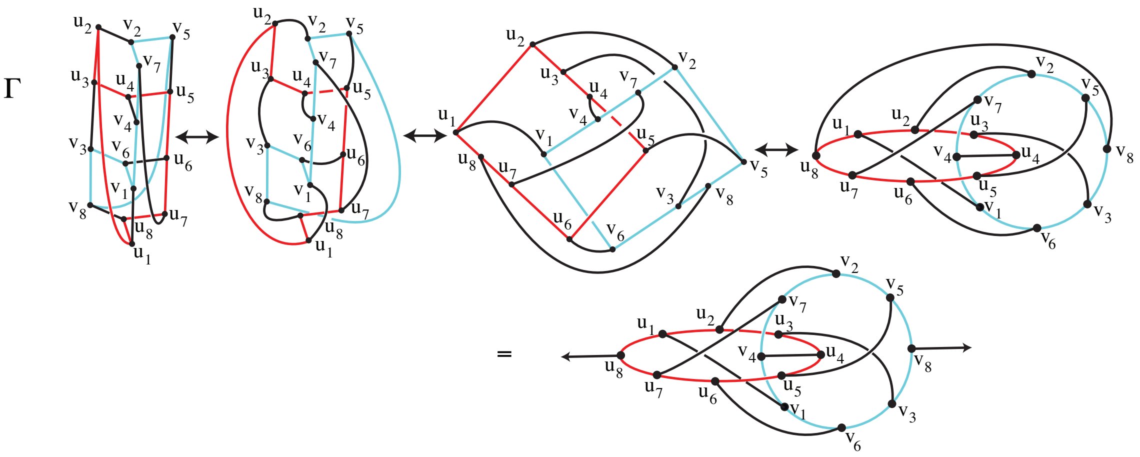


Figure 6: An embedding Γ of $P(8,3)$ with $\mathrm{TSG}_+(\Gamma) = \mathrm{Aut}(P(8,3))$.

the axis through ∞ containing the edges $\overline{u_4v_4}$ and $\overline{u_8v_8}$. Since $\mathrm{Aut}(P(8,3)) = \langle \mu, \beta, \gamma\rangle$, we now have $\mathrm{TSG}_+(\Gamma) = \mathrm{Aut}(P(8,3))$. Hence we have proved the following.

Proposition 6.1 $\mathrm{GL}(2,3) \rtimes \mathbb{Z}_2 = \mathrm{Aut}(P(8,3))$ *is positively realizable and hence realizable for* $P(8,3)$.

All of the proper subgroups of $\mathrm{GL}(2,3) \rtimes \mathbb{Z}_2$ are contained in the maximal groups D_{12}, $D_8 \rtimes \mathbb{Z}_2$, $\mathrm{GL}(2,3)$, or $\mathrm{SL}(2,3) \rtimes \mathbb{Z}_2$, which are addressed individually below.

Proposition 6.2 D_{12} *and all of its subgroups are positively realizable and hence realizable for* $P(8,3)$.

Proof Let $\beta = (u_1u_7)(u_2u_6)(u_3u_5)(v_1v_7)(v_2v_6)(v_3v_5)$ and

$$\beta_1 = (u_1u_2u_3v_3v_8v_5u_5u_6u_7v_7v_4v_1)(u_4v_6u_8v_2).$$

Then $D_{12} = \langle \beta_1, \beta\rangle$ leaves the set of vertices $U = \{u_4, v_6, u_8, v_2\}$ invariant, but $\mathrm{Aut}(P(8,3))$ does not leave U invariant. We create an embedding Γ_{12} by adding identical 3_1 knots to the edges of Γ that include a vertex in U. Then $\langle \beta_1, \beta\rangle \le \mathrm{TSG}_+(\Gamma_{12})$. Since D_{12} is a maximal subgroup of $\mathrm{GL}(2,3) \rtimes \mathbb{Z}_2$, we have $\mathrm{TSG}_+(\Gamma_{12}) = D_{12}$.

Suppose that the edge $\overline{v_5u_5}$ of Γ_{12} is pointwise fixed by an orientation-preserving homeomorphism h of (S^3, Γ_{12}). Since $\overline{v_5v_2}$ is knotted and $\overline{v_5v_8}$ is not, both of these edges must also be pointwise fixed. Now by Theorem 1.3, there is an embedding Γ'_{12} of $P(8,3)$ and a finite-order orientation-preserving homeomorphism f of (S^3, Γ'_{12}) inducing the same automorphism of $P(8,3)$ as h. But by Smith theory [14], a triad cannot be pointwise fixed by f unless f is the identity. Thus $\overline{v_5u_5}$ is not pointwise fixed by any nontrivial element of $\mathrm{TSG}_+(\Gamma_{12})$. Applying Theorem 1.7, we conclude that all of the subgroups of D_{12} are positively realizable for $P(8,3)$. □

The result below follows from Theorem 3.1.

Proposition 6.3 $D_8 \rtimes \mathbb{Z}_2$ *and all of its subgroups are positively realizable and hence realizable for* $P(8,3)$.

Proposition 6.4 $\mathrm{GL}(2,3)$ *is positively realizable and hence realizable for* $P(8,3)$.

Proof Let $A = \{u_i, v_j \mid i \text{ even}, j \text{ odd}\}$ and $B = \{u_i, v_j \mid i \text{ odd}, j \text{ even}\}$. Starting with the embedding Γ from Figure 6, we add the noninvertible knot 8_{17} to each edge between a vertex in A and a vertex in B oriented from A to B. This gives us an embedding Γ_{GL}. Now $\mathrm{GL}(2,3) = \langle \delta_1, \delta_2\rangle$ for

$$\delta_1 = (u_2v_1u_8)(u_3v_6v_8)(u_4u_6v_5)(u_7v_2v_4) \quad \text{and} \quad \delta_2 = (u_1u_3v_4v_2u_5u_7v_8v_6)(u_2u_4v_7v_5u_6u_8v_3v_1).$$

Observe that A and B are each invariant under $\langle \delta_1, \delta_2\rangle$. Thus $\langle \delta_1, \delta_2\rangle \le \mathrm{TSG}_+(\Gamma_{\mathrm{GL}})$. On the other hand, $\gamma = (u_1u_2u_3u_4u_5u_6u_7u_8)(v_1v_2v_3v_4v_5v_6v_7v_8)$ does not leave A and B invariant. Thus γ cannot be induced by a homeomorphism of $(S^3, \Gamma_{\mathrm{GL}})$, and hence $\gamma \notin \mathrm{TSG}_+(\Gamma_{\mathrm{GL}})$. Since $\langle \delta_1, \delta_2\rangle = \mathrm{GL}(2,3)$ is a maximal subgroup of $\mathrm{Aut}(P(8,3)) = \mathrm{GL}(2,3) \rtimes \mathbb{Z}_2$, we have $\mathrm{TSG}_+(\Gamma_{\mathrm{GL}}) = \mathrm{GL}(2,3)$. □

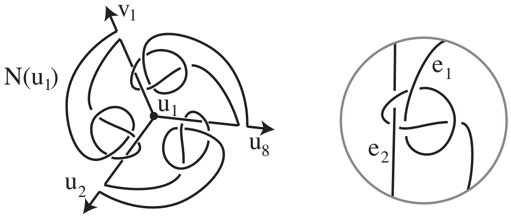


Figure 7: We embed the edges near u_1 as on the left. We say e_1 is *knotted around* e_2 if there is a tangle as on the right.

Proposition 6.5 $\mathrm{SL}(2,3) \rtimes \mathbb{Z}_2$ *and all of its subgroups are positively realizable and hence realizable for* $P(8,3)$.

Proof We start with the embedding Γ of $P(8,3)$ illustrated in Figure 6, which has

$$\mathrm{TSG}_+(\Gamma) = \mathrm{Aut}(P(8,3)) = \mathrm{GL}(2,3) \rtimes \mathbb{Z}_2.$$

Now by Theorem 1.5, there is an embedding Γ' such that $\mathrm{Aut}(P(8,3)) \le \mathrm{TSG}_+(\Gamma') \le \mathrm{Aut}(P(8,3))$ which is induced by an isomorphic group G of orientation-preserving isometries of (S^3, Γ').

Now we modify Γ in a neighborhood of the vertex u_1 so that it looks like the neighborhood $N(u_1)$ illustrated on the left in Figure 7. We say that an edge e_1 is *knotted around* an edge e_2, if there is a ball whose boundary intersects the graph in four points giving us the tangle on the right in Figure 7, where the 3_1 knot in one string can be replaced by any nontrivial knot which is linked with the other string as indicated. Thus after changing Γ in the neighborhood $N(u_1)$, we see that $\overline{u_1u_2}$ is knotted around $\overline{u_1u_8}$ which is knotted around $\overline{u_1v_1}$ which is knotted around $\overline{u_1u_2}$. Since none of the reverse knottings hold, this gives an *order to the edges* around u_1.

Observe that $\mathrm{SL}(2,3) \rtimes \mathbb{Z}_2 = \langle \rho_1, \rho_2, \rho_3 \rangle$ where

$$\begin{aligned} \rho_1 &= (u_2v_1u_8)(u_3v_6v_8)(u_4u_6v_5)(u_7v_2v_4), \\ \rho_2 &= (u_1v_3u_5v_7)(u_2v_8u_6v_4)(u_3v_5u_7v_1)(u_4v_2u_8v_6), \\ \rho_3 &= (u_1u_7v_8)(u_2v_7v_5)(u_3v_4u_5)(u_6v_3v_1). \end{aligned}$$

Since $\langle \rho_1, \rho_2, \rho_3 \rangle$ acts transitively on the vertices, we can apply the isometry group G to the neighborhood $N(u_1)$ to modify a neighborhood around every vertex of Γ'. To check that this is well-defined, we used Sage to determine that the only nontrivial automorphisms in $\langle \rho_1, \rho_2, \rho_3 \rangle$ which fix a vertex are of order 3. Hence no edge is pointwise fixed by any nontrivial element of $\langle \rho_1, \rho_2, \rho_3 \rangle$. It follows that no automorphism in $\langle \rho_1, \rho_2, \rho_3 \rangle$ changes the order of the edges sharing a vertex. Thus our modification of the neighborhoods around vertices is well-defined, and hence we have a well-defined embedding $\Gamma_{\mathrm{SL}2}$ of $P(8,3)$ such that the edges around every vertex have an order. Hence

$$\mathrm{SL}(2,3) \rtimes \mathbb{Z}_2 = \langle \rho_1, \rho_2, \rho_3 \rangle \le \mathrm{TSG}_+(\Gamma_{\mathrm{SL}2}).$$

Now $\beta = (u_1u_7)(u_2u_6)(u_3u_5)(v_1v_7)(v_2v_6)(v_3v_5) \in \mathrm{Aut}(P(8,3) = \mathrm{GL}(2,3) \rtimes \mathbb{Z}_2$ does not preserve the order of edges around u_4 since β fixes $\overline{u_4u_5}$ and interchanges u_3 and u_5. Thus

$$\mathrm{TSG}_+(\Gamma_{\mathrm{SL2}}) \neq \mathrm{Aut}(P(8,3)).$$

Since $\mathrm{SL}(2,3) \rtimes \mathbb{Z}_2$ is a maximal subgroup of $\mathrm{Aut}(P(8,3))$, it follows that $\langle \rho_1, \rho_2, \rho_3\rangle = \mathrm{TSG}_+(\Gamma_{\mathrm{SL2}})$. Now since no edge of $P(8,3)$ is pointwise fixed by any nontrivial element of $\langle \rho_1, \rho_2, \rho_3\rangle$, we can apply Theorem 1.7 to conclude that all of the subgroups of $\mathrm{SL}(2,3) \rtimes \mathbb{Z}_2$ are also positively realizable. □

In summary, we have proved the following.

Theorem 6.6 $\mathrm{Aut}(P(8,3)) = \mathrm{GL}(2,3) \rtimes \mathbb{Z}_2$ *and all of its subgroups are positively realizable and hence realizable for* $P(8,3)$.

7 The exceptional case $P(10,2)$

Frucht proved that $\mathrm{Aut}(P(10,2)) = A_5 \times \mathbb{Z}_2$ [11]. The embedding Γ of $P(10,2)$ as the 1-skeleton of a regular dodecahedron (see the left image of Figure 8) will be the basis for all of the embeddings in this section. The group $G = A_5$ of rotations of a solid dodecahedron consists of six order-5 rotations about an axis through the centers of opposite faces, ten order-3 rotations about an axis through opposite vertices, and fifteen order-2 rotations about an axis through midpoints of opposite edges.

The automorphism $\alpha = (v_2v_8)(u_2u_8)(v_6v_4)(u_6u_4)(v_3v_7)(u_3u_7)(v_1v_9)(u_1u_9)$ is not induced by a rotation of the solid dodecahedron, and hence is not in G. However, it is induced by a rotation h of S^3 around the equator of Γ containing $\overline{u_{10}v_{10}}$ and $\overline{u_5v_5}$ which interchanges the inside and outside of the dodecahedron. Thus $H = \langle G, h\rangle = A_5 \times \mathbb{Z}_2$ is a group of rotations of (S^3, Γ) which induces $\mathrm{Aut}(P(10,2)) = A_5 \times \mathbb{Z}_2$. Hence we have proved the following.

Proposition 7.1 $\mathrm{Aut}(P(10,2)) = A_5 \times \mathbb{Z}_2$ *is positively realizable and hence realizable for* $P(10,2)$.

By Theorem 2.1, we have the following.

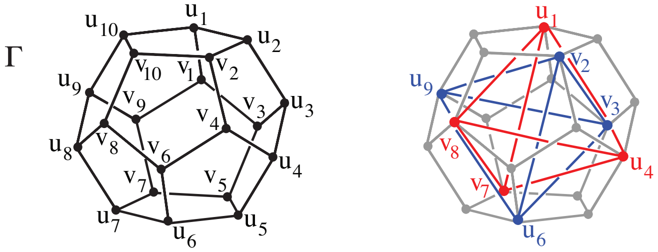


Figure 8: On the left, Γ is an embedding of $P(10,2)$ with $\mathrm{TSG}_+(\Gamma) = A_5 \times \mathbb{Z}_2$. On the right, red and blue inscribed tetrahedra which will be used in the proof of Proposition 7.3.

Proposition 7.2 *D_{10} and all of its subgroups are positively realizable and hence realizable for $P(10,2)$.*

Proposition 7.3 *$A_4 \times \mathbb{Z}_2$ and all of its subgroups are positively realizable and hence realizable for $P(10,2)$.*

Proof Let $T_1 = \{u_1, u_4, v_7, v_8\}$ and $T_2 = \{u_6, u_9, v_2, v_3\}$ be the red and blue sets of vertices, respectively, illustrated on the right in Figure 8. The vertices in each T_i are equidistant from each other and form the corners of a solid tetrahedron inscribed in the solid dodecahedron. Let G' be the subgroup of the group G of rotations of the solid dodecahedron which takes each of these solid tetrahedron to itself inducing the group of rotations A_4 on both tetrahedra. The rotation h of S^3 around the equator of Γ containing $\overline{u_{10}v_{10}}$ and $\overline{u_5v_5}$ interchanges the vertices in T_1 and T_2. Thus $H' = \langle G', h\rangle$ is a group of rotations of (S^3, Γ) which induces $A_4 \times \mathbb{Z}_2$ on Γ.

Now we modify the embedding Γ by adding the 4_1 knot to each edge which has a vertex in $T_1 \cup T_2$ to get an embedding Γ_{A42} which is invariant under H'. Thus $A_4 \times \mathbb{Z}_2 \leq \mathrm{TSG}_+(\Gamma_{A42})$, and every homeomorphism of (S^3, Γ_{A42}) leaves $T_1 \cup T_2$ setwise invariant. Now observe that no automorphism of order 5 of $P(10,2)$ leaves $T_1 \cup T_2$ setwise invariant. Thus no order-5 automorphism is in $\mathrm{TSG}_+(\Gamma_{A42})$. Since $A_4 \times \mathbb{Z}_2$ is a maximal subgroup of $A_5 \times \mathbb{Z}_2$, we have $A_4 \times \mathbb{Z}_2 = \mathrm{TSG}_+(\Gamma_{A42})$ induced by H'.

Finally, suppose that the edge $\overline{v_2v_4}$ is pointwise fixed by a nontrivial element of $\mathrm{TSG}_+(\Gamma_{A42})$. Then there is an element $f \in H'$ which pointwise fixes $\overline{v_2v_4}$. Since $\overline{v_4u_4}$ has a knot in it and $\overline{v_4v_6}$ does not, both of these edges must also be pointwise fixed by f. Since all the elements of H' are rotations, this means that f must be trivial. Thus $\overline{v_2v_4}$ is not pointwise fixed by any nontrivial element of $\mathrm{TSG}_+(\Gamma_{A42})$. Hence we can apply Theorem 1.7 to conclude that all of the subgroups of $A_4 \times \mathbb{Z}_2$ are positively realizable for $P(10,2)$. □

Proposition 7.4 *A_5 and its subgroups are positively realizable and hence realizable for $P(10,2)$.*

Proof Starting with Γ we embed the edges around a neighborhood of u_1 as in $N(u_1)$ in Figure 7. Now as in the proof of Proposition 6.5, we apply G to $N(u_1)$ to modify a neighborhood around every vertex. Since G is the group of rotations of the solid dodecahedron, no nontrivial element of G changes the order of edges around any vertex. Hence this gives us a well-defined embedding Γ_{A5} of $P(10,2)$ which is invariant under G. Observe that α changes the order of edges around u_9 and hence $\alpha \notin \mathrm{TSG}_+(\Gamma_{A5})$. Because A_5 is a maximal subgroup of $\mathbb{Z}_5 \times \mathbb{Z}_2$, it follows that $\mathrm{TSG}_+(\Gamma_{A5}) = A_5$.

Finally, since no nontrivial element of G pointwise fixes any edge of $\mathrm{TSG}_+(\Gamma_{A5})$, we can apply Theorem 1.7 to conclude that every subgroup of A_5 is positively realizable for $P(10,2)$. □

Proposition 7.5 *D_6 is positively realizable and hence realizable for $P(10,2)$.*

Proof We again start with the embedding Γ of $P(10,2)$ from Figure 8. Let h_1 be a rotation of (S^3, Γ) of order 2 around the equator of Γ containing the edges $\overline{u_{10}v_{10}}$ and $\overline{u_5v_5}$. Then h_1 induces the automorphism

$$\alpha_1 = (u_1u_9)(u_2u_8)(u_3u_7)(u_4u_6)(v_1v_9)(v_2v_8)(v_3v_7)(v_4v_6).$$

Let h_2 be a rotation of (S^3, Γ) of order 2 around an axis that passes through the midpoints of the edges $\overline{u_7u_8}$ and $\overline{u_2u_3}$. Then h_2 induces the automorphism

$$\alpha_2 = (u_2u_3)(u_6u_9)(u_7u_8)(u_5u_{10})(u_4u_1)(v_2v_3)(v_6v_9)(v_7v_8)(v_5v_{10})(v_4v_1).$$

It follows that $h = h_2h_1$ is an isometry of (S^3, Γ) inducing the automorphism

$$\alpha' = \alpha_2\alpha_1 = (u_1u_6)(u_2u_7)(u_3u_8)(u_4u_9)(u_5u_{10})(v_1v_6)(v_2v_7)(v_3v_8)(v_4v_9)(v_5v_{10}).$$

Let f be an order-3 rotation of (S^3, Γ) around an axis that passes through vertices u_1 and u_6. Then f induces the automorphism

$$\beta = (u_2u_{10}v_1)(u_3v_{10}v_9)(u_4v_8v_7)(u_5v_6u_7)(u_8v_5v_4)(u_9v_3v_2).$$

Finally, let g be the order-2 rotation of (S^3, Γ) around an axis that passes through the midpoints of edges $\overline{u_3u_4}$ and $\overline{u_8u_9}$. Then g induces the automorphism

$$\gamma = (u_1u_6)(u_2u_5)(u_3u_4)(u_7u_{10})(u_8u_9)(v_1v_6)(v_2v_5)(v_3v_4)(v_7v_{10})(v_8v_9).$$

The pair $\{u_1, u_6\}$ is setwise fixed by α', β, and γ. Now we add 4_1 knots to the six edges containing u_1 or u_6 to obtain an embedding Γ_{D6} such that the isometries h, g, and f leave Γ_{D6} setwise invariant. Then $\langle h, g, f\rangle$ induces $\langle \alpha', \beta, \gamma\rangle$ on Γ_{D6}. Since no nontrivial finite-order orientation-preserving isometry of S^3 can pointwise fix Γ_{D6}, the isometry group $\langle h, g, f\rangle$ is isomorphic to the automorphism group $\langle \alpha', \beta, \gamma\rangle$. Furthermore, because of the 4_1 knots on the edges of the triads centered at u_1 and u_6, every element of $\langle h, g, f\rangle$ leaves this pair of triads setwise invariant. Since no nontrivial finite-order orientation-preserving isometry can pointwise fix a triad, $\langle h, g, f\rangle$ induces an isomorphic action on this pair of triads.

However, the action of $\langle \alpha', \beta, \gamma\rangle$ on the two triads is D_6 because α' interchanges the two triads, β rotates both triads, and $\alpha'\gamma$ flips over each triad fixing the edges $\overline{u_6v_6}$ and $\overline{u_1v_1}$. Thus $D_6 \leq \mathrm{TSG}_+(\Gamma_{D6})$. On the other hand, because of the 4_1 knots, not every element of $\mathrm{Aut}(P(10,2))$ can be induced on the embedding Γ_{D6}. It follows that $\mathrm{TSG}_+(\Gamma_{D6})$ is a proper subgroup of $\mathrm{Aut}(P(10,2)) = A_5 \times \mathbb{Z}_2$. Since D_6 a maximal in $A_5 \times \mathbb{Z}_2$, this means that $D_6 = \mathrm{TSG}_+(\Gamma_{D6})$. □

In summary, we have proven the following.

Theorem 7.6 $\mathrm{Aut}(P(10,2)) = A_5 \times \mathbb{Z}_2$ *and all of its subgroups are positively realizable and hence realizable for* $P(10,2)$.

8 The exceptional case $P(10, 3)$

The group $\mathrm{Aut}(P(10,3)) = S_5 \times \mathbb{Z}_2 = \langle \alpha, \beta\rangle$ where

$$\alpha = (u_1u_2u_3u_4u_5u_6u_7u_8u_9u_{10})(v_1v_2v_3v_4v_5v_6v_7v_8v_9v_{10}),$$
$$\beta = (u_1v_4)(u_2u_4)(u_5v_2)(u_6v_9)(u_7u_9)(u_{10}v_7).$$

The isomorphism classes of proper nontrivial subgroups of $S_5 \times \mathbb{Z}_2$ are $A_5 \times \mathbb{Z}_2$, S_5, A_5, $S_4 \times \mathbb{Z}_2$, $(\mathbb{Z}_5 \rtimes \mathbb{Z}_2) \times \mathbb{Z}_2$, $S_3 \times \mathbb{Z}_2^2$, $A_4 \times \mathbb{Z}_2$, S_4, D_{10}, $\mathbb{Z}_5 \rtimes \mathbb{Z}_2$, $D_4 \times \mathbb{Z}_2$, $\mathbb{Z}_6 \times \mathbb{Z}_2$, D_6, A_4, $\mathbb{Z}_{10}$, D_5, $\mathbb{Z}_2^3$, D_4, $\mathbb{Z}_4 \times \mathbb{Z}_2$, $\mathbb{Z}_6$, S_3, $\mathbb{Z}_5$, $\mathbb{Z}_2^2$, $\mathbb{Z}_4$, $\mathbb{Z}_3$ and $\mathbb{Z}_2$. We will see that all of these subgroups are realizable, but not all are positively realizable for $P(10,3)$. We begin by determining which subgroups are not positively realizable.

Proposition 8.1 *$\mathbb{Z}_5 \rtimes \mathbb{Z}_4$, $\mathbb{Z}_{10} \rtimes \mathbb{Z}_4$ and $S_5 \times \mathbb{Z}_2$ are not positively realizable for $P(10,3)$.*

Proof Suppose that $\mathbb{Z}_5 \rtimes \mathbb{Z}_4 \leq \mathrm{TSG}_+(\Lambda)$ for some embedding Λ of $P(10,3)$ in S^3. Then by Theorem 1.5, for some embedding Γ of $P(10,3)$ in S^3, the group $\mathbb{Z}_5 \rtimes \mathbb{Z}_4$ is induced on Γ by an isomorphic group of orientation-preserving isometries G.

Using Sage, we determined the elements of $\mathrm{Aut}(P(10,3))$. In particular, every involution in $\mathbb{Z}_5 \rtimes \mathbb{Z}_4$ fixes either two or four vertices. Thus for every involution $g \in G$ we have $\mathrm{fix}(g) = S^1$. Since no element of G of order 4 or 5 fixes any vertices, no other element of G has the same fixed-point set as an involution. Hence G satisfies the hypothesis of Theorem 1.6, but $\mathbb{Z}_5 \rtimes \mathbb{Z}_4$ is not one of the groups in the conclusion of the theorem. Thus $\mathbb{Z}_5 \rtimes \mathbb{Z}_4$ is not contained in $\mathrm{TSG}_+(\Lambda)$ for any embedding Λ of $P(10,3)$ in S^3. Now since $\mathbb{Z}_{10} \rtimes \mathbb{Z}_4$ and $S_5 \times \mathbb{Z}_2$ each contain $\mathbb{Z}_5 \rtimes \mathbb{Z}_4$ as a subgroup, they also cannot be positively realizable. □

Lemma 8.2 *The following automorphisms are not positively realizable for $P(10,3)$.*

- *An order-2 automorphism with only six 2-cycles.*
- *An order-4 automorphism with 2-cycles which contain adjacent vertices.*
- *An order-6 automorphism with 3-cycles.*

Proof Suppose that an automorphism of $P(10,3)$ is positively realizable. Then by Theorem 1.3, that automorphism is induced on some embedding by a finite-order orientation-preserving homeomorphism of S^3. By Smith theory [14], such a homeomorphism either pointwise fixes an S^1 or is fixed-point free.

All order-2 automorphisms that contain only six 2-cycles are conjugate to

$$(u_1 v_4)(u_2 u_4)(u_5 v_2)(u_6 v_9)(u_7 u_9)(u_{10} v_7),$$

which pointwise fixes the edges $\overline{v_3 u_3}$, $\overline{v_3 v_{10}}$, $\overline{v_3 v_6}$, $\overline{v_8 u_8}$, $\overline{v_8 v_5}$, $\overline{v_8, v_1}$. Since these six edges form two triads, they cannot be contained in an S^1. Thus no such order-2 automorphism can be positively realizable.

An order-4 automorphism that contains 2-cycles with adjacent vertices, flips over the edges between such vertices, pointwise fixing their midpoints. Suppose that such an automorphism is induced on some embedding by a finite-order orientation-preserving homeomorphism h of S^3. Then $\mathrm{fix}(h)$ is an S^1 and so is $\mathrm{fix}(h^2)$. But $\mathrm{fix}(h) \subseteq \mathrm{fix}(h^2)$ and hence these sets are equal. This is impossible since h^2 pointwise fixes edges which are flipped over by h.

All order-6 automorphisms that contain 3-cycles are conjugate to the automorphism

$$\rho = (u_1, v_2, u_3)(u_4 u_{10} v_5 v_3 v_1 v_9)(u_5 v_{10} v_8 v_6 v_4 u_9)(u_6 v_7 u_8).$$

If ρ were positively realizable, then ρ^3 would be as well. But ρ^3 is an order-2 automorphism that contains only six 2-cycles, which we saw is not positively realizable. □

Proposition 8.3 *$\mathbb{Z}_6 \times \mathbb{Z}_2$, $D_6 \times \mathbb{Z}_2$, S_4, S_5, $S_4 \times \mathbb{Z}_2$, $\mathbb{Z}_4 \times \mathbb{Z}_2$, and $D_4 \times \mathbb{Z}_2$ are not positively realizable for $P(10,3)$.*

Proof Using Sage, we determined that there are two conjugacy classes of subgroups of $\mathrm{Aut}(P(10,3))$ that are isomorphic to S_4. One contains an order-2 automorphism with only six 2-cycles while the other contains an order-4 automorphism with a 2-cycle with adjacent vertices. Thus by Lemma 8.2, neither is positively realizable. Since S_5 and $S_4 \times \mathbb{Z}_2$ both contain S_4, they too are not positively realizable.

We also determined with Sage that $\mathrm{Aut}(P(10,3))$ has only one conjugacy class isomorphic to $\mathbb{Z}_6 \times \mathbb{Z}_2$ and it contains an order-6 automorphism that includes 3-cycles. Thus by Lemma 8.2, it is not positively realizable. Since $D_6 \times \mathbb{Z}_2$ contains $\mathbb{Z}_6 \times \mathbb{Z}_2$, it too is not positively realizable.

Finally, we determined that there is only one conjugacy class of subgroups isomorphic to $\mathbb{Z}_4 \times \mathbb{Z}_2$ and it contains an order-4 element with 2-cycles with adjacent vertices. Thus by Lemma 8.2, it is not positively realizable. Since $D_4 \times \mathbb{Z}_2$ contains $\mathbb{Z}_4 \times \mathbb{Z}_2$, it too is not positively realizable. □

Below we determine the positively realizable subgroups of $\mathrm{Aut}(P(10,3))$. By Theorem 2.1 we have Proposition 8.4.

Proposition 8.4 *D_{10} and all of its subgroups are positively realizable and hence realizable for $P(10,3)$.*

Proposition 8.5 *$\mathbb{Z}_4$ and D_4 are positively realizable and hence realizable for $P(10,3)$.*

Proof Since $\mathbb{Z}_4 \le \mathbb{Z}_{10} \rtimes \mathbb{Z}_4 = B(10,3)$, it follows from Theorem 4.2 that $\mathbb{Z}_4$ is positively realizable. Now we consider $D_4 = \langle \nu_1, \nu_2 \rangle$, where

$$\nu_1 = (u_1 u_5)(u_2 v_5 v_1 u_4)(u_3 v_2 v_8 v_4)(u_6 u_{10})(u_7 v_{10} v_6 u_9)(u_8 v_7 v_3 v_9),$$
$$\nu_2 = (u_1 u_6)(u_2 u_7)(u_3 v_7)(u_4 v_{10})(u_5 u_{10})(u_8 v_2)(u_9 v_5)(v_1 v_6)(v_3 v_4)(v_8 v_9).$$

Let Γ_{D4} be the embedding of $P(10,3)$ in a standardly embedded solid torus T in S^3, illustrated in Figure 9, where the edges which are on ∂T are green on the flat torus and the edges of Γ_{D4} which go into the interior of T are red. The edges $\overline{u_{10} u_1}$ and $\overline{u_5 u_6}$ are diametrically opposed arcs on the core of the solid torus illustrated in the center. The vertices in a given cycle of ν_1 are the same color. Two special meridians are colored purple and orange on the right.

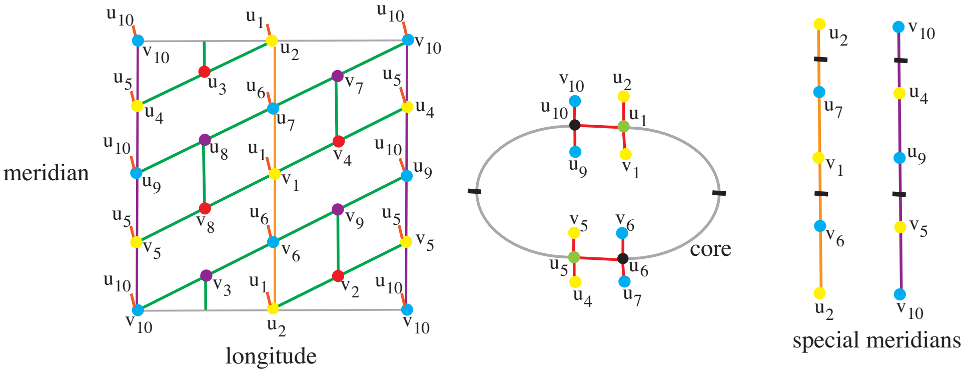


Figure 9: On the left and center is an embedding Γ_{D_4} of $P(10,3)$ on the boundary and core of a standardly embedded solid torus such that $\mathrm{TSG}_+(\Gamma_{D_4}) = D_4$. On the right, we display two special meridians.

The automorphism ν_1 is induced by the glide rotation h_{ν_1} which rotates T meridionally by $\pi/2$ around its core and rotates T longitudinally by π around the core of a complementary solid torus. The automorphism ν_2 is induced by a rotation h_{ν_2} by π around an axis which pierces ∂T in two points on the orange meridian, two points on the purple meridian, and two points on the core of T, as indicated by black dashes in in the center and right images of Figure 9. Now $\langle h_{\nu_1}, h_{\nu_2}\rangle$ induces $D_4 = \langle \nu_1, \nu_2\rangle$ on Γ_{D4}. Thus $D_4 \leq \mathrm{TSG}_+(\Gamma_{D_4})$. If $\mathrm{TSG}_+(\Gamma_{D_4}) \neq D_4$, then $\mathrm{TSG}_+(\Gamma_{D_4})$ would contain a group which has D_4 as a maximal subgroup. However, the only subgroups of $\mathrm{Aut}(P(10,3)) = S_5 \times \mathbb{Z}_2$ which contain D_4 as a maximal subgroup are $D_4 \times \mathbb{Z}_2$ and S_4, and we saw in the proof of Proposition 8.3 that each of these groups contains an automorphism which is not positively realizable for $P(10,3)$. Thus $\mathrm{TSG}_+(\Gamma_{D_4}) = D_4$. □

Proposition 8.6 *$A_5 \times \mathbb{Z}_2$ and A_5 are positively realizable and hence realizable for $P(10,3)$.*

Proof Let Ω denote a regular 4-simplex embedded in S^3. The group G of orientation-preserving isometries of (S^3, Ω) is A_5. We embed half of the vertices of $P(10,3)$ as midpoints of the edges of Ω (as illustrated on the left of Figure 10) and the other half as center points of the faces of Ω. Then we connect vertices with straight edges to obtain an embedding Λ of $P(10,3)$. We illustrate Λ in an unfolded picture of Ω in the center of Figure 10. Since Λ is symmetrically embedded in Ω, it is setwise invariant under any isometry of Ω, in particular under G.

We define the dual 4-simplex Ω' in S^3 as follows. We embed a vertex of Ω' as the center point of each tetrahedron of Ω and we embed each edge of Ω' as a straight segment joining two vertices. Then each edge of Ω' will pass through the center point of a face of Ω. Next we embed each face of Ω' as the triangle bounded by three pairwise adjacent edges of Ω'. Then the midpoint of each edge of Ω will pass through the center point of a face of Ω'. Finally, we define a tetrahedron of Ω' as the solid bounded by four pairwise adjacent faces of Ω'. Thus each vertex of Ω will be the center point of a tetrahedron of Ω'. In Figure 10, we illustrate Ω' in blue on the right. We label each vertex of Ω' by the tetrahedron of Ω

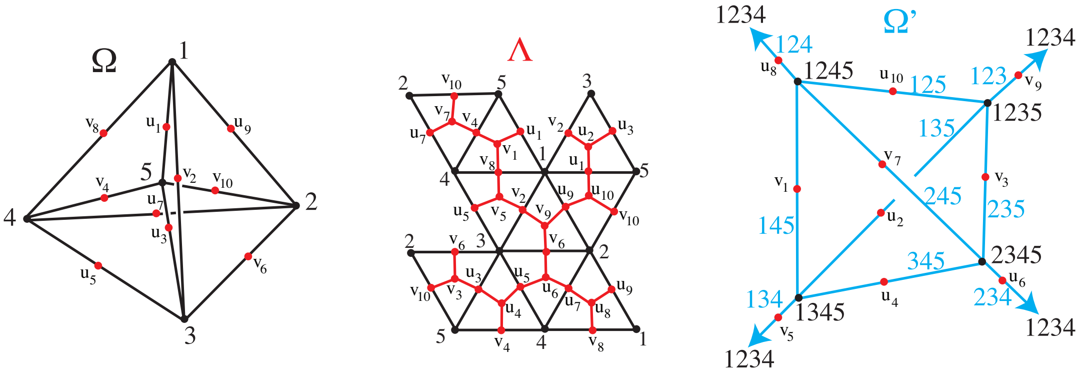


Figure 10: The 4-simplex on the left is Ω. The red graph in the center is the embedding Λ of $P(10,3)$ in an unfolded picture of Ω such that $\mathrm{TSG}_+(\Lambda) = A_5 \times \mathbb{Z}_2$. On the right is the 4-simplex Ω' which is dual to Ω.

which it is the center of, and we label each edge of Ω' by the face of Ω which it intersects. In Figure 10 the vertex 1234 of Ω' is at the point at ∞. We see from the center image that half of the vertices of Λ are on edges of Ω and the other half are on edges of Ω'. Also, all of the edges of Λ go between Ω and Ω'.

Now there is a glide rotation h of S^3 which interchanges Ω and Ω', interchanging each vertex $i \in \Omega$ with the vertex $jklm \in \Omega'$ such that $i \notin \{j,k,l,m\}$, and h takes Λ to itself inducing the automorphism

$$(u_1u_6)(u_2u_7)(u_3u_8)(u_4u_9)(u_5u_{10})(v_1v_6)(v_2v_7)(v_3v_8)(v_4v_9)(v_5v_{10}).$$

Since h commutes with every element of the group G of isometries, $H = \langle G, h\rangle = A_5 \times \mathbb{Z}_2$, and H leaves Λ setwise invariant. Hence $A_5 \times \mathbb{Z}_2 \le \mathrm{TSG}_+(\Lambda)$. Now since $A_5 \times \mathbb{Z}_2$ is a maximal subgroup of $S_5 \times \mathbb{Z}_2$ and $S_5 \times \mathbb{Z}_2$ is not positively realizable for $P(10,3)$, we have $A_5 \times \mathbb{Z}_2 = \mathrm{TSG}_+(\Lambda)$. Thus $A_5 \times \mathbb{Z}_2$ is positively realizable.

We obtain a new embedding Λ' by adding the noninvertible knot 8_{17} to every edge of Λ such that the orientation of the knots goes from the vertices on edges of Ω towards the vertices on edges of Ω'. Then Λ' is setwise invariant under the group G of orientation-preserving isometries of (S^3, Ω), but no element of $\mathrm{TSG}_+(\Lambda)$ interchanges the sets of vertices on edges of Ω with those on edges of Ω'. It follows that $\mathrm{TSG}_+(\Lambda') = A_5$. Hence, A_5 is positively realizable for $P(10,3)$. □

Proposition 8.7 *D_6 and all of its subgroups are positively realizable and hence realizable for $P(10,3)$.*

Proof $D_6 = \langle \theta_1, \theta_2\rangle$ where

$$\theta_1 = (u_1u_4v_{10}u_6u_9v_5)(u_2v_4v_3u_7v_9v_8)(u_3v_7v_6u_8v_2v_1)(u_5u_{10}),$$
$$\theta_2 = (u_9u_1)(u_4u_6)(u_2u_8)(u_3u_7)(v_9v_1)(v_4v_6)(v_2v_8)(v_3v_7).$$

Let Γ_{D6} be the embedding in Figure 11 with u_{10} at ∞. The automorphism θ_1 is induced by an order-6 glide rotation h_1 which rotates Γ_{D6} by $\pi/3$ around the vertical axis through u_5 and u_{10}, while rotating by

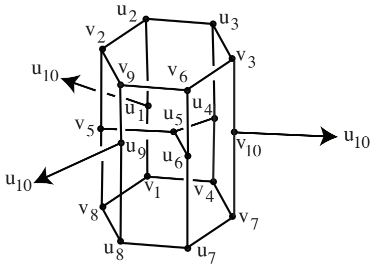


Figure 11: The embedding Γ_{D6} of $P(10,3)$ such that $\mathrm{TSG}_+(\Gamma'_{D6}) = D_6$, where v_{10} is the point at ∞.

π around a meridian of the hexagonal tube interchanging u_5 and u_{10}. The automorphism θ_2 is induced by the involution h_2 which rotates Γ_{D6} by π around an axis through u_{10}, v_{10}, u_5, and v_5.

Now we create a new embedding Γ'_{D6} by adding the 4_1 knot to the edges of the hexagons at the top and bottom of Γ_{D6}. Then $D_6 = \langle\theta_1, \theta_2\rangle \le \mathrm{TSG}_+(\Gamma'_{D6})$ and every homeomorphism of (S^3, Γ'_{D6}) leaves this pair of hexagons setwise invariant. The only positively realizable subgroup of $\mathrm{Aut}(P(10,3))$ containing D_6 as a proper subgroup is $A_5 \times \mathbb{Z}_2$. However, $A_5 \times \mathbb{Z}_2$ contains an element of order 5, which cannot leave a pair of hexagons setwise invariant. Thus $D_6 = \mathrm{TSG}_+(\Gamma'_{D6})$, and hence D_6 is positively realizable.

Finally, observe that the edge $\overline{v_2v_5}$ is not fixed by any nontrivial element of $\mathrm{TSG}_+(\Gamma'_{D6})$. Thus we can apply Theorem 1.7 to conclude that every subgroup of D_6 is positively realizable for $P(10,3)$. □

Proposition 8.8 *$A_4 \times \mathbb{Z}_2$ and all of its subgroups are positively realizable and hence realizable for $P(10,3)$.*

Proof Recall that $\mathrm{TSG}_+(\Lambda) = A_5 \times \mathbb{Z}_2$ is induced by the group H of orientation-preserving isometries of $(S^3, \Omega \cup \Omega')$. We create an embedding Λ_T by adding the 4_1 knot to the edges $\overline{v_8v_1}$, $\overline{u_9u_{10}}$, $\overline{v_6v_3}$, $\overline{v_2u_2}$, $\overline{u_5u_4}$, $\overline{u_7v_7}$ of Λ. Then $\mathrm{TSG}_+(\Lambda_T)$ is the subgroup of $\mathrm{TSG}_+(\Lambda)$ that takes the set $\{\overline{v_8v_1}, \overline{u_9u_{10}}, \overline{v_6v_3}, \overline{v_2u_2}, \overline{u_5u_4}, \overline{u_7v_7}\}$ to itself. Hence $\mathrm{TSG}_+(\Lambda_T)$ is induced by the subgroup $H_T \le H$ taking the pair of vertices $\{5, 1234\}$ of $\Omega \cup \Omega'$ to itself. It follows that H_T is the set of orientation-preserving isometries of the pair of tetrahedra Ω_T and Ω'_T with vertices 1, 2, 3, and 4, and 1345, 1245, 1235, and 2345, respectively. Thus $H_T = A_4 \times \mathbb{Z}_2$.

Now observe that if both vertices of the edge $\overline{u_1u_2}$ were fixed by a nontrivial automorphism in $\mathrm{TSG}_+(\Lambda_T)$, then the automorphism would be induced by an element of H_T which interchanges $\overline{u_2v_2}$ and $\overline{u_2u_3}$. But $\overline{u_2v_2}$ contains a knot and $\overline{u_2u_3}$ does not. As this is impossible, we can apply Theorem 1.7 to conclude that every subgroup of $A_4 \times \mathbb{Z}_2$ is positively realizable for $P(10,3)$. □

Next we prove that all of the subgroups of $\mathrm{Aut}(P(10,3))$ that are not positively realizable for $P(10,3)$ are in fact realizable.

Proposition 8.9 *$\mathrm{Aut}(P(10,3)) = S_5 \times \mathbb{Z}_2$ and S_5 are realizable for $P(10,3)$.*

Proof The embedding Λ, in Figure 10, is invariant under the group $S_5 \times \mathbb{Z}_2$ of isometries of the pair of dual 4-simplices Ω and Ω', including a reflection taking each 4-simplex to itself. Since $\mathrm{Aut}(P(10,3)) = S_5 \times \mathbb{Z}_2$, we have $\mathrm{TSG}(\Lambda) = S_5 \times \mathbb{Z}_2$.

Recall from the proof of Proposition 8.6 that the embedding Λ' was obtained from Λ by adding the knot 8_{17} to every edge of Λ oriented from vertices on edges of Ω to vertices on edges of Ω'. Then Λ' is invariant under the isometries of (S^3, Ω), and no homeomorphism of (S^3, Λ') can interchange Ω and Ω'. Thus $\mathrm{TSG}(\Lambda) = S_5$. □

Proposition 8.10 *$S_4 \times \mathbb{Z}_2$ and S_4 are realizable for $P(10,3)$.*

Proof Recall from the proof of Proposition 8.8 that Λ_T was obtained from Λ by adding a 4_1 knot to the edges going between the tetrahedron Ω_T with vertices 1, 2, 3, 4 and the dual tetrahedron Ω'_T with vertices 1245, 1345, 1235, 1245. Also $\mathrm{TSG}_+(\Lambda_T)$ is induced by the group $A_4 \times \mathbb{Z}_2$ of orientation-preserving isometries of $(S^3, \Omega_T \cup \Omega'_T)$. Now there is a reflection of $(S^3, \Omega_T \cup \Omega'_T)$ which pointwise fixes the sphere containing the vertices u_9, v_{10}, u_1, u_5 and u_8, v_7, u_2, u_6. Thus the full group of isometries of $(S^3, \Omega_T \cup \Omega'_T)$ including reflections is $S_4 \times \mathbb{Z}_2$. Since these isometries induce a faithful action on Λ_T, it follows that $S_4 \times \mathbb{Z}_2 \leq \mathrm{TSG}(\Lambda_T)$.

Let $\alpha = (u_1u_2u_3u_4u_5u_6u_7u_8u_9u_{10})(v_1v_2v_3v_4v_5v_6v_7v_8v_9v_{10})$. Then $\alpha \in \mathrm{Aut}(P(10,3))$ taking the edge $\overline{v_1v_8}$ to the edge $\overline{v_2v_9}$. However, in the embedding Λ_T, the edge $\overline{v_1v_8}$ contains a knot while the edge $\overline{v_2v_9}$ does not. Thus α is not contained in $\mathrm{TSG}(\Lambda_T)$, and hence $\mathrm{TSG}(\Lambda_T) \neq \mathrm{Aut}(P(10,3)) = S_5 \times \mathbb{Z}_2$. Since, $S_4 \times \mathbb{Z}_2$ is a maximal subgroup of $S_5 \times \mathbb{Z}_2$ it follows that $\mathrm{TSG}(\Lambda_T) = S_4 \times \mathbb{Z}_2$.

Finally, we obtain Γ_{S4} from Λ_T by replacing the invertible 4_1 knots by noninvertible 8_{17} knots which are oriented from Ω_T to Ω'_T. Because of the orientation, no homeomorphism of (S^3, Γ_{S4}) can interchange Ω_T and Ω'_T. But all of the other elements of $\mathrm{TSG}(\Lambda_T)$ are also elements of $\mathrm{TSG}(\Gamma_{S4})$. Thus $\mathrm{TSG}(\Gamma_{S4}) = S_4$. □

Proposition 8.11 *$D_4 \times \mathbb{Z}_2$ and $\mathbb{Z}_4 \times \mathbb{Z}_2$ are realizable for $P(10,3)$.*

Proof We again start with the embedding Λ_T, but now we replace the 4_1 knots on the edges $\overline{u_7v_7}$ and $\overline{u_2v_2}$ by the achiral and invertible knot 5_2 to obtain an embedding Γ_{D4} such that $\mathrm{TSG}(\Gamma_{D4})$ leaves $\{\overline{v_7u_7}, \overline{u_2v_2}\}$ and $\{\overline{v_8v_1}, \overline{u_9u_{10}}, \overline{v_6v_3}, \overline{u_5u_4}\}$ setwise invariant. Then any homeomorphism of $(S^3, \Omega_T \cup \Omega'_T)$ which leaves Γ_{D4} setwise invariant either interchanges the squares $S = \overline{1234} \subseteq \Omega_T$ and $S' = \overline{1245, 1235, 2345, 1345} \subseteq \Omega'_T$ or leaves each square setwise invariant.

Consider the following automorphisms of $P(10,3)$:

$$\begin{aligned}
\gamma_1 &= (u_5v_8u_9v_6)(u_1v_{10}u_3v_4)(u_7v_2)(v_5u_8v_9u_6)(v_1u_{10}v_3u_4)(u_2v_7),\\
\gamma_2 &= (u_1v_{10})(u_2v_7)(u_3v_4)(u_6v_5)(u_7v_2)(u_8v_9)(v_1v_3)(v_6v_8),\\
\gamma_3 &= (v_3v_8)(v_1v_6)(v_9v_4)(v_7v_2)(v_5v_{10})(u_3u_8)(u_1u_6)(u_9u_4)(u_7u_2)(u_5u_{10}).
\end{aligned}$$

Since Ω_T and Ω'_T are dual regular tetrahedra in S^3, there is a reflection composed with an order-4 rotation g of $(S^3, \Omega_T, \Omega'_T)$ which rotates S and S' interchanging the edges $\overline{13}$ and $\overline{24}$ and interchanging the edges $\overline{245}$ and $\overline{135}$. Also, there is an order-2 rotation f of $(S^3, \Omega_T, \Omega'_T)$, which turns over the square S interchanging the edges $\overline{14}$ and $\overline{23}$, and turns over the square S' interchanging the edges $\overline{145}$ and $\overline{235}$. Now, g and f leave Γ_{D4} setwise invariant inducing γ_1 and γ_2, respectively. Since g rotates S_1 and S_2 and f turns S_1 and S_2 over, $\langle g, f\rangle = D_4$. Finally, since the induced action on Γ_{D4} is faithful, we have $\langle \gamma_1, \gamma_2\rangle = D_4$.

Recall from the proof of Proposition 8.6 that there is an order-2 glide rotation h of S^3 which interchanges Ω and Ω', interchanging each vertex $i \in \Omega$ with the vertex $jklm \in \Omega'$ where $i \notin \{j,k,l,m\}$. Observe that h interchanges S and S' and leaves Γ_{D4} setwise invariant inducing γ_3. Also, γ_3 commutes with γ_1 and γ_2, and hence $\langle \gamma_1, \gamma_2, \gamma_3\rangle = D_4 \times \mathbb{Z}_2 \le \mathrm{TSG}(\Gamma_{D4})$. However, since $\mathrm{TSG}(\Gamma_{D4})$ is a proper subgroup of $\mathrm{TSG}(\Lambda_T) = S_4 \times \mathbb{Z}_2$, and $D_4 \times \mathbb{Z}_2$ is maximal in $S_4 \times \mathbb{Z}_2$, it follows that $\mathrm{TSG}(\Gamma_{D4}) = D_4 \times \mathbb{Z}_2$.

Next, we replace a neighborhood of u_4 by one where each edge incident to u_4 is knotted around the next such edge as illustrated in Figure 7, except that now the chiral 3_1 knot in the figure is replaced by the achiral 4_1 knot. As explained in the proof of Proposition 6.5, this gives an order to the edges around u_4. We then apply the isometry group $\langle g, h\rangle$ to the neighborhood $N(u_4)$, so that the edges in the orbit of $N(u_4)$ also have an order. Since g rotates S_1 and S_2, and h is an involution interchanging S_1 and S_2, no vertex in the orbit of u_4 is fixed by a nontrivial element of $\langle g, h\rangle$. Thus, replacing the original neighborhoods of these vertices by the new ones yields a well-defined embedding Γ_{Z4} of $P(10,3)$ such that $\mathbb{Z}_4 \times \mathbb{Z}_2 = \langle \gamma_1, \gamma_3\rangle \le \mathrm{TSG}(\Gamma_{Z4})$.

Now suppose that a homeomorphism f' of (S^3, Γ_{Z4}) induces γ_2. Then f' fixes the edge $\overline{u_5u_4}$ and interchanges $\overline{u_4u_3}$ and $\overline{u_4v_4}$. But this means that f' reverses the order of the edges incident to u_4, which is impossible since each of these edges is knotted around the next one. Thus $\mathrm{TSG}(\Gamma_{Z4})$ is a proper subgroup of $\mathrm{TSG}(\Gamma_{D4})$. Since $\mathbb{Z}_4 \times \mathbb{Z}_2$ is maximal in $D_4 \times \mathbb{Z}_2$, we must have

$$\mathrm{TSG}(\Gamma_{Z4}) = \mathbb{Z}_4 \times \mathbb{Z}_2. \qquad \square$$

Proposition 8.12 *$D_6 \times \mathbb{Z}_2$ and $\mathbb{Z}_6 \times \mathbb{Z}_2$ are realizable for $P(10,3)$.*

Proof We begin with the embedding Γ'_{D_6} and automorphisms θ_1 and θ_2 from the proof of Proposition 8.7 such that $\mathrm{TSG}_+(\Gamma'_{D_6}) = \langle \theta_1, \theta_2\rangle = D_6$. Now let

$$\theta_3 = (u_3v_4)(v_3v_7)(v_6u_7)(v_9u_8)(v_2v_8)(u_2v_1)(u_3v_4).$$

Then θ_3 commutes with θ_1 and θ_2, and hence $\langle \theta_1, \theta_2, \theta_3\rangle = D_6 \times \mathbb{Z}_2$. Also, θ_3 is induced on Γ'_{D_6} by a reflection through the sphere containing the set of vertices $\{u_{10}, u_5, u_4, u_6, v_5, u_9, u_1\}$. Thus

$$D_6 \times \mathbb{Z}_2 \le \mathrm{TSG}(\Gamma'_{D_6}).$$

Because of the 4_1 knots in the edges of the top and bottom hexagons, every homeomorphism of (S^3, Γ'_{D6}) takes this pair of hexagons to itself. Thus $\mathrm{TSG}(\Gamma'_{D_6})$ is a proper subgroup of $S_5 \times \mathbb{Z}_2 = \mathrm{Aut}(P(10,3))$. However, since $D_6 \times \mathbb{Z}_2$ is a maximal subgroup of $S_5 \times \mathbb{Z}_2$, we have $\mathrm{TSG}(\Gamma'_{D_6}) = D_6 \times \mathbb{Z}_2$.

Next we replace the 4_1 knots in the edges of the top and bottom hexagons of Γ'_{D6} by the noninvertible knot 8_{17} oriented consistently around the two hexagons to get a new embedding Γ_{Z6}. Now no homeomorphism of (S^3, Γ_{Z6}) can turn either hexagon over. Hence the automorphism θ_2 is not in $\mathrm{TSG}(\Gamma_{Z_6})$, but all of the other elements of $\mathrm{TSG}(\Gamma'_{D_6})$ are in $\mathrm{TSG}(\Gamma_{Z_6})$. Thus $\mathbb{Z}_6 \times \mathbb{Z}_2 \leq \mathrm{TSG}(\Gamma_{Z_6})$. Now since $\mathbb{Z}_6 \times \mathbb{Z}_2$ is a maximal subgroup of $D_6 \times \mathbb{Z}_2$, we have $\mathbb{Z}_6 \times \mathbb{Z}_2 = \mathrm{TSG}(\Gamma_{Z_6})$. □

Since $3^2 \equiv -1 \pmod{10}$, by Theorem 4.3 we have the following.

Proposition 8.13 $\mathbb{Z}_{10} \rtimes \mathbb{Z}_4$ *and all of its subgroups are realizable for* $P(10,3)$.

The following is a summary of our results for $P(10,3)$.

Theorem 8.14 $\mathrm{Aut}(P(10,3)) = S_5 \times \mathbb{Z}_2$ *and all of its subgroups are realizable for* $P(10,3)$. *Furthermore*:

1. $\mathbb{Z}_5 \rtimes \mathbb{Z}_4$, $\mathbb{Z}_{10} \rtimes \mathbb{Z}_4$, $S_5 \times \mathbb{Z}_2$, $\mathbb{Z}_6 \times \mathbb{Z}_2$, $D_6 \times \mathbb{Z}_2$, S_4, S_5, $S_4 \times \mathbb{Z}_2$, $\mathbb{Z}_4 \times \mathbb{Z}_2$, *and* $D_4 \times \mathbb{Z}_2$ *are not positively realizable.*
2. D_{10}, $\mathbb{Z}_{10}$, D_5, $\mathbb{Z}_5$, D_2, $\mathbb{Z}_2$, $A_5 \times \mathbb{Z}_2$, A_5, $A_4 \times \mathbb{Z}_2$, A_4, $\mathbb{Z}_3$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, D_4, $\mathbb{Z}_4$, D_6, $\mathbb{Z}_6$, *and* D_3 *are positively realizable.*

9 Conclusion

In Table 1 we summarize our results for $P(n,k)$. A check mark ✓ in the "TSG(Γ) complete" column means that every subgroup of the automorphism group is realizable. A check mark ✓ in the "$\mathrm{TSG}_+(\Gamma)$

$P(n,k)$	$\mathrm{Aut}(P(n,k))$	$\mathrm{TSG}_+(\Gamma)$ complete	$\mathrm{TSG}(\Gamma)$ complete
$k^2 \neq \pm 1 \bmod n$	D_n	✓	✓
$k^2 = 1 \bmod n$	$D_n \rtimes \mathbb{Z}_2$	✓	✓
$k^2 = -1 \bmod n$	$\mathbb{Z}_n \rtimes \mathbb{Z}_4$	X	✓
$P(4,1)$	$S_4 \times \mathbb{Z}_2$	✓	✓
$P(5,2)$	S_5	X	X
$P(8,3)$	$\mathrm{GL}(2,3) \rtimes \mathbb{Z}_2$	✓	✓
$P(10,2)$	$A_5 \times \mathbb{Z}_2$	✓	✓
$P(10,3)$	$S_5 \times \mathbb{Z}_2$	X	✓

Table 1: This table summarizes our results. A check mark means that all of the subgroups are realized. The first three rows are in the nonexceptional cases.

complete" column means that every subgroup of the automorphism group is positively realizable. An X means that some subgroups are not realizable or not positively realizable. The first three rows are in the nonexceptional cases. The final five rows are all the exceptional cases except for $P(12,5)$ and $P(24,5)$, which will be analyzed in a subsequent paper.

Acknowledgements

We thank AIM for its support of the REUF program which enabled us to start this project and supported our meeting at the JMM. We also thank ICERM for hosting our second meeting. Flapan was supported in part by NSF Grant DMS-1607744. Finally, we thank the referee for helpful suggestions.

References

[1] **F Bonahon**, *Arborescent knots and links*, from "Encyclopedia of Knot Theory", CRC Press, Boca Raton, FL (2021) 121–146

[2] **D Chambers**, **E Flapan**, *Topological symmetry groups of small complete graphs*, Symmetry 6 (2014) 189–209 MR

[3] **D Chambers**, **E Flapan**, **D Heath**, **E Lawrence**, **C Thatcher**, **R Vanderpool**, *Topological symmetry groups of the Petersen graph*, J. Knot Theory Ramifications 30 (2021) art. id. 2141004 MR

[4] **J H Conway**, **C M Gordon**, *Knots and links in spatial graphs*, J. Graph Theory 7 (1983) 445–453 MR

[5] **E Flapan**, *Rigidity of graph symmetries in the* 3*-sphere*, J. Knot Theory Ramifications 4 (1995) 373–388 MR

[6] **E Flapan**, **E D Lawrence**, *Topological symmetry groups of Möbius ladders*, J. Knot Theory Ramifications 23 (2014) art. id. 1450077 MR

[7] **E Flapan**, **B Mellor**, **R Naimi**, *Complete graphs whose topological symmetry groups are polyhedral*, Algebr. Geom. Topol. 11 (2011) 1405–1433 MR

[8] **E Flapan**, **B Mellor**, **R Naimi**, **M Yoshizawa**, *Classification of topological symmetry groups of* K_n, Topology Proc. 43 (2014) 209–233 MR

[9] **E Flapan**, **R Naimi**, **J Pommersheim**, **H Tamvakis**, *Topological symmetry groups of graphs embedded in the* 3*-sphere*, Comment. Math. Helv. 80 (2005) 317–354 MR

[10] **E Flapan**, **R Naimi**, **H Tamvakis**, *Topological symmetry groups of complete graphs in the* 3*-sphere*, J. London Math. Soc. 73 (2006) 237–251 MR

[11] **R Frucht**, **J E Graver**, **M E Watkins**, *The groups of the generalized Petersen graphs*, Proc. Cambridge Philos. Soc. 70 (1971) 211–218 MR

[12] **J W Morgan**, **F T-H Fong**, *Ricci flow and geometrization of* 3*-manifolds*, University Lecture Series 53, Amer. Math. Soc., Providence, RI (2010) MR

[13] **J Simon**, *Topological chirality of certain molecules*, Topology 25 (1986) 229–235 MR

[14] **P A Smith**, *Transformations of finite period*, Ann. of Math. 39 (1938) 127–164 MR

[15] **R Von Frucht**, *Die Gruppe des Petersenschen Graphen und der Kantensysteme der regulären Polyeder*, Comment. Math. Helv. 9 (1936) 217–223 MR

AA: *Department of Mathematics, Embry–Riddle Aeronautical University*
Prescott, AZ, United States

EF: *Department of Mathematics and Statistics, Pomona College*
Claremont, CA, United States

MH: *Department of Mathematics, Winston-Salem State University*
Winston-Salem, NC, United States

JH, EL: *Department of Mathematics and Statistics, University of San Francisco*
San Francisco, CA, United States

PL: *Department of Chemistry, Mathematics, and Physics, Houston Christian University*
Houston, TX, United States

CP: *Department of Mathematics and Statistics, Smith College*
Northampton, MA, United States

RV: *School of Interdisciplinary Arts and Sciences, University of Washington, Tacoma*
Tacoma, WA, United States

`angelynn.alvarez@erau.edu`, `eflapan@pomona.edu`, `hunnellm@wssu.edu`, `jhutchens@usfca.edu`, `edlawrence@usfca.edu`, `plewis@hc.edu`, `cprice@smith.edu`, `rvanderp@uw.edu`

Received: 3 April 2024 Revised: 15 July 2024

Algebraic & Geometric Topology **25**:8 (2025) 4949–5012
DOI: 10.2140/agt.2025.25.4949
Published: 20 November 2025

Crushing surfaces of positive genus

BENJAMIN A BURTON
THIAGO DE PAIVA
ALEXANDER HE
CONNIE ON YU HUI

The operation of crushing a normal surface has proven to be a powerful tool in computational 3-manifold topology, with applications both to triangulation complexity and to algorithms. The main difficulty with crushing is that it can drastically change the topology of a triangulation, so applications to date have been limited to relatively simple surfaces: 2-spheres, discs, annuli, and closed boundary-parallel surfaces. We give the first detailed analysis of the topological effects of crushing closed essential surfaces of positive genus. To showcase the utility of this new analysis, we use it to prove some results about how triangulation complexity interacts with JSJ decompositions and satellite knots; although similar applications can also be obtained using techniques of Matveev, our approach has the advantage that it avoids the machinery of almost simple spines and handle decompositions.

57K30, 57Q15

1 Introduction

The idea of crushing a normal surface was first developed by Jaco and Rubinstein [18] as part of a broader program of giving a theory of “efficient” 3-manifold triangulations. This led to new insights on minimal triangulations [18], and has also been the key to developing “efficient” (in various senses of the word, depending on the particular application) algorithms to solve a number of fundamental problems in low-dimensional topology [2; 3; 5; 6; 7; 8; 9; 13].

The key obstacle in developing new applications of crushing is that this operation can drastically alter the topology of a triangulation. This difficulty was initially compounded by the complicated formulation of crushing that was originally given by Jaco and Rubinstein; although they were able to give a number of applications, these required intricate arguments about the topological effects of crushing 2-spheres, discs and closed boundary-parallel surfaces [18]. More recent applications rely on simpler formulations of crushing that are easier to understand and use:

- Following unpublished ideas of Casson, Fowler [9] used the language of special spines to understand the effect of crushing 2-spheres.
- Burton introduced a way to break crushing down into a sequence of simple atomic moves, and used this atomic approach to describe the topological effects of crushing 2-spheres and discs [3]; this has proven to be extremely useful for turning crushing into an accessible algorithmic tool for working with 3-manifolds [2; 3; 5; 6; 7; 8]. This atomic approach has also recently been applied to crushing certain types of properly embedded annuli [13].

We emphasise that although it is, in principle, possible to crush any normal surface, the applications to date have only involved 2-spheres, discs, annuli and closed boundary-parallel surfaces. Probably the main reason for this is that as the surfaces get more complicated, the topological effects of crushing also appear to get more complicated. Nevertheless, we demonstrate in this paper that it is possible to push through this challenge by building upon the atomic approach to crushing from [3].

To be precise, we use the atomic approach to understand the topological effects of crushing closed normal surfaces of positive genus; in particular, we are able to crush essential surfaces, not just boundary-parallel ones. This work is distributed across two sections of this paper. First, in Section 3, we carefully work through the necessary details to extend the atomic approach to crushing. Then, in Section 4, we apply the work from Section 3 to actually understand the effect of crushing a surface of positive genus.

To state the main theorem from Section 4, we require some notation and terminology which we now outline (see Section 4 for the precise definitions). Given a normal surface S in a 3-manifold triangulation $\mathcal{T}$, our goal is to triangulate a submanifold X of $\mathcal{T}$ that is "cut out" by the surface S. More precisely, we fix a component of $\mathcal{T} - S$, which we call the *chosen region*, and then take X to be the closure of the chosen region. After crushing S to obtain a new triangulation $\mathcal{T}'$, each component of $\mathcal{T} - S$ "falls apart" to yield some subset of the components of $\mathcal{T}'$. In particular, the chosen region yields some subset $\mathcal{T}^*$ of $\mathcal{T}'$, and our hope is that (a component of) $\mathcal{T}^*$ actually gives a triangulation of X.

For our purposes, it turns out to be important to "push" or "expand" S as far into the chosen region as possible, to obtain what we call a *maximal* surface. We show in Lemma 14 that we can always, without loss of generality, assume that S is maximal. With this groundwork, together with our analysis from Section 3, we are able to prove the following theorem in Section 4:

Theorem 1 *Suppose that X is irreducible, ∂-irreducible and anannular, and that it contains no two-sided properly embedded Möbius bands. Also, suppose S is maximal. Then $\mathcal{T}^*$ is a valid triangulation such that:*

- *One of its components is an ideal triangulation of X.*
- *Every other component is a triangulation of the 3-sphere.*

In Section 5, we apply Theorem 1 to study the *triangulation complexity* $\Delta(\mathcal{M})$: the minimum number of tetrahedra required to triangulate some particular 3-manifold $\mathcal{M}$. In particular, we obtain the following general result as a relatively straightforward consequence of Theorem 1:

Theorem 2 *Let $\mathcal{M}$ be a compact 3-manifold with no 2-sphere boundary components. Suppose $\mathcal{M}$ contains a (possibly disconnected) closed incompressible surface S with no 2-sphere components, no projective plane components, and no boundary-parallel components. Let $\mathcal{R}$ be a component obtained after cutting $\mathcal{M}$ along S. If $\mathcal{R}$ is irreducible, ∂-irreducible, anannular, and does not contain any two-sided properly embedded Möbius bands, then $\Delta(\mathcal{R}) < \Delta(\mathcal{M})$.*

We continue in Section 5 by specialising Theorem 2 to the particularly interesting setting where S is a collection of essential tori. This gives various nice results about how triangulation complexity interacts with JSJ decompositions and satellite knots.

The applications that we obtain in Section 5 are not entirely new, since they can also be obtained by combining various pieces of machinery from Matveev's book [28] (we discuss this in a little more detail in Section 5). Nevertheless, our applications demonstrate that crushing normal surfaces of positive genus has nontrivial consequences for objects that are of independent interest. This provides hope that future refinements of our techniques could lead to further applications, such as new algorithms involving decompositions along surfaces of positive genus.

It is worth noting that whilst Matveev's techniques use almost simple spines and handle decompositions, our work does not require such machinery; instead, our analysis of crushing only uses triangulations and cell decompositions. Some readers might therefore find our approach more accessible than that of Matveev. Moreover, in contrast to handle decompositions, crushing has the advantage that it is well established in software such as Regina [2; 4]; thus, our approach is probably more amenable for practical algorithmic applications.

Acknowledgements

This project is the result of a conversation that began at the *PhD Student Symposium*: *Graduate Talks in Geometry and Topology Get-Together*, or $(GT)^3$, hosted by MATRIX in July 2022; we would therefore like to thank the MATRIX Institute and the organisers of the symposium.

Alexander He was supported by an Australian Federal Government Research Training Program Scholarship.

We would like to thank the reviewer for their many suggestions, which greatly helped to improve the readability of this paper.

2 Preliminaries

The main purpose of this section is to review all the definitions that we will require for our analysis of crushing.

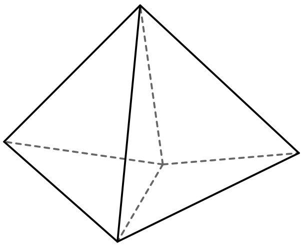


Figure 1: Two tetrahedra glued together along a single pair of triangular faces.

As a convention that we will use throughout this paper, except where we explicitly state otherwise, all 3-manifolds will be compact. We will call a (compact) 3-manifold *closed* if it has empty boundary, and *bounded* if it has nonempty boundary.

Also, whenever we are working with an object X (such as a knot or a surface) embedded in a 3-manifold $\mathcal{M}$, we will often refer to ambient isotopies of X in $\mathcal{M}$ simply as isotopies of X. For example, when we speak of isotoping a knot K (embedded in the 3-sphere S^3), we really mean that we are applying an *ambient* isotopy to the embedding of K in S^3.

2.1 Triangulations and cell decompositions

A (*generalised*) *triangulation* $\mathcal{T}$ consists of finitely many (abstract) tetrahedra with some or all of their triangular faces *glued* together in pairs via affine identifications (Figure 1 illustrates a single such gluing); denote the number of tetrahedra in $\mathcal{T}$ by $|\mathcal{T}|$. We allow faces from the same tetrahedron to be glued together, which means that $\mathcal{T}$ need not be a simplicial complex; indeed, generalised triangulations can usually be made much smaller than topologically equivalent simplicial complexes, which is often important for computational purposes.

In this paper, we also work with cell decompositions, which generalise the triangulations that we just defined. We build gradually towards a definition of cell decompositions, starting with an explanation of how we generalise tetrahedra to obtain a larger class of "building blocks".

Topologically, we can think of a tetrahedron as a 3-ball whose boundary 2-sphere is divided into triangles by an embedding of the complete graph on four vertices. To generalise this, consider a topological 3-ball Δ with a multigraph Γ embedded in $\partial\Delta$. We call Δ an (*abstract*) 3-*cell* if:

- Γ has no degree one vertices.
- The closure of each component of $(\partial\Delta) - \Gamma$ forms an embedded disc, which we call a *face* of Δ, whose boundary circle contains two or more vertices of Γ.

Assuming that these conditions are indeed satisfied, we refer to the vertices and edges of Γ as *vertices* and *edges*, respectively, of the 3-cell Δ. Intuitively, each face of an abstract 3-cell forms a curvilinear

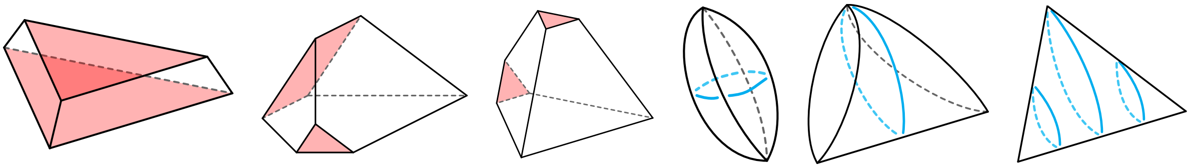


Figure 2: Some examples of the nontetrahedron cells that we will encounter.

polygon with two or more edges; indeed, depending on the number of edges, we will often describe 3-cell faces as bigons, triangles, quadrilaterals, and so on.

There are infinitely many types of 3-cells. However, for our purposes, we will only need to deal with a finite number of these; some examples are shown in Figure 2. For details on precisely which types of 3-cells we need, see Definitions 3 and Section 2.5.

We now explain how we glue 3-cells together to obtain a cell decomposition. Endow every edge e of a 3-cell with an affine structure — a homeomorphism from e to the interval $[0,1]$. We *glue* two distinct faces of two (not necessarily distinct) 3-cells via a homeomorphism that

- maps vertices to vertices;
- maps edges to edges; and
- restricts to an affine map on each edge.

A *cell decomposition* is a collection of finitely many 3-cells with some or all of their faces glued together in pairs; we emphasise again that we allow faces from the same 3-cell to be glued together. Since triangulations are a special case of cell decompositions, all of the subsequent definitions for cell decompositions apply to triangulations too.

Let $\mathcal{D}$ denote a cell decomposition. The gluings that define $\mathcal{D}$ give an equivalence relation on the faces of the 3-cells of $\mathcal{D}$; call each equivalence class a *face* or 2*-cell* of $\mathcal{D}$. More explicitly, a face of $\mathcal{D}$ is either:

- A pair of 3-cell faces that have been glued together, in which case we say that the face is *internal*.
- A single 3-cell face that has been left unglued, in which case we say that the face is *boundary*.

The *boundary* of $\mathcal{D}$ is the (possibly empty) union of all its boundary faces.

The gluings that define $\mathcal{D}$ also merge vertices and edges of the 3-cells into equivalence classes; call each such vertex class a *vertex* or 0*-cell* of $\mathcal{D}$, and call each such edge class an *edge* or 1*-cell* of $\mathcal{D}$. For each $k \in \{0,1,2\}$, define the k*-skeleton* of $\mathcal{D}$, denoted by $\mathcal{D}^{(k)}$, to be the union of all n-cells of $\mathcal{D}$, where n runs over all dimensions up to and including k.

In general, if we consider the quotient topology arising from the face gluings that define a cell decomposition $\mathcal{D}$, the resulting topological space might fail to be a 3-manifold. Specifically, although nothing goes wrong in the interiors of 3-cells and the interiors of faces, we need to be careful with vertices and with midpoints of edges.

We begin by considering the midpoint p of an edge e. If e lies entirely in the boundary of $\mathcal{D}$, then the frontier of a small regular neighbourhood of p is a disc; in this case, nothing goes wrong, and we say that e is *boundary*. However, if e does not lie in the boundary, then we have two possibilities:

- If e is identified with itself in reverse, then the frontier of a small regular neighbourhood of p is a projective plane; this cannot occur in a 3-manifold. In this case, we say that e is *invalid*.
- Otherwise, the frontier of a small regular neighbourhood of p is a 2-sphere. In this case, nothing goes wrong and we say that e is *internal*.

We also say that e is *valid* if it is either boundary or internal.

For a vertex v, consider the surface given by the frontier of a small regular neighbourhood of v; we call this surface the *link* of v. When v lies in the boundary of $\mathcal{D}$, its link is a surface with boundary. If the link is a disc, then nothing goes wrong and we say that v is *boundary*; otherwise, if the link is any other surface with boundary, then we say that the vertex is *invalid*.

On the other hand, when v does not lie in the boundary, its link is a closed surface. If the link is a 2-sphere, then nothing goes wrong and we say that v is *internal*; otherwise, if the link is any other closed surface, then we say that v is *ideal*.

A cell decomposition is *valid* if it has no invalid edges or vertices, and *invalid* otherwise. Given a (possibly invalid) cell decomposition $\mathcal{D}$, we often find it useful to *truncate* a vertex v by deleting a small open regular neighbourhood of v. In particular, by truncating each ideal or invalid vertex in $\mathcal{D}$, we obtain a pseudomanifold $\mathcal{P}$ that we call the *truncated pseudomanifold* of $\mathcal{D}$; the reason $\mathcal{P}$ is a pseudomanifold (and not necessarily a manifold) is that midpoints of invalid edges in $\mathcal{D}$ would give nonmanifold points in $\mathcal{P}$.

Observe that if $\mathcal{D}$ has no invalid edges, then the truncated pseudomanifold $\mathcal{P}$ is actually a (compact) 3-manifold. In this case, we will often refer to $\mathcal{P}$ as the *truncated* 3*-manifold* of $\mathcal{D}$, and we will say that $\mathcal{D}$ *represents* the 3-manifold $\mathcal{P}$; when $\mathcal{D}$ happens to be a triangulation, we will also often say that $\mathcal{D}$ *triangulates* $\mathcal{P}$. Moreover, in the case where $\mathcal{D}$ is valid and has no ideal vertices, since we do not need to truncate any vertices to obtain the truncated 3-manifold $\mathcal{P}$, we will sometimes find it more natural to refer to $\mathcal{P}$ as the *underlying* 3*-manifold* of $\mathcal{D}$.

If we assume that $\mathcal{D}$ is actually valid (so it has neither invalid edges nor invalid vertices), then the boundary components of the truncated 3-manifold $\mathcal{P}$ come in two possible types, namely

- *ideal* boundary components, which are the boundary components that arise from truncating the ideal vertices; and
- *real* boundary components, which are built from boundary faces of $\mathcal{D}$.

In this case, it will be convenient to distinguish the following special types of cell decompositions:

- A valid cell decomposition is *closed* if every vertex is internal. For a closed cell decomposition, the truncated 3-manifold is a closed 3-manifold.
- A valid cell decomposition is *bounded* if it has at least one boundary vertex, and has no ideal vertices. For a bounded cell decomposition, the truncated 3-manifold is a bounded 3-manifold whose boundary components are all real.
- A valid cell decomposition is *ideal* if it has at least one ideal vertex, and has no boundary vertices. For an ideal cell decomposition, the truncated 3-manifold is again a bounded 3-manifold, but this time the boundary components are all ideal.

Remark When we have an ideal cell decomposition $\mathcal{D}$, we use the notion of the truncated 3-manifold to turn $\mathcal{D}$ into a *compact* 3-manifold $\mathcal{M}$. A very common alternative (which we do not use in this paper) is to turn $\mathcal{D}$ into a *noncompact* 3-manifold $\mathcal{M}'$ by simply deleting (rather than truncating) each ideal vertex. Observe that $\mathcal{M}'$ is homeomorphic to the interior of $\mathcal{M}$, so this distinction is not too important.

Remark Suppose $\mathcal{T}$ is either a closed or ideal triangulation, and let $\mathcal{M}$ denote the truncated 3-manifold of $\mathcal{T}$. Since we do not truncate the internal vertices of $\mathcal{T}$, observe that $\mathcal{M}$ is a 3-manifold with no 2-sphere boundary components. For this reason, we will often find it convenient to make the mild assumption that a 3-manifold has no 2-sphere boundary components.

2.2 Decomposing along curves and surfaces

The goal in this section is to introduce some terminology that will streamline our descriptions of the topological effects of crushing. The idea is that crushing often changes the truncated 3-manifold or pseudomanifold by "decomposing along" a properly embedded surface; we will build gradually towards defining precisely what we mean by this. We start by going one dimension down, and defining what we mean by decomposing a surface along an embedded curve; this is useful in its own right, since it will help us describe how crushing changes the links of vertices.

Consider an embedded closed curve γ in a compact surface S. Let $S^\dagger$ denote the surface obtained from S by *cutting along* γ — that is, removing a small open regular neighbourhood of γ from S. If γ is a two-sided curve in S, then we have two new copies of γ in $\partial S^\dagger$; on the other hand, if γ is one-sided, then we have a single new curve in $\partial S^\dagger$. Call each of these new curves in $\partial S^\dagger$ a *remnant* of γ; see Figure 3 Consider the surface S' given by *filling* each remnant of γ with a disc; we say that S' is obtained from S by *decomposing along* γ.

We now aim to define similar terminology for truncated pseudomanifolds. Consider a (possibly disconnected) properly embedded surface S in a truncated pseudomanifold $\mathcal{P}$. Let $\mathcal{P}^\dagger$ denote the pseudomanifold obtained from $\mathcal{P}$ by *cutting along* S — similar to before, this means that we obtain $\mathcal{P}^\dagger$ by removing a small open regular neighbourhood of S from $\mathcal{P}$. For each two-sided component E of S, we have two

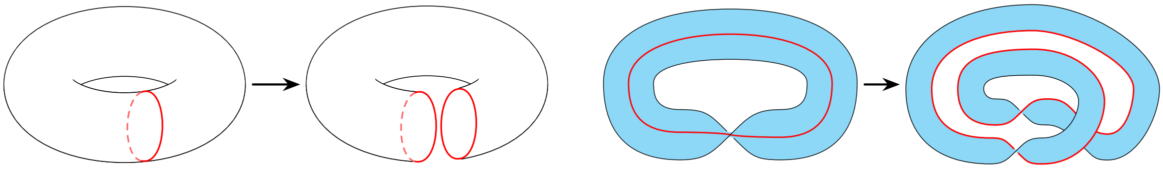


Figure 3: Cutting along an embedded closed curve in a surface. Left: cutting along a two-sided curve yields a pair of remnants. Right: cutting along a one-sided curve yields a single remnant.

new copies of E in $\partial\mathcal{P}^\dagger$; on the other hand, for each one-sided component E of S, we have a single new double cover of E in $\partial\mathcal{P}^\dagger$. Call each of these new pieces in $\partial\mathcal{P}^\dagger$ a *remnant* of S.

For our purposes, it will be useful to have a notion of "decomposing along" S when S is one of the following seven types of (properly embedded) surface:

- A 2-sphere — which means that cutting along S yields a pair of 2-sphere remnants.
- A two-sided annulus — which means that cutting along S yields a pair of annulus remnants.
- A one-sided annulus — which means that cutting along S yields a single annulus remnant.
- A two-sided projective plane — which means that cutting along S yields a pair of projective plane remnants.
- A one-sided projective plane — which means that cutting along S yields a single 2-sphere remnant.
- A two-sided Möbius band — which means that cutting along S yields a pair of Möbius band remnants.
- A one-sided Möbius band — which means that cutting along S yields a single annulus remnant.

Notice that for these types of surface, the remnants are always either 2-spheres, annuli, projective planes or Möbius bands.

Similar to what we did with curves on surfaces, we construct the result of "decomposing along" S by "filling" the remnants of S. To do this for projective plane and Möbius band remnants, we use the following terminology: define an *invalid cone* to be a pseudomanifold given by taking a cone over a projective plane. With this in mind, let $S^\dagger$ denote a remnant of S in $\mathcal{P}^\dagger$, and suppose $S^\dagger$ is either a 2-sphere, annulus, projective plane or Möbius band. We define the operation of *filling* $S^\dagger$ as follows:

- If $S^\dagger$ is a 2-sphere, then filling means attaching a 3-ball B by identifying $S^\dagger$ with the 2-sphere boundary of B.
- If $S^\dagger$ is an annulus, then filling means attaching a thickened disc $D \times [0,1]$ by identifying $S^\dagger$ with the annulus $(\partial D) \times [0,1]$.
- If $S^\dagger$ is a projective plane, then filling means attaching an invalid cone $\mathcal{C}$ by identifying $S^\dagger$ with the projective plane boundary of $\mathcal{C}$.
- If $S^\dagger$ is a Möbius band, then filling means attaching an invalid cone $\mathcal{C}$ by choosing a small open disc D in $\partial\mathcal{C}$, and identifying $S^\dagger$ with the Möbius band given by $(\partial\mathcal{C}) - D$.

Putting everything together, suppose S is one of the seven types of surface listed above, and let $\mathcal{P}'$ denote the pseudomanifold obtained from $\mathcal{P}^{\dagger}$ by filling each remnant of S. We say that $\mathcal{P}'$ is obtained from $\mathcal{P}$ by *decomposing along* S.

2.3 Normal surfaces

A *normal surface* in a triangulation $\mathcal{T}$ is a (possibly disconnected) properly embedded surface that

- is disjoint from the vertices of $\mathcal{T}$;
- meets the edges and faces of $\mathcal{T}$ transversely; and
- intersects each tetrahedron Δ of $\mathcal{T}$ in a (possibly empty) disjoint union of finitely many discs, called *elementary discs*, where each such disc forms a curvilinear triangle or quadrilateral whose vertices lie on different edges of Δ.

Two normal surfaces are *normally isotopic* if they are related by a *normal isotopy* — that is, an ambient isotopy that preserves each vertex, edge, face and tetrahedron of the triangulation. Up to normal isotopy, the elementary discs in each tetrahedron Δ come in seven possible types,

- four *triangle types*, each of which separates one vertex of Δ from the other three, as shown in Figure 4 (left image); and
- three *quadrilateral types*, each of which separates a pair of opposite edges of Δ, as shown in Figure 4 (middle three images).

Observe that if a tetrahedron contains two elementary quadrilaterals of different types, then these two quadrilaterals will always intersect each other; since normal surfaces are embedded, this means that if a tetrahedron contains quadrilaterals, then these quadrilaterals must all be of the same type.

We call a normal surface *nontrivial* if it includes at least one elementary quadrilateral, and *trivial* otherwise. It is easy to see that trivial normal surfaces always exist, and that every component of such a surface is just a vertex link. The existence of nontrivial normal surfaces is less obvious. In fact, it is possible to prove that many "interesting" embedded surfaces appear as (nontrivial) normal surfaces; we will get a glimpse of why this is the case when we discuss the theory of barriers and normalisation in Section 2.4.

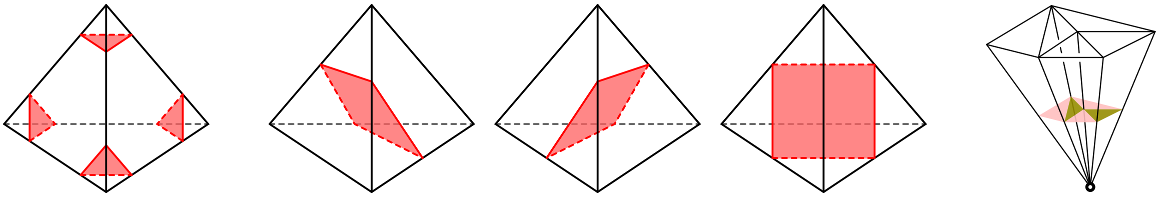


Figure 4: The seven types of elementary disc. Left image: the four triangle types. Middle three images: the three quadrilateral types. Right image: a portion of a normal surface built entirely out of triangles.

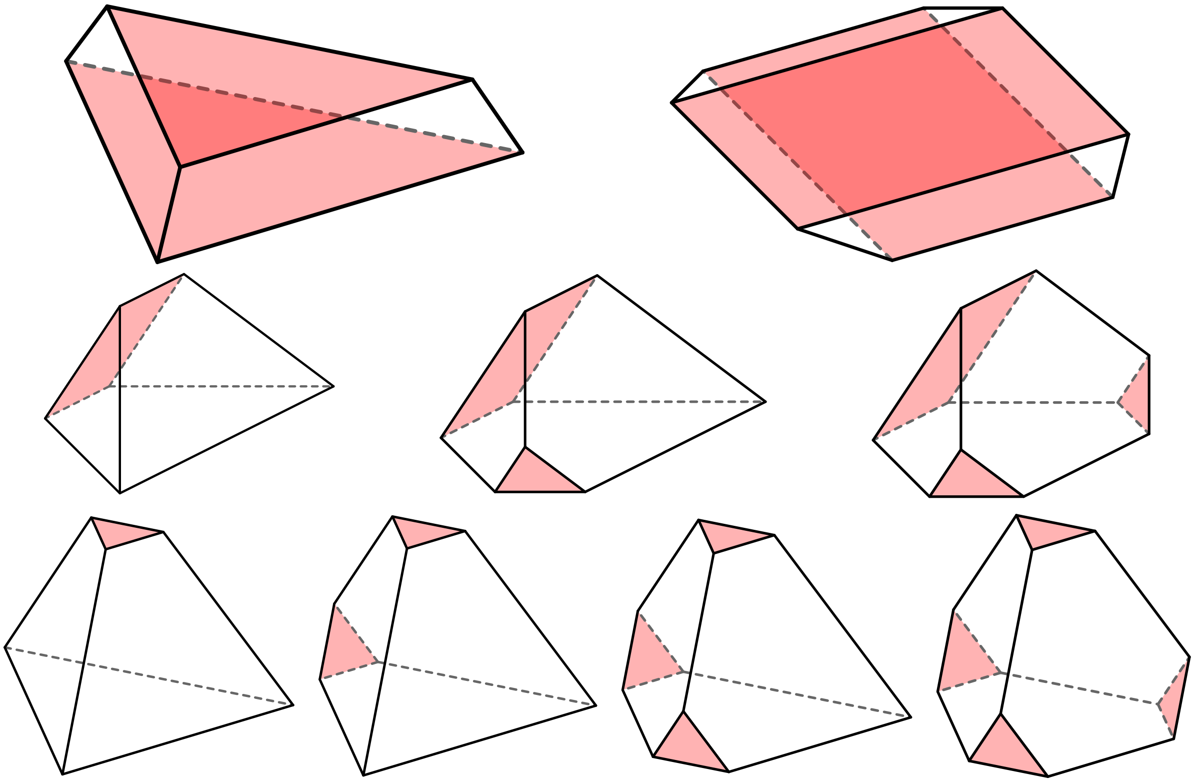


Figure 5: A normal surface S can induce parallel cells (two types), wedge cells (three types) and central cells (five types). The faces that lie inside S are shaded red. First row: a parallel triangular cell (left) and a parallel quadrilateral cell (right). All nonshaded faces are bridge faces. Second row: a wedge cell with no bridge faces (left), a wedge cell with one bridge face (middle) and a wedge cell with two bridge faces (right). Third row: four types of nontetrahedron central cell. A tetrahedron is the fifth type of central cell.

A normal surface naturally splits a triangulation into a finer cell decomposition. To describe this idea more precisely, we introduce the following definitions, which are partly based on some terminology used by Jaco and Rubinstein [18, page 91]:

Definitions 3 Let S be a normal surface in a triangulation $\mathcal{T}$. The surface S divides each tetrahedron Δ of $\mathcal{T}$ into a collection of *induced cells* of the following types:

- *Parallel cells* of two types (see Figure 5 (first row)):
  - *Parallel triangular cells* lie between two parallel triangles of S.
  - *Parallel quadrilateral cells* lie between two parallel quadrilaterals of S.
- *Nonparallel cells* of nine types:
  - *Corner cells* are tetrahedra that lie between a single triangle of S and a single vertex of Δ.

- *Wedge cells* of three types (see Figure 5 (second row)) only occur when S meets Δ in one or more quadrilaterals. In this case, if we ignore any parallel and corner cells in Δ, then the two cells left over are the wedge cells.
- *Central cells* of five types (see Figure 5 (third row)) only occur when S does not meet Δ in any quadrilaterals. In this case, if we ignore any parallel and corner cells in Δ, then the single cell left over is the central cell.

Amongst the faces of these induced cells, we will find it useful to distinguish the *bridge faces*, which are the quadrilateral faces that intersect S precisely in a pair of opposite edges. Note that bridge faces only appear in parallel and wedge cells (see Figure 5 (first and second rows)).

Let $\mathcal{P}$ denote the truncated pseudomanifold of $\mathcal{T}$, and let $\mathcal{P}^\dagger$ denote the pseudomanifold obtained from $\mathcal{P}$ by cutting along S. The induced cells naturally yield a cell decomposition $\mathcal{D}$ of $\mathcal{P}$, such that the surface S is given by a union of faces of $\mathcal{D}$. Moreover, ungluing the faces of $\mathcal{D}$ that lie inside S yields a cell decomposition $\mathcal{D}^\dagger$ of $\mathcal{P}^\dagger$. We say that the cell decompositions $\mathcal{D}$ and $\mathcal{D}^\dagger$, and any cell decompositions given by components of $\mathcal{D}$ and $\mathcal{D}^\dagger$, are *induced* by the normal surface S.

Since a tetrahedron can contain many parallel elementary discs, we could have arbitrarily many parallel cells. However, there are always at most six nonparallel cells per tetrahedron Δ:

- If Δ meets the normal surface in one or more quadrilaterals, then we have no central cells, exactly two wedge cells, and up to four corner cells.
- If Δ does not meet the normal surface in any quadrilaterals, then we have no wedge cells, exactly one central cell, and again up to four corner cells.

We will find this simple observation useful in Section 4.1.

2.4 Barriers and normalisation

We now review the theory of normalisation, which gives a procedure for transforming any properly embedded surface S into a normal surface (not necessarily isotopic to S). We also review the notion of a barrier surface, which gives a tool for "controlling" the result of the normalisation procedure. The material here is essentially an abridged and informal version of Section 3 of [18], focusing only on the details that are necessary for our purposes in this paper.

Throughout Section 2.4, let S, S' and B denote (possibly disconnected) surfaces that are properly embedded in a triangulation $\mathcal{T}$. Assume that these surfaces are disjoint from the vertices of $\mathcal{T}$, and transverse to the 2-skeleton of $\mathcal{T}$.

The idea of the normalisation procedure is to reduce the number of "anomalies" in a surface S until it becomes a normal surface. For instance, for S to be a normal surface, it cannot intersect any tetrahedron Δ in anomalous pieces such as

- a 2-sphere component that is *trivial* in the sense that it lies entirely inside Δ; or
- a disc component that is *trivial* in the sense that its boundary curve lies entirely inside a single boundary face, and its interior lies entirely inside Δ.

To keep track of these and other anomalous features of S, we use the following measures of "complexity":

- Define the *weight* $\mathrm{wt}(S)$ to be the number of times S meets the 1-skeleton $\mathcal{T}^{(1)}$: $\mathrm{wt}(S) = |S \cap \mathcal{T}^{(1)}|$.

In general, S could meet a tetrahedron Δ of $\mathcal{T}$ in a nonnormal piece that "doubles back" on itself to meet a single edge twice (for example, see Figure 6 (top right and bottom left)); the weight of S gives a proxy for counting the number of such anomalies.

- For each tetrahedron Δ of $\mathcal{T}$, let

$$x_\Delta = \sum_{c \neq S^2} (1 - \chi(c)),$$

where c runs over all components of $S \cap \Delta$ other than 2-spheres. Also define the *local Euler number*

$$\lambda(S) = \sum_{\Delta} x_\Delta.$$

Recall that a normal surface must, in particular, meet each tetrahedron of $\mathcal{T}$ in a disjoint union of discs; apart from trivial 2-spheres (which we handle separately), the local Euler number detects any anomalies that violate this requirement.

- Let $\sigma(S)$ denote the number of closed curves in which S intersects the internal faces of $\mathcal{T}$. A normal surface cannot have any such anomalous curves.
- Let $\tau(S)$ denote the number of components of S that form trivial 2-spheres or trivial discs.

Define the *complexity* of S, denoted $C(S)$, to be the tuple

$$(\mathrm{wt}(S), \lambda(S), \sigma(S), \tau(S)).$$

We will consider S to have smaller complexity than some other surface S' if $C(S)$ occurs before $C(S')$ in the lexicographical ordering. As suggested earlier, normalisation consists of a series of steps, each of which reduces the complexity.

Before we define the steps involved in normalisation, we introduce some useful terminology. Call a disc D an *edge-compression disc* for S if it is embedded so that

- the interior of D lies entirely in the interior of a tetrahedron Δ of $\mathcal{T}$; and
- the boundary of D consists of two arcs α and γ that intersect each other only at their endpoints, such that $\alpha = D \cap S$ and γ is a subarc of an edge e of Δ.

Examples of edge-compression discs are shown in Figure 6 (top right and bottom left). Call an edge-compression disc *internal* if it meets an internal edge of $\mathcal{T}$, and *boundary* if it meets a boundary edge of $\mathcal{T}$; notice that a boundary edge-compression disc is, in particular, a ∂-compression disc for S.

With all the preceding setup in mind, the normalisation procedure proceeds by performing the following *normal moves* on a surface S:

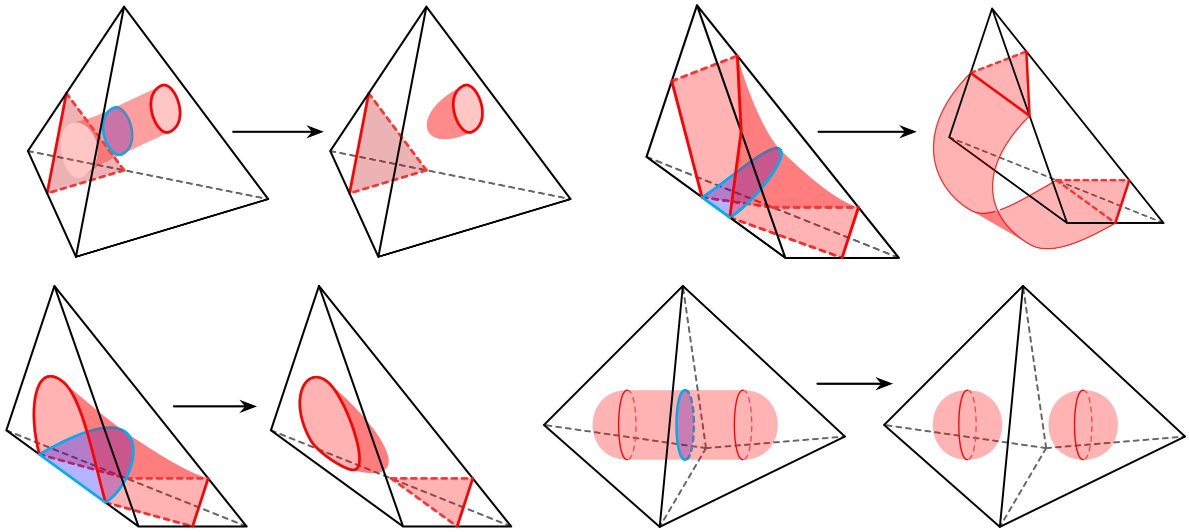


Figure 6: Examples of normal moves. First row, left images: type (1), a compression of a surface (shaded red) along a compression disc (shaded blue) lying entirely inside a tetrahedron. First row, right images: type (2), an isotopy of a surface (shaded red) along an internal edge-compression disc (shaded blue). Second row, left images: type (3), a ∂-compression of a surface (shaded red) along a boundary edge-compression disc (shaded blue). Second row, right images: type (4), a compression of a surface (shaded red) along a compression disc (shaded blue) lying entirely inside an internal face.

(1) **Compressions along discs that lie entirely in the interior of a tetrahedron** (see Figure 6 (top left)) Each such compression reduces the complexity $C(S)$ because it leaves the weight $\mathrm{wt}(S)$ unchanged and reduces the local Euler number $\lambda(S)$. These compressions can be performed until S meets each tetrahedron of $\mathcal{T}$ in a union of 2-spheres and discs, at which point $\lambda(S) = 0$. We assume for the rest of the normal moves that we have already reduced $\lambda(S)$ to 0 in this way.

(2) **Isotopies along internal edge-compression discs** (see Figure 6 (top right)) Each such isotopy reduces the complexity $C(S)$ because it reduces the weight $\mathrm{wt}(S)$.

(3) **∂-compressions along boundary edge-compression discs** (see Figure 6 (bottom left)) Like the isotopies in the previous step, each such ∂-compression reduces the complexity $C(S)$ because it reduces the weight $\mathrm{wt}(S)$. For the remaining two normal moves, we assume that we have performed all possible isotopies and ∂-compressions along edge-compression discs, which ensures that S meets each tetrahedron Δ of $\mathcal{T}$ in a union of

- elementary discs;
- trivial 2-spheres; and
- discs whose boundary curves lie entirely in the interior of some face of Δ.

(4) **Compressions along discs that lie entirely in the interior of an internal face** (see Figure 6 (bottom right)) Each such compression reduces the complexity $C(S)$ because it leaves $\mathrm{wt}(S)$ and $\lambda(S)$ unchanged, and reduces $\sigma(S)$. After performing these compressions until no more such moves are possible, S meets each tetrahedron of $\mathcal{T}$ in a union of elementary discs, trivial 2-spheres, and trivial discs; we assume for the final normal move that this has already been done.

(5) **Deletion of trivial 2-sphere and disc components** This final "clean-up" step reduces the complexity $C(S)$ because it leaves $\mathrm{wt}(S)$, $\lambda(S)$ and $\sigma(S)$ unchanged, and reduces $\tau(S)$ to zero. At the end of this step, S is a normal surface.

For a complete explanation of why normalisation works as we have claimed, see Section 3.2 of [18]. In general, the normal surface that we obtain might not be isotopic to the original surface, because of the steps where we perform compressions and ∂-compressions. However, if we assume that the original surface was incompressible and ∂-incompressible, and also that the ambient 3-manifold is irreducible and ∂-irreducible, then normalising must produce a normal surface with one component isotopic to the original surface.

We can get even more control over the result of normalisation using the notion of a barrier surface; we now review the aspects of barrier surfaces that we require for our purposes. Given a properly embedded surface B in $\mathcal{T}$, let $\mathcal{N}$ denote a fixed but arbitrary component of $\mathcal{T} - B$. Call B a *barrier* for $\mathcal{N}$ if any surface S that is properly embedded in $\mathcal{N}$ can actually be normalised inside $\mathcal{N}$; that is, the discs along which we compress, isotope or ∂-compress always lie entirely inside $\mathcal{N}$, and at every stage the surface S remains properly embedded in $\mathcal{N}$.

In Theorem 3.2 from [18], Jaco and Rubinstein list a number of examples of barrier surfaces. For our purposes, we will need part (5) of this theorem, which we restate here:

Theorem 4 *Consider a (compact) 3-manifold $\mathcal{M}$ with no 2-sphere boundary components. If $\mathcal{M}$ is closed, let $\mathcal{T}$ be a closed triangulation of $\mathcal{M}$; otherwise, if $\mathcal{M}$ is bounded, let $\mathcal{T}$ be an ideal triangulation of $\mathcal{M}$. Let S be a normal surface in $\mathcal{T}$, and let A be a subcomplex of the cell decomposition of $\mathcal{M}$ induced by S. The boundary B of a small regular neighbourhood of $S \cup A$ is a barrier surface for any component of $\mathcal{M} - B$ that does not meet $S \cup A$.*

2.5 Crushing via atomic moves

The main purpose of this section is to review the atomic formulation of crushing that was introduced by Burton [3]. We augment this with some new terminology, as this will be useful for our purposes in Section 4. To begin, we state a version of Definition 1 from [3]:

Definitions 5 (crushing procedure) Let S be a normal surface in a triangulation $\mathcal{T}$. Each of the following operations builds on the previous one:

(1) Cut along S, and let $\mathcal{D}$ denote the resulting induced cell decomposition.

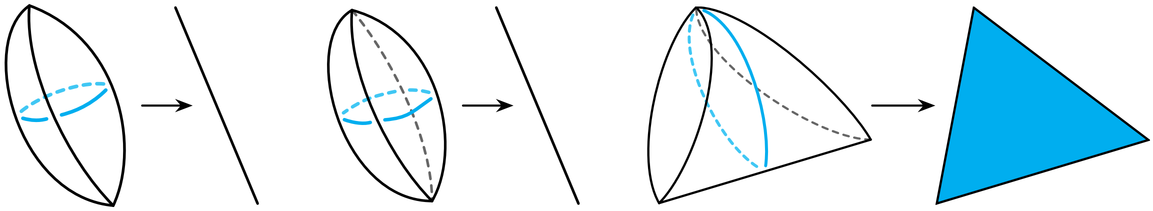


Figure 7: In addition to tetrahedra, a destructible cell decomposition $\mathcal{D}^*$ can contain three other types of 3-cells. To recover a triangulation from $\mathcal{D}^*$, we need to flatten the nontetrahedron cells. Left: flattening a 3-sided football to an edge. Middle: flattening a 4-sided football to an edge. Right: flattening a triangular purse to a triangular face.

(2) Using the quotient topology, collapse each remnant of S to a point. This turns $\mathcal{D}$ into a new cell decomposition $\mathcal{D}'$ with 3-cells of the following four possible types (see Figure 7):

- 3-*sided footballs*, which are obtained from corner cells and parallel triangular cells;
- 4-*sided footballs*, which are obtained from parallel quadrilateral cells;
- *triangular purses*, which are obtained from wedge cells; and
- *tetrahedra*, which are obtained from central cells.

We say that $\mathcal{D}'$ is obtained by *nondestructively crushing* S. Also, if a cell decomposition $\mathcal{D}^*$ is built entirely from 3-cells of the four types listed above (even if it was not directly obtained by nondestructive crushing), then we call $\mathcal{D}^*$ a *destructible* cell decomposition.

(3) To recover a triangulation from a destructible cell decomposition $\mathcal{D}^*$, we first build an intermediate cell complex $\mathcal{C}^*$ by using the quotient topology to flatten

- all 3-sided and 4-sided footballs to edges; and
- all triangular purses to triangular faces.

This is illustrated in Figure 7. Since triangulations are defined only by face gluings between tetrahedra, there are two ways in which $\mathcal{C}^*$ might fail to form a triangulation:

(a) $\mathcal{C}^*$ could contain vertices, edges and/or triangles that are *isolated*, meaning that they do not belong to any tetrahedra.

(b) $\mathcal{C}^*$ could contain vertices or edges that are *pinched*, meaning that they include identifications that are independent of any face gluings. In contrast, recall that every vertex of a triangulation is an equivalence class of vertices of tetrahedra, where all vertex identifications arise as consequences of face gluings. Similarly, every edge of a triangulation is given by edge identifications arising solely as consequences of face gluings.

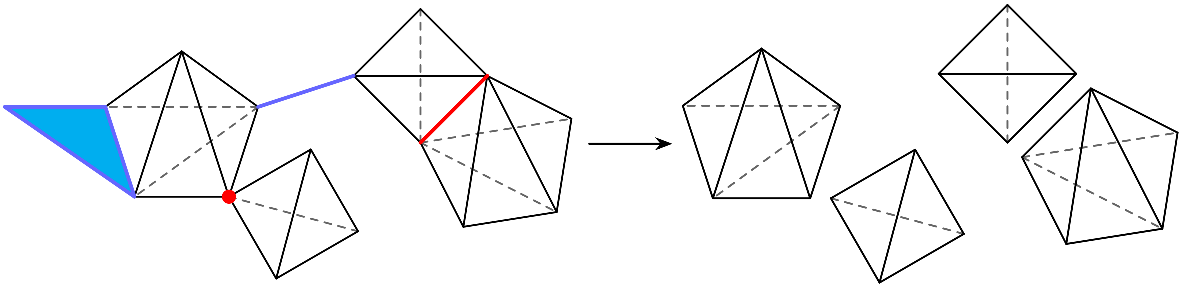


Figure 8: Extracting a triangulation by deleting isolated edges and triangles (highlighted in blue), and separating pinched vertices and edges (highlighted in red).

Thus, as illustrated in Figure 8, we need to perform the following two operations to *extract* a triangulation $\mathcal{T}^*$ from $\mathcal{C}^*$:

(a) Delete all isolated vertices, edges and triangles.

(b) Separate pieces of the cell complex that are only joined together along pinched vertices or edges (thereby ensuring that all vertex and edge identifications arise solely as consequences of face gluings).

We say that $\mathcal{T}^*$ is obtained by *flattening* $\mathcal{D}^*$. Consider the triangulation $\mathcal{T}'$ obtained by flattening the cell decomposition $\mathcal{D}'$ that results from nondestructively crushing S; we say that $\mathcal{T}'$ is obtained by (*destructively*) *crushing* S.

It is not too difficult to see what happens if we crush a *trivial* normal surface S in a triangulation $\mathcal{T}$. Cutting along S yields one central cell per tetrahedron, together with some number of corner and parallel triangular cells. All the corner and parallel cells together form components that do not contain any central cells, so after the nondestructive crushing and flattening steps, these components become isolated edges that do not appear in the final triangulation. For the central cells, observe that nondestructive crushing turns these into tetrahedra that are glued together in the same way as the original triangulation. The upshot is that crushing a trivial normal surface always leaves the triangulation unchanged.

Suppose now that S is a *nontrivial* normal surface in a triangulation $\mathcal{T}$, and let $\mathcal{T}'$ denote the triangulation obtained by crushing S. As before, each tetrahedron of $\mathcal{T}'$ comes from a central cell in the cell decomposition $\mathcal{D}$ given by cutting along S. However, this time, at least one tetrahedron of $\mathcal{T}$ contains an elementary quadrilateral, which means that not every tetrahedron of $\mathcal{T}$ gives rise to a central cell in $\mathcal{D}$. Thus, we see that crushing has the following useful feature:

Observation 6 *Let $\mathcal{T}$ be a triangulation, and let $\mathcal{T}'$ denote the triangulation obtained by crushing a nontrivial normal surface in $\mathcal{T}$. Then $|\mathcal{T}'| < |\mathcal{T}|$.*

The difficulty with crushing a nontrivial normal surface is that this operation could drastically change the topology of our triangulations. In particular, the triangulations before and after crushing could represent

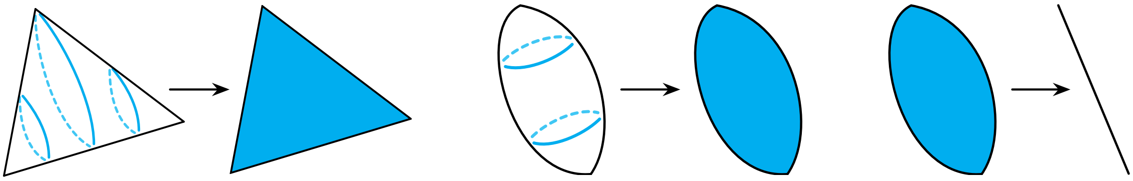


Figure 9: The three atomic moves for flattening a destructible cell decomposition. Left: flattening a triangular pillow. Middle: flattening a bigon pillow. Right: flattening a bigon face.

different 3-manifolds, assuming they even represent 3-manifolds at all. In [18], Jaco and Rubinstein work through this difficulty using a complicated global analysis of their version of the crushing procedure.

In contrast, the formulation of crushing given in Definitions 5 is simpler to work with. This is because the process of flattening a destructible cell decomposition can always be realised by a sequence consisting of atomic moves of three types. The following lemma [3, Lemma 3] gives a precise statement of this idea:

Lemma 7 (crushing lemma) *Let $\mathcal{T}^*$ be the triangulation given by flattening some destructible cell decomposition $\mathcal{D}^*$. Then $\mathcal{T}^*$ can be obtained from $\mathcal{D}^*$ by performing a sequence of zero or more of the following **atomic moves** (see Figure 9), one at a time, in some order:*

- *flattening a **triangular pillow** to a triangular face;*
- *flattening a **bigon pillow** to a bigon face; and*
- *flattening a bigon face to an edge.*

*Since our cell decompositions are defined only by face gluings between 3-cells, after each atomic move we implicitly **extract** a cell decomposition by*

- *deleting all **isolated** vertices, edges, bigons and triangles that do not belong to any 3-cells; and*
- *separating pieces of the cell complex that are only joined together along **pinched** vertices or edges.*

As part of the proof of the crushing lemma, Burton showed [3] that if we are careful about the order in which we perform the atomic moves, then we only ever encounter cell decompositions with 3-cells of the following seven types:

- 3-sided footballs;
- 4-sided footballs;
- triangular purses;
- tetrahedra;
- triangular pillows;
- bigon pillows; and
- *bigon pyramids* (see Figure 10).

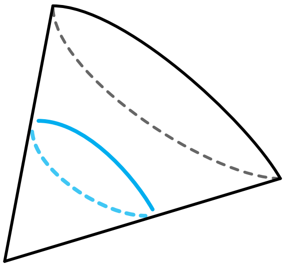


Figure 10: A bigon pyramid.

The crushing lemma allows us to understand the topological effects of crushing by examining atomic moves one at a time. In particular, Burton proved the following result [3, Lemma 4] (which, among other things, paved the way for a practical algorithm for nonorientable prime decomposition [2; 3]):

Lemma 8 *Let $\mathcal{D}_0$ be a valid cell decomposition with no ideal vertices. If the underlying* 3*-manifold $\mathcal{M}_0$ contains no two-sided projective planes, then performing one of the atomic moves of Lemma* 7 *will yield a* (*valid*) *cell decomposition of a* 3*-manifold $\mathcal{M}_1$ such that one of the following holds*:

- $\mathcal{M}_0 = \mathcal{M}_1$.
- *We flattened a triangular pillow, and $\mathcal{M}_1$ is obtained from $\mathcal{M}_0$ by deleting a single component $\mathcal{C}$, where $\mathcal{C}$ is either a* 3*-ball, a* 3*-sphere or a copy of the lens space $L_{3,1}$.*
- *We flattened a bigon pillow, and $\mathcal{M}_1$ is obtained from $\mathcal{M}_0$ by deleting a single component $\mathcal{C}$, where $\mathcal{C}$ is either a* 3*-ball, a* 3*-sphere, or a copy of real projective space $\mathbb{R}P^3$.*
- *We flattened a bigon face, and $\mathcal{M}_1$ is related to $\mathcal{M}_0$ in one of the following ways*:
  - (i) *$\mathcal{M}_1$ is obtained by cutting along a properly embedded disc in $\mathcal{M}_0$.*
  - (ii) *$\mathcal{M}_1$ is obtained by filling a boundary* 2*-sphere of $\mathcal{M}_0$ with a* 3*-ball.*
  - (iii) *$\mathcal{M}_1$ is obtained by decomposing along an embedded* 2*-sphere in $\mathcal{M}_0$.*
  - (iv) *$\mathcal{M}_0 = \mathcal{M}_1 \# \mathbb{R}P^3$ — that is, $\mathcal{M}_1$ removes a single $\mathbb{R}P^3$ summand from the connected sum decomposition of $\mathcal{M}_0$.*

One of our main goals in this paper is to extend Lemma 8 to cell decompositions that may be invalid and may have ideal vertices; this, in particular, allows us to study the topological effects of crushing a closed surface of positive genus, since nondestructively crushing such a surface produces a cell decomposition with ideal vertices. To do this, we will find it helpful to have "flattening maps" that keep track of how the points in a cell decomposition are affected by an atomic move. Although an atomic move "looks like" a quotient operation, the corresponding quotient map does not account for the implicit operation of extracting a cell decomposition, so a little care is required to define "flattening maps" appropriately:

Definitions 9 Let $\mathcal{D}_1$ be a cell decomposition obtained by performing a single atomic move on some cell decomposition $\mathcal{D}_0$. In Lemma 7, each atomic move implicitly finishes with the operation of extracting a cell decomposition; consider the intermediate cell complex $\mathcal{C}$ that we obtain by performing the atomic move *without* subsequently extracting a cell decomposition. Note that $\mathcal{C}$ is obtained as a quotient of $\mathcal{D}_0$, so we have a quotient map $q\colon \mathcal{D}_0 \to \mathcal{C}$.

We use q to construct a map $\hat{\varphi}_0\colon \mathcal{D}_0 \to 2^{\mathcal{D}_1}$ (here, 2^X denotes the *power set* of a set X) that acts on points p in $\mathcal{D}_0$ as follows:

- If $q(p)$ is part of an isolated vertex, edge, bigon or triangle — which means that $q(p)$ is deleted when we extract a cell decomposition — then take $\hat{\varphi}_0(p)$ to be the empty set.
- If $q(p)$ is part of a pinched edge or vertex — which means that $q(p)$ gets separated into multiple points when we extract a cell decomposition — then take $\hat{\varphi}_0(p)$ to be the set of points in $\mathcal{D}_1$ that originate from $q(p)$.
- Otherwise, $q(p)$ remains untouched when we extract a cell decomposition, in which case we take $\hat{\varphi}_0(p) = \{q(p)\}$ (here, by an abuse of notation, we are viewing $q(p)$ as a point in $\mathcal{D}_1$).

Intuitively, $\hat{\varphi}_0$ keeps track of how points in $\mathcal{D}_0$ are affected when we perform an atomic move.

Observe that the nonempty sets in the image of $\hat{\varphi}_0$ give a partition of the points in $\mathcal{D}_1$. Thus, we can construct a map $\hat{\varphi}_1\colon \mathcal{D}_1 \to 2^{\mathcal{D}_0}$ as follows: for each point p in $\mathcal{D}_1$, let U be the (unique) set in the image of $\hat{\varphi}_0$ that contains p, and define $\hat{\varphi}_1(p)$ to be the set $\hat{\varphi}_0^{-1}(U)$. Intuitively, $\hat{\varphi}_1$ keeps track of how points in $\mathcal{D}_1$ would be affected if we perform an atomic move in reverse.

For each $i \in \{0, 1\}$, define a map $\varphi_i\colon 2^{\mathcal{D}_i} \to 2^{\mathcal{D}_{1-i}}$ that sends any subset S of $\mathcal{D}_i$ (when we actually use the ideas defined here, S will usually be a vertex, edge, face or 3-cell of $\mathcal{D}_i$) to the set

$$\bigcup_{p \in S} \hat{\varphi}_i(p) \subseteq \mathcal{D}_{1-i}.$$

We call φ_0 the *flattening map* associated to the atomic move, and φ_1 the *inverse flattening map* (although, strictly speaking, these maps are not actually inverses of each other).

3 Atomic moves on cell decompositions with ideal vertices

Let S be a normal surface in a triangulation $\mathcal{T}$. When S is either a 2-sphere or a disc, nondestructively crushing S creates new vertices whose links are either 2-spheres or discs. Thus, if the vertices of $\mathcal{T}$ are all either internal or boundary, then the topological effect of destructively crushing S only depends on how atomic moves affect cell decompositions whose vertices are all either internal or boundary; this was the motivation for Lemma 8 in [3].

Our main goal in this section is to extend this atomic approach to crushing beyond the case where S is a 2-sphere or disc. This requires us to study atomic moves on cell decompositions that are allowed to have ideal or invalid vertices. A similarly general understanding of atomic moves is necessary to understand crushing if we allow the initial triangulation $\mathcal{T}$ to have ideal or invalid vertices. Moreover, when $\mathcal{T}$ triangulates a nonorientable 3-manifold, it turns out to be possible for an atomic move to create an invalid *edge*. The upshot is that, for a completely general analysis of atomic moves, we should not restrict the links of the vertices involved, and we should not exclude the possibility of invalid edges.

Of the three atomic moves, flattening a triangular pillow and flattening a bigon pillow are relatively straightforward to understand. We study these two atomic moves in full generality in Section 3.1.

We then devote Section 3.2 to understanding the topological effects of flattening a bigon face. In contrast to the other two atomic moves, there would be a tediously large number of cases to consider if we wanted to give a complete analysis. Thus, for the sake of brevity and clarity, we will focus mainly on flattening bigon faces in valid cell decompositions whose vertices are all either internal or ideal. This is sufficient to understand the effects of crushing S if the following conditions are satisfied:

- S is a closed surface.
- $\mathcal{T}$ is valid and has no boundary vertices.
- The truncated 3-manifold of $\mathcal{T}$ contains no two-sided properly embedded projective planes or Möbius bands (which is true, in particular, for all orientable 3-manifolds).

We leave to whomever may require them in future work the details of the cases we did not cover.

3.1 Flattening triangular and bigon pillows

Lemma 10 (flattening triangular pillows) *Let $\mathcal{D}_0$ be a (possibly invalid) cell decomposition, and let $\mathcal{D}_1$ be the cell decomposition obtained by flattening a triangular pillow F in $\mathcal{D}_0$. One of the following holds:*

(a) *The two triangular faces of F are not identified and not both boundary, in which case the truncated pseudomanifolds of $\mathcal{D}_0$ and $\mathcal{D}_1$ are homeomorphic.*

(b) *F forms a (bounded) cell decomposition of a 3-ball, in which case $\mathcal{D}_1$ is obtained from $\mathcal{D}_0$ by deleting this 3-ball component.*

(c) *F forms a (closed) cell decomposition of either S^3 (the 3-sphere) or $L_{3,1}$ (a lens space), in which case $\mathcal{D}_1$ is obtained from $\mathcal{D}_0$ by deleting this closed component.*

(d) *F forms a two-vertex component $\mathcal{C}$ of $\mathcal{D}_0$ with exactly one invalid edge e; one of the vertices is incident to e and has 2-sphere link, while the other vertex is not incident to e and has projective plane link. In this case, $\mathcal{D}_1$ is obtained from $\mathcal{D}_0$ by deleting this invalid component $\mathcal{C}$.*

Proof Throughout this proof, let t and t' denote the triangular faces that bound the triangular pillow F. We have several cases to consider, depending on how t and t' are glued to other faces of $\mathcal{D}_0$ (if at all).

First, suppose t and t' are not glued to each other. In this case, the triangular pillow F forms a 3-ball. If t and t' are not both boundary, then this ball lives inside some larger component of $\mathcal{D}_0$, and flattening F does not change the truncated pseudomanifold; this corresponds to case (a). On the other hand, if t and t'

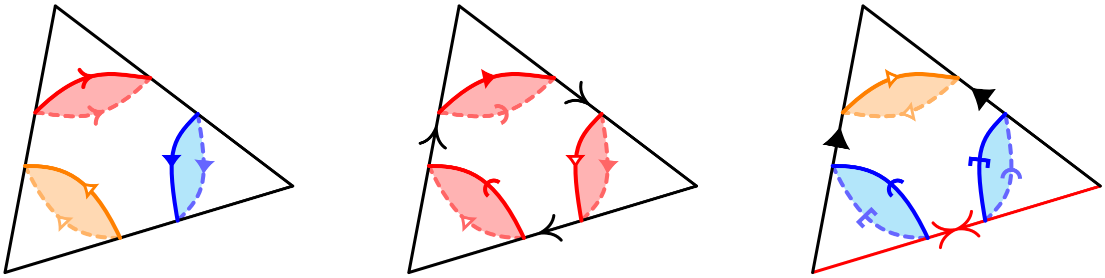


Figure 11: Cases where the two faces of a triangular pillow are glued to each other. Left: a triangular pillow that forms a (closed) cell decomposition of S^3. Middle: a triangular pillow that forms a (closed) cell decomposition of $L_{3,1}$. Right: a triangular pillow that forms a component with an invalid edge (highlighted red).

are both boundary, then F forms the entirety of a 3-ball component of $\mathcal{D}_0$, and flattening F deletes this 3-ball component; this corresponds to case (b).

With that out of the way, suppose t and t' *are* glued to each other. Up to symmetry, there are two possibilities for an orientation-reversing gluing:

- If t and t' are glued without a twist, then F forms a cell decomposition of S^3 (see Figure 11 (left)).
- If t and t' are glued with a twist, then F forms a cell decomposition of $L_{3,1}$ (see Figure 11 (middle)).

In either case, we see that F forms a closed component of $\mathcal{D}_0$. Moreover, flattening F has the effect of deleting this closed component. This corresponds to case (c).

For an orientation-preserving gluing of t and t', there is only one possibility up to symmetry. With this gluing, F forms a two-vertex component $\mathcal{C}$ of $\mathcal{D}_0$ with exactly one invalid edge e (see Figure 11 (right)). One of the vertices of $\mathcal{C}$ is given by identifying the two endpoints of e, and has 2-sphere link. The other vertex of $\mathcal{C}$ is given by the vertex of F disjoint from e, and has projective plane link. This corresponds to case (d). □

Lemma 11 (flattening bigon pillows) *Let $\mathcal{D}_0$ be a (possibly invalid) cell decomposition, and let $\mathcal{D}_1$ be the cell decomposition obtained by flattening a bigon pillow F in $\mathcal{D}_0$. One of the following holds:*

(a) *The two bigon faces of F are not identified and not both boundary, in which case the truncated pseudomanifolds of $\mathcal{D}_0$ and $\mathcal{D}_1$ are homeomorphic.*

(b) *F forms a (bounded) cell decomposition of a 3-ball, in which case $\mathcal{D}_1$ is obtained from $\mathcal{D}_0$ by deleting this 3-ball component.*

(c) *F forms a (closed) cell decomposition of either S^3 (the 3-sphere) or $\mathbb{R}P^3$ (real projective space), in which case $\mathcal{D}_1$ is obtained from $\mathcal{D}_0$ by deleting this closed component.*

(d) *F forms an ideal cell decomposition of $\mathbb{R}P^2 \times [0,1]$, in which case $\mathcal{D}_1$ is obtained from $\mathcal{D}_0$ by deleting this ideal component.*

(e) *F forms a one-vertex component of $\mathcal{D}_0$ with exactly two invalid edges, in which case $\mathcal{D}_1$ is obtained from $\mathcal{D}_0$ by deleting this invalid component.*

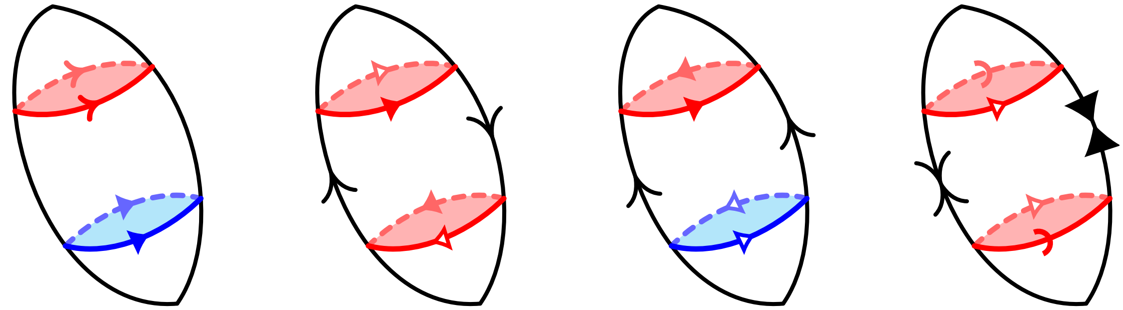


Figure 12: Cases where the two faces of a bigon pillow are glued to each other. First image: a bigon pillow that forms a (closed) cell decomposition of S^3. Second image: a bigon pillow that forms a (closed) cell decomposition of $\mathbb{R}P^3$. Third image: a bigon pillow that forms an ideal cell decomposition of $\mathbb{R}P^2 \times [0,1]$. Fourth image: a bigon pillow that forms a component with two invalid edges.

Proof Throughout this proof, let b and b' denote the bigon faces that bound the bigon pillow F. We have several cases to consider, depending on how b and b' are glued to other faces of $\mathcal{D}_0$ (if at all).

First, suppose b and b' are not glued to each other. In this case, the bigon pillow F forms a 3-ball. If b and b' are not both boundary, then this ball lives inside some larger component of $\mathcal{D}_0$, and flattening F does not change the truncated pseudomanifold; this corresponds to case (a). On the other hand, if b and b' are both boundary, then F forms the entirety of a 3-ball component of $\mathcal{D}_0$, and flattening F deletes this 3-ball component; this corresponds to case (b).

With that out of the way, suppose b and b' *are* glued to each other. There are two possibilities for an orientation-reversing gluing:

- If b and b' are glued without a twist, then F forms a cell decomposition of S^3 (see Figure 12 (first image)).
- If b and b' are glued with a twist, then F forms a cell decomposition of $\mathbb{R}P^3$ (see Figure 12 (second image)).

In either case, we see that F forms a closed component of $\mathcal{D}_0$. Moreover, flattening F has the effect of deleting this closed component. This corresponds to case (c).

Finally, for an orientation-preserving gluing, we again have two possibilities:

- One of these gluings causes the two edges of F to be identified together, and does not create invalid edges. In this case, F forms an ideal cell decomposition of $\mathbb{R}P^2 \times [0,1]$ (see Figure 12 (third image)), and flattening F has the effect of deleting this ideal component. This corresponds to case (d).
- The other orientation-preserving gluing causes each edge of F to be identified with itself in reverse, so that F forms a one-vertex component of $\mathcal{D}_0$ with exactly two invalid edges (see Figure 12 (fourth image)). Flattening F has the effect of deleting this invalid component. This corresponds to case (e). □

3.2 Flattening bigon faces

We now study the effect of flattening a bigon face F. As mentioned earlier, our main goal is to give a detailed analysis in the case where F belongs to a valid cell decomposition whose vertices are all either internal or ideal. Our arguments only rely on the following properties:

(a) F is an internal face.
(b) Each edge incident to F is internal.
(c) Each vertex incident to F is either internal or ideal.

Provided these properties hold, our analysis will apply even if F belongs to an invalid cell decomposition.

With this in mind, we assume throughout Section 3.2 that F is an internal bigon face. However, for the sake of generality, we do *not* assume that conditions (b) and (c) are satisfied; instead, we carefully enumerate the cases where these conditions hold, and for each such case we subsequently give a detailed description of the effect of flattening F.

We present our analysis in four parts. First, in Section 3.2.1, we give a brief user guide for the reader seeking to apply our results. Then, in Section 3.2.2, we make some preliminary observations by examining how flattening F interacts with the vertices incident to F. Finally, we partition the main analysis into two broad cases that we handle separately in Sections 3.2.3 and 3.2.4.

Before we dive into the details, we make some general comments about our proof strategy, and we introduce some notation and terminology to support this. One of the key ideas throughout our analysis is that, under our assumption that F is internal, flattening F has the side-effect that we lose the face-gluing along F. This means that flattening F has the same result as the following two-step procedure:

(1) Undo the gluing along F, which yields two new boundary bigons $F_0^\dagger$ and $F_1^\dagger$.
(2) Flatten $F_0^\dagger$ and $F_1^\dagger$; since these are boundary faces, flattening these faces has no side-effects (unlike the original face F).

We will see that step (1) often corresponds to cutting along a properly embedded surface S, and that step (2) often corresponds to filling the remnants of S, so that the overall topological effect of flattening F is often to decompose along S (as defined in Section 2.2). With this in mind, we introduce the following notation (also see Figure 13), which we will use throughout the rest of this section:

Notation A As above, let F be an internal bigon face in a (possibly invalid) cell decomposition $\mathcal{D}_0$. Let $\mathcal{D}_1$ be the cell decomposition obtained by flattening F, and let φ denote the associated flattening map. For each $i \in \{0,1\}$, let V_i denote the set of ideal and invalid vertices in $\mathcal{D}_i$, and let $\mathcal{P}_i$ denote the truncated pseudomanifold of $\mathcal{D}_i$; recall that $\mathcal{P}_i$ is obtained from $\mathcal{D}_i$ by truncating the vertices in V_i.

As in step (1) above, let $F_0^\dagger$ and $F_1^\dagger$ denote the two new boundary bigons that we obtain after undoing the face-gluing along F, and let $\mathcal{D}^\dagger$ denote the cell decomposition that we obtain after undoing this gluing.

$$\begin{tikzcd}
\mathcal{D}_0 \arrow[r, rightsquigarrow, "{g^{-1}, \text{ unglue } F}"] \arrow[rr, dashed, bend left, "{\varphi, \text{ flatten } F}"] \arrow[d, "\text{truncate } V_0"] & \mathcal{D}^{\dagger} \arrow[r, dashed, "\text{flatten } F_0^{\dagger} \text{ and } F_1^{\dagger}"] & \mathcal{D}_1 \arrow[d, "\text{truncate } V_1"'] \\
\mathcal{P}_0 & & \mathcal{P}_1
\end{tikzcd}$$

Figure 13: The protagonists introduced in Notation A.

Let $g : \mathcal{D}^{\dagger} \to \mathcal{D}_0$ be the quotient map associated to the operation of regluing $F_0^{\dagger}$ and $F_1^{\dagger}$ to recover the original bigon face F.

We also introduce the following terminology, which will be useful not only for flattening bigon faces, but also for proving our main theorem in Section 4:

Definitions 12 Let $\mathcal{D}$ be a (possibly invalid) cell decomposition, and let $\mathcal{P}$ be the truncated pseudomanifold of $\mathcal{D}$. Since $\mathcal{P}$ is obtained from $\mathcal{D}$ by truncating the ideal and invalid vertices of $\mathcal{D}$, we can view $\mathcal{P}$ as a subset of $\mathcal{D}$; using this viewpoint, the *truncated bigon* associated to a bigon face B in $\mathcal{D}$ is given by $B \cap \mathcal{P}$ (see Figure 14); we will see that in many cases, the truncated bigon forms a properly embedded surface in $\mathcal{P}$.

For some positive integer n, consider an embedded curve γ in $\mathcal{D}$ that

- starts at the midpoint of an edge e_0;
- ends at the midpoint of an edge e_n (possibly equal to e_0, to allow for the possibility that γ is a closed curve); and
- passes through the midpoints of a sequence $e_0, \dots, e_n$ of edges, such that for each $i \in \{0, \dots, n-1\}$, the edges e_i and e_{i+1} together bound a single bigon face B_i that is bisected by γ.

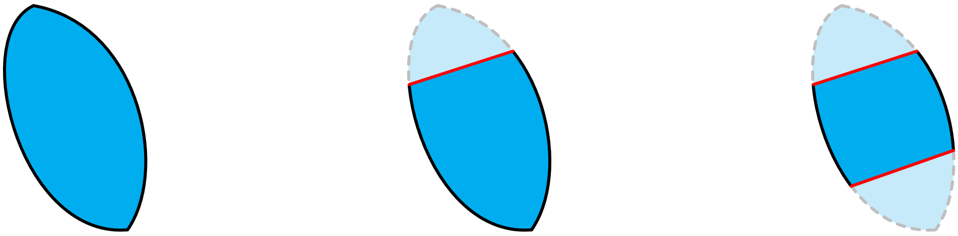


Figure 14: There are three possibilities for the truncated bigon associated to a bigon face B. The portion of B that lives outside the truncated bigon is indicated by dashed edges and faint shading. Left: the case where both vertices of B are internal or boundary (either forming two distinct vertices, or identified to form a single such vertex). Middle: the case where one vertex of B is ideal or invalid, while the other is internal or boundary. Right: the case where both vertices of B are ideal or invalid (either forming two distinct vertices, or identified to form a single such vertex).

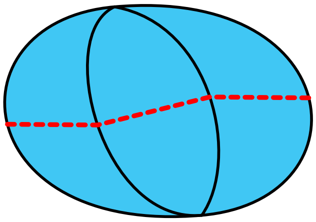


Figure 15: Three bigon faces bisected by a curve (drawn as a dashed red line) passing through midpoints of edges. These bigon faces together form a bigon path of length 3.

This is illustrated in Figure 15. We call the union $\mathcal{U} := B_0 \cup \dots \cup B_{n-1}$ a *bigon path* of *length* n in $\mathcal{D}_0$, and we call the edges e_0 and e_n the *ends* of $\mathcal{U}$. If the bigon faces $B_0, \dots, B_{n-1}$ are all boundary, then we say that $\mathcal{U}$ is *boundary*; similarly, if $B_0, \dots, B_{n-1}$ are all internal, then we say that $\mathcal{U}$ is *internal*.

For each $i \in \{0, \dots, n-1\}$, let S_i denote the truncated bigon associated to B_i. We call the union $S_0 \cup \dots \cup S_{n-1}$ the *truncated bigon path* associated to $\mathcal{U}$; similar to individual truncated bigons, truncated bigon paths often form properly embedded surfaces in $\mathcal{P}$.

We mentioned earlier that we divide our analysis into two cases that we handle separately in Sections 3.2.3 and 3.2.4. We now have the terminology to describe these two cases. Specifically, after ungluing F, the two new boundary bigons $F_0^\dagger$ and $F_1^\dagger$ could either

- share at least one common edge, so that they together form a single boundary bigon path of length two; or
- have no common edges, in which case they form two separate boundary bigon paths of length one.

There is no technical reason for dividing our analysis according to these two cases; we make this choice simply to help organise our analysis into smaller, more manageable pieces.

3.2.1 User guide We split the effects of flattening F into several parts:

- The effect on the vertices incident to F is described in Claim B.
- The effect on the edges incident to F is described in Claims D and E.
- The effect on the truncated pseudomanifold $\mathcal{P}_0$ is described in Claims D.1, D.2 and D.4, and in Claims E.1 and E.2.

Claims D, D.1, D.2 and D.4 all deal with the case where $F_0^\dagger$ and $F_1^\dagger$ form a single boundary bigon path, so they can be found in Section 3.2.3; on the other hand, Claims E, E.1 and E.2 all deal with the case where $F_0^\dagger$ and $F_1^\dagger$ form two separate boundary bigon paths, so they can be found in Section 3.2.4. The intended way to use all these results is to begin by referring to Claims D and E, as these two overarching claims will indicate which of the other claims are relevant for any given application.

The only other result that we prove is Claim C. This is a useful tool for our proofs in Sections 3.2.3 and 3.2.4, but it is otherwise not a crucial part of our description of the effects of flattening F. Having

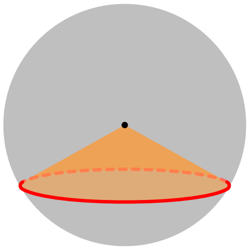


Figure 16: A simple example of a v-cone (orange) over a subset S of a vertex link L. Here, L is a 2-sphere, and S is an embedded closed curve (red) in L.

said this, Claim C might be useful for the reader seeking to extend our analysis of flattening F to the cases that we do not study in detail.

3.2.2 Interaction with vertices Let v denote a vertex incident to F, and consider a small regular neighbourhood N of v. To describe how flattening F interacts with v, we will find it useful to view N as a cone over the link L of v; that is, we view N as a union of lines, with each point in L being joined to v by one such line, and with any two such lines intersecting only at the vertex v. Under this viewpoint, any subset S of L defines a subset C_S of N consisting of the lines joining S to v (for example, see Figure 16); we will call C_S the *v-cone* over S. We will use this notion of v-cones to prove two claims:

- In Claim B, we describe how flattening F changes the vertex v.
- In Claim C, we give conditions under which we can, in some sense, "push F away from v"; we will give a more precise formulation of this later. Roughly, the purpose of this is that it gives us a unified method to deal with some of the more inconvenient ways in which flattening F interacts with v; this will become clearer when we see Claim C in action in Sections 3.2.3 and 3.2.4.

Since we are interested in the effect of flattening F, we devote particular attention to the subset of L given by $F \cap L$. Assuming that each edge incident to F is internal, we have the following possibilities (see Figure 17):

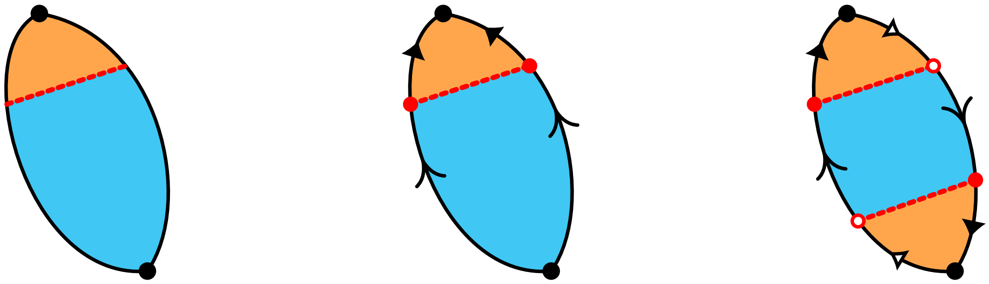


Figure 17: The three possibilities for a component γ (dotted red) of $F \cap L$. In each case, the v-cone (orange) over γ forms part of $F \cap N$. Left: when F forms a disc, γ forms an arc. Middle: when F forms a 2-sphere, γ forms a closed curve. Right: when F forms a projective plane, γ forms a closed curve.

- Suppose the edges of F are not identified, so that F forms a disc. In this case, $F \cap L$ consists of either one or two arcs:
  - If the two vertices of F form two distinct vertices in $\mathcal{D}_0$, then v is one of these two vertices, and $F \cap L$ consists of a single arc in L.
  - If the two vertices of F are identified to form a single vertex in $\mathcal{D}_0$, then $F \cap L$ consists of two disjoint arcs in L.
- Suppose the edges of F are identified to form a single edge e, and suppose this identification causes F to form a 2-sphere. In this case, $F \cap L$ consists of either one or two closed curves:
  - If the two vertices of F form two distinct vertices in $\mathcal{D}_0$, then v is one of these two vertices, and $F \cap L$ consists of a single closed curve in L.
  - If the two vertices of F are identified to form a single vertex in $\mathcal{D}_0$, then $F \cap L$ consists of two disjoint closed curves in L.
- Suppose the edges of F are identified to form a single edge e, and suppose this identification causes F to form a projective plane. In this case, the two vertices of F are identified to form a single vertex in $\mathcal{D}_0$, and $F \cap L$ consists of a single closed curve in L.

In each of the above cases, observe that the v-cone over $F \cap L$ coincides exactly with $F \cap N$. Intuitively, this means that flattening F changes N in a way that "respects the cone structure". This idea allows us to give a fairly straightforward description of how flattening F affects the vertex v:

Claim B *Assume that each edge incident to F is internal. Let v be a vertex incident to F, and let L denote the link of v. We have the following possibilities*:

(a) *If the edges of F are not identified, then $\varphi(v)$ consists of a single vertex whose link is topologically equivalent to L.*

(b) *If the edges of F are identified, then $F \cap L$ consists of either one or two closed curves in L. Let $L'_0, \dots, L'_{k-1}$ denote the components of the surface obtained by decomposing L along the curves in $F \cap L$; there could be up to three such components* (*that is*, *we have $1 \leqslant k \leqslant 3$*). *After flattening F, the image $\varphi(v)$ consists of k vertices $v'_0, \dots, v'_{k-1}$ such that for each $i \in \{0, \dots, k-1\}$, the vertex v'_i has link L'_i.*

Proof As above, let N denote a small regular neighbourhood of v, and view N as the v-cone over the vertex link L. We first consider the case where the two edges of F are not identified, so that $F \cap L$ consists of either one or two arcs in L. For each such arc γ, flattening F has the effect of collapsing γ to a single point p_γ, which leaves L topologically unchanged; the corresponding effect on N is to flatten the v-cone over γ to a single line joining p_γ to v, which means that we can continue to view N as the v-cone over L. As a result, we see that $\varphi(v)$ consists of a single vertex whose link is topologically equivalent to L. This proves case (a).

In the case where the edges of F *are* identified, recall that $F \cap L$ consists of either one or two closed curves in L. This time, we study the effect of flattening F by first ungluing F, and then flattening $F_0^{\dagger}$ and $F_1^{\dagger}$:

(1) Ungluing F changes L by cutting along the curves in $F \cap L$. Since $F \cap L$ could have up to two components, each of which could possibly form a *separating* curve in L, cutting along $F \cap L$ could split L into up to three components; let k be the number of such components, and denote these components by $L'_0, \dots, L'_{k-1}$. The corresponding change to N is to cut along the v-cone over $F \cap L$, which has the following effects:

- v gets split into k new vertices $v'_0, \dots, v'_{k-1}$.
- N gets split into k components $N'_0, \dots, N'_{k-1}$ such that for each $i \in \{0, \dots, k-1\}$, N'_i forms the v'_i-cone over L'_i.

At this stage, the vertices v'_i and the surfaces L'_i are not yet the same as the corresponding objects that appear in the claim statement. However, this situation will soon be fixed when we flatten $F_0^{\dagger}$ and $F_1^{\dagger}$, which will have the consequence of modifying the vertices v'_i and the surfaces L'_i; to keep notation as simple as possible, we will continue to use the same notation for these objects even after they are modified. Likewise, the neighbourhoods N'_i will also be modified, but we will continue to denote the modified neighbourhoods by N'_i.

(2) For each $i \in \{0, \dots, k-1\}$, flattening $F^{\dagger_0}$ and $F^{\dagger_1}$ modifies L'_i by collapsing each remnant γ of $F \cap L$ to a single point p_γ, which is topologically equivalent to filling γ with a disc; the corresponding effect on N'_i is to flatten the v'_i-cone over γ to a single line joining p_γ to v'_i, which means that we can continue to view N'_i as the v'_i-cone over L'_i. The end result of all this is that $\varphi(v)$ consists precisely of the (now modified) vertices $v'_0, \dots, v'_{k-1}$, and that the links of these vertices are given by the (now modified) surfaces $L'_0, \dots, L'_{k-1}$, respectively. We also note that, topologically, $L'_0, \dots, L'_{k-1}$ form the components of the surface obtained by decomposing L along $F \cap L$.

This proves case (b). □

We now turn our attention to the idea of "pushing F away from v"; we will build up to this idea in a slightly roundabout way. Assume that the two edges of F are identified to form a single internal edge. As observed earlier, this means that $F \cap L$ consists of either one or two closed curves in L. Suppose a component γ of $F \cap L$ forms a separating curve that bounds a disc in L. Under these conditions, rather than beginning the process of flattening F by ungluing the entirety of F all at once, we will find it useful to follow a more fine-grained procedure for flattening F (see Figure 18):

(i) Cut along the subset of F given by the v-cone C_γ over γ. Since γ is a separating curve in L, this has the following effects:

- v gets split into two new vertices v'_0 and v'_1.
- C_γ gets split into two remnants $C_0^{\dagger}$ and $C_1^{\dagger}$ such that for each $i \in \{0, 1\}$, $C_i^{\dagger}$ forms the v'_i-cone over γ.

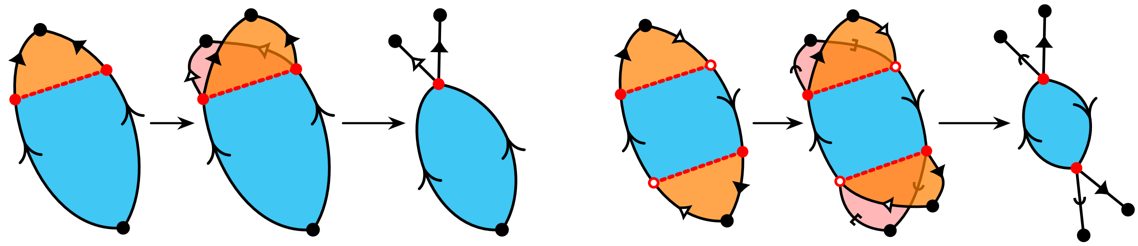


Figure 18: Flattening C_γ (orange) by first cutting along it, and then flattening the two remnants (orange and pink). This collapses γ (dotted red) to a point p_γ that we temporarily view as a vertex; it also causes $F - C_\gamma$ (blue) to become a new bigon F'. (These illustrations do not accurately reflect how everything is embedded in $\mathcal{D}_0$.) Left images: the case where F forms a 2-sphere. In this case, the new bigon F' also forms a 2-sphere, and one of its two vertices is given by p_γ. Right images: the case where F forms a projective plane. In this case, the new bigon F' also forms a projective plane, and its two vertices are identified together to form p_γ.

We postpone ungluing or cutting along $F - C_\gamma$ (ie the rest of F) until later; as a result, the curve γ does not yet fall apart into two pieces because it is still "held together" by $F - C_\gamma$.

(ii) Flatten γ to a single point p_γ, and for each $i \in \{0,1\}$ flatten the remnant $C_i^\dagger$ to a single line α_i joining p_γ to v_i'. Intuitively, the lines α_0 and α_1 will eventually form segments of the edges in $\varphi(F)$. Viewing p_γ as a temporary vertex, this step causes $F - C_\gamma$ to become a new bigon F'.

(iii) Treating F' as if it were an internal bigon face in a cell decomposition, flatten F' by first cutting along it, and then flattening each of its remnants. Since γ was originally a separating curve in L, this step splits the temporary vertex p_γ into a pair of new points. We no longer view these new points as vertices (which is why we called p_γ a "temporary" vertex); instead, each of these new points will occur in the interior of an edge in $\varphi(F)$.

Together, we refer to steps (i) and (ii) as the operation of *flattening* C_γ. We emphasise that after performing this operation, the intermediate object that we obtain from $\mathcal{D}_0$ might not be a cell decomposition anymore; however, this problem is only temporary, since we will recover a cell decomposition once we complete step (iii).

In describing the operation of flattening C_γ, we used the assumption that γ is a separating curve, but made no mention of the assumption that γ bounds a disc in L. The purpose of the latter assumption is that it allows us to show that flattening C_γ leaves the topology of $\mathcal{D}_0$ unchanged, which means that flattening F is topologically equivalent to the operation of flattening the new bigon F'. Moreover, we can view F' as the bigon that results from "pushing F away from v":

Claim C *Assume that each edge incident to F is internal. Let v be a vertex incident to F, and let L denote the link of v. Suppose a component γ of $F \cap L$ forms a separating curve that bounds a disc E in L, and suppose the interior of E is disjoint from F. Consider the subset of F given by the v-cone C_γ*

over γ. Flattening C_γ creates a new internal vertex without changing the topology of $\mathcal{D}_0$, and reduces the operation of flattening F to the operation of flattening a new bigon F' that is

- *topologically equivalent to a bigon obtained from F by an isotopy that takes C_γ to E; and*
- *incident to a temporary* ***internal*** *vertex given by the point p_γ that results from flattening γ.*

Proof To see how flattening C_γ affects $\mathcal{D}_0$ topologically, we first claim that since γ bounds a disc E in L, we can find a 3-ball $\mathcal{B}$ such that $C_\gamma - v$ lies in the interior of $\mathcal{B}$. If v were internal or boundary, we could simply take $\mathcal{B}$ to be a small regular neighbourhood of the v-cone C_E over E. However, to account for the possibility that v is ideal or invalid, we instead construct $\mathcal{B}$ as follows:

(a) Consider a regular neighbourhood N^* of v that is "large enough" so that C_E lies entirely in the interior of N^*.

(b) Slightly isotope the disc E so that it lies in the frontier of N^*, and then enlarge this disc slightly so that the v-cone over this disc forms the desired 3-ball $\mathcal{B}$ (see Figure 19 (left images)).

The operation of flattening C_γ leaves everything outside of $\mathcal{B}$ untouched, so we just need to understand how this operation affects $\mathcal{B}$ topologically. We follow steps (i) and (ii) from above:

(i) As illustrated in Figure 19 (middle images), cutting along C_γ has the following effects:

- The vertex v gets split into two new vertices v'_0 and v'_1. One of the new vertices, say v'_0, has link given by L minus a disc. The other new vertex v'_1 has link given by the disc E, as illustrated in Figure 19 (bottom middle).
- The v-cone C_γ gets split into two remnants $C^\dagger_0$ and $C^\dagger_1$ such that for each $i \in \{0, 1\}$, $C^\dagger_i$ forms the v'_i-cone over γ. These two remnants bound a newly created void inside our 3-ball $\mathcal{B}$.

(ii) As illustrated in Figure 19 (right images) flattening $C^{\dagger_0}$ and $C^{\dagger_1}$ has the following effects:

- The link of v'_0 gets "closed up" so that it becomes topologically equivalent to L. Thus, we can equate v'_0 with the original vertex v.
- The link of v'_1 gets "closed up" to become the 2-sphere corresponding to $E/\partial E$, as illustrated in Figure 19 (bottom right). Thus, we can view v'_1 as a newly created internal vertex.
- The void that we created in the previous step gets filled in, so that we once again have a 3-ball $\mathcal{B}'$.
- The curve γ gets flattened to a single point p_γ that we temporarily view as a vertex. (Recall that p_γ is only a "temporary" vertex because after performing step (iii), p_γ gets split into two new points that we no longer view as vertices.) Since p_γ lies in the interior of the 3-ball $\mathcal{B}'$, we can think of p_γ as an *internal* vertex.

Topologically, all we have done is replaced the 3-ball $\mathcal{B}$ with another 3-ball $\mathcal{B}'$, so we have not changed the topology of $\mathcal{D}_0$.

To finish this proof, consider $F - C_\gamma$; this is the part of F that is being left "unflattened". Recall that after flattening C_γ, this unflattened part of F becomes a new bigon F', and that the operation of flattening F is

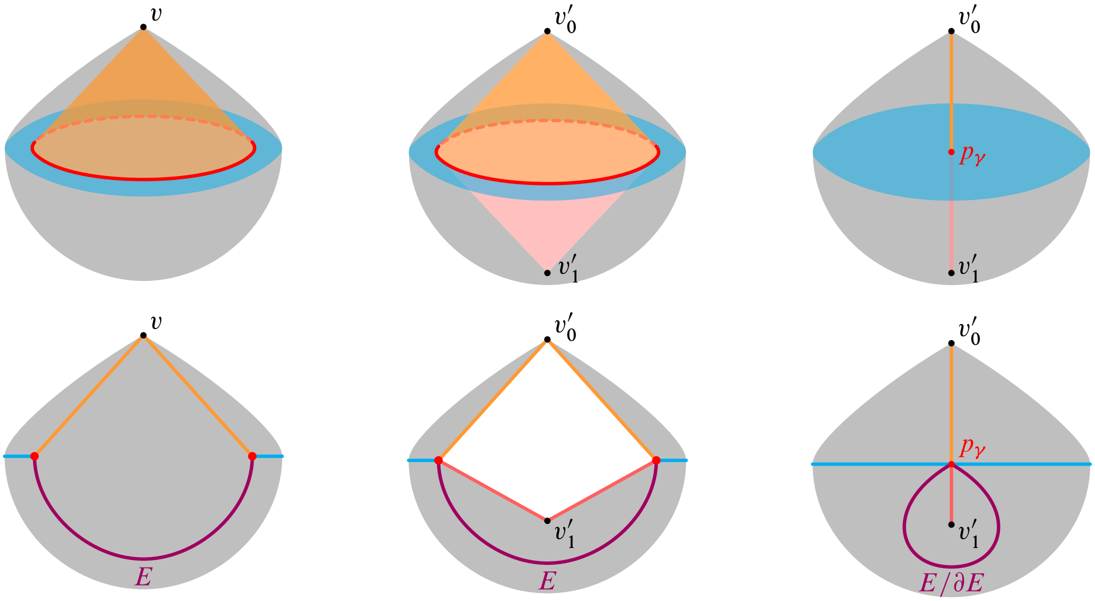


Figure 19: Since γ (red, top images) bounds a disc in L, flattening C_γ has no topological effect on $\mathcal{D}_0$. The intersection of the 3-ball $\mathcal{B}$ (grey) with the "unflattened" part of F is shaded blue. Top left: the 3-ball $\mathcal{B}$ (grey) contains the entirety of the v-cone C_γ (orange), as well as a portion of $F - C_\gamma$ (blue). Top middle: cutting along the v-cone C_γ yields two remnants (orange and pink), and creates a void. Top right: flattening the remnants of C_γ fills the void back in, so we recover a 3-ball. Bottom row: schematic cross-sections of the 3-dimensional pictures in the top row. Here we also include the disc E (purple), which is not shown in the 3-dimensional pictures; the vertex v'_1 is repositioned slightly to accommodate this inclusion.

reduced to the operation of flattening F'. Topologically, observe that F' is equivalent to a bigon obtained from F by an isotopy that replaces C_γ with the disc E; this can be seen by equating the vertices v and v'_0, and then comparing how F intersects the grey 3-ball in Figure 19 (top left) with how F' intersects the grey 3-ball in Figure 19 (top right). □

3.2.3 The case where ungluing gives a single boundary bigon path We are now ready to present the main analysis of the effect of flattening F. We first consider the case where $F_0^\dagger$ and $F_1^\dagger$ together form a single boundary bigon path $\mathcal{F}^\dagger$. Depending on how the ends of $\mathcal{F}^\dagger$ are identified (if at all), $\mathcal{F}^\dagger$ could form a 2-sphere, projective plane or disc in the boundary of $\mathcal{D}^\dagger$. We refine this list of cases as follows:

Claim D *If $F_0^\dagger$ and $F_1^\dagger$ together form a single boundary bigon path $\mathcal{F}^\dagger$, then one of the following holds:*

- *The boundary bigon path $\mathcal{F}^\dagger$ forms a 2-sphere, in which case the result $\varphi(F)$ of flattening F is a single internal edge in $\mathcal{D}_1$. Letting $e_0^\dagger$ and $e_1^\dagger$ denote the edges incident to $\mathcal{F}^\dagger$, we have four cases depending on the behaviour of the gluing map g (see Figure 20):*

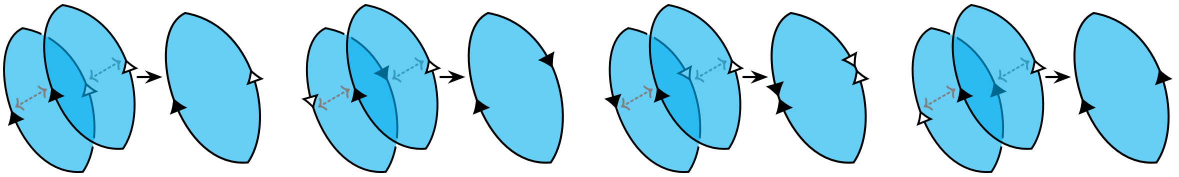


Figure 20: The four ways to glue $F_0^\dagger$ and $F_1^\dagger$ together when $\mathcal{F}^\dagger$ forms a 2-sphere. They are gluings that result in F forming a disc (first image), F forming a projective plane (second image), F being incident to two invalid edges (third image) and F forming a 2-sphere (fourth image).

(1) *The map g realises an orientation-reversing gluing of $F_0^\dagger$ and $F_1^\dagger$ such that for each $i \in \{0,1\}$, the edge $e_i^\dagger$ gets identified with itself to form an internal edge of $\mathcal{D}_0$. In this case, F forms a* ***disc*** *in $\mathcal{D}_0$. (See Claim D.1 for details about the effect of flattening F in this case.)*

(2) *The map g realises an orientation-reversing gluing of $F_0^\dagger$ and $F_1^\dagger$ that causes $e_0^\dagger$ and $e_1^\dagger$ to be identified together to form a single internal edge of $\mathcal{D}_0$. In this case, F forms a* ***projective plane*** *in $\mathcal{D}_0$. (See Claim D.2 for details about the effect of flattening F in this case.)*

(3) *The map g realises an orientation-preserving gluing of $F_0^\dagger$ and $F_1^\dagger$ such that for each $i \in \{0,1\}$, the edge $e_i^\dagger$ gets identified with itself in reverse to form an invalid edge of $\mathcal{D}_0$.*

(4) *The map g realises an orientation-preserving gluing of $F_0^\dagger$ and $F_1^\dagger$ that causes $e_0^\dagger$ and $e_1^\dagger$ to be identified together to form a single internal edge of $\mathcal{D}_0$. In this case, F forms a* ***2-sphere*** *in $\mathcal{D}_0$. (See Claim D.4 for details about the effect of flattening F in this case.)*

- *The boundary bigon path $\mathcal{F}^\dagger$ forms a projective plane, in which case F is incident to an invalid edge in $\mathcal{D}_0$.*
- *The boundary bigon path $\mathcal{F}^\dagger$ forms a disc, in which case F is incident to a boundary edge in $\mathcal{D}_0$.*

Proof We first consider the case where the ends of $\mathcal{F}^\dagger$ are identified in such a way that $\mathcal{F}^\dagger$ forms a 2-sphere in the boundary of $\mathcal{D}^\dagger$. In this case, observe that after flattening $F_0^\dagger$ and $F_1^\dagger$, the image $\varphi(F)$ is a single internal edge in $\mathcal{D}_1$. Let $e_0^\dagger$ and $e_1^\dagger$ denote the two edges of $\mathcal{D}^\dagger$ that are incident to $\mathcal{F}^\dagger$. As illustrated in Figure 20, there are four ways to glue $F_0^\dagger$ and $F_1^\dagger$ together to recover the bigon face F, depending on whether the gluing is orientation-reversing, and on whether the gluing causes $e_0^\dagger$ and $e_1^\dagger$ to be identified together:

- The two orientation-reversing gluings are shown in Figure 20 (first and second images); these correspond to cases (1) and (2), respectively.
- The two orientation-preserving gluings are shown in Figure 20 (third and fourth images); these correspond to cases (3) and (4), respectively.

Suppose now that the ends of $\mathcal{F}^\dagger$ are identified in such a way that $\mathcal{F}^\dagger$ forms a projective plane in the boundary of $\mathcal{D}^\dagger$. Let $e_0^\dagger$ and $e_1^\dagger$ denote the two edges of $\mathcal{D}^\dagger$ that are incident to $\mathcal{F}^\dagger$. Up to symmetry, there are two ways to glue $F_0^\dagger$ and $F_1^\dagger$ back together, depending on whether the gluing causes $e_0^\dagger$ and $e_1^\dagger$ to be identified together. As illustrated in Figure 21, F is incident to an invalid edge in both cases.

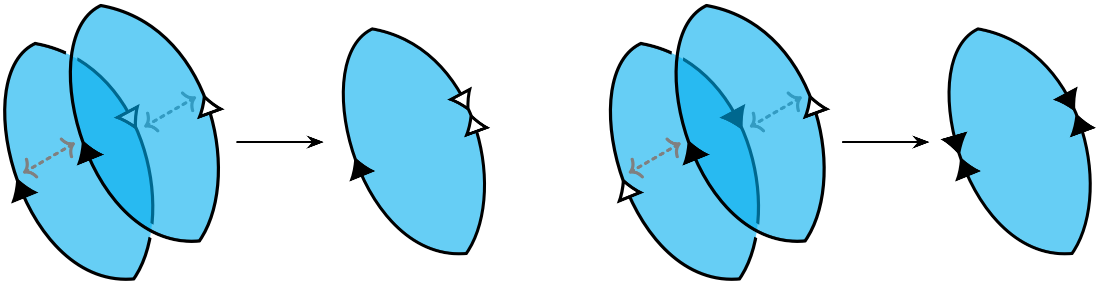


Figure 21: The two ways to glue $F_0^\dagger$ and $F_1^\dagger$ together when $\mathcal{F}^\dagger$ forms a projective plane. They are gluings that result in F being incident to one internal edge and one invalid edge (left) and F being incident to an invalid edge (right).

Finally, suppose the ends of $\mathcal{F}^\dagger$ are not identified, so that $\mathcal{F}^\dagger$ forms a disc in the boundary of $\mathcal{D}^\dagger$. Observe that each end of $\mathcal{F}^\dagger$ must be incident to a boundary face of $\mathcal{D}^\dagger$ that is not part of $\mathcal{F}^\dagger$. This means that regardless of how we glue $F_0^\dagger$ and $F_1^\dagger$ back together, the bigon face F will always be incident to an edge lying in the boundary of $\mathcal{D}_0$. □

As mentioned earlier, we only give a detailed analysis of the effect of flattening F in the cases where F is not incident to any boundary or invalid edges. This corresponds to cases (1), (2) and (4) of Claim D.

Claim D.1 (disc) *In case* (1) *of Claim D, the truncated pseudomanifolds* $\mathcal{P}_0$ *and* $\mathcal{P}_1$ *are homeomorphic.*

Proof Recall that in case (1) of Claim D, the edges of F are not identified, so that F forms a disc. In this case, we can flatten F without changing the topology of $\mathcal{D}_0$; in particular, as we saw in Claim B, the links of the vertices incident to F remain unchanged after flattening F. Thus, we see that $\mathcal{P}_0$ and $\mathcal{P}_1$ are homeomorphic. □

Claim D.2 (projective plane) *In case* (2) *of Claim D, the two vertices of* F *are identified to form a single vertex* v*, and one of the following holds*:

(a) *The vertex* v *is internal, in which case* F *forms a one-sided properly embedded projective plane in* $\mathcal{P}_0$*, and* $\mathcal{P}_1$ *is obtained from* $\mathcal{P}_0$ *by decomposing along this projective plane.*

(b) *The vertex* v *is ideal, in which case the truncated bigon associated to* F *forms a one-sided properly embedded Möbius band* S *in* $\mathcal{P}_0$*; the boundary curve* γ *of* S *forms a two-sided curve in* $\partial\mathcal{P}_0$*. In this case, flattening* F *has one of the following effects*:
  - *If* γ *bounds a disc* E *in* $\partial\mathcal{P}_0$*, then* $\mathcal{P}_1$ *is obtained from* $\mathcal{P}_0$ *by decomposing along a one-sided projective plane given by isotoping* $S \cup E$ *slightly off the boundary of* $\mathcal{P}_0$.
  - *If* γ *does not bound a disc in* $\partial\mathcal{P}_0$*, then* $\mathcal{P}_1$ *is obtained from* $\mathcal{P}_0$ *by first decomposing along the Möbius band* S*, and then filling any new* 2*-sphere boundary components with* 3*-balls.*

(c) *The vertex* v *is boundary or invalid.*

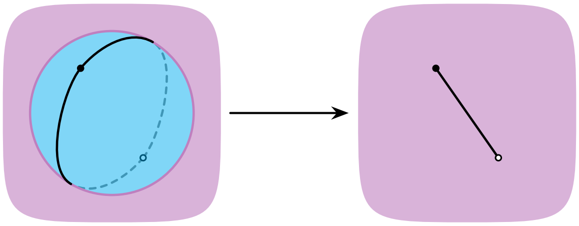


Figure 22: When $\mathcal{F}^\dagger$ forms a 2-sphere, flattening $\mathcal{F}^\dagger$ is equivalent to filling it with a 3-ball.

Proof Recall that in case (2) of Claim D, the edges of F are identified so that F forms a projective plane. In particular, this means that the two vertices of F are identified to form a single vertex v in $\mathcal{D}_0$. We start by getting the easy cases out of the way:

- If v is boundary or invalid, then we are in case (c).
- If v is internal, then F forms an embedded projective plane that lies entirely in the interior of $\mathcal{P}_0$. In this case, observe that ungluing F corresponds topologically to cutting along this projective plane; this yields a single 2-sphere remnant, corresponding precisely to the boundary bigon path $\mathcal{F}^\dagger$. This means, in particular, that F forms a *one-sided* projective plane in $\mathcal{P}_0$. Moreover, as illustrated in Figure 22, flattening $F_0^\dagger$ and $F_1^\dagger$ is topologically equivalent to filling the 2-sphere $\mathcal{F}^\dagger$ with a 3-ball. Altogether, we see that flattening F is topologically equivalent to decomposing along F. This proves case (a).

The rest of this proof is devoted to the case where v is ideal; we need to prove all the conclusions stated in case (b). For this, we first observe that the truncated bigon associated to F forms a properly embedded Möbius band S in $\mathcal{P}_0$. Consider the pseudomanifold $\mathcal{P}^\dagger$ obtained from $\mathcal{D}^\dagger$ by truncating the vertices in $g^{-1}(V_0)$; viewing $\mathcal{P}^\dagger$ as a subset of $\mathcal{D}^\dagger$, let $S^\dagger$ denote the annulus in $\partial\mathcal{P}^\dagger$ given by $\mathcal{F}^\dagger \cap \mathcal{P}^\dagger$. Topologically, observe that $\mathcal{P}^\dagger$ is obtained from $\mathcal{P}_0$ by cutting along the Möbius band S; as shown in Figure 23, this yields a single remnant — namely, the annulus $S^\dagger$ — so S must be a *one-sided* Möbius band in $\mathcal{P}_0$.

Consider the ideal boundary component L of $\mathcal{P}_0$ given by truncating the vertex v, and let γ denote the boundary curve of S. To see that γ forms a two-sided curve in L, observe that cutting along S splits γ into the two disjoint curves that bound the annulus $S^\dagger$.

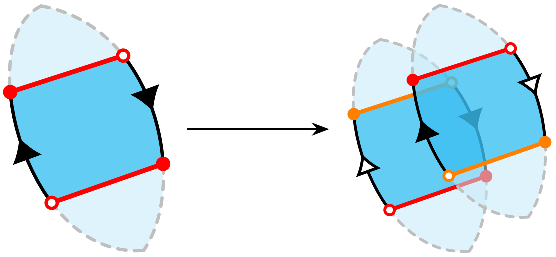


Figure 23: The truncated bigon associated to F forms a one-sided Möbius band S. Cutting along S yields a single annulus remnant.

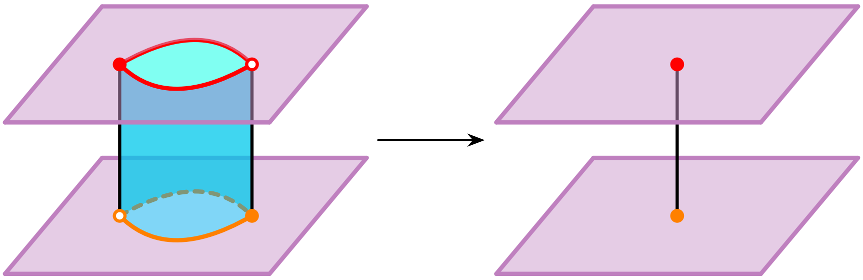


Figure 24: Flattening $\mathcal{F}^\dagger$ has the effect of filling the annulus remnant $S^\dagger$ with a thickened disc.

All that remains is to understand the overall topological effect of flattening F. We begin with the case where γ bounds a disc E in L. In this case, we use Claim C to flatten the v-cone over γ. This reduces the operation of flattening F to the operation of flattening a new bigon F' given by pushing F slightly away from v; topologically, F' is equivalent to a projective plane given by isotoping $S \cup E$ slightly off the boundary of $\mathcal{P}_0$. Since the vertices of F' are identified to form the temporary internal vertex that results from flattening the curve γ, flattening F' has the same topological effect as flattening F in the case where v is internal (case (a)). That is, F' forms a *one-sided* projective plane in $\mathcal{P}_0$, and $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along this projective plane. This completes the case where γ bounds a disc in L.

For the case where γ does *not* bound a disc in L, we flatten F by first ungluing F, and then flattening $F_0^\dagger$ and $F_1^\dagger$. Earlier, we observed that ungluing F has the effect of cutting $\mathcal{P}_0$ along S, which yields a single annulus remnant $S^\dagger$ in a new pseudomanifold $\mathcal{P}^\dagger$. With this in mind, consider the pseudomanifold $\mathcal{P}^*$ obtained from $\mathcal{D}_1$ by truncating the vertices in $\varphi(V_0)$. Topologically, observe that $\mathcal{P}^*$ is obtained from $\mathcal{P}^\dagger$ by filling the annulus $S^\dagger$ with a thickened disc; see Figure 24. In other words, $\mathcal{P}^*$ is obtained from $\mathcal{P}_0$ by decomposing along the Möbius band S.

To see how $\mathcal{P}^*$ is related to $\mathcal{P}_1$, we need to compare the truncated vertex sets $\varphi(V_0)$ and V_1. The only way these vertex sets can differ is if $\varphi(v)$ contains a vertex that is neither ideal nor invalid — such a vertex would be in $\varphi(V_0)$ but not in V_1. We can use Claim B to determine the composition of $\varphi(v)$, and hence determine the relationship between $\mathcal{P}^*$ and $\mathcal{P}_1$:

- If γ is a nonseparating curve in L, then decomposing L along γ gives a single new closed surface L^*, and $\varphi(v)$ consists of a single vertex whose link is given by L^*. If L^* is not a 2-sphere, then $\varphi(v)$ is an ideal vertex, so $\mathcal{P}^*$ is homeomorphic to $\mathcal{P}_1$. However, if L^* is a 2-sphere, then $\varphi(v)$ is an internal vertex; topologically, L^* corresponds to a 2-sphere boundary component of $\mathcal{P}^*$, and we need to fill this 2-sphere with a 3-ball to recover $\mathcal{P}_1$ from $\mathcal{P}^*$.
- If γ is a separating curve in L, then decomposing along γ gives two new closed surfaces, and $\varphi(v)$ consists of two vertices whose links are given by these two new surfaces. Since γ does not bound a disc in L, both vertices in $\varphi(v)$ are ideal, so $\mathcal{P}^*$ is homeomorphic to $\mathcal{P}_1$.

This completes the proof of case (b). □

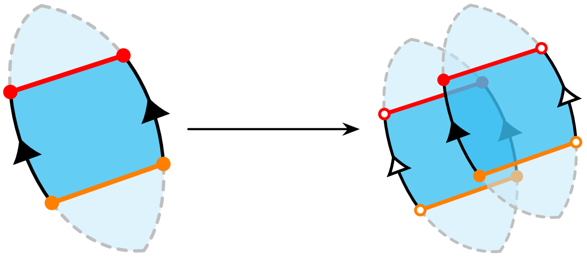


Figure 25: The truncated bigon associated to F forms a one-sided annulus S. Cutting along S yields a single annulus remnant.

Claim D.4 (2-sphere) *In case* (4) *of Claim D, each vertex incident to F is either ideal or invalid. Moreover, the truncated bigon associated to F forms a one-sided properly embedded annulus S in $\mathcal{P}_0$, and each boundary curve of S forms a one-sided curve in $\partial\mathcal{P}_0$.*

Suppose F is only incident to ideal vertices (*the two vertices of F could either be identified to form a single ideal vertex, or they could form two distinct ideal vertices*)*. In this case, $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by first decomposing along the annulus S, and then filling any new* 2*-sphere boundary components with* 3*-balls.*

Proof Recall that in case (4) of Claim D, the edges of F are identified so that F forms a 2-sphere. The two vertices of F could either be identified to form a single vertex of $\mathcal{D}_0$, or they could form two distinct vertices of $\mathcal{D}_0$. Let L denote the union of the links of the vertices incident to F, and consider the two curves γ_0 and γ_1 in which F meets L. For each $i \in \{0,1\}$, observe that ungluing F causes the curve γ_i to "unravel" to form a single new curve; thus, γ_i forms a one-sided curve in L. In particular, the vertices incident to F must have nonorientable vertex links, which implies that these vertices must be either ideal or invalid. This means that the truncated bigon associated to F forms a properly embedded annulus S in $\mathcal{P}_0$.

To see that S is one-sided, consider the pseudomanifold $\mathcal{P}^\dagger$ obtained from $\mathcal{D}^\dagger$ by truncating the vertices in $g^{-1}(V_0)$; viewing $\mathcal{P}^\dagger$ as a subset of $\mathcal{D}^\dagger$, let $S^\dagger$ denote the annulus in $\partial\mathcal{P}^\dagger$ given by $\mathcal{F}^\dagger \cap \mathcal{P}^\dagger$. Topologically, $\mathcal{P}^\dagger$ is obtained from $\mathcal{P}_0$ by cutting along the annulus S; as shown in Figure 25, this yields a single remnant — namely, the annulus $S^\dagger$ — which tells us that S is a one-sided annulus in $\mathcal{P}_0$.

Suppose now that the vertices incident to F are all ideal. Consider the pseudomanifold $\mathcal{P}^*$ obtained from $\mathcal{D}_1$ by truncating the vertices in $\varphi(V_0)$. Topologically, $\mathcal{P}^*$ is obtained from $\mathcal{P}^\dagger$ by filling the annulus $S^\dagger$ with a thickened disc; see Figure 24. In other words, $\mathcal{P}^*$ is obtained from $\mathcal{P}_0$ by decomposing along the annulus S.

To see how $\mathcal{P}^*$ and $\mathcal{P}_1$ are related, we need to compare the truncated vertex sets $\varphi(V_0)$ and V_1. For this, we first note that since F is only incident to ideal vertices, each component of L must be a closed surface other than a 2-sphere. Let L^* denote the (possibly disconnected) surface obtained by decomposing L along γ_0 and γ_1; each component of L^* must be a closed surface, but could possibly be a 2-sphere.

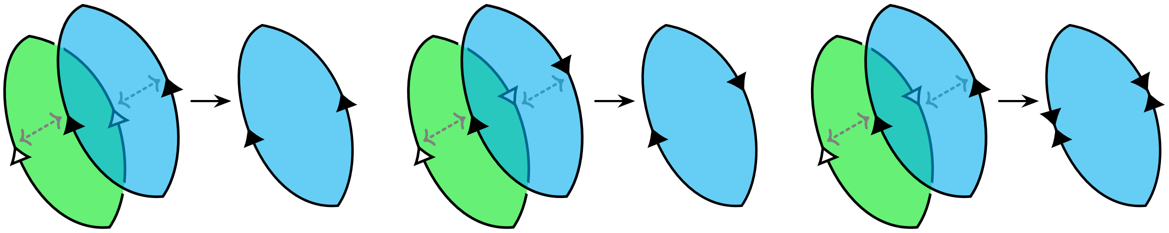


Figure 26: Gluing $F_0^\dagger$ and $F_1^\dagger$ together when each forms a separate 2-sphere or projective plane in the boundary of $\mathcal{D}^\dagger$. Left: gluing two separate 2-spheres results in F forming a 2-sphere. Middle: gluing two separate projective planes results in F forming a projective plane. Right: gluing a 2-sphere and a projective plane results in F being incident to an invalid edge.

By Claim B, the components of L^* correspond precisely to the boundary components of $\mathcal{P}^*$ given by truncating the vertices in $\varphi(v)$. The only way $\mathcal{P}^*$ can differ from $\mathcal{P}_1$ is if L^* has 2-sphere components; we need to fill each such 2-sphere with a 3-ball to recover $\mathcal{P}_1$ from $\mathcal{P}^*$. □

3.2.4 The case where ungluing gives two separate boundary bigon paths We now consider the case where $F_0^\dagger$ and $F_1^\dagger$ form two separate boundary bigon paths. We have the following cases:

Claim E *If $F_0^\dagger$ and $F_1^\dagger$ form two separate boundary bigon paths, then one of the following holds (see Figure 26):*

1. *For each $i \in \{0,1\}$, the boundary bigon path $F_i^\dagger$ forms a 2-sphere. In this case, $\varphi(F)$ consists of two distinct internal edges in $\mathcal{D}_1$. Moreover, the edges of F are identified to form a single internal edge in $\mathcal{D}_0$, and F itself forms a **2-sphere** in $\mathcal{D}_0$. (See Claim E.1 for details about the effect of flattening F in this case.)*
2. *For each $i \in \{0,1\}$, the boundary bigon path $F_i^\dagger$ forms a projective plane. In this case, $\varphi(F)$ consists of two distinct invalid edges in $\mathcal{D}_1$. Moreover, the edges of F are identified to form a single internal edge in $\mathcal{D}_0$, and F itself forms a **projective plane** in $\mathcal{D}_0$. (See Claim E.2 for details about the effect of flattening F in this case.)*
3. *For some $i \in \{0,1\}$, the boundary bigon path $F_i^\dagger$ forms a 2-sphere, but the boundary bigon path $F_{1-i}^\dagger$ forms a projective plane. In this case, the edges of F are identified to form a single invalid edge in $\mathcal{D}_0$.*
4. *For some $i \in \{0,1\}$, the boundary bigon path $F_i^\dagger$ forms a disc, in which case (regardless of the behaviour of $F_{1-i}^\dagger$) F is incident to a boundary edge in $\mathcal{D}_0$.*

Proof For each $i \in \{0,1\}$, depending on how the ends of $F_i^\dagger$ are identified (if at all), $F_i^\dagger$ could form a 2-sphere, projective plane or disc in the boundary of $\mathcal{D}^\dagger$. Observe that:

- If $F_i^\dagger$ forms a 2-sphere, then flattening $F_i^\dagger$ yields a single internal edge.
- If $F_i^\dagger$ forms a projective plane, then flattening $F_i^\dagger$ yields a single invalid edge.

With this in mind, we have the following four cases, which correspond precisely to the cases stated in the claim:

(1) Suppose that $F_0^\dagger$ and $F_1^\dagger$ both form 2-spheres. After the gluing these bigon faces back together, the edges of F are identified so that F forms a 2-sphere, as illustrated in Figure 26 (left).

(2) Suppose that $F_0^\dagger$ and $F_1^\dagger$ both form projective planes. After gluing these bigon faces back together, the edges of F are identified so that F forms a projective plane, as illustrated in Figure 26 (middle).

(3) Suppose that one of $F_0^\dagger$ or $F_1^\dagger$ forms a 2-sphere, whilst the other forms a projective plane. After gluing these bigon faces back together, the edges of F are identified to form a single invalid edge, as illustrated in Figure 26 (right).

(4) Suppose that for some $i \in \{0, 1\}$, $F_i^\dagger$ forms a disc in the boundary of $\mathcal{D}^\dagger$; each end of $F_i^\dagger$ must therefore be incident to a boundary face other than $F_0^\dagger$ or $F_1^\dagger$. Thus, regardless of how we glue $F_0^\dagger$ and $F_1^\dagger$ back together, the bigon face F will always be incident to at least one boundary edge of $\mathcal{D}_0$. □

As before, we only give a detailed analysis of the effect of flattening F in the cases where F is not incident to any boundary or invalid edges. This corresponds to cases (1) and (2) of Claim E.

Claim E.1 (2-sphere) *In case (1) of Claim E, one of the following holds:*

(a) *The bigon face F is only incident to internal vertices (the two vertices of F could either be identified to form a single internal vertex, or they could form two distinct internal vertices). In this case, $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along a properly embedded 2-sphere given by pushing F slightly away from its incident vertices.*

(b) *The bigon face F is incident to one internal vertex and one ideal vertex. In this case, the truncated bigon associated to F forms a properly embedded disc S in $\mathcal{P}_0$; the boundary curve γ of S is a two-sided curve in $\partial\mathcal{P}_0$. Flattening F has one of the following effects:*

- *If γ bounds a disc E in $\partial\mathcal{P}_0$, then $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along a properly embedded 2-sphere given by isotoping $S \cup E$ slightly off the boundary.*
- *If γ does not bound a disc in $\mathcal{P}_0$, then $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by first cutting along the disc S, and then filling any new 2-sphere boundary components with 3-balls.*

(c) *The bigon face F is only incident to ideal vertices (the two vertices of F could either be identified to form a single ideal vertex, or they could form two distinct ideal vertices). In this case, the truncated bigon associated to F forms a two-sided properly embedded annulus S in $\mathcal{P}_0$; the boundary curves γ_0 and γ_1 of S form two-sided curves in $\partial\mathcal{P}_0$. Flattening F has one of the following effects:*

- *If γ_0 and γ_1 respectively bound discs E_0 and E_1 in $\partial\mathcal{P}_0$, then either these discs are disjoint or one of these discs lies entirely in the interior of the other; choose $i \in \{0, 1\}$ so that E_i either lies entirely inside or entirely outside E_{1-i}. In this case, $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along a properly embedded 2-sphere S^* constructed as follows:*

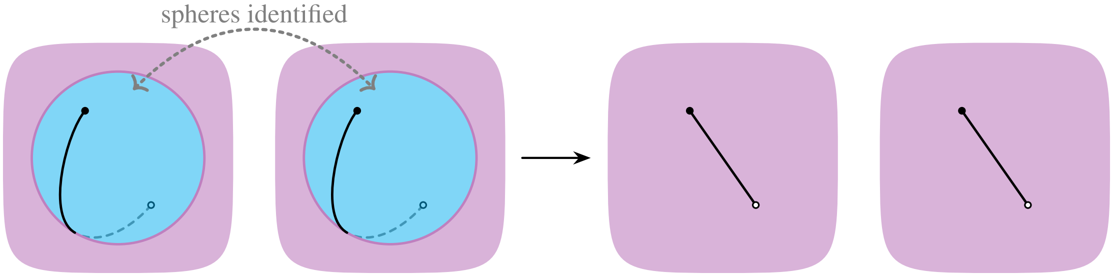


Figure 27: When F' forms a properly embedded 2-sphere, flattening F' is equivalent to decomposing along this 2-sphere.

  (i) *Isotope* $S \cup E_i$ *slightly off the boundary to obtain a properly embedded disc* S' *in* $\mathcal{P}_0$*.*
  (ii) *Isotope* $S' \cup E_{1-i}$ *slightly off the boundary to obtain the desired* 2*-sphere* S^**.*

- *If for some* $i \in \{0,1\}$*, the curve* γ_i *bounds a disc* E_i *in* $\partial\mathcal{P}_0$*, but* γ_{1-i} *does not bound a disc in* $\partial\mathcal{P}_0$*, then* $\mathcal{P}_1$ *is obtained from* $\mathcal{P}_0$ *by first cutting along a disc given by isotoping* $S \cup E_i$ *slightly away from* E_i*, and then filling any new* 2*-sphere boundary components with* 3*-balls.*
- *If neither* γ_0 *nor* γ_1 *bounds a disc in* $\partial\mathcal{P}_0$*, then* $\mathcal{P}_1$ *is obtained from* $\mathcal{P}_0$ *by first decomposing along the annulus* S*, and then filling any new* 2*-sphere boundary components with* 3*-balls.*

(d) *There is a boundary or invalid vertex incident to* F*.*

Proof Recall that in case (1) of Claim E, the edges of F are identified so that F forms a 2-sphere. The two vertices of F could either be identified to form a single vertex of $\mathcal{D}_0$, or they could form two distinct vertices of $\mathcal{D}_0$. Let L denote the union of the links of the vertices incident to F, and let γ_0 and γ_1 denote the two curves in which F meets L. For each $i \in \{0,1\}$, let v_i denote the vertex of F that is cut off by the curve γ_i, and let L_i denote the link of v_i; if v_i is ideal, then we also think of L_i as the ideal boundary component of $\mathcal{P}_0$ given by truncating v_i. If v_0 and v_1 are identified, then $L = L_0 = L_1$; otherwise, L is the disjoint union of L_0 and L_1. With all this setup in mind, we start by getting the easy cases out of the way:

- If there is a boundary or invalid vertex incident to F, then we are in case (d).
- If the vertices incident to F are all internal, then F forms a 2-sphere in $\mathcal{P}_0$. This 2-sphere might not be embedded, since the two vertices of F could be identified to form a single internal vertex. Thus, to ensure that we have a properly embedded 2-sphere, we use Claim C to flatten the v_0-cone over γ_0 and the v_1-cone over γ_1, one at a time. This reduces the operation of flattening F to the operation of flattening a new bigon F' given by pushing F slightly away from its incident vertices. Since the vertices of F' form two distinct temporary internal vertices, F' forms the desired properly embedded 2-sphere in $\mathcal{P}_0$. As shown in Figure 27, flattening F' is topologically equivalent to decomposing along this 2-sphere. This proves case (a).

We now consider the case where F is incident to one internal vertex and one ideal vertex; without loss of generality, suppose that v_0 is the internal vertex and v_1 is the ideal vertex. We need to prove

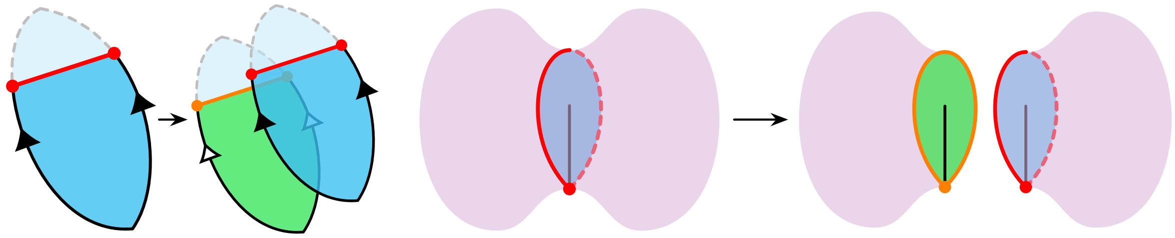


Figure 28: The truncated bigon associated to F forms a properly embedded disc S. Cutting along S yields two disc remnants.

the conclusions stated in case (b). Observe that the truncated bigon associated to F forms a properly embedded disc S in $\mathcal{P}_0$. Consider the pseudomanifold $\mathcal{P}^\dagger$ obtained from $\mathcal{D}^\dagger$ by truncating the vertices in $g^{-1}(V_0)$; viewing $\mathcal{P}^\dagger$ as a subset of $\mathcal{D}^\dagger$, for each $i \in \{0,1\}$ let $S_i^\dagger$ denote the disc in $\partial\mathcal{P}^\dagger$ given by $\mathcal{F}_i^\dagger \cap \mathcal{P}^\dagger$. Topologically, $\mathcal{P}^\dagger$ is obtained from $\mathcal{P}_0$ by cutting along the disc S; the two discs $S_0^\dagger$ and $S_1^\dagger$ form the remnants of cutting along S. This is illustrated in Figure 28. We also note that the boundary of the disc S is given by the curve γ_1; since cutting along S splits γ_1 into two remnants, one bounding each of the discs $S_0^\dagger$ and $S_1^\dagger$, we see that γ_1 is a two-sided curve in L_1. It remains to describe how flattening F changes $\mathcal{P}_0$. This depends on whether γ_1 bounds a disc in L_1:

- Suppose that γ_1 bounds a disc E in L_1. We can use Claim C to flatten the v_1-cone over γ_1. This reduces the operation of flattening F to the operation of flattening a new bigon F' such that
  - one of the vertices of F' is the temporary internal vertex that results from flattening γ_1; and
  - the other vertex of F' is the internal vertex v_0.

Topologically, F' is equivalent to a properly embedded 2-sphere given by isotoping $S \cup E$ slightly off the boundary of $\mathcal{P}_0$. Moreover, by analogy with case (a), we see that $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along this 2-sphere F'.

- Suppose that γ_1 does not bound a disc in L_1. Consider the pseudomanifold $\mathcal{P}^*$ obtained from $\mathcal{D}_1$ by truncating the vertices in $\varphi(V_0)$. As shown in Figure 29, $\mathcal{P}^*$ is obtained from $\mathcal{P}^\dagger$ by collapsing $S_0^\dagger$ and $S_1^\dagger$ to arcs, which has no topological effect; in other words, $\mathcal{P}^*$ is homeomorphic to $\mathcal{P}^\dagger$. To see how $\mathcal{P}^*$ is related to $\mathcal{P}_1$, consider the surface L^* obtained from L_1 by decomposing along γ_1. Using Claim B, we

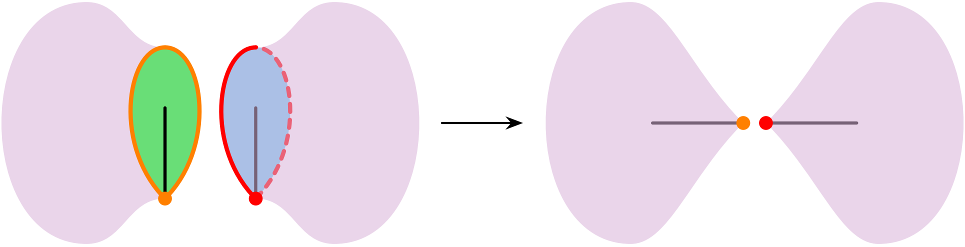


Figure 29: Collapsing the discs $S_0^\dagger$ and $S_1^\dagger$ to arcs has no topological effect.

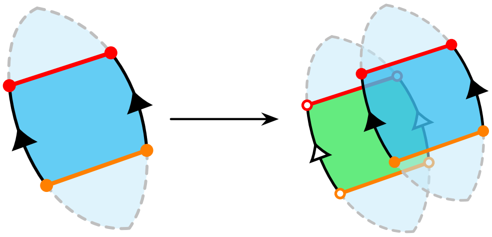


Figure 30: The truncated bigon associated to F forms a two-sided annulus S. Cutting along S yields two annulus remnants.

see that each component of L^* corresponds to a boundary component of $\mathcal{P}^*$ given by truncating one of the vertices in $\varphi(v_1)$. Thus, if $\varphi(v_1)$ contains any internal vertices (since γ is two-sided and does not bound a disc in L_1, this is only possible if L_1 is a torus or a Klein bottle), then we need to fill each corresponding 2-sphere boundary component of $\mathcal{P}^*$ with a 3-ball to recover $\mathcal{P}_1$.

This proves case (b).

All that remains is to consider the case where the vertices incident to F are all ideal; we need to prove the conclusions stated in case (c). This time, the truncated bigon associated to F forms a properly embedded annulus S in $\mathcal{P}_0$. Similar to before, consider the pseudomanifold $\mathcal{P}^\dagger$ obtained from $\mathcal{D}^\dagger$ by truncating the vertices in $g^{-1}(V_0)$; viewing $\mathcal{P}^\dagger$ as a subset of $\mathcal{D}^\dagger$, for each $i \in \{0,1\}$ let $S_i^\dagger$ denote the annulus in $\partial\mathcal{P}^\dagger$ given by $\mathcal{F}_i^\dagger \cap \mathcal{P}^\dagger$. Topologically, $\mathcal{P}^\dagger$ is obtained from $\mathcal{P}_0$ by cutting along the annulus S; as shown in Figure 30, this yields a pair of remnants — namely, the annuli $S_0^\dagger$ and $S_1^\dagger$ — which means that S is a *two-sided* annulus in $\mathcal{P}_0$. We also note that for each $i \in \{0,1\}$, the curve γ_i forms a two-sided curve in L_i, since cutting along S causes this curve to split into two remnants, one meeting $S_0^\dagger$ and the other meeting $S_1^\dagger$.

Depending on whether γ_0 and γ_1 bound discs in $\partial\mathcal{P}_0$, we have the following possibilities for how flattening F changes $\mathcal{P}_0$:

- Suppose that for some $i \in \{0,1\}$, γ_i bounds a disc E_i in L_i. We can assume without loss of generality that the interior of E_i is disjoint from F. To see why, note that the only way this assumption can fail is if γ_{1-i} lies in the interior of E_i; in this case, γ_{1-i} bounds a "smaller" disc, so we can simply exchange the roles of γ_i and γ_{1-i}. This allows us to use Claim C to flatten the v_i-cone over γ_i. This reduces the operation of flattening F to the operation of flattening a new bigon F' such that
  - one of the vertices of F' is given by the temporary internal vertex that results from flattening γ_i; and
  - the other vertex of F' is given by the ideal vertex v_{1-i}.

Thus, flattening F' has the same topological effect as flattening F in case (b). In more detail, after truncating the ideal vertex v_{1-i}, we see that F' becomes a properly embedded disc S' in $\mathcal{P}_0$; we can view γ_{1-i} as the boundary curve of this disc S'. Topologically, S' is obtained by isotoping $S \cup E_i$ slightly off the boundary of $\mathcal{P}_0$, and the effect of flattening F' depends on whether γ_{1-i} bounds a disc in L_{1-i}:

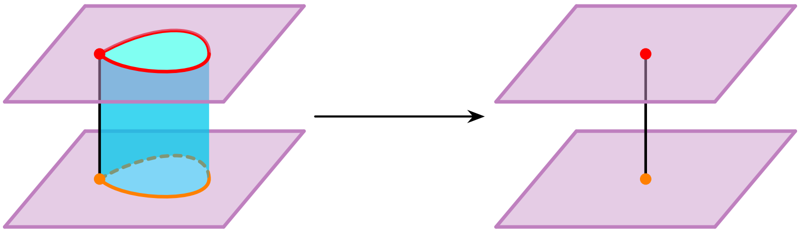


Figure 31: For each $i \in \{0,1\}$, flattening $F_i^\dagger$ has the effect of filling the annulus remnant $S_i^\dagger$ with a thickened disc.

- If γ_{1-i} bounds a disc E_{1-i} in L_{1-i}, then $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along a properly embedded 2-sphere given by isotoping $S' \cup E_{1-i}$ slightly off the boundary of $\mathcal{P}_0$.
- If γ_{1-i} does not bound a disc in L_{1-i}, then $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by first cutting along the disc S', and then filling any new 2-sphere boundary components with 3-balls.

- Suppose that neither γ_0 nor γ_1 bounds a disc in $\partial\mathcal{P}_0$. Consider the pseudomanifold $\mathcal{P}^*$ obtained from $\mathcal{D}_1$ by truncating the vertices in $\varphi(V_0)$. Topologically, $\mathcal{P}^*$ is obtained from $\mathcal{P}^\dagger$ by filling the annuli $S_0^\dagger$ and $S_1^\dagger$ with thickened discs; see Figure 31. In other words, $\mathcal{P}^*$ is obtained from $\mathcal{P}_0$ by decomposing along the annulus S. To see how $\mathcal{P}^*$ is related to $\mathcal{P}_1$, consider the surface L^* obtained from $L_0 \cup L_1$ by decomposing along γ_0 and γ_1. Using Claim B, we see that each component of L^* corresponds to a boundary component of $\mathcal{P}^*$ given by truncating one of the vertices in $\varphi(v_0)$ or $\varphi(v_1)$. Thus, if there are any internal vertices in $\varphi(v_0)$ or $\varphi(v_1)$, then we need to fill each corresponding 2-sphere boundary component of $\mathcal{P}^*$ with a 3-ball to recover $\mathcal{P}_1$.

This proves case (c). □

Claim E.2 (projective plane) *In case* (2) *of Claim E, the two vertices of F are identified to form a single vertex v, and one of the following holds*:

(a) *The vertex v is internal, in which case F forms a two-sided properly embedded projective plane in $\mathcal{P}_0$, and $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along this projective plane.*

(b) *The vertex v is ideal, in which case the truncated bigon associated to F forms a two-sided properly embedded Möbius band S in $\mathcal{P}_0$; the boundary curve γ of S forms a two-sided curve in $\partial\mathcal{P}_0$. In this case, flattening F has one of the following effects*:
  - *If γ bounds a disc E in $\partial\mathcal{P}_0$, then $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along a two-sided projective plane given by isotoping $S \cup E$ slightly off the boundary of $\mathcal{P}_0$.*
  - *If γ does not bound a disc in $\partial\mathcal{P}_0$, then $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by first decomposing along the Möbius band S, and then filling any new 2-sphere boundary components with 3-balls.*

(c) *The vertex v is boundary or invalid.*

Before we prove Claim E.2, we need to understand the topological effect of flattening a boundary bigon face whose corresponding truncated bigon forms either a projective plane or a Möbius band. This is the purpose of the following lemma:

Lemma 13 *Let B be a boundary bigon face, with edges identified so that B forms a projective plane. Thus, the associated truncated bigon B' forms either a projective plane or a Möbius band. Topologically, flattening B has the effect of filling B' with an invalid cone, as described in Section 2.2.*

Proof Let $\mathcal{P}$ denote the truncated pseudomanifold corresponding to the cell decomposition containing B. We first consider the case where the truncated bigon B' forms a projective plane boundary component of $\mathcal{P}$; in other words, we are considering the case where the vertex incident to B does not get truncated, and hence $B' = B$. We produce a copy B_0 of B by an isotopy of B that

- fixes the vertex; and
- pushes the rest of B slightly into the interior of $\mathcal{P}$.

Let $\mathcal{R}$ denote the region that is swept out by this isotopy, and let $\mathcal{C}$ denote the quotient of $\mathcal{R}$ obtained by flattening B.

By construction, deleting $\mathcal{R} - B_0$ preserves the ambient space $\mathcal{P}$ up to homeomorphism, and replaces the boundary component B with its copy B_0. Thus, we see that flattening B has the same topological effect as attaching a copy of $\mathcal{C}$ to B. It therefore suffices to show that $\mathcal{C}$ is actually an invalid cone. Our strategy for this is to express $\mathcal{R}$ as a union of lines such that after flattening B, these lines exhibit the structure of a cone over the projective plane B_0.

To this end, begin with $\mathcal{Q} := [0,1]^3$. Let Λ denote the set of lines of the form $\{x\}\times\{y\}\times[0,1]$, so that $\mathcal{Q}$ is a (disjoint) union of the lines in Λ. Take $\sim$ to be the minimal equivalence relation satisfying the following:

- $(x,0,0) \sim (x',0,0)$ for all $x, x' \in [0,1]$.
- $(x,1,0) \sim (x',1,0)$ for all $x, x' \in [0,1]$.
- $(x,y,1) \sim (x,y',1)$ for all $x, y, y' \in [0,1]$.
- $(0,y,z) \sim (1,1-y,z)$ for all $y, z \in [0,1]$.

In the quotient $\mathcal{Q}/{\sim}$, the $z=0$ rectangle becomes a projective plane P_0, and the $y=0$ and $y=1$ rectangles become a pair of triangles that together form a projective plane P_1; see Figure 32. Indeed, we can identify this quotient with $\mathcal{R}$ in such a way that P_0 is identified with B_0 and P_1 is identified with B.

With this identification, we can express $\mathcal{R}$ as a union of the lines given by $\Lambda/{\sim}$. Each of these lines joins a point in B_0 to a point on the curve $\{(x,*,1)\}$. Moreover, we observe the following:

- For each point of the form $(x,*,1)$, we have exactly two lines that join this point to the vertex of B_0. The union of these lines is precisely the projective plane B.
- For every point in B_0 other than the vertex, there is a unique line that joins this point to the curve $\{(x,*,1)\}$.

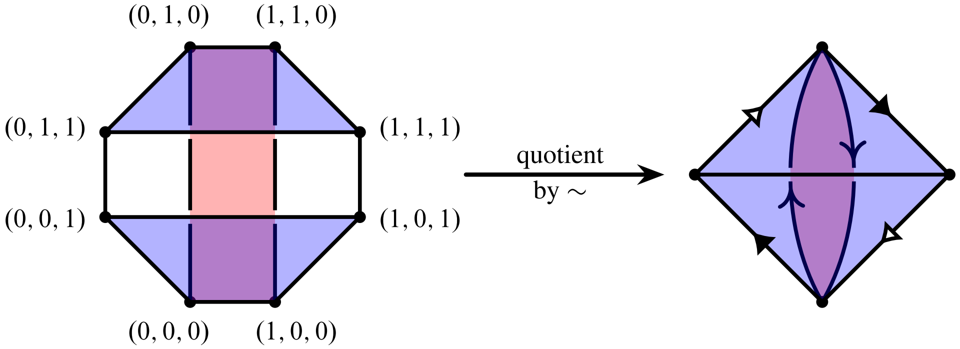


Figure 32: The region $\mathcal{R}$ can be constructed as a quotient of the box $\mathcal{Q} := [0,1]^3$. The $z = 0$ rectangle becomes the projective plane B_0 (shaded red). The $y = 0$ and $y = 1$ rectangles together become the projective plane B (shaded blue).

After flattening B, the curve $\{(x, *, 1)\}$ gets collapsed to a single "apex" point a, and each point in the projective plane B_0 is now joined to a by a unique line. This exhibits precisely the required cone structure on $\mathcal{C}$, and hence we conclude that flattening B is topologically equivalent to filling B' with this invalid cone $\mathcal{C}$.

What remains is the case where B' forms a Möbius band in $\partial\mathcal{P}$. In this case, flattening B is still equivalent to attaching an invalid cone, except we need to account for the fact that the vertex incident to B is truncated. This truncation only removes a small neighbourhood of the vertex, and this removed neighbourhood is disjoint from the apex of the cone. Thus, we still have an invalid cone $\mathcal{C}$, but the truncation means that there is a small disc $D \subset \partial\mathcal{C}$ that lies inside $\partial\mathcal{P}$, and $\mathcal{C}$ is attached by identifying the Möbius bands B' and $\partial\mathcal{C} - D$. In other words, we again conclude that flattening B is topologically equivalent to filling B' with the invalid cone $\mathcal{C}$. □

Proof of Claim E.2 Recall that in case (2) of Claim E, the edges of F are identified so that F forms a projective plane. We will see that the proof is almost identical to the proof of Claim D.2; the main difference is that here, we end up working with embedded surfaces that are two-sided rather than one-sided.

Since F forms a projective plane, the two vertices of F are identified to form a single vertex v in $\mathcal{D}_0$. When v is not ideal, the claim is easy to prove:

- If v is boundary or invalid, then we are in case (c).
- If v is internal, then F forms an embedded projective plane in the interior of $\mathcal{P}_0$. Ungluing F yields two projective plane remnants corresponding to the boundary bigon paths $F_0^\dagger$ and $F_1^\dagger$, which tells us that F forms a *two-sided* projective plane in $\mathcal{P}_0$. Moreover, by Lemma 13, flattening $F_0^\dagger$ and $F_1^\dagger$ corresponds to filling these two projective plane boundary components with invalid cones. Altogether, we see that flattening F is topologically equivalent to decomposing along F. This proves case (a).

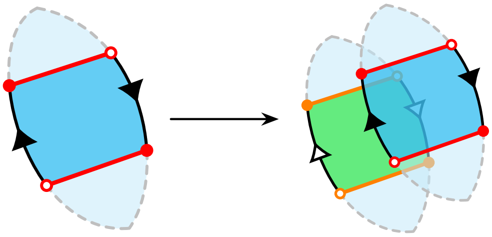


Figure 33: The truncated bigon associated to F forms a two-sided Möbius band S. Cutting along S yields two Möbius band remnants.

With that out of the way, suppose for the rest of this proof that v is ideal; we need to prove all the conclusions stated in case (b). Observe that the truncated bigon associated to F forms a properly embedded Möbius band S in $\mathcal{P}_0$. Consider the pseudomanifold $\mathcal{P}^\dagger$ obtained from $\mathcal{D}^\dagger$ by truncating the vertices in $g^{-1}(V_0)$; viewing $\mathcal{P}^\dagger$ as a subset of $\mathcal{D}^\dagger$, for each $i \in \{0, 1\}$ let $S_i^\dagger$ denote the Möbius band in $\partial\mathcal{P}^\dagger$ given by $F_i^\dagger \cap \mathcal{P}^\dagger$. Topologically, $\mathcal{P}^\dagger$ is obtained from $\mathcal{P}_0$ by cutting along the Möbius band S; as shown in Figure 33, this yields two remnants — namely, the Möbius bands $S_0^\dagger$ and $S_1^\dagger$ — so S must be a *two-sided* Möbius band in $\mathcal{P}_0$.

Consider the ideal boundary component L of $\mathcal{P}_0$ given by truncating the vertex v, and let γ denote the boundary curve of S; we need to show that γ forms a two-sided curve in L. For this, it suffices to observe that cutting along S has the effect of splitting γ into two curves, one bounding the Möbius band $S_0^\dagger$ and the other bounding the Möbius band $S_1^\dagger$.

All that remains is to understand the overall effect of flattening F. We begin with the case where γ bounds a disc E in L. In this case, we use Claim C to flatten the v-cone over γ. This reduces the operation of flattening F to the operation of flattening a new bigon F' given by pushing F slightly away from v; topologically, F' is equivalent a properly embedded projective plane given by isotoping $S \cup E$ slightly off the boundary of $\mathcal{P}_0$. Since the vertices of F' are identified to form a single temporary internal vertex, flattening F' has the same topological effect as flattening F in the case where v is internal (case (a)). In other words, F' forms a *two-sided* projective plane in $\mathcal{P}_0$, and $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along this projective plane. This completes the case where γ bounds a disc in L.

For the case where γ does *not* bound a disc in L, consider the pseudomanifold $\mathcal{P}^*$ obtained from $\mathcal{D}_1$ by truncating the vertices in $\varphi(V_0)$. Topologically, by Lemma 13, $\mathcal{P}^*$ is obtained from $\mathcal{P}^\dagger$ by filling the Möbius bands $S_0^\dagger$ and $S_1^\dagger$ with invalid cones; in other words, $\mathcal{P}^*$ is obtained from $\mathcal{P}_0$ by decomposing along the Möbius band S. To see how $\mathcal{P}^*$ is related to $\mathcal{P}_1$, we compare the truncated vertex sets $\varphi(V_0)$ and V_1. For this, let L^* denote the surface obtained by decomposing L along γ. Claim B tells us that the components of L^* correspond to boundary components of $\mathcal{P}^*$ given by truncating the vertices in $\varphi(v)$. The only way $\mathcal{P}^*$ can differ from $\mathcal{P}_1$ is if L^* has 2-sphere components; we need to fill each such 2-sphere with a 3-ball to recover $\mathcal{P}_1$ from $\mathcal{P}^*$. This completes the proof of case (b). □

4 Crushing surfaces of positive genus

Consider a normal surface S in a 3-manifold $\mathcal{M}$. Roughly, our goal in this section is to give sufficient conditions under which crushing S gives an ideal triangulation of a component of $\mathcal{M}-S$. For this, we fix the following notation throughout this section:

- Let $\mathcal{M}$ be a (compact) 3-manifold with no 2-sphere boundary components. If $\mathcal{M}$ is closed, let $\mathcal{T}$ be a closed triangulation of $\mathcal{M}$; otherwise, if $\mathcal{M}$ is bounded, let $\mathcal{T}$ be an ideal triangulation of $\mathcal{M}$.
- Let S be a (possibly disconnected) separating normal surface in $\mathcal{T}$. (Since $\mathcal{T}$ has no real boundary components, note that S must be a closed surface.) Assume that every component of S is two-sided, and that none of these components are 2-spheres.
- Fix any particular component of $\mathcal{M}-S$ that meets each component of S on exactly one side, and call it the *chosen region* for S; also, for any (not necessarily normal) surface E isotopic to S, call the corresponding component of $\mathcal{M}-E$ the chosen region for E. Let X denote the compact 3-manifold given by the closure of the chosen region for S.

The assumptions that we have made on S and on the chosen region are not as restrictive as they might appear at first glance. If S has a component C that is either nonseparating or one-sided (or both), then we can always "repair" S as follows: build a new surface Σ by replacing C with the frontier of a regular neighbourhood of C. Up to homeomorphism, each component of $\mathcal{M}-S$ appears as a component of $\mathcal{M}-\Sigma$, so we lose nothing by repairing S in this way. A similar trick allows us to deal with components of $\mathcal{M}-S$ that meet a component of S on both sides.

With this in mind, let $\mathcal{T}'$ denote the triangulation obtained by (destructively) crushing S. As we hinted earlier, our goal is to give sufficient conditions for $\mathcal{T}'$ to include an ideal triangulation of X as one of its components (the precise statement is given in Theorem 1).

To this end, consider the cell decomposition $\mathcal{D}'$ given by *nondestructively* crushing S. One of the components $\mathcal{D}^*$ of $\mathcal{D}'$ gives a destructible ideal cell decomposition of X; see Figure 34. Call a 3-cell in $\mathcal{D}'$ *benign* if and only if it belongs to the component $\mathcal{D}^*$.

From the formulation of crushing given in Section 2.5 (in particular, recall Definitions 5 and Lemma 7), $\mathcal{T}'$ is obtained from $\mathcal{D}'$ by a finite sequence of atomic moves. Since the atomic moves only either destroy or

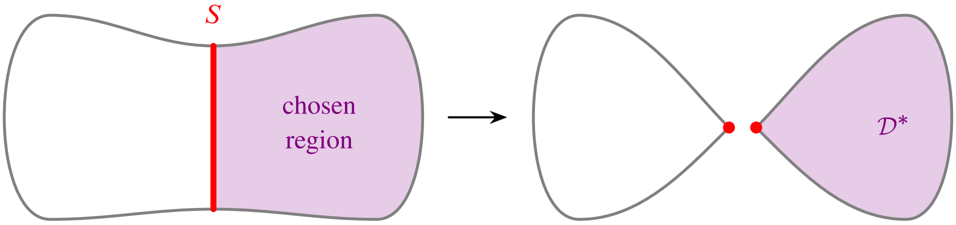


Figure 34: Schematic illustration of nondestructively crushing S, in the case where S is connected.

$$\begin{tikzcd}
\mathcal{T} \arrow[r, "\text{nondestructively}", "\text{crushing } S"'] \arrow[rr, bend left, "\text{destructively crushing } S"] & \mathcal{D}' \arrow[r, "\text{flattening}"] & \mathcal{T}' \\
& \mathcal{D}^* \arrow[u, phantom, "\cup" description] \arrow[r, "\text{flattening}"'] & \mathcal{T}^* \arrow[u, phantom, "\cup" description] \\
& & \textit{benign}
\end{tikzcd}$$

Figure 35: Some notation that we use throughout Section 4.

modify the existing 3-cells in $\mathcal{D}'$, we can naturally speak about benign 3-cells in all of the intermediate cell decompositions that we encounter as we perform the atomic moves.

We will also call a component of a cell decomposition *benign* if this component is built entirely from benign 3-cells. Whilst $\mathcal{D}'$ initially has exactly one benign component, namely $\mathcal{D}^*$, this number can change as we perform atomic moves.

With this in mind, let $\mathcal{T}^*$ denote the triangulation consisting only of the benign components of $\mathcal{T}'$; see Figure 35 for a visual summary of all the notation we have just introduced. We can now give a more precise statement of our main goal in this section: we want to show that one of the components of $\mathcal{T}^*$ gives an ideal triangulation of X.

The proof boils down to checking that we do not make "drastic" topological changes when performing atomic moves on an ideal cell decomposition of X. In all but one of the cases, it is enough to require that X satisfies the following conditions:

- It is irreducible and ∂-irreducible.
- It contains no essential annuli and no two-sided properly embedded Möbius bands.

The one difficult case is when we flatten a bigon whose corresponding truncated bigon forms a boundary-parallel annulus in X; we will discuss how we circumvent this difficulty in Section 4.1. We then put everything together to prove Theorem 1 in Section 4.2.

4.1 Avoiding bad bigon paths

Throughout the rest of Section 4, call a bigon path (recall Definitions 12) *bad* if it is internal and its corresponding truncated bigon path forms a boundary-parallel annulus. Using this terminology, the difficult case that we mentioned earlier is when we flatten a bad bigon path of length one. In Section 4.2, we will see that the only way to have a bad bigon path of length one is if the cell decomposition $\mathcal{D}^*$ initially contained a bad bigon path (of some arbitrary length). With this in mind, our goal is to cut this problem off at the source: we will give conditions on the surface S that will ensure that $\mathcal{D}^*$ does not contain any bad bigon paths.

Roughly, the idea is that if $\mathcal{D}^*$ contains a bad bigon path, then we can "push" or "expand" S further into the chosen region; thus, we would like to ensure that S cannot be "expanded" in this way. To make this

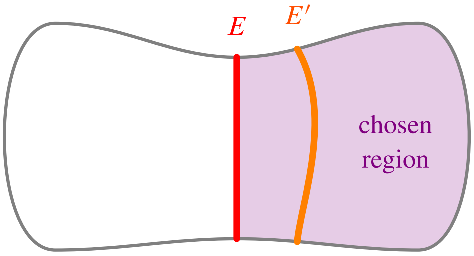


Figure 36: Let E and E' be two disjoint normal surfaces in the same isotopy class. If E' cannot be *normally* isotoped to lie outside the chosen region for E, then it is an expansion of E.

idea precise, we introduce the following terminology for any normal surface E in the isotopy class of S:

- Let E' be any normal surface that is, up to normal isotopy, disjoint from E. We call E' an *expansion* of E if it is isotopic to E, but cannot be normally isotoped to lie entirely outside the chosen region for E; see Figure 36.
- Call E *maximal* if it does not admit such an expansion.

In Lemma 15, we will show that to avoid bad bigon paths in $\mathcal{D}^*$, it is enough to assume that X is irreducible, and that S is incompressible and maximal. Before we do so, it is worth noting that the maximality assumption is not very restrictive. Specifically, the following result says that, up to isotopy, we can always choose S to be maximal:

Lemma 14 *There is a maximal normal surface in the isotopy class of S.*

One way to prove Lemma 14 would be to appeal to Kneser's finiteness theorem (as is done, for instance, in [18, pages 160–161]); however, we do not actually need the full strength of this theorem. The following simple proof distils precisely the part of Kneser's finiteness theorem that is necessary:

Proof of Lemma 14 If S is itself maximal, then there is nothing to prove; thus, assume for the rest of this proof that S is not maximal. Let $E_0 = S$, and consider any sequence $E_0, \dots, E_n$ of normal surfaces such that for each $i \in \{1, \dots, n\}$, E_i is an expansion of E_{i-1}. The idea is to show that such a sequence cannot be extended indefinitely, which means that after extending as much as possible, the final entry of the sequence will be maximal.

To this end, consider the surface $E_0 \cup \dots \cup E_n$; this is a normal surface, since $E_0, \dots, E_n$ are mutually disjoint, up to normal isotopy. Let $\mathcal{C}$ denote the induced cell decomposition obtained by cutting along $E_0 \cup \dots \cup E_n$. For each $i \in \{1, \dots, n\}$, let B_i denote the component of $\mathcal{C}$ given by the trivial I-bundle between E_i and E_{i-1}. Since E_i and E_{i-1} are not normally isotopic, B_i must contain at least one nonparallel cell. However, each tetrahedron of $\mathcal{T}$ gives rise to at most six nonparallel cells, so we see that $n \leqslant 6|\mathcal{T}|$. This implies that there is a sequence $E_0, \dots, E_k$ of expansions whose length k is maximum among all such sequences; the surface E_k must therefore be a maximal normal surface in the isotopy class of S, otherwise we would be able to extend the sequence by adding an expansion of E_k. □

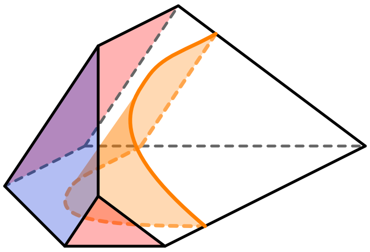


Figure 37: An example of an induced cell Δ that meets the chosen region for S_1. The red faces are parts of S_1 that originated from S, and the blue bridge face is a part of S_1 that originated from A. The intersection of S_2 with Δ is shaded orange. In this example, the orange piece forms an elementary triangle, which means that this piece is preserved by the normalisation procedure; thus, the final surface S_4 will include this orange triangle among its elementary discs.

Lemma 15 *If X is irreducible, S is maximal, and the remnants of S are incompressible in X, then $\mathcal{D}^*$ contains no bad bigon paths.*

Proof Suppose $\mathcal{D}^*$ contains a bad bigon path $\mathcal{F}$. We will show that S cannot be maximal by using $\mathcal{F}$ to construct an expansion of S.

To do this, let $\mathcal{D}$ denote the cell decomposition of X induced by S, and let $q\colon \mathcal{D} \to \mathcal{D}^*$ be the quotient map given by nondestructively crushing S. Observe that $q^{-1}(\mathcal{F})$ realises an annulus A that

- consists of bridge faces (as defined in Definitions 3) in the 2-skeleton of $\mathcal{D}$; and
- is parallel to an annulus $A^{\parallel}$ lying entirely inside a component of S.

We now aim to build a sequence S_1, S_2, S_3, S_4 of surfaces isotopic to S, each of which is "expanded further" into the chosen region than the last. Our goal is for S_4 to be the required expansion of S. We achieve this as follows:

(1) Let S_1 be the surface obtained from S by replacing $A^{\parallel}$ with A.

(2) Consider the boundary B of a regular neighbourhood of $S \cup A$, and let S_2 be the union of the components of B that lie inside the chosen region for S_1; see Figure 37. Since the chosen region for S_1 meets each component of S_1 on exactly one of its two sides, observe that S_2 is isotopic to S. By Theorem 4, B is a barrier for any component of $\mathcal{M} - B$ that does not meet $S \cup A$; observe that the chosen region $\mathcal{N}$ for S_2 is one such component of $\mathcal{M} - B$.

(3) Let S_3 be the surface given by isotoping S_2 slightly into $\mathcal{N}$.

(4) Using the barrier B, normalise S_3 to obtain a normal surface E in $\mathcal{N}$. Since the remnants of S are incompressible in X, we know that S_3 must be incompressible in $\mathcal{N}$. This, together with the fact that X is irreducible, implies that after deleting any 2-sphere components of E, we must be left with a normal surface S_4 isotopic to S. By construction, S_4 is disjoint from S, and it cannot be normally isotoped to lie outside the chosen region for S, so it is an expansion of S. □

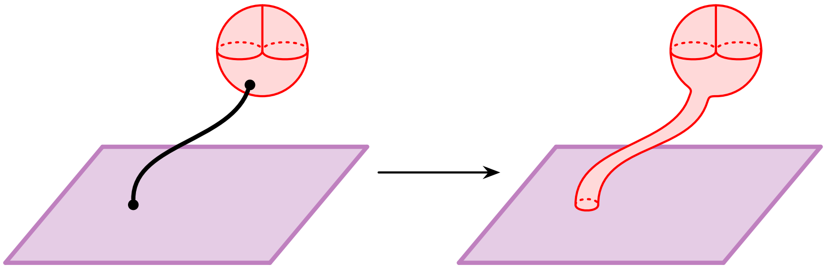


Figure 38: Schematic illustration of using a tube to turn a two-sided projective plane into a two-sided Möbius band.

4.2 Crushing the benign components

In Theorem 1 below, we give sufficient conditions so that after crushing S, one of the components of the triangulation $\mathcal{T}^*$ gives an ideal triangulation of X. Specifically, our proof relies on the following:

- We require that S is maximal. As discussed in Section 4.1, this is not a serious restriction, thanks to Lemma 14.

- We require that X is irreducible, ∂-irreducible and *anannular* (ie X contains no essential annuli). These are quite common "niceness" conditions for 3-manifolds. It is worth noting that we do *not* need to assume that X is *atoroidal* (ie X contains no essential tori).

- We require that X contains no two-sided properly embedded Möbius bands. This condition holds for all orientable 3-manifolds, but sometimes fails for nonorientable 3-manifolds.

- We require that X contains no two-sided properly embedded projective planes. This follows from the previous condition, together with the fact that X has at least one boundary component (given by a remnant of S). To see why, suppose X contains a two-sided projective plane E. Consider a path γ that starts at a point p_0 in E and ends at a point p_1 in ∂X. Remove a small disc around p_0 from E, and replace it with a thin tube that "follows" the path γ and ends with a curve that bounds a small disc around p_1 in ∂X; see Figure 38. This turns E into a two-sided properly embedded Möbius band in X.

Theorem 1 *Suppose that X is irreducible, ∂-irreducible and anannular, and that it contains no two-sided properly embedded Möbius bands. Also, suppose S is maximal. Then $\mathcal{T}^*$ is a valid triangulation such that:*

- *One of its components is an ideal triangulation of X.*
- *Every other component is a triangulation of the 3-sphere.*

Proof Throughout this proof, call a cell decomposition *acceptable* if:

- One of its components is an ideal cell decomposition of X.
- Every other component is a (closed) cell decomposition of the 3-sphere.

$$\begin{tikzcd}
\mathcal{D}_0 \arrow[rr, bend left, dashed, "{\varphi, \text{ flatten } F}"] \arrow[d, "\text{truncate}"'] & & \mathcal{D}_1 \arrow[ll, bend left, dashed, "\psi"] \arrow[d, "\text{truncate}"] \\
\mathcal{P}_0 & & \mathcal{P}_1
\end{tikzcd}$$

Figure 39: Notation for the inductive step in Theorem 1.

Recall from Lemma 7 that the procedure of flattening $\mathcal{D}^*$ to get $\mathcal{T}^*$ can be realised by a finite sequence of atomic moves. Thus, our strategy will be to inductively prove that each atomic move preserves the property of being acceptable.

For the proof to work, we actually need to prove slightly more than this. The problem is that flattening a bad bigon path (as defined in Section 4.1) has the topological effect of decomposing along a boundary-parallel annulus, which could potentially yield a cell decomposition that is no longer acceptable. To circumvent this problem, we will show by induction that performing any number of atomic moves on $\mathcal{D}^*$ always yields an acceptable cell decomposition *that contains no bad bigon paths*.

For the base case, consider the initial cell decomposition $\mathcal{D}^*$ (which is obtained by performing zero atomic moves). Recall that $\mathcal{D}^*$ is an ideal cell decomposition of X (with no extra 3-sphere components), so it is acceptable. By Lemma 15, we also know that $\mathcal{D}^*$ contains no bad bigon paths.

For the inductive step, assume that we have some acceptable cell decomposition $\mathcal{D}_0$ that contains no bad bigon paths. We need to show that performing any atomic move on $\mathcal{D}_0$ yields a new acceptable cell decomposition $\mathcal{D}_1$ that has no bad bigon paths. In particular, to show that $\mathcal{D}_1$ remains acceptable, we will show that an atomic move always either: has no topological effect on the truncated pseudomanifold, or changes the truncated pseudomanifold by adding or removing a 3-sphere component.

Throughout the rest of this proof, let F denote the triangular pillow, bigon pillow or bigon face in $\mathcal{D}_0$ that we flatten to obtain $\mathcal{D}_1$. Let φ denote the flattening map associated to this atomic move, and let ψ denote the inverse flattening map. For each $i \in \{0, 1\}$, let $\mathcal{P}_i$ denote the truncated pseudomanifold of $\mathcal{D}_i$. This notation is summarised in Figure 39.

We first consider the case where F is a triangular pillow. Recall from Lemma 10 that the effect of flattening F depends on whether the two triangular faces of F are identified:

- If the faces of F are *not* identified, then we are in case (a) of Lemma 10. Thus, the truncated pseudomanifolds of $\mathcal{D}_0$ and $\mathcal{D}_1$ are homeomorphic, which implies that $\mathcal{D}_1$ is acceptable. Suppose for the sake of contradiction that $\mathcal{D}_1$ contains a bad bigon path B_1. Observe that B_1 meets the triangle $\varphi(F)$ in some (possibly empty) subset of the edges of this triangle, which implies that $\psi(B_1)$ is a bad bigon path in $\mathcal{D}_0$; this violates the inductive hypothesis, and hence shows that $\mathcal{D}_1$ cannot contain a bad bigon path.

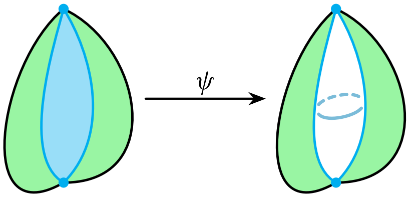


Figure 40: If F is a bigon pillow such that the bigon face $\varphi(F)$ forms part of a bad bigon path B_1, then $\psi(B_1)$ contains a (bad) bigon path that is topologically equivalent to B_1.

- If the faces of F *are* identified, then we are in case (c) of Lemma 10: F forms a (closed) cell decomposition of either S^3 or $L_{3,1}$. Since the only closed components of $\mathcal{D}_0$ are 3-spheres, $\mathcal{D}_1$ must be obtained from $\mathcal{D}_0$ by deleting the 3-sphere component given by F; thus, $\mathcal{D}_1$ is acceptable. Moreover, since the ideal component of $\mathcal{D}_0$ is left entirely untouched by the operation of flattening F, we see that $\mathcal{D}_1$ cannot contain any bad bigon paths.

This completes the inductive step for the case where F is a triangular pillow.

The case where F is a bigon pillow is similar, but slightly more involved. Recall from Lemma 11 that the effect of flattening F depends on whether the two bigon faces of F are identified:

- If the faces of F are *not* identified, then we are in case (a) of Lemma 11. Thus, the truncated pseudomanifolds of $\mathcal{D}_0$ and $\mathcal{D}_1$ are homeomorphic, which implies that $\mathcal{D}_1$ is acceptable. Suppose for the sake of contradiction that $\mathcal{D}_1$ contains a bad bigon path B_1. If B_1 is disjoint from the interior of the bigon $\varphi(F)$, then observe that $\psi(B_1)$ is a bad bigon path in $\mathcal{D}_0$. This would violate the inductive hypothesis, so we conclude that the bigon $\varphi(F)$ must be part of the bigon path B_1; this situation is illustrated in Figure 40. Consider the internal bigon path B_0 in $\mathcal{D}_0$ given by $\psi(B_1) - F$; the ends of B_0 are precisely the two edges incident to the bigon pillow F. Observe that augmenting B_0 with one of the two bigon faces of F gives a bad bigon path in $\mathcal{D}_0$. This again violates the inductive hypothesis, so we conclude that $\mathcal{D}_1$ cannot contain a bad bigon path.

- If the faces of F *are* identified, then we are in either case (c) or case (d) of Lemma 11. Actually, F cannot form an ideal cell decomposition of $\mathbb{R}P^2 \times [0,1]$ (case (d)) because such a component would contain a two-sided projective plane. Thus, we must be in case (c): F must form a (closed) cell decomposition of either S^3 or $\mathbb{R}P^3$. By assumption, the only closed components of $\mathcal{D}_0$ are 3-spheres, so $\mathcal{D}_1$ must be obtained from $\mathcal{D}_0$ by deleting the 3-sphere component given by F; this shows that $\mathcal{D}_1$ is acceptable. We also see that flattening F leaves the ideal component of $\mathcal{D}_0$ entirely untouched, which implies that $\mathcal{D}_1$ contains no bad bigon paths.

This completes the inductive step for the case where F is a bigon pillow.

With the pillow cases out of the way, all that remains is to consider the case where F is a bigon face. As in Sections 3.2.3 and 3.2.4, we divide our study into cases depending on whether the two new boundary bigons given by ungluing F form a single boundary bigon path or two separate boundary bigon paths.

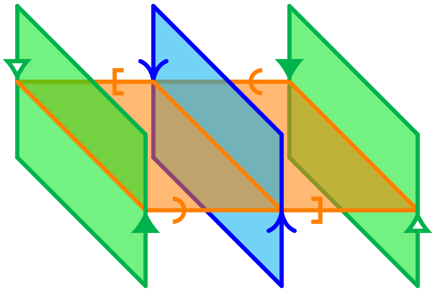


Figure 41: When the truncated bigon associated to F forms a one-sided Möbius band S (blue), the frontier of a small regular neighbourhood of S forms an annulus A (green). If A admits an essential compression disc E, then the boundary curve of E coincides with the boundary of the Möbius band B (orange).

First, suppose ungluing F yields a single boundary bigon path. Since $\mathcal{D}_0$ is valid and has no boundary edges, Claim D tells us that $\varphi(F)$ is a single internal edge in $\mathcal{D}_1$. Validity of $\mathcal{D}_0$ also tells us that we are in either case (1), (2) or (4) of Claim D. Actually, cases (2) and (4) are both impossible:

- Consider case (2) of Claim D. In this case, F forms a *projective plane* in $\mathcal{D}_0$, and Claim D.2 tells us that the two vertices of F are identified to form a single vertex v. Moreover, since $\mathcal{D}_0$ is valid and has no boundary vertices, we must be in either case (a) or case (b) of Claim D.2:

(a) If v is internal, then Claim D.2 tells us that F forms a one-sided properly embedded projective plane in $\mathcal{P}_0$. In fact, F lies in X, since the 3-sphere components of $\mathcal{P}_0$ cannot contain an embedded projective plane. Observe that a small regular neighbourhood N of F is homeomorphic to $\mathbb{R}P^3$ minus a small open 3-ball, and that the frontier of N forms a properly embedded 2-sphere E in X. Notice that E does not bound a 3-ball in X: the region on the "inside" of E is N, and the region on the "outside" contains all the boundary components of X.

(b) If v is ideal, then by Claim D.2 the truncated bigon associated to F forms a one-sided properly embedded Möbius band S in $\mathcal{P}_0$. In fact, S lies in X, since the 3-sphere components of $\mathcal{P}_0$ have empty boundary. Consider the annulus A given by the frontier of a small regular neighbourhood N of S. Since X is anannular, A must be either compressible or boundary-parallel. We claim that neither case is possible:

- If A is compressible, then consider an essential compression disc E for A. Up to isotopy, the boundary curve of E coincides with the boundary curve of a Möbius band B in N given by thickening the core curve of S; see Figure 41. Observe that $E \cup B$ forms an embedded projective plane in X. By assumption, this projective plane cannot be two-sided. However, it also cannot be one-sided, since this would contradict the fact that X is irreducible, by the same argument as before. Thus, we conclude that A cannot be compressible.
- If A is boundary-parallel, then isotoping A into the boundary shows that the entire component X is homeomorphic to a regular neighbourhood of the Möbius band S. But this means that X is a solid torus, which contradicts the assumption that X is ∂-irreducible.

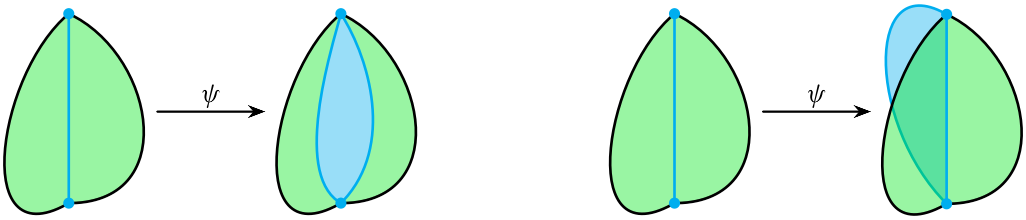


Figure 42: The two cases when the bigon face F is a disc that forms part of $\psi(B_1)$. Left: if $\psi(B_1)$ is a bigon path, then it is topologically equivalent to B_1. Right: if there is a bigon path B_0 such that $B_0 \cap F$ is a single edge and $B_0 \cup F = \psi(B_1)$, then B_0 is topologically equivalent to B_1.

The upshot is that, under our assumptions on X, case (2) of Claim D can never occur.

- Consider case (4) of Claim D. In this case, F forms a 2-*sphere* in $\mathcal{D}_0$, and Claim D.4 tells us that the truncated bigon associated to F forms a one-sided properly embedded annulus in $\mathcal{P}_0$. But this gives an essential annulus in $\mathcal{P}_0$, which is impossible since X is anannular and none of the 3-sphere components of $\mathcal{P}_0$ can contain properly embedded annuli.

We are left with case (1) of Claim D. In this case, F forms a *disc* in $\mathcal{D}_0$, and Claim D.1 tells us that the truncated pseudomanifolds $\mathcal{P}_0$ and $\mathcal{P}_1$ are homeomorphic, and hence that $\mathcal{D}_1$ is acceptable. Suppose for the sake of contradiction that $\mathcal{D}_1$ contains a bad bigon path B_1. We note that B_1 must contain the edge $\varphi(F)$; otherwise, flattening F would leave B_1 untouched, which would imply that $\psi(B_1)$ is a bad bigon path in $\mathcal{D}_0$, contradicting the inductive hypothesis. Thus, $\psi(B_1)$ must contain the bigon face F, and we are left with the following two possibilities:

- If $\psi(B_1)$ itself forms an internal bigon path in $\mathcal{D}_0$, then observe that flattening F reduces the length of this bigon path by one, but has no topological effect; see Figure 42 (left). Thus, $\psi(B_1)$ is bad bigon path in $\mathcal{D}_0$, contradicting the inductive hypothesis.
- Otherwise, there must be an internal bigon path B_0 in $\mathcal{D}_0$ such that $B_0 \cap F$ is a single edge and $B_0 \cup F = \psi(B_1)$; see Figure 42 (right). In this case, we have $\varphi(B_0) = B_1$, and flattening F essentially leaves B_0 untouched, which means that B_0 is a bad bigon path in $\mathcal{D}_0$, again contradicting the inductive hypothesis.

Thus, we conclude that $\mathcal{D}_1$ contains no bad bigon paths. This completes the inductive step for the case where ungluing F yields a single boundary bigon path.

Suppose now that ungluing F yields two separate boundary bigon paths. Since $\mathcal{D}_0$ is valid and has no boundary edges, we are in either case (1) or case (2) of Claim E. Actually, case (2) is impossible, since all three possibilities in Claim E.2 contradict the assumption that $\mathcal{D}_0$ is acceptable:

- Possibility (a) requires $\mathcal{P}_0$ to contain a two-sided properly embedded projective plane.
- Possibility (b) requires $\mathcal{P}_0$ to contain a two-sided properly embedded Möbius band.
- Possibility (c) requires $\mathcal{D}_0$ to contain either a boundary vertex or an invalid vertex.

We are left with case (1) of Claim E. In this case, F forms a 2-*sphere* in $\mathcal{D}_0$, and $\varphi(F)$ consists of two distinct internal edges in $\mathcal{D}_1$. Since $\mathcal{D}_0$ is valid and has no boundary vertices, we are in either case (a), (b) or (c) of Claim E.1. In cases (a) and (b), it is relatively easy to see that $\mathcal{D}_1$ is acceptable and contains no bad bigon paths:

- Consider case (a) of Claim E.1. In this case, F is only incident to internal vertices, and $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along a properly embedded 2-sphere E. Since X is irreducible and every other component of $\mathcal{P}_0$ is a 3-sphere, we see that E bounds a 3-ball, which implies that flattening F only changes the truncated pseudomanifold by creating a new 3-sphere component. Thus, $\mathcal{D}_1$ remains acceptable. Moreover, observe that $\varphi(F)$ is not incident to any ideal vertices of $\mathcal{D}_1$, which means that any bad bigon path B_1 in $\mathcal{D}_1$ must be disjoint from $\varphi(F)$; no such B_1 can exist, otherwise $\psi(B_1)$ would be a bad bigon path in $\mathcal{D}_0$.

- Consider case (b) of Claim E.1. In this case, F is incident to one internal vertex and one ideal vertex, which means that the truncated bigon associated to F forms a properly embedded disc S in X. Since X is ∂-irreducible, the boundary curve of S must bound a disc lying entirely in ∂X, in which case Claim E.1 tells us that $\mathcal{P}_1$ is obtained from $\mathcal{P}_0$ by decomposing along a properly embedded 2-sphere in X. As before, by irreducibility of X, we conclude that flattening F only changes the truncated pseudomanifold by creating a new 3-sphere component. Thus, $\mathcal{D}_1$ remains acceptable. To see that $\mathcal{D}_1$ contains no bad bigon paths, we first note that $\varphi(F)$ is incident to exactly one ideal vertex v. With this in mind, suppose for the sake of contradiction that $\mathcal{D}_1$ contains a bad bigon path B_1. Observe that $B_1 \cap \varphi(F)$ is either empty or consists only of the ideal vertex v, which means that flattening F essentially leaves B_1 untouched. This implies that $\psi(B_1)$ is a bad bigon path in $\mathcal{D}_0$, contradicting the inductive hypothesis.

All that remains is to consider case (c) of Claim E.1. In this case, F is only incident to ideal vertices, which means that the truncated bigon associated to F forms a properly embedded annulus S in X; let γ_0 and γ_1 denote the two boundary curves of S. Observe that S cannot be boundary-parallel, because this would mean that F itself forms a bad bigon path in $\mathcal{D}_0$. Combining this with the assumption that X is anannular, we see that S must be a compressible annulus.

Given what we know about the 3-manifold X and the annulus S, we claim that flattening F only changes the truncated pseudomanifold by creating a new 3-sphere component. To prove this, we start by compressing S along an essential compression disc, which yields two properly embedded discs E_0 and E_1 such that for each $i \in \{0,1\}$, the boundary curve of E_i is γ_i. Since X is ∂-irreducible, each curve γ_i must therefore bound a disc E_i' lying entirely in the boundary of X; see Figure 43 (left). Thus, by case (c) of Claim E.1, flattening F corresponds to decomposing along a properly embedded 2-sphere in X. Since X is irreducible, this 2-sphere bounds a 3-ball, so the only topological effect is to create a new 3-sphere component. This shows that $\mathcal{D}_1$ remains acceptable.

To finish, we just need to verify that $\mathcal{D}_1$ contains no bad bigon paths. If a bad bigon path B_1 in $\mathcal{D}_1$ is disjoint from $\varphi(F)$ or meets $\varphi(F)$ only in (ideal) vertices, then observe that $\psi(B_1)$ would be a bad bigon

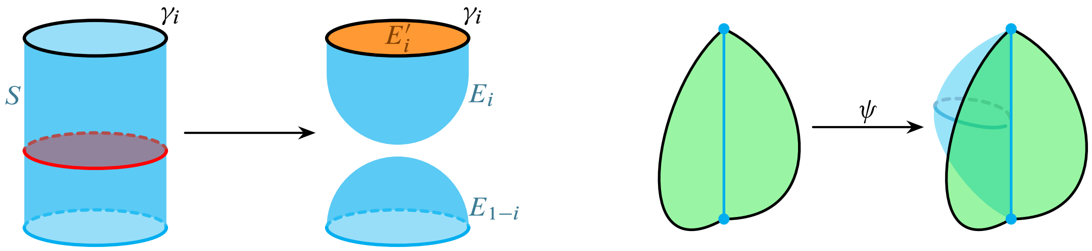


Figure 43: Left: for each $i \in \{0,1\}$, the curve γ_i bounds a properly embedded disc E_i in X given by compressing the annulus S. The fact that X is ∂-irreducible therefore implies that γ_i bounds a disc E_i' that lies entirely in ∂X. Right: if F is a bigon face that forms a 2-sphere, and if there is a bigon path B_0 such that $B_0 \cap F$ is a single edge and $B_0 \cup F = \psi(B_1)$, then B_0 is topologically equivalent to B_1.

path in $\mathcal{D}_0$, which is impossible. The only other possibility is that B_1 meets $\varphi(F)$ in an edge, in which case there would exist an internal bigon path B_0 in $\mathcal{D}_0$ such that $B_0 \cap F$ is a single edge and $B_0 \cup F = \psi(B_1)$; see Figure 43 (right). Observe that B_0 would be a bad bigon path in $\mathcal{D}_0$, which is again impossible.

In summary, we have shown that in every possible case, performing an atomic move on $\mathcal{D}_0$ gives a new acceptable cell decomposition $\mathcal{D}_1$ (and also that $\mathcal{D}_1$ contains no bad bigon paths). By induction, this shows that after performing however many atomic moves we need to flatten $\mathcal{D}^*$, the triangulation $\mathcal{T}^*$ that results from flattening will be acceptable. □

5 Triangulation complexity of 3-dimensional submanifolds

We wish to showcase some applications of Theorem 1 to a notion — namely, *triangulation complexity* — whose significance is independent from crushing. There has been substantial effort devoted to finding upper and lower bounds on triangulation complexity for various families of 3-manifolds; for instance, see [1; 14; 15; 16; 17; 19; 20; 21; 22; 26; 27; 32; 33]. For this paper, the triangulation complexity for a closed 3-manifold $\mathcal{M}$ will refer to the minimum number of tetrahedra in any (closed) triangulation of $\mathcal{M}$, and the triangulation complexity for a bounded 3-manifold $\mathcal{M}$ will refer to the minimum number of tetrahedra in any *ideal* triangulation of $\mathcal{M}$; in either case, we will denote this quantity by $\Delta(\mathcal{M})$.

Our main application of Theorem 1 is to prove that, under quite general conditions, the triangulation complexity of a 3-manifold $\mathcal{M}$ is strictly bigger than the triangulation complexity of a 3-dimensional submanifold of $\mathcal{M}$ bounded by surfaces of positive genus. The precise statement is given in Theorem 2 below.

To our knowledge, Theorem 2 has never previously been written down in the literature. However, as mentioned in Section 1, it is important to note that a similar result can also be obtained by combining, from Matveev's book [28],

- ideas about the duality of triangulations and special spines (from Section 1.1);
- ideas about the conversion of almost simple spines into special spines (from Section 2.1.1); and

- results about how the complexity of almost simple spines interacts with the operation of cutting along normal surfaces in handle decompositions (from Section 4.2).

In any case, a good reason to explicitly write down Theorem 2 is that its assumptions are relatively easy to check. This gives a way to streamline some applications by avoiding the need to directly use either our crushing machinery or Matveev's spine machinery. We demonstrate this in Sections 5.1 to 5.3 by giving some straightforward applications of Theorem 2 to JSJ decompositions and satellite knots.

Theorem 2 *Let $\mathcal{M}$ be a compact 3-manifold with no 2-sphere boundary components. Suppose $\mathcal{M}$ contains a (possibly disconnected) closed incompressible surface S with no 2-sphere components, no projective plane components, and no boundary-parallel components. Let $\mathcal{R}$ be a component obtained after cutting $\mathcal{M}$ along S. If $\mathcal{R}$ is irreducible, ∂-irreducible, anannular, and does not contain any two-sided properly embedded Möbius bands, then $\Delta(\mathcal{R}) < \Delta(\mathcal{M})$.*

Proof Let $\mathcal{T}$ be a closed (if $\mathcal{M}$ is closed) or ideal (if $\mathcal{M}$ has boundary) triangulation of $\mathcal{M}$ such that $|\mathcal{T}| = \Delta(\mathcal{M})$. Our goal is to find an ideal triangulation of $\mathcal{R}$ with strictly fewer tetrahedra than $\mathcal{T}$. We do this by constructing a suitable normal surface S' in $\mathcal{T}$, and using Theorem 1 to ensure that crushing S' yields the desired triangulation of $\mathcal{R}$.

In detail, let $N(S)$ denote a closed tubular neighbourhood of S in $\mathcal{M}$. Viewing $\mathcal{R}$ as the submanifold of $\mathcal{M}$ given by deleting the interior of $N(S)$, let S' be the union of all the components of $\partial N(S)$ that meet $\mathcal{R}$. Since S has no 2-sphere or projective plane components, observe that S' cannot have any 2-sphere components. Also, each component of S' is two-sided, since it meets $N(S)$ on one side and $\mathcal{R}$ on the other side; moreover, since $\mathcal{R}$ meets each component of S' on exactly one side, we can take the interior of $\mathcal{R}$ to be the *chosen region* (as defined at the beginning of Section 4) for S'. This already establishes most of what we require to apply Theorem 1; what remains is to show that S' is an essential surface in $\mathcal{M}$, which will allow us to assume that S' is a nontrivial normal surface with respect to $\mathcal{T}$.

Since S' is a closed (but possibly disconnected) surface, showing that S' is essential entails verifying that every component of S' is incompressible, and that at least one component of S' is not boundary-parallel; in fact, we will be able to show that every component of S' is not boundary-parallel, which is stronger than we require. To this end, consider any particular component C' of S'. Since C' lies in the boundary of a closed tubular neighbourhood $N(C)$ of some component C of S, we have the following two cases:

- If C is two-sided, then C' is isotopic to C, so the fact that C' is incompressible and not boundary-parallel follows from the assumption that these conditions are satisfied by every component of S.
- If C is one-sided, then $C' = \partial N(C)$. To see that C' is incompressible, consider a compression disc D for C' in $\mathcal{M}$; we need to show that D cannot be an *essential* compression disc. We have the following cases:
  - If $D \subset \mathcal{R}$, then D is a compression disc for $\mathcal{R}$. Since $\mathcal{R}$ is assumed to be ∂-irreducible, D cannot be essential.

- If $D \subset N(C)$, then we can use a standard fundamental group argument. Note that $\pi_1(N(C)) \cong \pi_1(C)$, and that the double-covering map $p\colon \partial N(C) \to C$ induces an injective homomorphism $p_*\colon \pi_1(\partial N(C)) \to \pi_1(C)$. Since ∂D is homotopically trivial in $N(C)$, injectivity of p_* tells us that ∂D must also be homotopically trivial in $C' = \partial N(C)$. Thus, we again see that D cannot be essential.

The upshot is that C' does not admit an essential compression disc, so it is incompressible. To see that C' is not boundary-parallel, suppose instead that this is false. The isotopy of C' into $\partial\mathcal{M}$ defines a product region P in $\mathcal{M} - C'$. Note that C' meets two components of $\mathcal{M} - C'$: the interior of $N(C)$, and the interior of $\mathcal{R}$. The product region P must coincide with $\mathcal{R}$. However, this would contradict the assumption that $\mathcal{R}$ is anannular, so we conclude that C' cannot be boundary-parallel.

As mentioned above, this suffices to show that S' is a closed essential surface in $\mathcal{M}$.

By incompressibility of S', we can use the normalisation procedure (recall Section 2.4) to ensure that S' is normal with respect to $\mathcal{T}$. Moreover, since links of ideal vertices of $\mathcal{T}$ correspond to boundary-parallel surfaces, the fact that S' is not boundary-parallel ensures that we have a *nontrivial* normal surface in $\mathcal{T}$. By Lemma 14, we may further assume that this normal surface S' is maximal.

To recap, we are now in the setting laid out at the beginning of Section 4: we have a suitable normal surface S', together with a suitable chosen region given by the interior of $\mathcal{R}$. By assumption, we have that $\mathcal{R}$ is irreducible, ∂-irreducible and anannular, and also that $\mathcal{R}$ contains no two-sided properly embedded Möbius bands. Thus, all the prerequisites for Theorem 1 are satisfied, and applying this theorem tells us that after crushing S', one of the benign components (as defined in Section 4) forms an ideal triangulation $\mathcal{T}^*$ of $\mathcal{R}$. Since S' is a *nontrivial* normal surface, we have $|\mathcal{T}^*| < |\mathcal{T}|$. Hence $\Delta(\mathcal{R}) \leqslant |\mathcal{T}^*| < |\mathcal{T}| = \Delta(\mathcal{M})$, as required. □

5.1 Application: hyperbolic JSJ pieces

Let $\mathcal{M}$ be an irreducible and ∂-irreducible 3-manifold with no 2-sphere boundary components. Recall that by work of Jaco and Shalen [23; 24], and independently by Johannson [25], there is a canonical collection $\{S_i\}$ of finitely many disjoint essential tori in $\mathcal{M}$ such that each piece resulting from cutting along $\bigcup_i S_i$ is either atoroidal or Seifert fibred; formal statements of this result can also be found in [10, Theorem 1.9; 31, Theorem 8.23]. This collection of tori is called the *JSJ decomposition* (or the *torus decomposition*) of $\mathcal{M}$; it is closely related to (but not exactly the same as) the decomposition along tori described by the Thurston–Perelman geometrisation theorem. Theorem 2 almost immediately yields the following consequence for JSJ decompositions:

Theorem 16 *Let $\mathcal{M}$ be an orientable 3-manifold with no 2-sphere boundary components. If $\mathcal{M}$ is irreducible, ∂-irreducible and has nonempty JSJ decomposition $\{S_i\}$, then any hyperbolic component $\mathcal{H}$ that results from cutting $\mathcal{M}$ along $\bigcup_i S_i$ satisfies $\Delta(\mathcal{H}) < \Delta(\mathcal{M})$.*

Proof Suppose that after cutting along the tori in the JSJ decomposition of $\mathcal{M}$, (at least) one of the resulting pieces $\mathcal{H}$ is hyperbolic. By Thurston's hyperbolisation theorem, $\mathcal{H}$ is irreducible, ∂-irreducible and anannular. Moreover, orientability of $\mathcal{M}$ implies that $\mathcal{H}$ is orientable, and hence that $\mathcal{H}$ contains no two-sided properly embedded Möbius bands. Thus, by Theorem 2, we have $\Delta(\mathcal{H}) < \Delta(\mathcal{M})$. □

5.2 Application: satellite knots

Recall that the *exterior* of a knot or link L in S^3 is the 3-manifold obtained by deleting an open regular neighbourhood of L from S^3. We take the *triangulation complexity* of a link L, denoted $\Delta(L)$, to mean the triangulation complexity of the exterior of L.

Our goal now is to present an easy consequence of Theorem 2 concerning the triangulation complexity of satellite knots; see Theorem 18 below. For this, we first review the definition of a satellite knot:

Definitions 17 Let V^* denote the solid torus given by the exterior of an unknot U^*, let C^* denote the core circle of V^*, and let $e\colon V^* \to S^3$ be an embedding such that the image of C^* under e is a nontrivial knot C. Consider a knot K^* in the interior of V^* such that

- K^* is not isotopic (inside V^*) to C^*; and
- every meridional disc of V^* meets K^* at least once.

The image of K^* under e is a nontrivial knot K called a *satellite knot*. We call the knot C a *companion* of K, and we call the torus $e(\partial V^*)$ a *companion torus* of K. We also call the link $K^* \cup U^*$ a *pattern* of K.

Theorem 18 *Let K be a satellite knot, and let L denote either a companion or a pattern of K. If the exterior of L is anannular, then $\Delta(L) < \Delta(K)$.*

It is worth noting that every hyperbolic link is anannular. Thus, Theorem 18 applies to a very large class of satellite knots; Figure 44 shows one example of such a satellite knot.

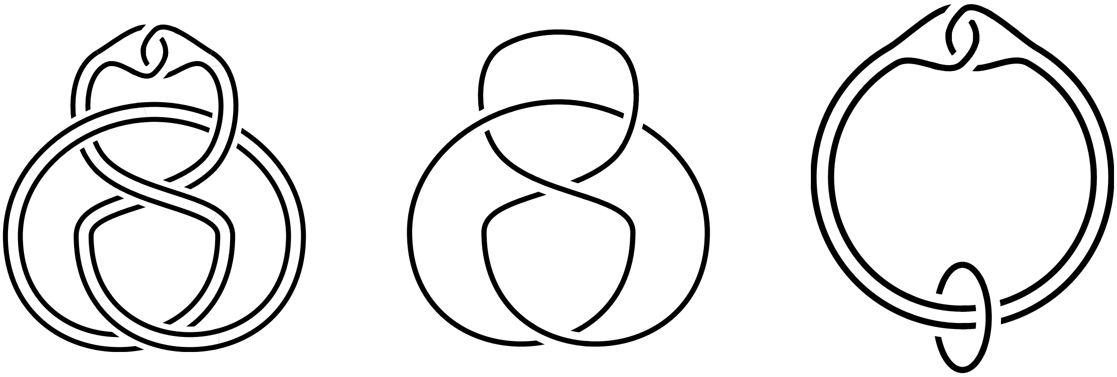


Figure 44: An example of a satellite knot, together with a companion and a pattern; in this case, both the companion and the pattern are hyperbolic, and hence anannular. Left: a satellite knot K, the untwisted Whitehead double of the figure-eight knot. Middle: a companion of K, the figure-eight knot. Right: a pattern of K, the Whitehead link.

Proof of Theorem 18 Let $\bar{K}$ and $\bar{L}$ denote the exteriors of K and L, respectively. Since $\bar{L}$ is one of the components given by cutting the exterior of $\bar{K}$ along a companion torus (which is an essential torus), we can apply Theorem 2 provided that $\bar{L}$

- is irreducible, ∂-irreducible and anannular; and
- contains no two-sided properly embedded Möbius bands.

We have assumed that $\bar{L}$ is anannular, and the fact that $\bar{L}$ is orientable implies that it contains no two-sided properly embedded Möbius bands. Moreover, in the case where L is a companion, irreducibility and ∂-irreducibility follow from the fact that L must be a nontrivial knot; on the other hand, when L is a pattern, irreducibility and ∂-irreducibility follow from the fact that L must be a nonsplit link. The upshot is that, by Theorem 2, we have $\Delta(L) = \Delta(\bar{L}) < \Delta(\bar{K}) = \Delta(K)$. □

Corollary 19 *Consider the connected sum* $K \# K'$ *of two nontrivial knots* K *and* K'. *If* K *is a hyperbolic knot, then* $\Delta(K) < \Delta(K \# K')$.

Proof Recall that the connected sum $K \# K'$ can be viewed as a satellite knot with companion given by either of its summands. Thus, if the summand K is hyperbolic, in which case the exterior of K is anannular, then it follows immediately from Theorem 18 that $\Delta(K) < \Delta(K \# K')$. □

5.3 Application: rod complements in the 3-torus

The crushing techniques developed in previous sections can also be used to study link exteriors in ambient spaces other than the 3-sphere, such as the 3-torus $\mathbb{T}^3$. Hui and Purcell [12] initiated the use of 3-dimensional geometry and topology to study rod packing structures in crystallography. In crystallographic chemistry, a rod packing is a packing of uniform cylinders (also called rods) that represent linear or zigzag chains of particles. Readers may refer to [29; 30] for examples of rod packing structures.

Many rod packing structures exhibit translational symmetry in each dimension of 3-dimensional Euclidean space. Taking a quotient by this symmetry allows such rod packings to be encoded as geodesic links in the 3-torus, which we can then study using tools from 3-manifold geometry and topology. Each component of such a geodesic link is called a *rod-shaped circle*, or often simply a *rod*, in $\mathbb{T}^3$; the complement of such a link in $\mathbb{T}^3$ is a 3-manifold called a *rod complement*.

For $\mathcal{M}$ belonging to a large family of rod complements, the JSJ decomposition gives a unique hyperbolic piece $\mathcal{H}$, and this hyperbolic piece is also a rod complement. Theorem 21 states this precisely, and then (by applying Theorem 16) relates the triangulation complexities of $\mathcal{M}$ and $\mathcal{H}$.

To study the JSJ decompositions of rod complements, we rely on work of Hui [11], which completely classified the geometry of all rod complements using simple linear algebra conditions. The linear algebra arises from the fact that a rod in $\mathbb{T}^3$ lifts to a straight line in the universal cover $\mathbb{R}^3$. This linear structure leads to the natural concepts of

- *linear independence* of rods in $\mathbb{T}^3$; and
- *linear isotopy* of rods in $\mathbb{T}^3$, which roughly means an isotopy along a planar annulus between two parallel rods.

These notions appear in the statements of Lemma 20 and Theorem 21 below; readers may refer to [11; 12] for definitions of these notions, as well as for explanations of other related terminology that we use in the proofs.

Lemma 20 *Let $\mathcal{M}$ be a toroidal rod complement in the 3-torus with at least three linearly independent rods. Any essential torus in $\mathcal{M}$ bounds a solid torus in $\mathbb{T}^3$ whose interior contains two or more linearly isotopic rods.*

Proof Since $\mathcal{M}$ is toroidal, by Proposition 3.8 in [11], either $\mathcal{M}$ is a rod complement with all rods spanning a plane torus, or there exist disjoint parallel rods that are linearly isotopic in the complement of the other rods (or possibly both). The assumption that $\mathcal{M}$ is a rod complement with three linearly independent rods thus implies the existence of at least two linearly isotopic parallel rods.

Let T_{e} be an essential torus in $\mathcal{M}$. Note that T_{e} is not in the homotopy class of a plane torus because there exist three linearly independent rods for $\mathcal{M}$. Hence, the torus T_{e}, essential in $\mathcal{M}$, has at least one generator in $\pi_1(T_{\mathrm{e}}) \cong \mathbb{Z} \times \mathbb{Z}$ represented by an essential loop that is homotopically trivial in $\mathbb{T}^3$.

By Lemma 3.7 in [11], there exists a compression disc D for T_{e} in $\mathbb{T}^3$. It then follows from the irreducibility of $\mathbb{T}^3$ that T_{e} is separating. As T_{e} is incompressible in $\mathcal{M}$, some rod must intersect D. Hence, the other generator of $\pi_1(T_{\mathrm{e}}) \cong \mathbb{Z} \times \mathbb{Z}$ is represented by a loop that is homotopically nontrivial in $\mathbb{T}^3$. By Lemma 3.9 in [11], T_{e} bounds a solid torus V_{e} in $\mathbb{T}^3$. Since T_{e} is not boundary-parallel, the interior of V_{e} contains two or more parallel rods that are linearly isotopic in the complement of the other rods. □

Theorem 21 *Let $\mathcal{M}$ be a toroidal rod complement in the 3-torus with at least three linearly independent rods. The JSJ decomposition of $\mathcal{M}$ gives a unique (up to homeomorphism) hyperbolic rod complement $\mathcal{H}$ with $\Delta(\mathcal{H}) < \Delta(\mathcal{M})$.*

Proof Without loss of generality, assume that $\mathcal{M}$ is the complement of an open neighbourhood of all the rods in the 3-torus.

We first show that $\mathcal{M}$ satisfies all the assumptions in Theorem 16. Note that $\mathcal{M}$ is an orientable 3-manifold with no 2-sphere boundary components. Since $\mathcal{M}$ is the complement of finitely many rods in the 3-torus, Proposition 3.6 in [11] tells us that $\mathcal{M}$ is irreducible and ∂-irreducible. We also know that $\mathcal{M}$ has nonempty JSJ decomposition because it is toroidal by assumption, and because it is not Seifert fibred by Theorem 4.1 in [11]. Thus, Theorem 16 applies to any hyperbolic component that results from cutting along the JSJ tori.

Let $\{T_i\}$ be the set of essential tori that gives the JSJ decomposition of $\mathcal{M}$; recall that this set of tori is unique up to isotopy, which means that the JSJ pieces (the 3-manifold components obtained by cutting along these tori) are unique up to homeomorphism. By Lemma 20, each essential torus T_i bounds a solid torus V_i in $\mathbb{T}^3$ whose interior contains two or more linearly isotopic rods. Thus, one of the JSJ pieces is a 3-manifold $\mathcal{H} := \mathcal{M} - \bigcup_i \operatorname{int}(V_i)$ whose interior is homeomorphic to a rod complement. We will show that $\mathcal{H}$ is the unique hyperbolic JSJ piece for $\mathcal{M}$.

To do this, we repeatedly appeal to the classification given by Theorem 4.1 in [11]. Since $\mathcal{M}$ has at least three linearly independent rods, observe that the same must be true for $\mathcal{H}$. Thus, by the classification, $\mathcal{H}$ cannot be Seifert fibred. This means that $\mathcal{H}$ must be atoroidal, since it is one of the JSJ pieces. Using the classification again, we therefore see that $\mathcal{H}$ cannot have a pair of linearly isotopic rods. This has two implications:

- First, we must have exactly one essential torus T_i for each linear isotopy class containing at least two rods from $\mathcal{M}$, and the corresponding solid torus V_i in $\mathbb{T}^3$ must contain *all* of the rods in this class.
- Second, by the classification, the interior of $\mathcal{H}$ admits a complete hyperbolic structure.

To see that $\mathcal{H}$ is the only hyperbolic JSJ piece, observe that each of the other JSJ pieces is obtained from one of the solid tori V_i by deleting a small neighbourhood of all the (linearly isotopic) rods inside V_i; in other words, all the other JSJ pieces are Seifert fibred, since they are homeomorphic to solid tori with at least two core curves removed.

Finally, since $\mathcal{H}$ is a hyperbolic piece obtained after cutting along the JSJ tori for $\mathcal{M}$, it follows from Theorem 16 that $\Delta(\mathcal{H}) < \Delta(\mathcal{M})$. □

References

[1] **M Bucher**, **R Frigerio**, **C Pagliantini**, *The simplicial volume of 3-manifolds with boundary*, J. Topol. 8 (2015) 457–475 MR

[2] **B A Burton**, *Computational topology with Regina: algorithms, heuristics and implementations*, from "Geometry and topology down under" (C D Hodgson, W H Jaco, M G Scharlemann, S Tillmann, editors), Contemp. Math. 597, Amer. Math. Soc., Providence, RI (2013) 195–224 MR

[3] **B A Burton**, *A new approach to crushing 3-manifold triangulations*, Discrete Comput. Geom. 52 (2014) 116–139 MR

[4] **B A Burton**, **R Budney**, **W Pettersson**, et al., *Regina: Software for low-dimensional topology* Available at `https://regina-normal.github.io`

[5] **B A Burton**, **A Coward**, **S Tillmann**, *Computing closed essential surfaces in knot complements*, from "Computational geometry" (G D da Fonseca, T Lewiner, L Peñaranda, editors), ACM, New York (2013) 405–413 MR

[6] **B A Burton**, **A He**, *Finding large counterexamples by selectively exploring the Pachner graph*, from "39th International Symposium on Computational Geometry" (E W Chambers, J Gudmundsson, editors), LIPIcs. Leibniz Int. Proc. Inform. 258, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern (2023) art. id. 21 MR

[7] **B A Burton**, **M Ozlen**, *A fast branching algorithm for unknot recognition with experimental polynomial-time behaviour*, preprint (2014) arXiv 1211.1079v3

[8] **B A Burton**, **S Tillmann**, *Computing closed essential surfaces in* 3*-manifolds*, J. Appl. Comput. Topol. 9 (2025) art. id. 18 MR

[9] **J Fowler**, *Finding* 0*-efficient triangulations of* 3*-manifolds*, senior honors thesis, Harvard University (2003)

[10] **A Hatcher**, *Notes on basic* 3*-manifold topology* Available at `https://pi.math.cornell.edu/~hatcher/3M/3Mdownloads.html`

[11] **C O Y Hui**, *A geometric classification of rod complements in the* 3*-torus*, Proc. Amer. Math. Soc. 153 (2025) 381–394 MR

[12] **C O Y Hui**, **J S Purcell**, *On the geometry of rod packings in the* 3*-torus*, Bull. Lond. Math. Soc. 56 (2024) 1291–1309 MR

[13] **K Ichihara**, **Y Nishimura**, **S Tani**, *The computational complexity of classical knot recognition*, J. Knot Theory Ramifications 32 (2023) art. id. 2350069 MR

[14] **A Jackson**, *Minimal triangulation size of Seifert fibered spaces with boundary*, preprint (2023) arXiv 2301.02085

[15] **W Jaco**, **J Johnson**, **J Spreer**, **S Tillmann**, *Bounds for the genus of a normal surface*, Geom. Topol. 20 (2016) 1625–1671 MR

[16] **W Jaco**, **H Rubinstein**, **J Spreer**, **S Tillmann**, *On minimal ideal triangulations of cusped hyperbolic* 3*-manifolds*, J. Topol. 13 (2020) 308–342 MR

[17] **W Jaco**, **H Rubinstein**, **S Tillmann**, *Minimal triangulations for an infinite family of lens spaces*, J. Topol. 2 (2009) 157–180 MR

[18] **W Jaco**, **J H Rubinstein**, 0*-efficient triangulations of* 3*-manifolds*, J. Differential Geom. 65 (2003) 61–168 MR

[19] **W Jaco**, **J H Rubinstein**, **J Spreer**, **S Tillmann**, $\mathbb{Z}_2$*-Thurston norm and complexity of* 3*-manifolds, II*, Algebr. Geom. Topol. 20 (2020) 503–529 MR

[20] **W Jaco**, **J H Rubinstein**, **J Spreer**, **S Tillmann**, *Complexity of* 3*-manifolds obtained by Dehn filling*, Algebr. Geom. Topol. 25 (2025) 301–327 MR

[21] **W Jaco**, **J H Rubinstein**, **S Tillmann**, *Coverings and minimal triangulations of* 3*-manifolds*, Algebr. Geom. Topol. 11 (2011) 1257–1265 MR

[22] **W Jaco**, **J H Rubinstein**, **S Tillmann**, $\mathbb{Z}_2$*-Thurston norm and complexity of* 3*-manifolds*, Math. Ann. 356 (2013) 1–22 MR

[23] **W Jaco**, **P B Shalen**, *A new decomposition theorem for irreducible sufficiently-large* 3*-manifolds*, from "Algebraic and geometric topology, part 2" (R J Milgram, editor), Proc. Sympos. Pure Math. 32, Amer. Math. Soc., Providence, RI (1978) 71–84 MR

[24] **W H Jaco**, **P B Shalen**, *Seifert fibered spaces in* 3*-manifolds*, Mem. Amer. Math. Soc. 220 (1979) MR

[25] **K Johannson**, *Homotopy equivalences of* 3*-manifolds with boundaries*, Lecture Notes in Math. 761, Springer (1979) MR

[26] **M Lackenby**, **J S Purcell**, *The triangulation complexity of elliptic and sol* 3*-manifolds*, Math. Ann. 390 (2024) 1623–1667 MR

[27] **M Lackenby**, **J S Purcell**, *The triangulation complexity of fibred* 3*-manifolds*, Geom. Topol. 28 (2024) 1727–1828 MR

[28] **S Matveev**, *Algorithmic topology and classification of* 3*-manifolds*, 2nd edition, Algorithms and Computation in Mathematics 9, Springer (2007) MR

[29] **M O'Keeffe**, **J Plévert**, **T Ogawa**, *Homogeneous cubic cylinder packings revisited*, Acta Crystallogr. Sect. A 58 (2002) 125–132 MR

[30] **M O'Keeffe**, **J Plévert**, **Y Teshima**, **Y Watanabe**, **T Ogama**, *The invariant cubic rod* (*cylinder*) *packings*: *symmetries and coordinates*, Acta Cryst. Sect. A 57 (2001) 110–111 MR

[31] **J S Purcell**, *Hyperbolic knot theory*, Graduate Studies in Math. 209, Amer. Math. Soc., Providence, RI (2020) MR

[32] **J H Rubinstein**, **J Spreer**, **S Tillmann**, *A new family of minimal ideal triangulations of cusped hyperbolic* 3*-manifolds*, from "2021–2022 MATRIX annals" (D R Wood, J de Gier, C E Praeger, editors), MATRIX Book Ser. 5, Springer (2024) 5–28 MR

[33] **A Y Vesnin**, **E A Fominykh**, *Exact values of complexity for Paoluzzi–Zimmermann manifolds*, Dokl. Math. (2011) 542–544 MR

BAB: *School of Mathematics and Physics, University of Queensland*
Brisbane, QLD, Australia

TdP, COYH: *School of Mathematics, Monash University*
Melbourne, VIC, Australia

AH: *Department of Mathematics, Oklahoma State University*
Stillwater, OK, United States

`bab@maths.uq.edu.au`, `thiago.depaivasouza@monash.edu`, `alex.he@okstate.edu`,
`onyu.hui@monash.edu`

`https://people.smp.uq.edu.au/BenjaminBurton/`,
`https://sites.google.com/view/thiago-de-paiva`,
`https://sites.google.com/view/alex-he`, `https://sites.google.com/view/oyhui`

Received: 10 April 2024 Revised: 9 October 2024

Algebraic & Geometric Topology **25**:8 (2025) 5013–5029
DOI: 10.2140/agt.2025.25.5013
Published: 20 November 2025

Annular links from Thompson's group T

LOUISA LILES

In 2014 Jones showed how to associate links in the 3-sphere to elements of Thompson's group F. We provide an analogue of this program for annular links and Thompson's group T. The main result is that any edge-signed graph embedded in the annulus is the Tait graph of an annular link built from an element of T. In analogy to the work of Aiello and Conti, we also show that the coefficients of certain unitary representations of T recover the Jones polynomial of annular links.

57K10, 57K14; 43A35

1 Introduction and statement of main results

Vaughan Jones introduced a method of constructing links in the 3-sphere from elements of the Thompson group F, which are piecewise linear orientation-preserving self-homeomorphisms of the unit interval; see [13; 14]. Jones proved that the Thompson group F gives rise to all link types in the 3-sphere, suggesting that it can be used as an analogue of braid groups for producing links [13, Theorem 5.3.1].

We provide a method for building links in the thickened annulus $\mathbb{A} \times I$ from Thompson's group T, which contains F and whose elements are piecewise-linear orientation-preserving self-homeomorphisms of S^1. This method recovers Jones' construction for the subgroup F, but differs from Jones' construction of links from T in [13]. Whereas Jones builds links in S^3 from T, we build links in $\mathbb{A} \times I$, the diagrams of which, under the inclusion $\mathbb{A} \hookrightarrow \mathbb{R}^2$, become the diagrams of the links arising from Jones' construction.

Given $g \in T$, one can follow the process introduced in Section 3 to build an annular link $\mathcal{L}_{\mathbb{A}}(g)$. On the other hand, given an edge-signed graph $\Gamma \hookrightarrow \mathbb{A}$, one can construct a diagram of an annular link $L_{\mathbb{A}}(g)$ in analogy with Tait's construction of links from planar graphs; see Figure 1. Jones proved that given any Tait graph $\Gamma \in \mathbb{R}^2$, there is some $g \in F$ which produces the same link as Γ. The following theorem states that the same is true for annular links and T:

Theorem 1.1 *Let $\Gamma \hookrightarrow \mathbb{A}$ be an edge-signed embedded graph. Then there exists some $g \in T$ such that $\mathcal{L}_{\mathbb{A}}(g)$ is isotopic in $\mathbb{A} \times I$ to $L_{\mathbb{A}}(\Gamma)$.*

The construction of links in the 3-sphere arose naturally in Jones' definition of certain unitary representations of F and T [13]. The Kauffman bracket and Jones polynomial of links in the 3-sphere were then shown to arise as coefficients of these unitary representations of F [2; 4] and T [3]; this was accomplished by proving that they are functions of positive type. We establish a similar result for annular links and T,

---

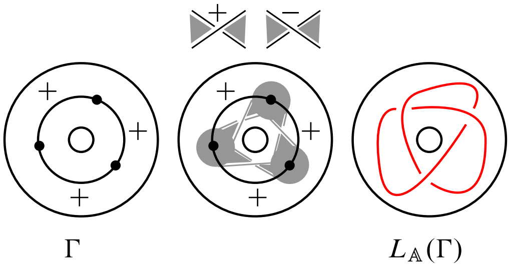


Figure 1: An annular link built from an edge-signed graph embedded in $\mathbb{A}$.

which follows from the construction of links outlined in Section 3, and from [3, Theorems 6.2 and 7.4]. We now introduce some notation necessary to state the precise result.

Elements of T can be specified by triples $(R, S; k)$, where R and S are trees and k is an integer; see Section 2 for more details. When R, S and k are relevant, we will use $\mathcal{L}_{\mathbb{A}}(R, S; k)$ to refer to the link resulting from the unique element $g \in T$ determined by $(R, S; k)$. Using this notation, we can establish the Jones polynomial of annular links as a function of positive type on $\vec{T}$, the oriented subgroup of T, which was first introduced by Jones in [13] and is further discussed in Section 3.

Corollary 1.2 *For $g = (R, S; k) \in \vec{T}$, let n be the number of leaves in R, and let $V^{\mathbb{A}}_{\mathcal{L}}(t)$ denote the Jones polynomial of an annular link $\mathcal{L}$, where unknotted curves wrapping once around $\mathbb{A}$ are equal to $(-t^{-1/2} - t^{1/2})$. Define $V^{\mathbb{A}}_g(t)\colon \vec{T} \to \mathbb{C}$ analogously to* [3], *that is,*

$$V^{\mathbb{A}}_g(t) := V^{\mathbb{A}}_{\mathcal{L}_{\mathbb{A}}(R,S;k)}(t)(-t^{-1/2} - t^{1/2})^{-n+1}.$$

Then, for $t \in \{1, i, e^{\pm\pi i/3}\}$, $V^{\mathbb{A}}_g(t)$ is a function of positive type on $\vec{T}$, and consequently the Jones polynomial of $\mathcal{L}_{\mathbb{A}}(g)$, evaluated at $t \in \{1, i, e^{\pm\pi i/3}\}$, is the coefficient of a unitary representation of $\vec{T}$.

Annular links arise naturally in the study of knot theory, categorification [5] and representation theory of planar algebras [10; 12; 15; 8]. The interplay between the Thompson group and link theory is an emerging subject, and its full interaction with categorification, planar algebras and representation theory is still being developed. Annular links will likely play an important role in this theory.

The paper proceeds as follows. Section 2 provides an overview of Thompson's groups F and T and outlines Jones' construction of links in S^3 from F. Section 3 introduces the construction of annular links from T and connects it to Jones' unitary representations. In this section the concept of *annular Thompson badness* is presented as an extension of Jones' concept of Thompson badness, which he uses to prove that F can produce all link types. Section 4 uses annular Thompson badness to prove Theorem 1.1.

Acknowledgments

The author is thankful for the guidance provided by her advisor Vyacheslav Krushkal, and for the constructive comments given by the referees.

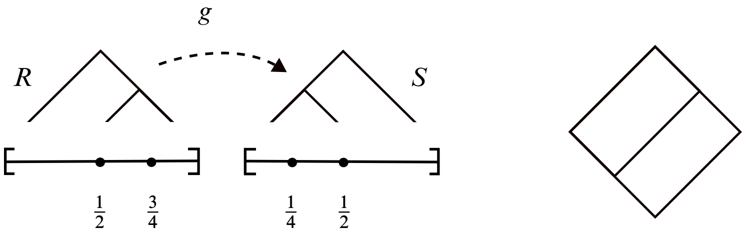


Figure 2: A pair of standard dyadic partitions, their corresponding trees R and S, and their associated element $g \in F$.

2 Thompson's groups F and T

2.1 Thompson's group F

The Thompson group F consists of piecewise linear orientation-preserving self-homeomorphisms of the unit interval [0, 1] such that all derivatives are powers of 2 and all points of nondifferentiability occur at dyadic numbers, that is, numbers of the form $a/2^b$ for $a, b \in \mathbb{Z}$. For example:

$$g(t) = \begin{cases} \frac{1}{2}t, & 0 \le t \le \frac{1}{2}, \\ t - \frac{1}{4}, & \frac{1}{2} \le t \le \frac{3}{4}, \\ 2t - 1, & \frac{3}{4} \le t \le 1. \end{cases}$$

A *standard dyadic partition* is a partition of the unit interval such that all subintervals are of the form $[a/2^b, (a+1)/2^b]$. Any ordered pair of standard dyadic partitions with the same number of parts determines an element of F, given by the function sending the first partition to the second. For example, the function g above is given by the ordered pair

$$(\{[0, \tfrac{1}{2}], [\tfrac{1}{2}, \tfrac{3}{4}], [\tfrac{3}{4}, 1]\}, \{[0, \tfrac{1}{4}], [\tfrac{1}{4}, \tfrac{1}{2}], [\tfrac{1}{2}, 1]\}).$$

Standard dyadic partitions can be represented as planar, rooted, binary trees, where each leaf represents an interval of the partition. Therefore an ordered pair of such trees (R, S) also determines an element of F. This pair of trees is often represented by taking the vertical reflection of S and attaching it to R along their leaves; see Figure 2.

Conversely, for every $g \in F$ there is a standard dyadic partition J such that $g(J)$ is standard dyadic. The pair $(J, g(J))$ therefore determines g, but this pair is not unique. For any refinement J' of J which also standard dyadic, $(J', g(J'))$ also represents g. In terms of trees, refining a pair of partitions corresponds to adding finitely many *canceling carets* to their pair of trees, as shown in Figure 3.

In fact, any two pairs of trees representing the same element of F must differ by the addition or deletion of finitely many canceling carets, and a pair of trees is called *reduced* if no carets can be canceled. Reduced pairs of planar, rooted, binary trees are therefore in bijection with elements of F; more details of this correspondence can be found in [6]. From now on, an ordered pair (R, S) will refer to both a pair of

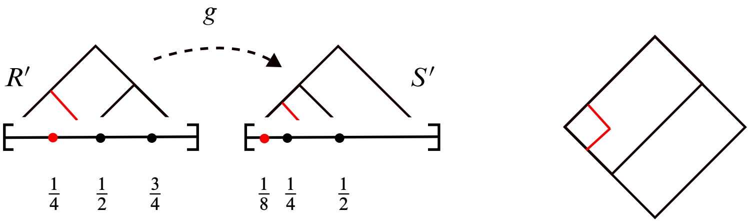


Figure 3: A pair-of-trees representation of the same element g from Figure 2, which differs from the pair in Figure 2 by a canceling caret.

standard dyadic partitions and its associated pair of trees, and elements of F will be specified by these pairs.

Pairs of trees corresponding to elements of F are part of a broader class of graphs called *strand diagrams*, introduced by Belk in [6]. A general strand diagram can be *reduced* according to moves of type I and II, which were independently found by [6; 11]. These moves are useful for visualizing the group operation in F. To compose g with f, place the pair of trees for g below that of f as in Figure 4 and then reduce the resulting strand diagram. This leads to the unique reduced pair-of-trees diagram representing $g \circ f$.

2.2 Link diagrams in the plane from F

Although Jones' original construction of links was formulated in terms of unitary representations of F, Jones provided two equivalent diagrammatic methods for building links [13; 14].

The first, pictured in Figure 5, turns a reduced pair of binary trees (R, S) into a reduced pair of ternary trees $(\phi(R), \phi(S))$, connects the two roots and the leaves from left to right, and then changes 4-valent vertices to crossings.

The second method, pictured in Figure 6, builds the Tait graph $\Gamma(g)$ of $\mathcal{L}(g)$. For $g = (R, S)$, one makes two graphs $\Gamma(R)$ and $\Gamma(S)$ which have the same number of edges. Specifically, $\Gamma(R)$ has one vertex for each leaf of R, and it is placed to immediately to the left of the leaf. The vertices for $\Gamma(S)$ are created in the same way from S. For every edge e in R (resp. S) that slopes up and to the right, $\Gamma(R)$ (resp. $\Gamma(S)$) will have one edge which transversely intersects e once and no other edges. $\Gamma(g)$ is then built by reflecting

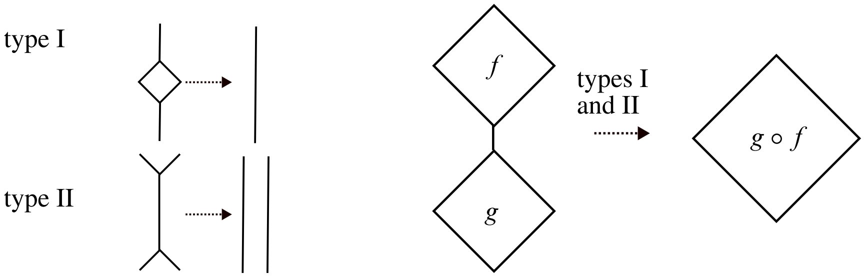


Figure 4: Diagrammatic composition in F as given by [6; 11].

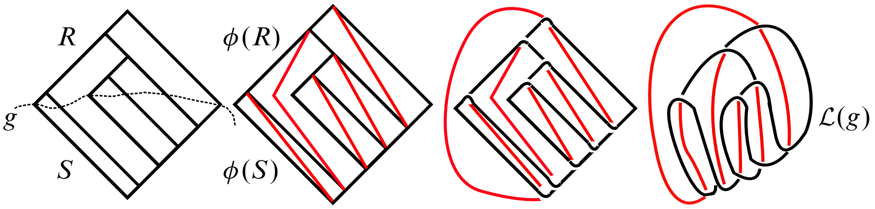


Figure 5: A Hopf link created from an element of F via the construction introduced by Jones [13].

$\Gamma(S)$ over the x-axis and identifying its leaves with those of $\Gamma(R)$. Edges of $\Gamma(g)$ originating from $\Gamma(R)$ are given a positive sign and edges originating from $\Gamma(S)$ are given a negative sign.

2.3 Thompson badness

To detect whether a general edge-signed planar graph Γ is equal to $\Gamma(g)$ for some $g \in F$, Jones introduced *Thompson badness*, a quantity which is zero exactly when $\Gamma = \Gamma(g)$. To calculate Thompson badness, first embed $\Gamma \hookrightarrow \mathbb{R}^2$ such that all vertices are on the x axis, the leftmost vertex is at the origin, and for each edge e, its interior, denoted $\mathrm{int}(e)$, is either entirely above or entirely below the x axis. Consider each edge to be oriented from left to right, so that its rightmost vertex is considered the *terminal vertex*. The formula for Thompson badness, which will be given momentarily, depends on the cardinality of the following sets:

$$\begin{aligned}
e_v^{\mathrm{in}} &:= \{e \in e(\Gamma) : v \text{ is the terminal vertex of } e\},\\
e^{\mathrm{up}} &:= \{e \in e(\Gamma) : \mathrm{int}(e) \text{ is in the upper half-plane}\},\\
e^{\mathrm{down}} &:= \{e \in e(\Gamma) : \mathrm{int}(e) \text{ is in the lower half-plane}\},\\
e_-^{\mathrm{up}} &:= \{e \in e(\Gamma) : \mathrm{int}(e) \text{ is in the upper half-plane and } e \text{ has sign } -\},\\
e_+^{\mathrm{down}} &:= \{e \in e(\Gamma) : \mathrm{int}(e) \text{ is in the lower half-plane and } e \text{ has sign } +\}.
\end{aligned}$$

Jones defines Thompson badness as

$$TB(\Gamma) = \sum_{v \in V(\Gamma)\setminus\{(0,0)\}} \left(\left|1 - |e_v^{\mathrm{in}} \cap e^{\mathrm{up}}|\right| + \left|1 - |e_v^{\mathrm{in}} \cap e^{\mathrm{down}}|\right| \right) + |e_-^{\mathrm{up}}| + |e_+^{\mathrm{down}}|$$

and shows that $TB(\Gamma) = 0$ if and only if $\Gamma = \Gamma(g)$ for some $g \in F$ [13, Sections 4 and 5].

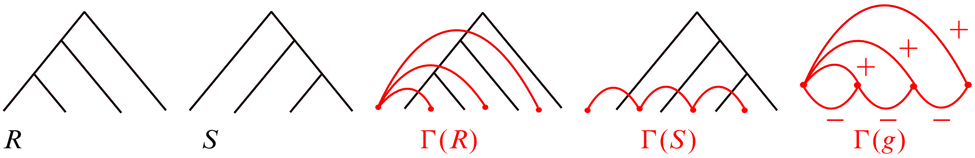


Figure 6: $\Gamma(g)$, the Tait graph for $\mathcal{L}(g)$, where g is specified by (R, S).

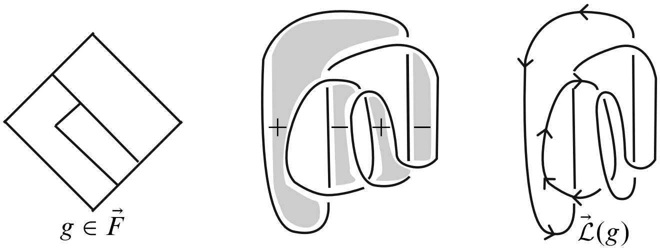


Figure 7: An oriented link $\vec{\mathcal{L}}(g)$ built from $g \in \vec{F}$.

2.4 The oriented subgroup $\vec{F}$

Jones defined $\vec{F}$ as the set of elements $g \in F$ whose link diagram $\mathcal{L}(g)$, when given the checkerboard shading, results in an orientable surface, ie a Seifert surface for $\mathcal{L}(g)$. Equivalently, this can be expressed in terms of the chromatic polynomial $\mathrm{Chr}_{\Gamma(g)}(Q)$:

$$\vec{F} = \{g \in F : \mathrm{Chr}_{\Gamma(g)}(2) = 2\}.$$

After its introduction by Jones, the subgroup $\vec{F}$ was further studied by Golan and Sapir in [9].

If one follows the convention that the leftmost face of the checkerboard surface is always positively oriented, each $g \in \vec{F}$ builds a link $\mathcal{L}(g)$ with a natural orientation, namely that induced by the orientation of the checkerboard surface as in Figure 7. It was shown in [1] that every oriented link can be built from this subgroup, giving an analogue of the Alexander Theorem for oriented links and $\vec{F}$.

2.5 Thompson's group T

T is the group of piecewise-linear orientation-preserving self-homeomorphisms of S^1, thought of as the unit interval with its endpoints identified, such that derivatives are powers of 2 and all points of nondifferentiability occur at dyadic numbers. F is the subgroup of T whose elements send $[1] \mapsto [1]$. Elements of T are given by triples $(R, S; k)$ where (R, S) is a pair of planar, rooted, binary trees and k is a positive integer between 1 and the number of leaves of R and S. The integer k indicates that the first part of R is sent to the k^{th} part of S, and this triple $(R, S; k)$ determines an element of T. Observe that $k = 1$ if and only if $g = (R, S; k) \in F$. To indicate the value of k in a pair-of-trees diagram, a decoration is placed on the k^{th} leaf of S, as in Figure 8.

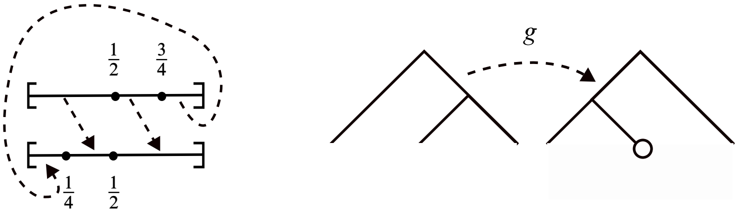


Figure 8: The reduced pair of trees and decorated leaf representing the element $g \in T$ which maps $\left[0, \frac{1}{2}\right] \mapsto \left[\frac{1}{4}, \frac{1}{2}\right]$, $\left[\frac{1}{2}, \frac{3}{4}\right] \mapsto \left[\frac{1}{2}, 1\right]$, and $\left[\frac{3}{4}, 1\right] \mapsto \left[0, \frac{1}{4}\right]$.

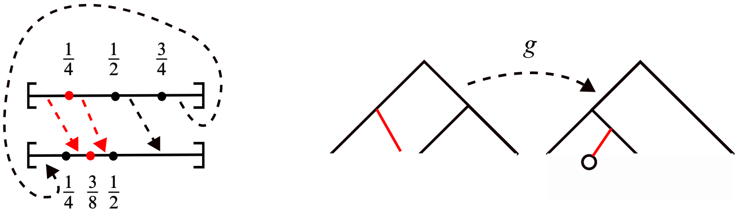


Figure 9: An unreduced triple representing the same element g as in Figure 8, which differs from the triple in Figure 8 by a canceling caret, shown in red.

As was the case for F, one can refine the partitions R and S to produce an unreduced triple $(R', S'; k')$ which differs from $(R, S; k)$ by canceling carets; see Figure 9. Canceling carets are slightly less obvious for diagrams in $T \setminus F$ due to the fact that interval corresponding to the first leaf of R is not mapped to the interval corresponding to the first leaf of S.

Section 2.1 introduced strand diagrams as a way to visualize the group operation in F. An analogue for T was developed by Belk and Matucci [7]. Specifically, every element of T corresponds to a unique reduced *cylindrical strand diagram*, which satisfies the same conditions as a strand diagram, but is now embedded in $S^1 \times [0, 1]$ rather than the unit square [7]. Following the definition in [7], isotopic cylindrical strand diagrams are considered equal, and isotopies are not required to fix the boundary circles. Therefore, cylindrical strand diagrams differing by Dehn twists are considered equal.

To associate a cylindrical strand diagram to an element $g = (R, S; k) \in T$, place the trees R and S in the cylinder as in Figure 10. Identify leaves so that the first leaf of R is sent to the k^{th} leaf of S, and then connect the rest of the leaves in unique way for which the graph remains embedded; see Figure 10.

In Figure 10, the rightmost picture differs from the picture to its left by the smoothing of edges. For the rest of this paper strand diagrams built from T will appear without smoothed edges, to indicate the pair of trees from which the diagram was created.

Cylindrical strand diagrams may be reduced according to local moves of type I and II as in Figure 4. As was the case for F, given cylindrical strand diagrams for $f, g \in T$, vertically stacking the cylinders and reducing using moves of type I and II results in the reduced cylindrical strand diagram for $g \circ f$ [7]. Just as strand diagrams are used to build links from F, this paper will use cylindrical strand diagrams to

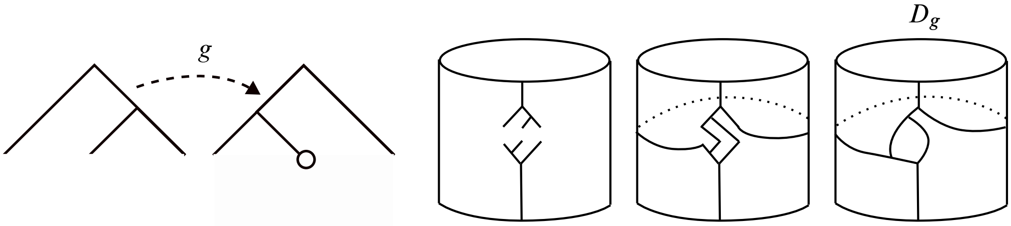


Figure 10: A cylindrical strand diagram D_g built from $g \in T$.

construct annular links from T. A forthcoming paper uses strand diagrams to relate Thompson's group F to Khovanov homology of links in 3-space [17; 16]; it may be possible to use cylindrical strand diagrams to give an analogue for the group T and annular link homology theories.

3 Building annular links from Thompson's group *T*

To construct annular links from T, this section introduces two equivalent methods analogous to those introduced by Jones for F.

The first method is pictured in Figure 11. Given $g = (R, S; k)$, consider the associated strand diagram D_g. Following the method for building links from F, add edges to R and S to make them ternary trees, and consider these new edges numbered from left to right. The edge above each root is considered to be numbered 0. Next, stack "empty" cylinders above and below D_g and connect numbered edges with noncrossing arcs according to the following rule: when the n^{th} edge of R is connected to the ${}^{\text{th}}m$ edge of S and $n > m$, the arc connecting them must wrap around the annulus. Otherwise, the arc does not wrap around the annulus. Note that this rule guarantees that the arc connecting the top root to another edge will never wrap around the cylinder, and the arc connecting the bottom root to another edge will always wrap around the cylinder, unless a single arc connects the two roots (in which case $g \in F$). Finally, all 4-valent vertices become crossings as before.

The second method for building annular links from T, pictured in Figure 12, involves building an edge-signed graph $\Gamma_{\mathbb{A}}(g) \hookrightarrow \mathbb{A}$ and defining $\mathcal{L}_{\mathbb{A}}(g) := L_{\mathbb{A}}(\Gamma_{\mathbb{A}}(g))$. $\Gamma_{\mathbb{A}}(g)$ is built from $\Gamma(R)$ and $\Gamma(S)$, which are created as in Section 2. However, the first vertex of $\Gamma(R)$ is now identified with the k^{th} vertex of $\Gamma(S)$, and edges of R attaching to edges to their left in S must wrap counterclockwise around $\mathbb{A}$; see Figure 12. This second construction is used in Section 4 to prove Theorem 1.1.

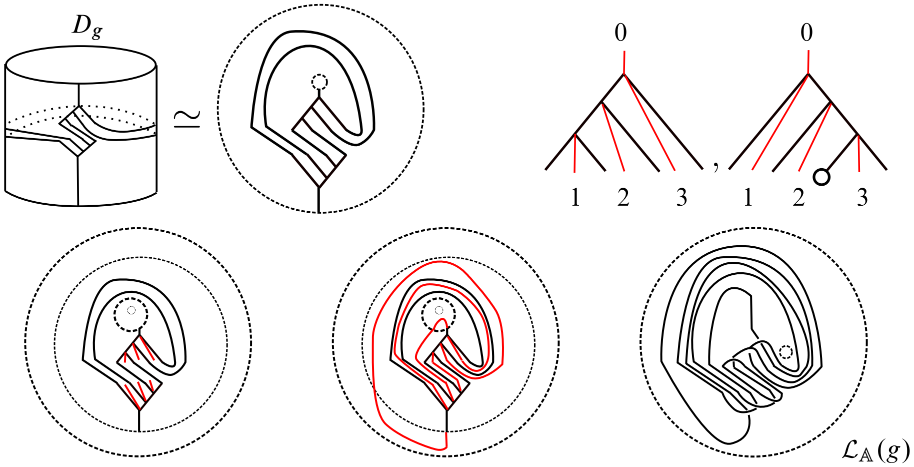


Figure 11: Building the annular link $\mathcal{L}_{\mathbb{A}}(g)$ from $g \in T$ via the strand diagram D_g.

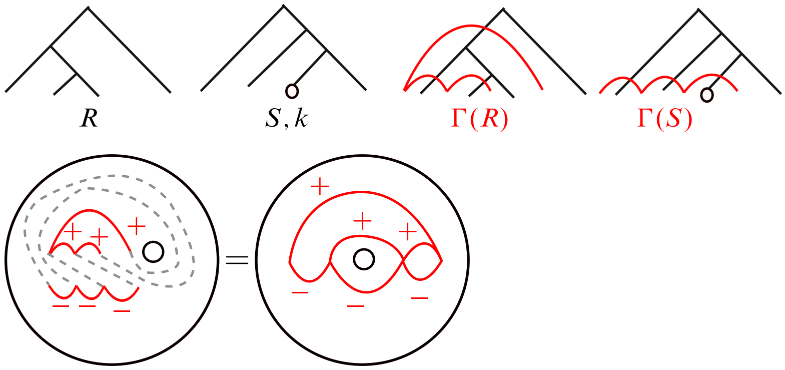


Figure 12: Building the graph $\Gamma_{\mathbb{A}}(g)$ from $g = (R, S; k)$. Dotted lines denote identification of vertices; they are not edges.

Annular links created from T are closely related to Jones' *planar* links built from T; see [13, Section 4.2]. By construction, $\Gamma(g)$ is the image of $\Gamma_{\mathbb{A}}(g)$ under the inclusion $\mathbb{A} \hookrightarrow \mathbb{R}^2$. Consequently, the diagram $\mathcal{L}(g)$ is the image of the diagram $\mathcal{L}_{\mathbb{A}}(g)$ under the same inclusion. From this we can relate the Kauffman bracket of $\mathcal{L}_{\mathbb{A}}(g)$ to that of $\mathcal{L}(g)$:

Proposition 3.1 *Let $g \in T$ be given by $(R, S; k)$. Consider $\mathcal{L}_{\mathbb{A}}(g)$ as an element of $\mathbb{C}[x]$, the Skein module $\mathcal{S}(\mathbb{A})$. Evaluating at $x = (-t^{-1/2} - t^{1/2})$ returns the Kauffman Bracket of $\mathcal{L}(g) \in S^3$.*

To discuss the analogous result for the Jones polynomial, we must first discuss the oriented subgroup $\vec{T}$, which was introduced by Jones in [13] and further studied by Nikkel and Ren in [18]. Defined analogously to $\vec{F}$,

$$\vec{T} := \{g \in T : \mathrm{Chr}_{\Gamma(g)}(2) = 2\}.$$

It follows that for $g \in \vec{T}$, $\mathcal{L}_{\mathbb{A}}(g)$ has a natural orientation.

The Jones polynomial of an annular link $\mathcal{L}$, denoted $V_{\mathcal{L}}^{\mathbb{A}}(t, h) \in \mathbb{Z}[t^{\pm 1/2}, h]$, can be evaluated using the usual skein relation, setting unknotted circles which do not wrap around the annulus equal to $(-t^{1/2} - t^{-1/2})$, and setting unknotted circles wrapping once around the annulus equal to h.

The following proposition relates the Jones polynomial of $\mathcal{L}(g)$ to that of $\mathcal{L}_{\mathbb{A}}(g)$.

Proposition 3.2 *Let $g \in \vec{T}$. Setting $h := (-t^{1/2} - t^{-1/2})$, the Jones polynomial of $\mathcal{L}_{\mathbb{A}}(g)$ is equal to that of $\mathcal{L}(g)$.*

Propositions 3.1 and 3.2, together with Aiello and Conti's proofs of [3, Theorems 6.2 and 7.4], imply Corollary 1.2.

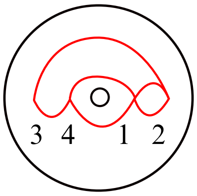
3.1 Annular Thompson badness

We now establish an annular analogue for Jones' Thompson badness. In this section and Section 4, we think of $\mathbb{A}$ as $\mathbb{D}^1 \setminus \{(0,0)\}$.

Definition Let $\Gamma \hookrightarrow \mathbb{A}$ be an edge-signed graph. We say Γ is ATB-*friendly* if

- Γ has no loops,
- all vertices lie on the x axis,
- all edges have interiors either entirely above, or entirely below the x axis.

Now suppose a graph $\Gamma \hookrightarrow \mathbb{A}$ is ATB-friendly. Define e_v^{in}, e^{up}, e^{down}, e_-^{up} and e_+^{down} as before. Let v_L describe the leftmost vertex and let v_1 describe the vertex immediately to the right of the origin. Label the N vertices by $\{1, \dots, N\}$ so that v_1 is labeled 1, the vertex immediately to its right is labeled 2, and so on, until the rightmost vertex is labeled $k-1$. Then label the leftmost vertex k and continue increasing left to right until the vertex immediately to the left of the origin is labeled N. For example:



Let $l(v)$ refer to the label of v. By construction $l(v_L) = k$. Define

$$e_v^< := \{e \in e(\Gamma) : e \text{ connects } v \text{ to some } w \text{ such that } l(w) < l(v)\}.$$

Now define annular Thompson badness, or ATB, as follows:

$$\mathrm{ATB}(\Gamma) := \sum_{v \in V(\Gamma) \setminus v_L} \bigl|1 - |e_v^{\mathrm{in}} \cap e^{\mathrm{down}}|\bigr| + \sum_{v \in V(\Gamma) \setminus v_1} \bigl|1 - |e_v^< \cap e^{\mathrm{up}}|\bigr| + |e_-^{\mathrm{up}}| + |e_+^{\mathrm{down}}|.$$

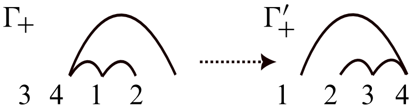
The following proposition motivates this definition as the correct analogue for Thompson badness.

Proposition 3.3 *Let $\Gamma \hookrightarrow \mathbb{A}$ be an* ATB*-friendly graph. Then* $\mathrm{ATB}(\Gamma) = 0$ *if and only if* $\Gamma = \Gamma_{\mathbb{A}}(g)$ *for some* $g \in T$.

Proof Suppose $\Gamma = \Gamma_{\mathbb{A}}(g)$, where $g = (R, S; k)$. By construction, $|e_-^{\mathrm{up}}| = |e_+^{\mathrm{down}}| = 0$. It remains to show $\sum_{v \in V(\Gamma) \setminus v_\ell} \bigl|1 - |e_v^{\mathrm{in}} \cap e^{\mathrm{down}}|\bigr| = \sum_{v \in V(\Gamma) \setminus v_1} \bigl|1 - |e_v^< \cap e^{\mathrm{up}}|\bigr| = 0$. Beginning with the first quantity, let Γ_- denote the subgraph of Γ whose edges have interiors in the lower half plane. Since $\Gamma_- = \Gamma(S)$ and $\Gamma(R, S)$ has Thompson badness equal to zero, $\sum_{v \in V(\Gamma) \setminus v_L} \bigl|1 - |e_v^{\mathrm{in}} \cap e^{\mathrm{down}}|\bigr| = 0$. It remains to show that

$$\sum_{v \in V(\Gamma) \setminus v_1} \bigl|1 - |e_v^< \cap e^{\mathrm{up}}|\bigr| = 0.$$

Define Γ_+ analogously to Γ_-. Because of the edges wrapping around the annulus, $\Gamma_+ \neq \Gamma(R)$, but we can recover $\Gamma(R)$ from Γ_+. This is accomplished by embedding Γ_+ in $\mathbb{R}^2$ so that labels of the edges increase from left to right. Call this embedding Γ_+' and observe that $\Gamma_+' = \Gamma(R)$:



Letting v'_L refer to the leftmost vertex of Γ'_+, we have

$$\sum_{v\in V(\Gamma'_+)\setminus\{v'_L\}} \big|1-|e^{\mathrm{up}}\cap e_v^{\mathrm{in}}|\big| = \sum_{v\in V(\Gamma)\setminus v_1} \big|1-|e^{\mathrm{up}}\cap e_v^{<}|\big|.$$

Since $\Gamma(R,S)$ has Thompson badness zero, the left hand side must be zero. Therefore, $\mathrm{ATB}(\Gamma_{\mathbb{A}}(g))=0$.

Conversely, let Γ be a graph with $\mathrm{ATB}(\Gamma)=0$. Let (R,S) be the unique pair of trees corresponding to (Γ'_+,Γ_-). Then $\Gamma=\Gamma_{\mathbb{A}}(g)$ where $g=(R,S;l(v_L))$. □

4 Proof of Theorem 1.1

This section uses annular Thompson badness to prove 1.1. Given a general graph Γ, we wish to find a graph Γ' such that $L_{\mathbb{A}}(\Gamma)\simeq L_{\mathbb{A}}(\Gamma')$, with $\mathrm{ATB}(\Gamma')<\mathrm{ATB}(\Gamma)$. For this we use Jones' definition of 2-equivalence [13].

4.1 2-equivalence

Two edge-signed planar graphs are defined to be 2-equivalent if they are related by a finite sequence of three moves, which Jones calls 2-moves. If two graphs Γ and Γ' are 2-equivalent then their associated links $L(\Gamma)$ and $L(\Gamma')$ are isotopic.

The first 2-move is the addition or deletion of a 1-valent vertex, which corresponds to a Reidemeister move of type I. The remaining two 2-moves, each corresponding to Reidemeister moves of type II, are shown in Figure 13.

Jones uses 2-moves to show that every link has a diagram whose Tait graph has Thompson badness zero, and thus can be built from an element of the Thompson group. The proof of Theorem 1.1 will use an analogous strategy to reduce annular Thompson badness of a given edge-signed graph $\Gamma\hookrightarrow\mathbb{A}$. To accomplish this we wish to calculate ATB of any graph Γ, but so far the definition of $\mathrm{ATB}(\Gamma)$ requires that Γ is ATB-friendly. The following lemma takes care of this.

Lemma 4.1 *Let $\Gamma\hookrightarrow\mathbb{A}$ be an edge-signed graph. Then Γ is 2-equivalent to an ATB-friendly graph.*

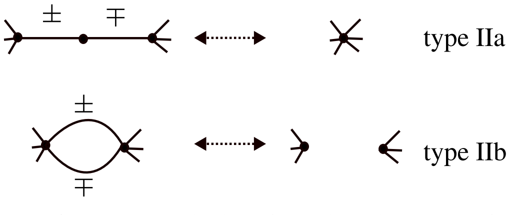


Figure 13: Moves of type IIa and IIb as introduced by Jones in [13].

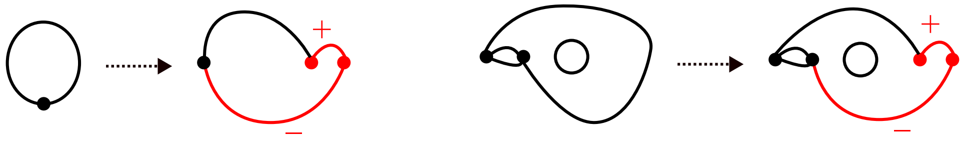


Figure 14: Using moves of type IIa to correct for a loop (left) and an edge whose interior is in both the upper and lower-half plane (right).

Proof Begin by arranging all vertices on the x-axis, which is always possible. At this point, if Γ is ATB-friendly we are done. Otherwise, both loops and edges whose interiors are in both the upper and lower half-plane can be corrected with moves of type IIa. Both cases are shown in Figure 14. $\square$

Therefore, for any graph $\Gamma \hookrightarrow \mathbb{A}$, we define $\mathrm{ATB}(\Gamma) := \mathrm{ATB}(\Gamma')$ where Γ' is obtained from Γ as in the proof of Lemma 4.1.

Remark The following, when applied inductively, proves Theorem 1.1.

Theorem 4.2 *Let* $\Gamma \hookrightarrow \mathbb{A}$ *be an edge-signed graph. If* $\mathrm{ATB}(\Gamma) \neq 0$, *there exists* Γ' *such that* Γ' *is* 2*-equivalent to* Γ, *and* $\mathrm{ATB}(\Gamma') < \mathrm{ATB}(\Gamma)$.

This proof can be thought of as an extension of Jones' proof of [13, Lemma 5.3.13] to the annular case.

Proof To begin, we split into three cases, based on what is causing $\mathrm{ATB}(\Gamma) > 0$:

(1) $\sum_{v \in v(\Gamma) \setminus v_L} \bigl|1 - |e_v^{\text{in}} \cap e^{\text{down}}|\bigr| + \sum_{v \in v(\Gamma) \setminus v_1} \bigl|1 - |e_v^{<} \cap e^{\text{up}}|\bigr| > 0$.

(2) The above quantity is zero but $|e_{-}^{\text{up}}| > 0$.

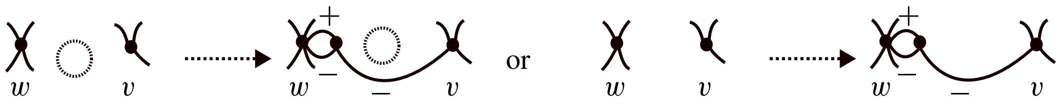
(3) The above quantities are zero but $|e_{+}^{\text{down}}| > 0$.

Case 1 The following four-step process will reduce

$$\sum_{v \in v(\Gamma) \setminus v_L} \bigl|1 - |e_v^{\text{in}} \cap e^{\text{down}}|\bigr| + \sum_{v \in v(\Gamma) \setminus v_1} \bigl|1 - |e_v^{<} \cap e^{\text{up}}|\bigr|$$

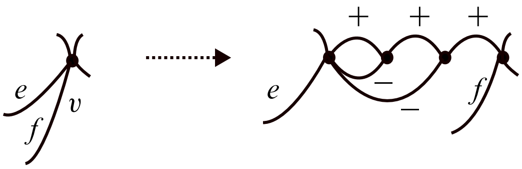
to zero while preserving 2-equivalence.

Case 1, step 1 For each $v \in v(\Gamma) \setminus v_L$ such that $|e_v^{\text{in}} \cap e^{\text{down}}| = 0$, let the vertex immediately to the left of v be called w and proceed as in [13] regardless of whether $v = v_1$:

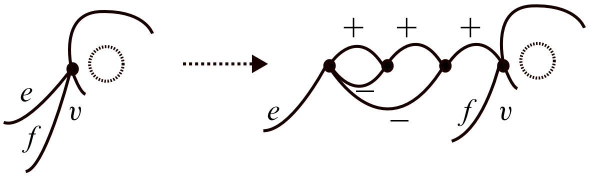


The new vertex does not impact ATB, and the vertex v now has $|e_v^{\text{in}} \cap e^{\text{down}}| = 1$. Each time this step is applied to a relevant vertex, $\sum_{v \in v(\Gamma) \setminus v_L} \bigl|1 - |e_v^{\text{in}} \cap e^{\text{down}}|\bigr|$ decreases by 1, and all other quantities remain unchanged, so ATB decreases by 1.

Case 1, step 2 For each $v \in v(\Gamma) \setminus v_L$ such that $|e_v^{\text{in}} \cap e_v^{\text{down}}| > 1$, proceed as in [13]:

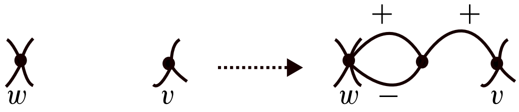


Step 2 ensures that $\sum_{v \in v(\Gamma') \setminus v_L} \bigl|1 - |e_v^{\text{in}} \cap e^{\text{down}}|\bigr| = 0$ for all relevant vertices, however it may increase the quantity $\sum_{v \in v(\Gamma') \setminus v_1} \bigl|1 - |e_v^{<} \cap e^{\text{up}}|\bigr|$. Take, for example, a vertex v with an outgoing edge stretching over the origin:

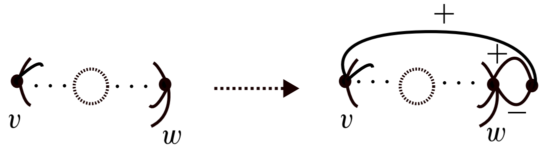


In this example, step 2 increases $|e_v^{<} \cap e^{\text{up}}|$ from 1 to 2. This will be addressed momentarily in step 4, but at this point steps 1 and 2 have reduced $\sum_{v \in v(\Gamma) \setminus v_L} \bigl|1 - |e_v^{\text{in}} \cap e^{\text{down}}|\bigr|$ to 0. If we also have that $\sum_{v \in v(\Gamma) \setminus v_1} \bigl|1 - |e_v^{<} \cap e^{\text{up}}|\bigr| = 0$, we are done. Otherwise proceed to step 3.

Case 1, step 3 We wish to deal with vertices $v \in V(\Gamma) \setminus v_1$ for which $|e_v^{<} \cap e_v^{\text{up}}| = 0$. If $v \neq v_L$, let w refer to the vertex immediately to the left of v and proceed as in [13]:



If $v = v_L$, modify the graph as follows:



In both cases, $\sum_{v \in v(\Gamma) \setminus v_1} \bigl|1 - |e_v^{<} \cap e^{\text{up}}|\bigr|$ decreases by 1 and all other quantities remain unchanged. Therefore each time this step is applied to a relevant vertex, ATB decreases by 1.

Case 1, step 4 We wish to deal with vertices $v \in v(\Gamma) \setminus v_1$ for which $|e_v^{<} \cap e^{\text{up}}| > 1$. This can happen one of four ways. In any case, proceed as in [13], see Figure 15. In each of the four cases, $\sum_{v \in v(\Gamma) \setminus v_1} \bigl|1 - |e_v^{<} \cap e^{\text{up}}|\bigr|$ decreases by 1 and all other quantities remain unchanged. Therefore each time this step is applied to a relevant vertex, ATB decreases by 1. After these four steps, we have

$$\sum_{v \in v(\Gamma) \setminus v_L} \bigl|1 - |e_v^{\text{in}} \cap e^{\text{down}}|\bigr| + \sum_{v \in v(\Gamma) \setminus v_1} \bigl|1 - |e_v^{<} \cap e^{\text{up}}|\bigr| = 0,$$

and $|e_-^{\text{up}}| + |e_+^{\text{down}}|$ remains unchanged. Therefore ATB has been reduced, and this concludes Case 1.

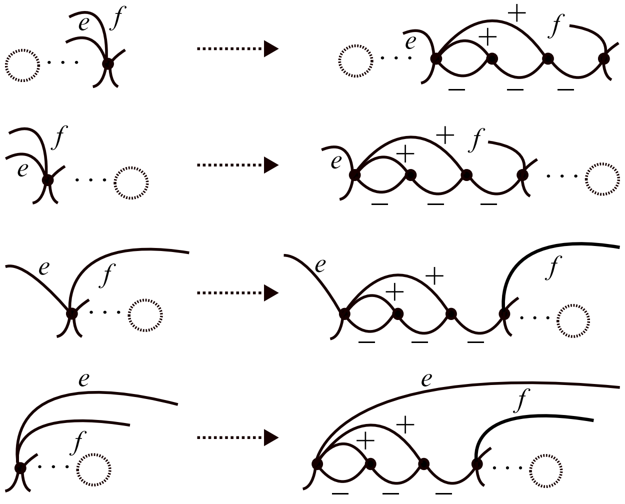


Figure 15: Reducing annular Thompson badness as in Case 1, step 4. Four possible modifications are shown, corresponding to four different ways a vertex v can have $|e_v^< \cap e^{\text{up}}| > 1$.

Case 2 We have $\sum_{v\in v(\Gamma)\setminus v_L}\big|1-|e_v^{\text{in}}\cap e^{\text{down}}|\big| + \sum_{v\in v(\Gamma)\setminus v_1}\big|1-|e_v^{<}\cap e^{\text{up}}|\big| = 0$ and $|e_+^{\text{down}}| > 0$. From now on, only edges in $|e_-^{\text{up}}| \cup |e_+^{\text{down}}|$ will be pictured with a sign; the rest are understood to be in $|e_-^{\text{down}}| \cup |e_+^{\text{up}}|$. Fix an edge in $e' \in e_+^{\text{down}}$ and call its terminal vertex v. We further split into two cases, depending on whether $v = v_1$. Let w refer to the vertex immediately to the left of v. If $v \neq v_1$, proceed as in [13]; see Figure 16.

If $v = v_1$, proceed as in Figure 17.

Note that m denotes the number of edges coming into v from the other side of the annulus. Distinguishing these edges in the picture is necessary because, by virtue of stretching over the origin, they are not in $e_v^<$ and will not affect ATB.

One may wonder why we treat $v = v_1$ differently from $v \neq v_1$ in Figures 16 and 17. Figure 18 depicts what would have happened if we did not. The vertex v_k now has $|e_{v_k}^< \cap e^{\text{up}}| = 2$ which increases ATB by 1 and necessitates a correction as in Case 1, step 4, bringing us back to Figure 17.

To see how Figure 16 preserves 2-equivalence use type IIb moves to remove canceling 2-cycles with opposite signs, then use type I moves to eliminate 1-valent vertices, and finally use type IIa moves to

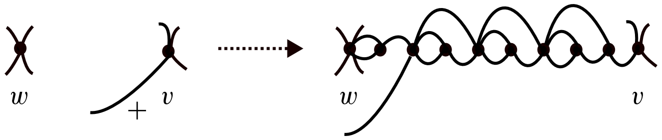


Figure 16: Reducing annular Thompson badness as in Case 2, when $v \neq v_1$.

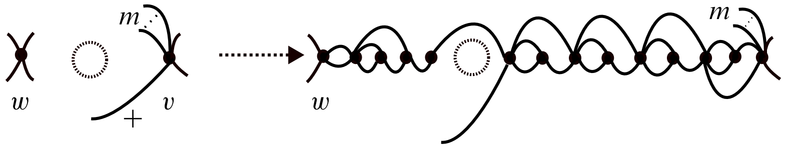
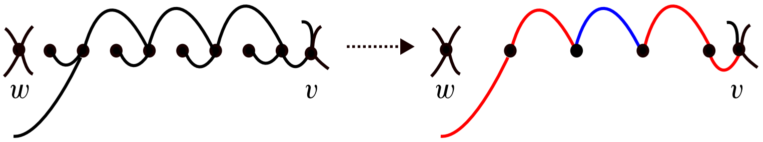


Figure 17: Reducing annular Thompson badness as in Case 2, when $v = v_1$.

collapse canceling edges, pictured in red, so that the original edge, pictured in blue, remains. A similar sequence of moves demonstrates the 2-equivalence for Figure 17:



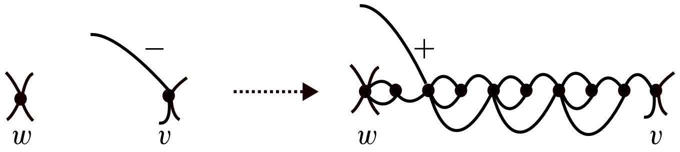
The modifications in Figures 16 and 17 decrease $|e_+^{\text{down}}|$ by 1 and do not impact the quantity

$$\sum_{v \in v(\Gamma)\setminus v_L} \bigl|1 - |e_v^{\text{in}} \cap e^{\text{down}}|\bigr| + \sum_{v \in v(\Gamma)\setminus v_1} \bigl|1 - |e_v^{<} \cap e^{\text{up}}|\bigr|$$

or $|e_-^{\text{up}}|$. Therefore each time this modification is applied to an edge in e_+^{down}, ATB decreases by 1 and any edge in e_+^{down} can be corrected.

Case 3 All other quantities relevant to $\text{ATB}(\Gamma)$ are zero, but $|e_-^{\text{up}}| > 0$. Once again we further split into cases, depending on which side of $\mathbb{A}$ contains the terminal vertex v of a problematic edge e'.

If v is on the left side of $\mathbb{A}$, let w refer to the vertex immediately to the left of v and proceed as in [13]:



A key fact that makes the above work is that as long as v is on the left side of $\mathbb{A}$, we have $|e_v^{\text{in}} \cap e^{\text{up}}| = 1$, which was assumed in [13]. When this is not true, a different correction will be required; specifically, if instead e' terminates at some v on the right side of $\mathbb{A}$, we may have that $|e_v^{\text{in}} \cap e^{\text{up}}| > 1$, due to any number of edges entering v from above which stretch over the origin. We must further divide into two cases, based on whether e' itself stretches over the origin.

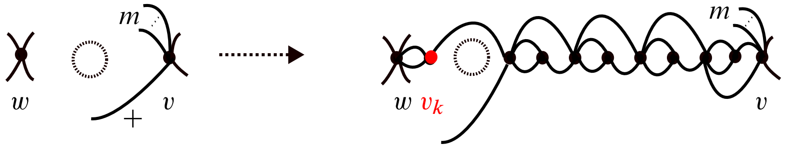
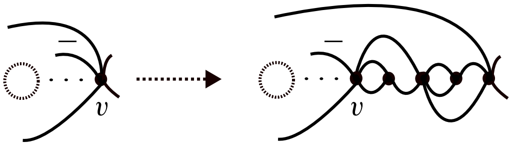
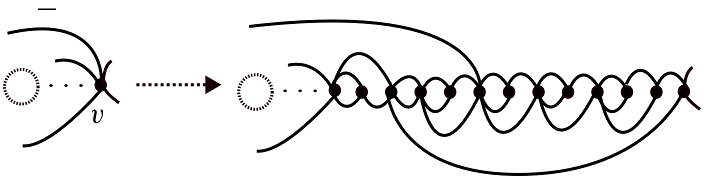


Figure 18: An illustration of why Case 2 must be treated differently when $v = v_1$.

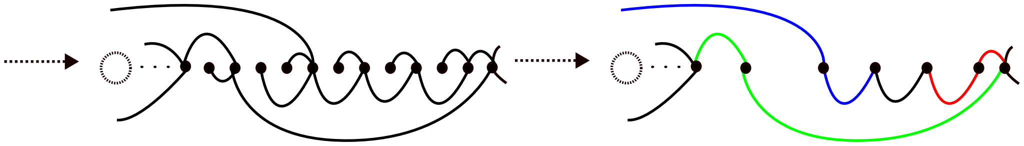
If not, perform the following combination of moves of type I, IIa and IIb, finitely many times until $|e_v^{\mathrm{in}} \cap e^{\mathrm{up}}| = 1$:



Then, we may proceed as in the previous case. If, on the other hand, e' does stretch over the origin, apply the above modification to isolate e' from all other edges in $|e_v^{\mathrm{in}} \cap e^{\mathrm{up}}|$ stretching over the origin. Once the problematic edge is isolated, one may modify the graph as follows:



To see 2-equivalence, use type IIb moves to remove canceling 2-cycles, then use type I moves to remove 1-valent vertices. Lastly apply type IIa moves to collapse the three pairs of canceling edges pictured below in red, blue and green:



The modifications in this step reduce $|e_-^{\mathrm{up}}|$ by 1 and do not affect any other quantities relevant to ATB. This concludes Case 3. These three cases demonstrate that any edge-signed graph embedded in $\mathbb{A}$ with nonzero ATB is 2-equivalent to a graph with lower ATB. □

4.2 Example: a positive trefoil embedded in $\mathbb{A} \times I$

Let $\Gamma \hookrightarrow \mathbb{A}$ be the graph in Figure 19, which corresponds to the positive trefoil embedded as in Figure 1. The graph $\Gamma \hookrightarrow \mathbb{A}$ has $|e_+^{\mathrm{down}}| = 1$, $|e_-^{\mathrm{up}}| = 0$, $|e_{v_3}^{\mathrm{in}} \cap e^{\mathrm{down}}| = 0$, $|e_{v_1}^{\mathrm{in}} \cap e^{\mathrm{down}}| = 1$, $|e_{v_2}^{<} \cap e^{\mathrm{up}}| = 0$ and $|e_{v_3}^{<} \cap e^{\mathrm{up}}| = 2$. Following the process outlined in Theorem 1.1 leads to the element $g = (R, S; k) \in T$ in Figure 19, and $\mathcal{L}_{\mathbb{A}}(g) = L_{\mathbb{A}}(\Gamma)$.

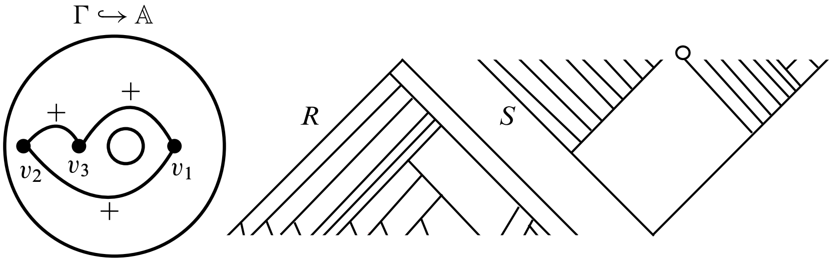


Figure 19: The triple $(R, S; k)$ such that $\mathcal{L}_{\mathbb{A}}(R, S; k) = L_{\mathbb{A}}(\Gamma)$.

References

[1] **V Aiello**, *On the Alexander theorem for the oriented Thompson group $\vec{F}$*, Algebr. Geom. Topol. 20 (2020) 429–438 MR

[2] **V Aiello**, **R Conti**, *Graph polynomials and link invariants as positive type functions on Thompson's group F*, J. Knot Theory Ramifications 28 (2019) art. id. 1950006 MR

[3] **V Aiello**, **R Conti**, *The Jones polynomial and functions of positive type on the oriented Jones–Thompson groups $\vec{F}$ and $\vec{T}$*, Complex Anal. Oper. Theory 13 (2019) 3127–3149 MR

[4] **V Aiello**, **R Conti**, **V F R Jones**, *The Homflypt polynomial and the oriented Thompson group*, Quantum Topol. 9 (2018) 461–472 MR

[5] **M M Asaeda**, **J H Przytycki**, **A S Sikora**, *Categorification of the Kauffman bracket skein module of I-bundles over surfaces*, Algebr. Geom. Topol. 4 (2004) 1177–1210 MR

[6] **J M Belk**, *Thompsons' group F*, PhD thesis, Cornell University (2004) Available at `https://www.proquest.com/docview/305213841`

[7] **J Belk**, **F Matucci**, *Conjugacy and dynamics in Thompson's groups*, Geom. Dedicata 169 (2014) 239–261 MR

[8] **S K Ghosh**, *Planar algebras: a category theoretic point of view*, J. Algebra 339 (2011) 27–54 MR

[9] **G Golan**, **M Sapir**, *On Jones' subgroup of R Thompson group F*, J. Algebra 470 (2017) 122–159 MR

[10] **J J Graham**, **G I Lehrer**, *The representation theory of affine Temperley–Lieb algebras*, Enseign. Math. 44 (1998) 173–218 MR

[11] **V Guba**, **M Sapir**, *Diagram groups*, Mem. Amer. Math. Soc. 620, Amer. Math. Soc., Providence, RI (1997) MR

[12] **V F R Jones**, *The annular structure of subfactors*, from "Essays on geometry and related topics, 1, 2", Monogr. Enseign. Math. 38, Enseignement Math., Geneva (2001) 401–463 MR

[13] **V Jones**, *Some unitary representations of Thompson's groups F and T*, J. Comb. Algebra 1 (2017) 1–44 MR

[14] **V F R Jones**, *On the construction of knots and links from Thompson's groups*, from "Knots, low-dimensional topology and applications" (C C Adams, C M Gordon, V F R Jones, L H Kauffman, S Lambropoulou, K C Millett, J H Przytycki, R Ricca, R Sazdanovic, editors), Springer Proc. Math. Stat. 284, Springer (2019) 43–66 MR

[15] **V F R Jones**, **S A Reznikoff**, *Hilbert space representations of the annular Temperley–Lieb algebra*, Pacific J. Math. 228 (2006) 219–249 MR

[16] **M Khovanov**, *A functor-valued invariant of tangles*, Algebr. Geom. Topol. 2 (2002) 665–741 MR

[17] **V Krushkal**, **L Liles**, **Y Luo**, *Thompson's group F, tangles, and link homology*, preprint (2024) arXiv 2403.16838

[18] **J Nikkel**, **Y Ren**, *On Jones' subgroup of R Thompson's group T*, Internat. J. Algebra Comput. 28 (2018) 877–903 MR

Department of Mathematics, University of Virginia
Charlottesville, VA, United States

`lml2tb@virginia.edu`

Received: 21 April 2024 Revised: 10 July 2024

Algebraic & Geometric Topology **25**:8 (2025) 5031–5057
DOI: 10.2140/agt.2025.25.5031
Published: 20 November 2025

Realizing pairs of multicurves as cylinders on translation surfaces

JULIET AYGUN
JANET BARKDOLL
AARON CALDERON
JENAVIE LORMAN
THEODORE SANDSTROM

Any pair of intersecting cylinders on a translation surface is "coherent," in that the geometric and algebraic intersection numbers of their core curves are equal (up to sign). In this paper, we investigate when a pair of multicurves can be simultaneously realized as the core curves of cylinders on some translation surface. Our main tools are surface topology and the "flat grafting" deformation introduced by Ser-Wei Fu.

57K20; 30F30

1 Introduction

Let S be a closed surface of genus g. The structure of a *translation surface* is an atlas of charts on S away from finitely many cone points whose transition functions are given by translations. Equivalently, a translation surface may be thought of as a collection of polygons in the plane with sides glued by translations (up to cut-and-paste equivalence) or as an identification of S with a Riemann surface X equipped with a holomorphic 1-form ω.

Given a simple closed curve c on S and a translation structure (X, ω) on S, say that c can be *realized as a cylinder* on (X, ω) if it is isotopic to the core curve of some embedded Euclidean cylinder on (X, ω). A *multicurve* γ on S is a union of pairwise disjoint simple closed curves; if each curve of γ is realized as a cylinder on (X, ω), then we say that the union of these cylinders is a *multicylinder.* A multicylinder is *parallel* if all of the core curves have the same slope (relative to the horizontal vector field). Cylinders and multicylinders have long been an important tool in the study of translation surfaces (see, eg, [11]), and recently new attention has been paid to their surface-topological aspects [1; 3; 7].

Given any multicurve γ on S, it is not hard to prove that γ can be realized as a multicylinder on some translation surface if and only if each of its curves is nonseparating: simply find a translation surface containing a multicylinder of the correct topological type (eg, using [13] or [2, Section 6.3]) and use a homeomorphism to take γ to that multicylinder. On the other hand, given a *pair* of multicurves α and β, their intersection pattern may obstruct the existence of a translation surface which simultaneously realizes them as a pair of multicylinders. This is because, given an orientation of the multicurves, the signs of the intersection points constrain the slopes of the realizing cylinders.

This sort of reasoning implies that if two oriented curves $\vec{a}$ and $\vec{b}$ are simultaneously realizable as cylinders on a translation surface, then their geometric intersection number must equal their algebraic intersection number (up to sign); see Lemma 2.2 below. Such a pair of curves is called *coherent*, and if $\vec{\alpha}$ and $\vec{\beta}$ are two oriented multicurves, we say they are *coherent* if the same property holds. For multicurves, coherence is stronger than the property of *pairwise coherence*, which just ensures that each pair of curves $\vec{a} \subset \vec{\alpha}$ and $\vec{b} \subset \vec{\beta}$ are coherent (see Example 5.1). Given unoriented multicurves α and β, we say they are *coherently orientable* if they can be oriented to be coherent.

We recall that a pair of multicurves *fills* S if the complement of their union is a collection of disks. Filling pairs are important for a number of reasons, not least because they can be used to build explicit examples of pseudo-Anosov homeomorphisms [8].

Theorem 1.1 *Let (α, β) be a pair of multicurves on S. The following are equivalent*:

1. *There exists a translation surface on which α and β are both realized as parallel multicylinders.*
2. *There is a coherently orientable filling pair (α', β') with $\alpha' \supset \alpha$ and $\beta' \supset \beta$.*
3. *The multicurves α and β are coherently orientable* ***and*** *no curve of α separates $S \setminus \beta$ and no curve of β separates $S \setminus \alpha$ (in particular, no curves of α and β are isotopic).*

That (2) implies (1) is well known, and that (1) implies (2) is not hard (see Lemma 2.9). Our main contribution, then, is showing the equivalence of these conditions with (3), which is easily checkable given two multicurves α and β. In fact, we prove a stronger theorem in which the orientations of α and β may be prescribed and where the core curves of the realizing multicylinders must all point the same direction. See Theorem 3.1.

Condition (3) implies that each curve in $\alpha \cup \beta$ must be nonseparating. Thus, as a special case of Theorem 1.1, we have the following corollary.

Corollary 1.2 *Two nonseparating curves a and b are jointly realizable as cylinders on a translation surface if and only if they are coherently orientable.*

Two multicurves both being realized as parallel multicylinders is a very strong condition, and one could weaken what it means for α and β to be realized. Say that a triple of unoriented multicurves $(\gamma_1, \gamma_2, \gamma_3)$ is coherently orientable if we can assign orientations so that simultaneously every pair $(\vec{\gamma}_i, \vec{\gamma}_j)$ is coherent.

Theorem 1.3 *Let (α, β) be a pair of multicurves on S. Then there exists a translation surface on which α is realizable as a parallel multicylinder and β is realizable as an arbitrary multicylinder if and only if the following conditions hold*:

1. *There exists a filling pair (α', γ) such that $\alpha \subset \alpha'$ and (α', γ, b) is coherently orientable for each curve $b \subset \beta$.*
2. *Each curve of α and of β is nonseparating.*

It remains an open question to give a characterization of (1) solely in terms of the combinatorial data of α and β on S. In particular, is (1) equivalent to being coherently orientable?

At the weakest end of the spectrum, one could drop all assumptions on parallelism and ask for α and β to be realized as any sort of multicylinders on some translation surface. In this setting,

1. coherence is not necessary (Example 5.1),
2. pairwise coherence is necessary, yet
3. pairwise coherence is not sufficient (Example 5.2).

A further open question is characterizing all the necessary and sufficient conditions to realize any pair of multicurves as arbitrary multicylinders on some translation surface.

1.1 Outline

In Section 2, we collect some relevant background about cylinders on translation surfaces. One of the most important definitions in this section is the “Thurston–Veech construction” that connects filling pairs of multicurves with translation surfaces. In Section 3, we prove Theorem 1.1 and its generalization, Theorem 3.1, using the Thurston–Veech construction and surface-topological arguments. Theorem 1.3 is proven in Section 4 using an explicit surgery called “horizontal grafting” which deforms a translation surface enough to realize certain concatenations of saddle connections as cylinders. Finally, in Section 5, we give a number of examples that show that the question of realizing a pair of multicurves as arbitrary multicylinders is more subtle than just (pairwise) coherence.

Acknowledgements

This project began at the Yale SUMRY REU program in summer 2021, and we would like to thank all of the participants of SUMRY for their support and attention to our project. We would also like to thank Marissa Loving, Nick Salter, and Jane Wang for helpful conversations as well as the rest of the visitors to the SUMRY program for contributing ideas and enthusiasm. Additional thanks goes to Luke Jeffreys for his helpful comments. Calderon acknowledges financial support from NSF grants DMS-2005328 and DMS-2202703.

2 Cylinders on translation surfaces

In this section, we discuss the basics of cylinders on translation surfaces. We first record a number of useful properties concerning intersections and coherence between geodesics on a translation surface. We then recall the Thurston–Veech construction, which realizes a filling pair of multicurves as multicylinders on a (half)-translation surface. This section also contains a proof that if this filling pair is coherent, then the construction yields a translation surface.

2.1 Translation surfaces

A *translation surface* is a collection of compact polygons in $\mathbb{C}$ in which each side (vector in $\mathbb{R}^2 \cong \mathbb{C}$) is identified with another by translation, considered up to cut-and-paste equivalence. Performing the gluings constructs a topological surface S, which inherits from $\mathbb{C}$ a metric that is Euclidean everywhere except on a finite set of cone points. Equivalently, a translation surface is a Riemann surface X equipped with a holomorphic differential ω. The cone points of the flat metric correspond to the zeros of ω. We denote a translation surface by the pair (X, ω).

The equivalence can be seen as follows. Surfaces obtained from gluing polygons in $\mathbb{C}$ have natural charts to $\mathbb{C}$ away from cone points. At a cone point of angle $2\pi(k+1)$, these charts are $z \mapsto z^k$. Translation maps are holomorphic, so we obtain the structure of a Riemann surface. There is the holomorphic differential dz on $\mathbb{C}$, which induces a holomorphic differential on the polygons, which then pulls back to one on the Riemann surface obtained after gluing.

Conversely, let $\Sigma = P_1, \dots, P_n$ be the zeros of ω. A geodesic arc whose endpoints are in Σ and is otherwise disjoint from Σ in its interior is called a *saddle connection*. For an oriented curve or arc $\vec{a}$ on S, we denote its *period* on the translation surface (X, ω) by

$$\mathrm{hol}(\vec{a}) := \int_{\vec{a}} \omega.$$

Equivalently, the period is the vector obtained by connecting the images of $\vec{a}(0)$ and $\vec{a}(1)$ under the developing map of the flat metric. When a is unoriented, its period is defined up to multiplication by -1.

Choose some basis $c_1, \dots, c_{2g+n-1}$ of $H_1(X, \Sigma; \mathbb{Z})$ consisting of saddle connections. Cutting along the c_i's, we obtain a collection of polygons where each of the two sides corresponding to one of the c_i has direction vector given by the complex number $\mathrm{hol}(c_i)$. Moreover, the vertices of these polygons glue up to be Σ, and the local horizontal vector field on the surface determined by ω corresponds to the horizontal vector field on $\mathbb{C}$. A different choice of basis for $H_1(X, \Sigma; \mathbb{Z})$ gives a different polygon up to cut-and-paste equivalence.

A *stratum* of translation surfaces is the subset of all translation surfaces with a fixed number and angle of cone points. The *periods* of a translation surface are the numbers $\{\mathrm{hol}(c_i)\}$ defined above. Adjusting the periods slightly corresponds to adjusting the edges of the polygons defining our translation surface, and doing this we still obtain a translation surface in the same stratum. In this way the periods of a translation surface establish local (orbifold) *period coordinates* for strata modeled by $H^1(X, \Sigma; \mathbb{C})$.

Similarly, a *half-translation surface* is a collection of polygons in $\mathbb{C}$ in which each side is identified with another by translation with possibly a rotation by π, considered up to cut-and-paste equivalence. Equivalently, a half-translation surface is a Riemann surface X equipped with a quadratic differential q. Sometimes, q may be the square of an abelian differential (ie, the sides of the polygon may all be glued by translation), so one can regard translation surfaces as a subset of half-translation surfaces. Strata of quadratic differentials also have period coordinates, but since we will not use them, we omit this discussion.

We refer the reader to [12] for a more thorough background on (half-)translation surfaces, period coordinates, and strata.

2.2 Geodesics and multicylinders

The *geometric intersection number* between two curves a and b is the minimal number of times a curve in the isotopy class of a can intersect a curve in the isotopy class of b. The *algebraic intersection number* between two oriented curves $\vec{a}$ and $\vec{b}$ is the sum over all of their intersection points p_i where $+1$ is added if $\vec{b}$ crosses $\vec{a}$ from right to left at p_i and -1 is added otherwise. In this paper, the form $\iota(\_,\_)$ denotes the geometric intersection number and the form $\hat{\iota}(\_,\_)$ denotes the algebraic intersection number. Two curves on a surface are said to be in *minimal position* if they realize the geometric intersection number for their isotopy classes. Throughout the paper, unless otherwise stated, we will assume that all curves are realized in minimal position.

A(n isotopy class of a) curve c on a (half-) translation surface (X,ω) is *realizable as a cylinder* if c is isotopic to the core curve of some embedded Euclidean cylinder on (X,ω). In other words, there is some geodesic representative of c that has constant slope on (X,ω) and does not contact a cone point. In fact, the family of geodesic representatives sweeps out the embedded cylinder, all have the same length, and all representatives except for the two boundary components of the cylinder are nonsingular. The circumference of the embedded cylinder is equal to the magnitude of $\mathrm{hol}(c)$ (equipped with either orientation).

Definition 2.1 A *multicylinder* on a (half-)translation surface (X,ω) is a collection of cylinders whose core curves are all disjoint and nonisotopic (for some, hence any, choice of nonsingular core curves). A multicylinder is a *parallel multicylinder* if each of the core curves of its constituent cylinders has the same slope.

More generally, every (half-)translation surface has a circle's worth of (singular) *directional foliations* by parallel lines of the same slope. The directional foliations of translation surfaces are always orientable, while those for half-translation surfaces are not necessarily so. If the directional foliation of a (half-)translation surface (X,ω) is a union of parallel multicylinders glued along parallel saddle connections, then (X,ω) is said to be *periodic* in that direction.

If α is a parallel multicylinder and b is a cylinder on a translation surface, then the intersection pattern is constrained. Recall that two multicurves are coherently orientable if we can orient them so that their geometric intersection number equals their algebraic intersection number.

Lemma 2.2 *Given a parallel multicylinder α and arbitrary multicylinder β on a translation surface, their core multicurves are coherently orientable.*

In particular, any two cylinders on a translation surface must have coherently orientable core curves.

Proof Let $\alpha = a_1 \cup \dots \cup a_n$ and $\beta = b_1 \cup \dots \cup b_m$. Rotating the surface as necessary, we may assume that α is horizontal. Orient the core curves of α to point in the $+x$ direction.

The core curve of each b_i has constant slope since it is realized as a cylinder. If some b_i does not intersect α, then $\iota(\alpha, b_i) = |\hat{\iota}(\alpha, b_i)| = 0$ and its orientation does not matter. If b_i intersects α, then it cannot be horizontal, and so we assign an orientation to b_i so that its geodesic representatives point with angle in $(0, \pi)$. Doing this for all b_i, we have thus oriented β so that all intersections with α are positive. This further implies there cannot be any bigons between the core curves, so the pair of oriented multicurves are in minimal position. We thus conclude they are coherent. □

2.3 Realizing topological multicurves as cylinders

Throughout the rest of this paper, we will be interested in when pairs of multicurves on a topological surface S of genus g can be realized as a pair of multicylinders on some (half-)translation surface (X, ω). However, there is no canonical way of identifying the curves on S with those on (X, ω). In order to make this identification, we must fix a marking.

A *marking* of a (half-)translation surface (X, ω) is a choice of homeomorphism φ from S to (X, ω), considered up to isotopy.[1] Markings allow us to compare curves on different flat surfaces. Usually one denotes a *marked translation surface* by the triple (X, ω, φ). However in this paper, any translation surface (X, ω) is assumed to be marked, so for simplicity of notation we will usually omit φ from the triple. The set of all marked half-translation surfaces is naturally identified with the cotangent bundle of the Teichmüller space of S, and the set of all marked translation surfaces (X, ω, φ) corresponds to a subbundle of the cotangent bundle.

Definition 2.3 A multicurve $\gamma = c_1 \cup \cdots \cup c_n$ on S *can be realized as a multicylinder on a (half-)translation surface* if there exists some marked (half-)translation surface (X, ω, φ) so that the curves $\varphi(c_i)$ are all isotopic to the core curves of a multicylinder on (X, ω). Similarly, we say that γ *can be realized as a parallel multicylinder* if $\varphi(c_i)$ can be isotoped to the core curves of a parallel multicylinder.

When a multicurve is oriented, we can also introduce a more restrictive criterion for cylindricity that forces the core curves to all point the same way (as opposed to just having the same slope).

Definition 2.4 An oriented multicurve $\vec{\gamma} = \vec{c}_1 \cup \cdots \cup \vec{c}_n$ *can be realized as a directional multicylinder* on a translation surface if there exists some marked translation surface (X, φ) and some direction $\theta \in [0, 2\pi)$ so that the curves of $\varphi(\vec{c}_i)$ are all isotopic (respecting orientations) to closed nonsingular curves of the directional foliation pointing in direction θ.

This definition only makes sense for translation surfaces, as the directional foliations of half-translation surfaces are not always orientable.

---

[1] If the translation surface has n cone points, one could also consider a marking $\varphi' \colon S_{g,n} \to X$ that records the location of the cone points. With this definition of marking, we cannot isotope curves over cone points because we treat cone points as punctures. This gives rise to a different notion of realizability as a multicylinder which depends heavily on the position of curves vis-à-vis cone points. See also Remark 4.5.

Equivalently, Definition 2.4 can also be stated in terms of the periods of the curves of $\vec{\gamma}$. Recall that the *period* of any oriented path $p\colon [0,1] \to X$ on a translation surface is the integral of the differential defining (X,ω) over the path p. An oriented multicurve $\vec{\gamma}$ is then realized as a directional multicylinder on (X,ω) if and only if it is realized as a parallel multicylinder and the periods of the constituent $\vec{c}_i \subset \vec{\gamma}$ are all *positive* real multiples of each other.

The property of realizing a given curve as a cylinder is open, as proven by the following lemma:

Lemma 2.5 *If a curve c is realized as a cylinder on some (half-)translation surface (X,ω), then there exists an open neighborhood N of (X,ω) in the ambient stratum of marked (half-)translation surfaces such that c is realized as a cylinder on every surface $(X',\omega') \in N$.*

Proof If c is realized as a cylinder on (X,ω), then it has a nonsingular geodesic representative of constant slope. There is some minimum distance, d, between the geodesic and any of the cone points in Σ. Varying slightly in period coordinates (equivalently, varying the edges of the defining polygons), (X,ω) can then be deformed into a new surface X' such that each point of Σ remains at least $d/2 > 0$ away from c, and on this surface, c is a geodesic line of constant slope. Thus, c is still realizable as a cylinder on (X',ω'). $\square$

The same argument shows that realizing a multicurve as a multicylinder is also an open condition, but realizing it as a parallel or directional multicylinder is not. We say more to this effect in Lemma 2.9.

2.4 The Thurston–Veech construction

In this subsection, we recall a method of finding a marked (half-)translation surface which will realize certain pairs of multicurves $\vec{\alpha}$ and $\vec{\beta}$ as (parallel) directional multicylinders.

A pair of multicurves (α,β) *fills* S if $S \setminus (\alpha \cup \beta)$ is a disjoint union of open disks. It is a standard result that a pair of multicurves (α,β) fills S if and only if for every curve $c \subset S$ that is not nullhomotopic, $\iota(c,\alpha) + \iota(c,\beta) > 0$. In the case that (α,β) is a filling pair of multicurves, Thurston and Veech independently identified a now standard construction which realizes α and β as cylinders on a half-translation surface. Heuristically, one can think of this procedure as thickening α and β into cylinders while contracting the remaining disks into cone points [10]. A more formal statement of the construction is given below.

Construction 2.6 (Thurston–Veech construction) Let (α,β) be a filling pair of nonempty multicurves on S and think of $\alpha \cup \beta$ as an embedded graph, with vertices corresponding to points of intersection and edges corresponding to strands of α and β running between intersection points. Then the dual graph of $\alpha \cup \beta$ partitions S into squares. Considering these to be flat unit squares, we obtain a *square-tiled surface* on which the thickened curves of α comprise the horizontal cylinders and the thickened curves of β comprise the vertical cylinders. See Figure 1. Because α is always horizontal and β is always vertical, edges of squares are only identified by translations and rotations by π. Thus, the resulting surface $\mathrm{TV}(\alpha,\beta)$ is a half-translation surface.

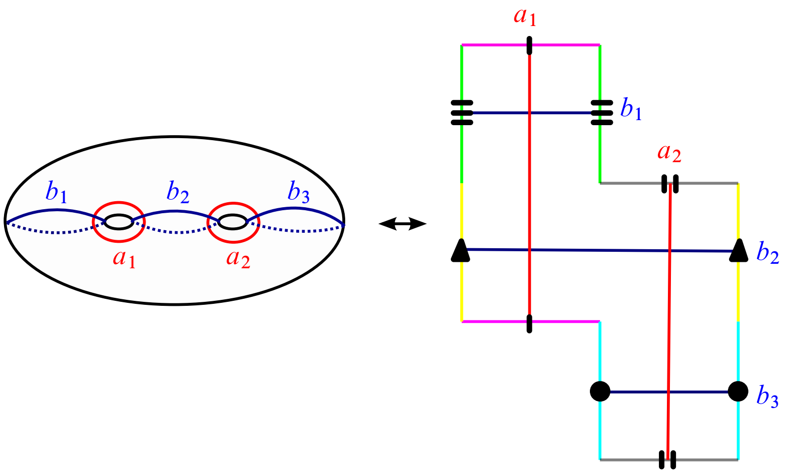


Figure 1: An example of the Thurston–Veech construction applied to the pair of multicurves $\alpha = a_1 \cup a_2$ and $\beta = b_1 \cup b_2 \cup b_3$.

We observe that by construction, any surface $\mathrm{TV}(\alpha, \beta)$ is periodic in both the horizontal and vertical directions and is tiled by squares. Such surfaces are sometimes called "origamis" in the literature.

In this paper, we restrict our attention to translation surfaces because half-translations are too flexible, as demonstrated in Lemma 2.7 below.

Recall that a *pants decomposition* of S is a multicurve α so that $S \setminus \alpha$ is a union of three-holed spheres. An *extension* of a multicurve α on S is a multicurve α' on S which contains α and preserves the orientations, if any, of the curves of α. A pair of multicurves (α, β) on S is *extended* to a pair of multicurves (α', β') such that α' is an extension of α and β' is an extension of β.

Lemma 2.7 *Every pair of multicurves α and β with no curves in common is realized as a pair of parallel multicylinders on some half-translation surface.*

Proof Choose two pants decompositions P_α and P_β extending α and β, respectively, that have no curves in common. This can be accomplished, for example, by extending both to arbitrary pants decompositions and then applying an "elementary move" to any curves they have in common. See [6].

We now claim that P_α and P_β jointly fill the surface. Indeed, since pairs of pants have no nonperipheral simple curves, if a curve c does not meet P_α, then it must be a curve of P_α. Since P_α and P_β share no curves, this means that c must meet P_β, and thus (P_α, P_β) fill the surface. Applying Construction 2.6 therefore produces a half-translation surface on which P_α and P_β, hence α and β, are both realized as parallel multicylinders. □

Adding the requirement that a pair of multicurves is coherently orientable guarantees Construction 2.6 will yield a translation surface.

Lemma 2.8 *Given a filling pair of multicurves* α *and* β *on* S, *the surface* $\mathrm{TV}(\alpha,\beta)$ *is a translation surface if and only if* α *and* β *are coherently orientable.*

This result has already been shown in [3; 7], but we include a short proof for completeness.

Proof The forward direction follows from Lemma 2.2.

In the backward direction, suppose that $\vec{\alpha}$ and $\vec{\beta}$ form a coherent filling pair of multicurves. Then $\mathrm{TV}(\vec{\alpha},\vec{\beta})$ can be represented by a collection of squares, each containing a single intersection of $\vec{\alpha}$ and $\vec{\beta}$. Rotating the squares so that $\vec{\alpha}$ runs left-to-right, the coherence condition ensures that, without loss of generality, $\vec{\beta}$ runs bottom-to-top. Therefore, top edges of squares are identified with bottom edges of squares, and right edges are identified with left edges. Thus, the resulting square-tiled surface is a translation surface. □

In particular, this lemma tells us that if a coherent pair of oriented multicurves $\vec{\alpha}$ and $\vec{\beta}$ can be extended to a coherent filling pair, then they can be realized as the horizontal and vertical cylinders of some square-tiled surface. It turns out the converse is also true:

Lemma 2.9 *Let* $(\vec{\alpha},\vec{\beta})$ *be a coherent pair of oriented multicurves on* S. *Then there exists a translation surface* (X,ω) *on which* $(\vec{\alpha},\vec{\beta})$ *are realized as directional multicylinders if and only if* $(\vec{\alpha},\vec{\beta})$ *can be extended to a coherent filling pair.*

Proof Suppose that $\vec{\alpha}=\vec{a}_1\cup\cdots\cup\vec{a}_m$ and $\vec{\beta}=\vec{b}_1\cup\cdots\cup\vec{b}_n$ are directional multicylinders on a translation surface (X,ω). Postcomposing with an element of $\mathsf{GL}_2(\mathbb{R})$ as necessary, we may assume that the geodesic representatives of each curve of $\vec{\alpha}$ points in the positive x-direction and each curve of $\vec{\beta}$ in the positive y-direction. That is,

$$(*)\qquad \mathrm{hol}(\vec{a}_i)\in\mathbb{R}_{>0} \quad\text{and}\quad \mathrm{hol}(\vec{b}_j)\in i\mathbb{R}_{>0}$$

for all $i=1,\dots,m$ and $j=1,\dots,n$.

Choose a local period coordinate chart for the ambient stratum around (X,ω); then the equations

$$\mathrm{Im}(\mathrm{hol}(\vec{a}_i))=0 \quad\text{and}\quad \mathrm{Re}(\mathrm{hol}(\vec{b}_j))=0$$

cut out a (nonzero) $\mathbb{R}$-linear subspace V of $H^1(S,\mathrm{Zeros}(\omega);\mathbb{C})$. The positivity conditions of $(*)$ further specify an intersection of open half-spaces inside of V, which is nonempty because it contains (X,ω). Let $U\subset V$ denote this intersection; note that U is a relatively open set inside of V.

Since V is cut out by integral-linear equations, rational points $H^1(S,\mathrm{Zeros}(\omega);\mathbb{Q}\oplus i\mathbb{Q})$ are dense in V. Since cylinders persist under small deformations (Lemma 2.5), this implies that there is some

$$(X',\omega')\in H^1(S,\mathrm{Zeros}(\omega);\mathbb{Q}\oplus i\mathbb{Q})\cap U$$

on which $\vec{\alpha}$ and $\vec{\beta}$ remain cylinders. In particular, (X', ω') is a square-tiled surface on which $\vec{\alpha}$ and $\vec{\beta}$ are horizontal and vertical cylinders. The entire horizontal and vertical multicurves of (X', ω'), oriented in the positive x- and y-directions, respectively, therefore constitute a (coherent) filling pair extending $\vec{\alpha}$ and $\vec{\beta}$.

The backwards direction is just Lemma 2.8. □

3 Coherent, filling pairs of multicurves

Given a pair of oriented multicurves $(\vec{\alpha}, \vec{\beta})$ on S, the strongest sense in which we could realize $\vec{\alpha}$ and $\vec{\beta}$ as multicylinders on a translation surface is as directional multicylinders. The goal of this section is to prove a refinement of Theorem 1.1, characterizing exactly when this is possible.

Let S be a surface and $\vec{c}$ an oriented simple closed curve on it. The surface $S \setminus \vec{c}$ is obtained by removing a small annular neighborhood of $\vec{c}$; we denote the two oriented boundary components of that neighborhood as well as the corresponding boundary components of $S \setminus \vec{c}$ by $\vec{c}_L$ and $\vec{c}_R$, depending on whether the surface is on the left- or right-hand side of the curve. Throughout this section, we assume that none of the constituent curves in a pair of multicurves are isotopic.

Theorem 3.1 *Let $(\vec{\alpha}, \vec{\beta})$ be a pair of coherent oriented multicurves such that no two curves of $\vec{\alpha} \cup \vec{\beta}$ are isotopic. Then $(\vec{\alpha}, \vec{\beta})$ can be simultaneously realized as directional multicylinders on a translation surface if and only if the following holds*:

($\star$) *For every multicurve $\vec{\gamma} \subseteq \vec{\alpha} \cup \vec{\beta}$ and every complementary subsurface W of $S \setminus \vec{\gamma}$, write*

$$\partial W = A_L \sqcup A_R \sqcup B_L \sqcup B_R,$$

where A_L denotes the set of boundary components arising from the left-hand sides of curves of $\vec{\alpha}$, and so on. Then $A_L \neq \varnothing$ if and only if $A_R \neq \varnothing$, and $B_L \neq \varnothing$ if and only if $B_R \neq \varnothing$.

In particular, if $(\vec{\alpha}, \vec{\beta})$ satisfies ($\star$) then none of the curves of $\vec{\alpha}$ or $\vec{\beta}$ are separating. For example, Figure 5 does not satisfy the criterion in Theorem 3.1 because after setting $\gamma = \vec{a}_1 \cup \vec{a}_2 \cup \vec{b}_1$ and W to be the subsurface in the upper left corner, $B_L = \varnothing$ but $B_R \neq \varnothing$.

The proof of this theorem is spread throughout the section. In Section 3.1, we give both geometric and topological proofs that ($\star$) is necessary. In light of the equivalence of Lemma 2.9, it is enough to prove that a coherent pair of multicurves $(\vec{\alpha}, \vec{\beta})$ satisfying the hypotheses of Theorem 3.1 can be extended to a coherent, filling pair. Our plan is to use an inductive "connected sum" construction to extend the pair $(\vec{\alpha}, \vec{\beta})$, building new curves out of, and informed by, existing ones, to decrease the complexity of $S \setminus (\alpha \cup \beta)$ at each step. However, it turns out curves of $\vec{\alpha}$ which do not meet any of the curves of $\vec{\beta}$ (or vice versa), called *singletons*, complicate this strategy. As such, in Section 3.2 we extend $(\vec{\alpha}, \vec{\beta})$ to a coherent pair $(\vec{\alpha}', \vec{\beta}')$ in which there are no singletons, that is, so that each curve of $\vec{\alpha}'$ intersects some curve of $\vec{\beta}'$, and vice versa. We then use connected sums in Section 3.3 to add curves to $\vec{\beta}'$ that are all coherent with $\vec{\alpha}'$, thereby completing the proof of Theorem 3.1. This section also contains an explanation of how Theorem 3.1 implies Theorem 1.1.

For the rest of the paper, let us assume (without loss of generality) that $\hat{\imath}(\vec{\alpha}, \vec{\beta}) \geq 0$, so that the curves of $\vec{\beta}$ cross the curves of $\vec{\alpha}$ from right to left.

3.1 Necessity

The first step to proving Theorem 3.1 is to show that $(\star)$ is necessary.

Lemma 3.2 *Let $(\vec{\alpha}, \vec{\beta})$ be a pair of multicurves on S with $\iota(\vec{\alpha}, \vec{\beta}) > 0$. Suppose that there exists a multicurve $\vec{\gamma} \subset \vec{\alpha} \cup \vec{\beta}$ and component W of $S \setminus \vec{\gamma}$ with $A_L \neq \varnothing$ but $A_R = \varnothing$. Then $(\vec{\alpha}, \vec{\beta})$ are not realizable as a pair of directional multicylinders on any translation surface.*

Proof Suppose that $\vec{\alpha} = \bigcup_i \vec{a}_i$ and $\vec{\beta} = \bigcup_j \vec{b}_j$ were realized as a pair of directional multicylinders on the marked translation surface (X, ω, φ). Because $\vec{\alpha}$ is realized as a directional multicylinder, all of the period vectors $\mathrm{hol}(\vec{a}_i)$ live in the same ray $\mathrm{hol}(\vec{a}_1) \cdot \mathbb{R}_{>0} \subset \mathbb{C}$. Similarly, there is a ray containing all of the period vectors $\mathrm{hol}(\vec{b}_j)$. Because $\vec{\alpha}$ and $\vec{\beta}$ intersect, the core curves of the corresponding cylinders are not parallel, and so the corresponding rays are not parallel.

By Stokes's theorem, we have

$$\sum_{\vec{a}_L \in A_L} \mathrm{hol}(\vec{a}_L) + \sum_{\vec{b}_L \in B_L} \mathrm{hol}(\vec{b}_L) - \sum_{\vec{b}_R \in B_R} \mathrm{hol}(\vec{b}_R) = \int_{\partial W} \omega = \int_W d\omega = 0,$$

because holomorphic forms on a Riemann surface are always closed. By linear independence, $\mathrm{hol}(\vec{a}_L) = 0$ for all $\vec{a}_L \in A_L$. Since cylinders never have 0 circumference, this is a contradiction. □

The same proof holds if one allows $\vec{\beta}$ to be empty, implying that no separating curve is realized as a cylinder on any translation surface.

While the following consequence of Lemma 2.9 is not strictly necessary for the proof of Theorem 3.1, it gives us another way to understand the obstruction from Lemma 3.2 in more topological terms.

Corollary 3.3 *Let (α, β) be a pair of unoriented multicurves on S. If some curve $a \subset \alpha$ separates $S \setminus \beta$, then α and β cannot be simultaneously realized as parallel multicylinders on a translation surface.*

Proof Suppose that α and β were both realized as parallel multicylinders on some translation surface. By Lemma 2.9, this implies that we can extend (α, β) to a coherently orientable filling pair (α', β').

Now if some curve $a \subset \alpha$ separates $S \setminus \beta$, there exists some subsurface W one of whose boundary components corresponds to $a \subset \alpha$ and the rest of which correspond to curves of β. Since β' fills with $\alpha' \supset \alpha$, this implies there is some curve $b \subset \beta'$ with $\iota(a,b) > 0$. But now since b is disjoint from the other curves of β', it cannot intersect any of the other boundary curves of W, so it must cross a at least twice: one time entering W, and one time escaping W. This is a contradiction with coherence, and so we see that (α,β) cannot be realized as a pair of parallel multicylinders. □

3.2 Removing singletons

We now begin to build towards a proof that the phenomenon described in Lemma 3.2 is the only obstruction to realizing a coherent pair of multicurves as a pair of directional multicylinders. The first step is to show that any coherent pair of oriented multicurves $(\vec{\alpha},\vec{\beta})$ satisfying $(\star)$ can be extended to a coherent pair $(\vec{\alpha}',\vec{\beta}')$ where every curve of $\vec{\alpha}'$ intersects a curve of $\vec{\beta}'$, and vice versa.

Definition 3.4 The *singleton set* $\mathbb{S}_{\vec{\beta}}(\vec{\alpha})$ of $\vec{\alpha}$ with respect to $\vec{\beta}$ is the set of curves of $\vec{\alpha}$ which do not intersect any curve of $\vec{\beta}$ when realized in minimal position.

Given $\vec{a} \in \mathbb{S}_{\vec{\beta}}(\vec{\alpha})$, we wish to construct a curve $\vec{b}$ intersecting $\vec{a}$ such that $(\vec{\alpha},\vec{\beta}\cup\vec{b})$ remains coherent. Our strategy is to concatenate arcs which are all coherent with $\vec{\alpha}$ but disjoint from $\vec{b}$. First, we need to show that there is a sufficient supply of such arcs (Lemma 3.6 below).

Let us first recall the notion of a connected sum of curves.

Definition 3.5 Let a be a simple curve or arc on a surface and let b be a disjoint simple closed curve. Let ε be an arc connecting a to b, disjoint from a and b except at its endpoints. The *connected sum* $a +_\varepsilon b$ is the curve (or arc) obtained by taking the boundary of a tubular neighborhood of $a \cup \varepsilon \cup b$.

If $\vec{a}$ and $\vec{b}$ are oriented and ε runs from the left-hand side of $\vec{a}$ to the left-hand side of $\vec{b}$, then moreover one can orient $\vec{a} +_\varepsilon \vec{b}$ so that it runs parallel to $\vec{a}$ and $\vec{b}$ away from ε. See Figure 2.

Lemma 3.6 *Let $(\vec{\alpha},\vec{\beta})$ be a coherent pair of oriented multicurves and let W be a component of $S \setminus (\vec{\beta} \cup \mathbb{S}_{\vec{\beta}}(\vec{\alpha}))$. Let A_L (A_R) denote the set of boundary components of W arising from the left-(right-)hand sides of curves of $\mathbb{S}_{\vec{\beta}}(\vec{\alpha})$. Then for every $\vec{a}_L \in A_L$ and $\vec{a}_R \in A_R$, there is an oriented arc on W traveling from $\vec{a}_L$ to $\vec{a}_R$ that crosses $\vec{\alpha}|_W$ from right to left.*

Proof Given any $\vec{a}_L \in A_L$ and any $\vec{a}_R \in A_R$, let $\vec{\gamma}$ be an arbitrary oriented arc on W connecting initial boundary component $\vec{a}_L$ to terminal boundary component $\vec{a}_R$. It is possible that $\vec{\gamma}$ intersects arcs of $\vec{\alpha}|_W$ from left to right (when realized in minimal position). Our strategy is to surger $\vec{\gamma}$ to remove these intersection points.

So suppose that $\vec{\gamma}$ crosses an arc $\mathfrak{a}$ of $\vec{\alpha}|_W$ from left to right. Here, $\vec{\alpha}|_W$ is the restriction of $\vec{\alpha}$ to W. Order all of the points $p_1,\dots,p_N$ of $\mathfrak{a} \cap \gamma$ as seen by the oriented arc $\vec{\gamma}$. We may now build a new arc $\vec{\gamma}'$

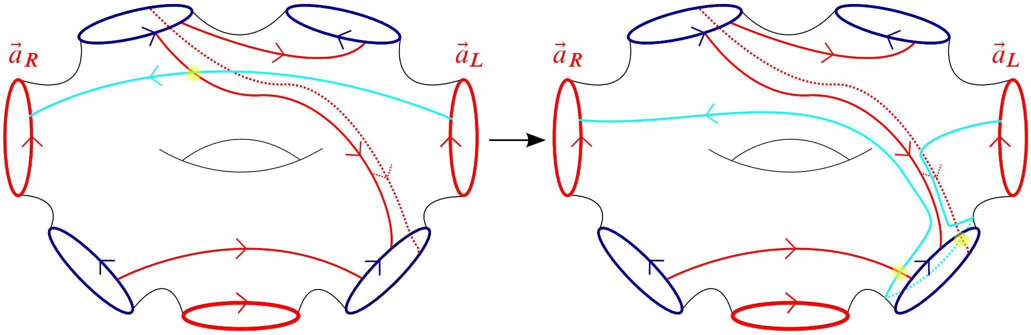


Figure 2: Replacing a noncoherent intersection with coherent ones.

obtained by following $\vec{\gamma}$ from $\vec{a}_L$ to p_1, then traveling along $\mathfrak{a}$ from p_1 to p_N, then following $\vec{\gamma}$ from p_N to $\vec{a}_R$. This arc has the same endpoints as $\vec{\gamma}$ but intersects $\mathfrak{a}$ at most once; if it is disjoint or crosses $\mathfrak{a}$ from right to left, then we are done.

Otherwise, $\vec{\gamma}'$ crosses $\mathfrak{a}$ from left to right exactly once. In this case, let ε denote the subarc of $\mathfrak{a}$ running from $\mathfrak{a} \cap \gamma'$ to the boundary curve $\vec{b}$ containing the terminal endpoint of $\mathfrak{a}$. Then we may take the connect sum of $\vec{\gamma}'$ and $\vec{b}$ along ε, adopting the orientations of $\vec{\gamma}'$ and of $\vec{b}$. This procedure has the effect of wrapping $\vec{\gamma}'$ around $\vec{b}$, thereby removing the left-to-right intersection of $\vec{\gamma}'$ with $\mathfrak{a}$. While the connected sum may introduce new intersections with $\vec{\alpha}|_W$, the coherence of $\vec{\alpha}$ and $\vec{\beta}$ guarantees that $\vec{\gamma}'$ crosses from right to left at these new intersections. See Figure 2.

In either case, we have built an arc with the same endpoints as $\vec{\gamma}$ with at least one fewer left-to-right intersection with $\vec{\alpha}|_W$. Repeating this procedure, we are left with an arc on W from $\vec{a}_L$ to $\vec{a}_R$ that crosses arcs of $\vec{\alpha}|_W$ only from right to left. $\square$

We now show that we can piece together these arcs coherently.

Lemma 3.7 *Let $(\vec{\alpha}, \vec{\beta})$ be a coherent pair of oriented multicurves satisfying $(\star)$. Given a singleton $\vec{a} \in \mathbb{S}_{\vec{\beta}}(\vec{\alpha})$, there is an oriented curve $\vec{b}$ that meets $\vec{a}$, is disjoint from $\vec{\beta}$, and whose intersections $\vec{\alpha}$ are all positive.*

Proof If $\vec{a}$ is nonseparating on $S \setminus \beta$, then a standard change-of-coordinates argument (see, for example, [4, Section 1.3.3]) implies that there must be an unoriented curve $b \subset S \setminus \beta$ meeting a exactly once, and we can choose its orientation to satisfy coherence.

Otherwise, let $W_1, \dots, W_n$ denote the complementary subsurfaces of $S \setminus (\vec{\beta} \cup \mathbb{S}_{\vec{\beta}}(\vec{\alpha}))$. Build a directed graph G whose vertices are the subsurfaces W_i and so that there is an edge from W_i to W_j if W_i is on the left and W_j is on the right of some curve $\vec{a}' \in \mathbb{S}_{\vec{\beta}}(\vec{\alpha})$. Connected components of this graph correspond to components of $S \setminus \beta$. See Figure 3.

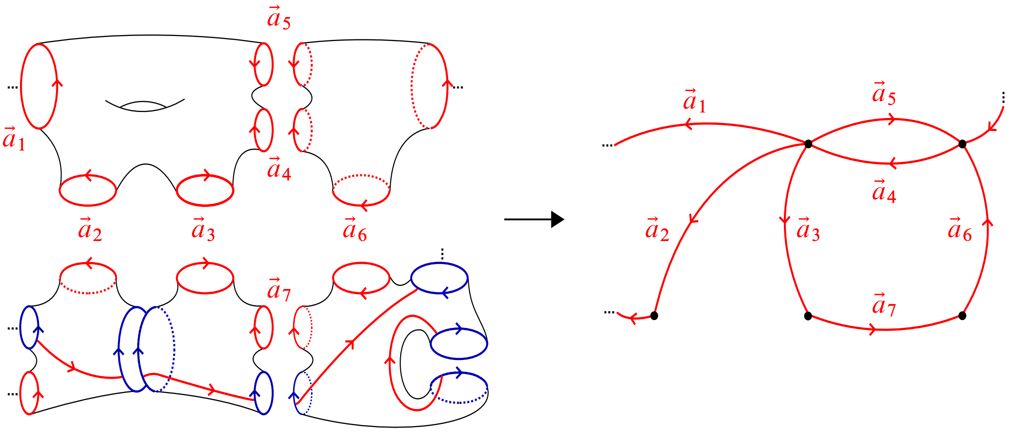


Figure 3: Building a graph out of components of $S \setminus (\vec{\beta} \cup \mathbb{S}_{\vec{\beta}}(\vec{\alpha}))$.

Claim 3.8 *Each connected component G_0 of G is strongly connected, ie, there is a directed path between any two vertices of G_0.*

Proof We prove an equivalent definition of strong connectivity: for any edge cut of G_0 (ie, any partition of its vertices into two sets), there are directed edges traveling from one set of the partition to the other and vice versa.

Consider an arbitrary partition

$$V(G_0) = V_1 \sqcup V_2$$

of the vertices of G_0. Let E_{12} denote set of oriented edges running from V_1 to V_2 and E_{21} the set of oriented edges from V_2 to V_1. Let $\vec{\alpha}_E \subset \mathbb{S}_{\vec{\beta}}(\vec{\alpha})$ be the oriented multicurve corresponding to $E = E_{12} \cup E_{21}$.

Now choose any component W_E of $S \setminus (\vec{\beta} \cup \vec{\alpha}_E)$; note that W_E is necessarily built by gluing subsurfaces corresponding to vertices of either V_1 or V_2, but not both. The components of ∂W_E are oriented depending on whether they correspond to edges from E_{12} or E_{21}. For example, if W_E consists of a union of subsurfaces corresponding to vertices of V_1, then the curves of ∂W_E arising from E_{12} are oriented with W_E on their left.

Because $(\vec{\alpha}, \vec{\beta})$ satisfies $(\star)$, if the boundary of a component of $S \setminus (\vec{\beta} \cup \vec{\alpha}_E)$ contains a curve of $\vec{\alpha}_E$ which has the component lying to its left (right), it also contains a curve of $\vec{\alpha}_E$ which has the component lying to its right (left). We therefore know that both E_{12} and E_{21} must be nonempty, hence there are edges running from V_1 to V_2 as well as edges running from V_2 to V_1. Since our partition was arbitrary, this allows us to deduce strong connectivity of G_0. □

Consider now the edge of G corresponding to our chosen singleton $\vec{a}$ and let W_i and W_t denote its initial and terminal vertices. Because each component of G is strongly connected, there is a directed (simple) path from W_t to W_i. This path corresponds to a sequence of subsurfaces

$$W_t = W_1, W_2, \dots, W_N = W_i$$

where W_j lies to the left and W_{j+1} to the right of some curve $\vec{a}_j \in \mathbb{S}_{\vec{\beta}}(\vec{\alpha})$.

On each W_j, use Lemma 3.6 to choose an oriented arc connecting $\vec{a}_j$ to $\vec{a}_{j+1}$ and crossing $\vec{\alpha}$ from right to left (where indices are interpreted mod N). We may then concatenate these arcs, possibly with a partial twist around the curves of $\vec{a}_j$ to ensure that endpoints match up. This yields a curve which crosses each $\vec{a}_j$ (in particular, crosses $\vec{a}$), is disjoint from $\vec{\beta}$, and each of its intersections with $\vec{\alpha}$ is positive. □

Iterating this lemma lets us remove the singletons of $\vec{\alpha}$ by adding curves to $\vec{\beta}$, none of which are themselves singletons. Then, we can swap the roles of $\vec{\alpha}$ and $\vec{\beta}$ to remove the singletons of $\vec{\beta}$ by adding curves to $\vec{\alpha}$. For later use, we record this as the following:

Proposition 3.9 *Let $(\vec{\alpha}, \vec{\beta})$ be a coherent pair of oriented multicurves satisfying $(\star)$. Then there are oriented multicurves $\vec{\alpha}' \supset \vec{\alpha}$ and $\vec{\beta}' \supset \vec{\beta}$ such that $(\vec{\alpha}', \vec{\beta}')$ is coherent, satisfies $(\star)$, and has no singletons. That is, every curve of $\vec{\alpha}'$ meets some curve of $\vec{\beta}'$ and vice versa.*

Moreover, if $\mathbb{S}_{\vec{\beta}}(\vec{\alpha}) = \varnothing$ then $\vec{\beta}'$ can be taken to equal $\vec{\beta}$ and if $\mathbb{S}_{\vec{\alpha}}(\vec{\beta}) = \varnothing$ then $\vec{\alpha}'$ can be taken to equal $\vec{\alpha}$.

Proof The only thing yet to prove is that if $\vec{b}$ is a curve obtained from Lemma 3.7 then the pair $(\vec{\alpha}, \vec{\beta} \cup \vec{b})$ still satisfies $(\star)$. This is also what allows us to iteratively apply Lemma 3.7.

So suppose that $\vec{\gamma} \subset \vec{\alpha} \cup \vec{\beta} \cup \{\vec{b}\}$ is a multicurve and W is a complementary subsurface of $S \setminus \vec{\gamma}$. Write

$$\partial W = A_L \sqcup A_R \sqcup B_L \sqcup B_R.$$

If none of the boundary components of W come from $\vec{b}$, then the conclusion of $(\star)$ follows because $(\vec{\alpha}, \vec{\beta})$ satisfies $(\star)$.

Otherwise, without loss of generality, we may assume that $\vec{b} \in B_R$. Now by our choice of $\vec{b}$, there is some curve $\vec{a}$ of $\vec{\alpha}$ crossing $\vec{b}$ from left to right. Since the pair $(\vec{\alpha}, \vec{\beta} \cup \vec{b})$ is coherent, and since $\vec{a}$ cannot cross either A_L or A_R, we see that when $\vec{a}$ exits W it must have done so by crossing a curve $\vec{b}'$. Coherence now implies that $\vec{b}' \in B_L$, and in particular B_L is nonempty. □

3.3 Extension to filling

We have reduced to the case where neither of our multicurves has any singletons with respect to the other; as a result, each component of $\vec{\alpha} \cup \vec{\beta}$ contains curves of both $\vec{\alpha}$ and $\vec{\beta}$. We may now extend the pair to fill the surface.

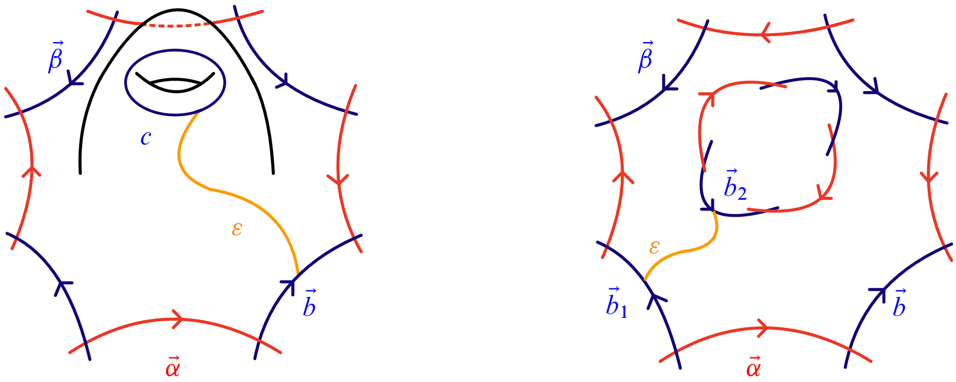


Figure 4: Adding curves to decrease the complexity of the complement of $\alpha \cup \beta$.

Lemma 3.10 *Suppose that* $(\vec{\alpha}, \vec{\beta})$ *is a coherent pair so that every curve of* $\vec{\alpha}$ *intersects a curve of* $\vec{\beta}$, *and vice versa. Then there exists a multicurve* $\vec{\beta}' \supset \vec{\beta}$ *so that* $(\vec{\alpha}, \vec{\beta}')$ *remains coherent and fills* S.

Proof Let Y denote the (possibly disconnected) subsurface obtained by removing $\vec{\alpha} \cup \vec{\beta}$. Our proof proceeds by iteratively adding curves to reduce the size of Y. The assumption that the curves of $\vec{\alpha}$ and $\vec{\beta}$ meet implies that each boundary component of Y is a concatenation of segments of both $\vec{\alpha}$ and $\vec{\beta}$. See Figure 4.

So long as Y is not an annulus or pair of pants, there is a nonboundary parallel simple closed curve c on Y. Pick an arbitrary segment of $\vec{\beta}$ that comprises ∂Y corresponding to a curve $\vec{b}$ of $\vec{\beta}$ and pick an arc ε connecting c to $\vec{b}$ (and otherwise disjoint from $\alpha \cup \beta \cup c$). We can then form the connect sum $\vec{b}' = \vec{b} +_{\varepsilon} c$ and orient it so that its orientation agrees with that of $\vec{b}$ as it is running parallel to $\vec{b}$. Since $\vec{b}$ crosses $\vec{\alpha}$ from right to left, so does $\vec{b}'$. Thus, the pair $(\vec{\alpha}, \vec{\beta} \cup \vec{b}')$ remains coherent and we see that the Euler characteristic of each component of $Y \setminus (c \cup \varepsilon)$ is strictly greater than that of Y. This follows because a neighborhood of $c \cup \varepsilon$ together with the relevant component of ∂Y is homeomorphic to a pair of pants.

In the case that Y is an annulus or pair of pants, all of its boundaries are concatenations of segments of $\vec{\alpha}$ and $\vec{\beta}$. The condition that $\vec{\alpha}$ and $\vec{\beta}$ are coherent further implies that each boundary component contains some segment of $\vec{\beta}$ that has Y on its right-hand side. Pick curves $\vec{b}_1$ and $\vec{b}_2$ (possibly equal) corresponding to these segments and ε an arc in Y connecting the segments. The connect sum $\vec{b}' = \vec{b}_1 +_{\varepsilon} \vec{b}_2$ is then naturally oriented and is coherent with $\vec{\alpha}$. In particular, $\vec{b}'$ has positive intersection number with $\vec{\alpha}$ and so is not nullhomotopic. The complement of $\vec{b}' \cap Y$ in Y is a disk and a surface homeomorphic to $Y \setminus \varepsilon$, so we see in these cases we can also add a curve to $\vec{\beta}$ and increase the Euler characteristic of each piece of the complement.

Therefore, we may iteratively add curves to $\vec{\beta}$ until every component of Y has Euler characteristic 1 (equivalently, until every component has no essential arcs), ie, until every component of Y is a disk. □

We can now put the pieces together to prove our main theorem.

Proof of Theorem 3.1 The necessity of ($\star$) was proven in Lemma 3.2.

To see that ($\star$) is sufficient, let $(\vec{\alpha}, \vec{\beta})$ be a coherent pair satisfying ($\star$). By Lemma 3.7, we may extend to a coherent pair $(\vec{\alpha}', \vec{\beta}')$ without any singleton curves. Applying Lemma 3.10, we can then find a $\vec{\beta}'' \supset \vec{\beta}'$ so that $(\vec{\alpha}', \vec{\beta}'')$ is a coherent filling pair. Applying the Thurston–Veech construction (Construction 2.6) to $(\vec{\alpha}', \vec{\beta}'')$, we obtain a square-tiled surface on which the curves of $\vec{\alpha}'$ (hence those of $\vec{\alpha}$) are horizontal cylinders and the curves of $\vec{\beta}''$ (hence those of $\vec{\beta}$) are vertical cylinders. □

Analyzing the steps in our proof, we can similarly give a criterion for when an oriented multicurve can be realized as the horizontal foliation of a translation surface, ie, realized as a horizontal multicylinder which covers all but a measure zero set of the translation surface. Thus, any horizontal trajectory beginning at any point away from a measure zero set of that translation surfaces closes up to become a core curve of the horizontal multicylinder. Consider β in Figure 1 for an example.

Corollary 3.11 *If $\vec{\alpha}$ is an oriented multicurve, there is a horizontally periodic translation surface (X, ω) with directional foliation $\vec{\alpha}$ (oriented in the $+x$-direction) if and only if for each $\vec{\gamma} \subseteq \vec{\alpha}$ and each complementary subsurface W of $S \setminus \vec{\gamma}$, we have that*

$$A_L \neq \varnothing \quad \text{and} \quad A_R \neq \varnothing,$$

where A_L (A_R) denotes the boundary components of W arising from left- (right-)hand sides of curves of $\vec{\gamma}$.

Proof The hypothesis allows us to apply Theorem 3.1 with $\vec{\beta}$ empty. The "only if" part of this statement is immediate. For the "if" part, we need to check that no curves are added to $\vec{\alpha}$ during the extension process. Since $\vec{\beta}$ is empty, it does not include any singletons, so neither Lemma 3.7 nor Lemma 3.10 add any new curves to $\vec{\alpha}$. Thus, we obtain a multicurve $\vec{\beta}$ such that $\vec{\alpha}$ and $\vec{\beta}$ are coherent and filling, hence $\mathrm{TV}(\vec{\alpha}, \vec{\beta})$ gives a translation surface on which $\vec{\alpha}$ constitutes the entire horizontal foliation. □

By assigning orientations, we can also use Theorem 3.1 to deduce our result about unoriented multicurves, which was stated in the Introduction as Theorem 1.1.

Corollary 3.12 *If α and β are a pair of (unoriented) multicurves on S with no curves in common, then they can be realized as a pair of parallel multicylinders if and only if the following hold:*

1. *The multicurves α and β are coherently orientable.*
2. *No single component of α separates $S \setminus \beta$.*
3. *No single component of β separates $S \setminus \alpha$.*

Proof Lemma 2.2 and Corollary 3.3 together prove that these conditions are necessary.

To prove that they are sufficient, let α and β be a pair of unoriented multicurves satisfying the conditions. Let $\alpha_0 = \alpha \setminus \mathbb{S}_\beta(\alpha)$ denote the nonsingletons of α and let $\beta_0 \subset \beta$ be defined similarly. By (1), we can orient (α_0, β_0) such that $(\vec{\alpha}_0, \vec{\beta}_0)$ are coherent. It remains to orient the singletons so that ($\star$) holds.

As in the proof of Lemma 3.7, consider the dual (undirected) graph G of the multicurve $\mathbb{S}_\beta(\alpha)$ on $S \setminus \beta$. By (2), each component G_0 of G is 2-edge-connected, that is, removing any edge does not separate G_0. Thus, by Robbins' theorem there exists a choice of orientation for the edges of G_0 so that the resulting directed graph is strongly connected. Choose one such orientation for each G_0, and orient each $a \in \mathbb{S}_\beta(\alpha)$ so that there is an edge from $W_i \subset S \setminus (\beta \cup \mathbb{S}_\beta(\alpha))$ to W_j if W_i is on the left and W_j lies on the right of $\vec{a}$. Orient the singletons of β similarly.

Now let $\vec{\gamma} \subseteq \vec{\alpha} \cup \vec{\beta}$ and let W be a complementary subsurface of $S \setminus \vec{\gamma}$. Let us prove that $A_L \neq \varnothing$ if and only if $A_R \neq \varnothing$; the proof for B_L and B_R is completely analogous. If $\vec{\gamma} \cap \vec{\alpha}_0$ is nonempty, then W must meet some curve of $\vec{\beta}_0$. Consider an arc of $\vec{\beta}_0|_W$; since $\vec{\beta}_0$ always crosses $\vec{\alpha}_0$ from right to left, this means that it must enter W when it meets a curve of A_L and leave W when it meets a curve of A_R. Thus A_L and A_R must both be nonempty.

Otherwise, suppose that $\vec{\gamma} \cap \vec{\alpha}_0 = \varnothing$. Of course, if $\vec{\gamma} \cap \mathbb{S}_{\vec{\beta}}(\vec{\alpha})$ is also empty then we are done, so assume there is some boundary component of W that is a singleton of α. In this case, then we see that W is a union of pieces of $S \setminus (\beta \cup \mathbb{S}_{\vec{\beta}}(\alpha))$, ie, it is a component of a cut of G_0. But now since G_0 is strongly connected, there are edges both entering and exiting this component, hence both A_L and A_R are nonempty.

Therefore, we have shown that we can orient α and β so that ($\star$) holds; applying Theorem 3.1 completes the proof of the corollary. □

4 Coherence and grafting

In the previous section, we characterized when two (un)oriented multicurves could be realized simultaneously as (parallel) directional multicylinders. In this one, we relax this requirement, giving necessary and sufficient conditions to simultaneously realize two multicurves as multicylinders with just one being (parallel) directional. As in the previous section, we prove a stronger statement for multicurves with prescribed orientations; compare Theorem 4.8 below.

See Figure 5 for an example showing that the conditions of Theorem 1.1 are too strong if one only requires a single multicylinder to be parallel.

As in the previous section, we will fix the convention that given two coherent multicurves $\vec{\alpha}$ and $\vec{\beta}$, the curves of $\vec{\beta}$ cross the curves of $\vec{\alpha}$ from right to left.

4.1 Geodesics on translation surfaces

We begin by recording a simple lemma about how cylinders intersect other geodesics on translation surfaces. We recall that if a curve b is *not* realized as a cylinder on (X, ω), then it has a unique geodesic representative which is a concatenation of saddle connections.

One of the handy things about cylinders is that their core curves are always in minimal position with respect to other geodesics.

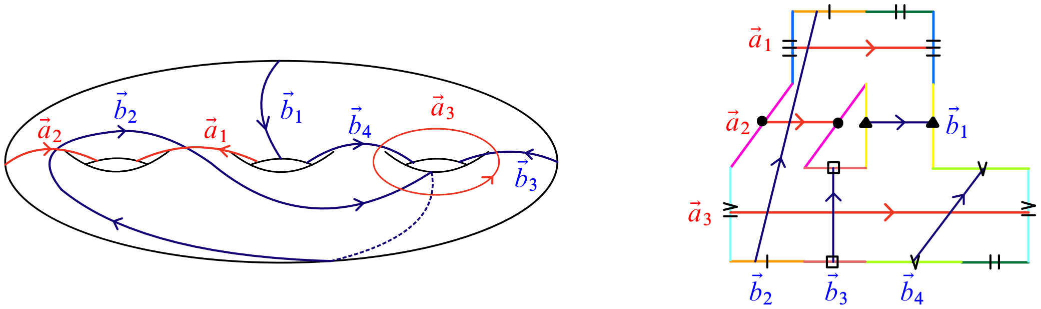


Figure 5: A pair of coherent multicurves that cannot be jointly realized as parallel multicylinders, together with a realization with one of the multicylinders parallel. Observe that even if one repartitions and adds the singleton of β to α, the pair is still not realizable as a pair of parallel multicylinders.

Lemma 4.1 *Let a be a nonsingular core curve of a cylinder on some translation surface (X,ω) and let b be the geodesic representative of a noncylinder curve on (X,ω). Then a and b are in minimal position.*

In particular, if the geometric intersection number of the isotopy classes of a and b is 0, then the actual geodesics a and b are disjoint.

Proof Because a is the core curve of a cylinder, it contains no saddle connections. The curve b is a concatenation of saddle connections, so a and b are transverse.

Assume for contradiction that a and b are not in minimal position. Then a and b bound a bigon [4, Proposition 1.7]. Consider the two arcs $c_a \subset a$ and $c_b \subset b$ constituting edges of this bigon; since a and b are transverse, $c_a \neq c_b$. Now c_a and c_b are isotopic rel endpoints, so the surgered curve $(b \setminus c_b) \cup c_a$ is isotopic to b. Because b is the unique shortest representative of its isotopy class, c_b must be strictly shorter than c_a. However, this implies that the curve $(a \setminus c_a) \cup c_b$ is isotopic to and shorter than a, which is a contradiction. Thus a and b are in minimal position. $\square$

Remark 4.2 Lemma 4.1 is not true for all pairs of geodesics on (X,ω). For example, there may be curves a and b which have geometric intersection number 0 but whose geodesic representatives share a saddle connection. Compare with the discussion in [9, Section 3].

If $I \subset [0,2\pi)$ is an interval (with any mix of open and closed conditions at its endpoints), then we say that $\vec{s}$ is an *I-saddle connection* if $\arg(\mathrm{hol}(\vec{s})) \in I$.

We now state and prove an analogue of Lemma 2.2 for noncylinder curves.

Corollary 4.3 *Suppose that (X,ω) is a horizontally periodic translation surface and let $\vec{\alpha}$ denote the core curves of the horizontal cylinders, oriented in the $+x$ direction. Let $\vec{b}$ be any oriented curve coherent with $\vec{\alpha}$ and suppose that $\vec{b}$ is not realized as a cylinder on (X,ω). Then the geodesic representative of $\vec{b}$ is a concatenation of $[0,\pi]$ saddle connections.*

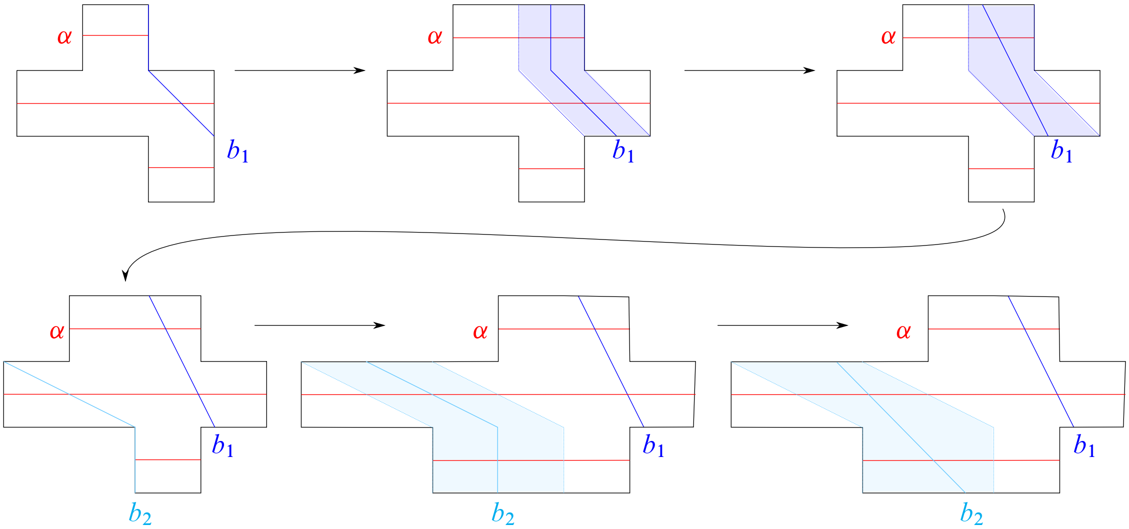


Figure 6: Example of grafting along two curves b_1 and b_2 to realize them as cylinders on a translation surface where α is realized as horizontal cylinders.

Proof Lemma 4.1 says $\vec{\alpha}$ and $\vec{b}$ realized on (X, ω) are in minimal position, so there are no bigons. Thus, the algebraic and geometric intersection numbers of the isotopy classes of $\vec{\alpha}$ and $\vec{b}$ agree with those of their geodesic representatives. Since (X, ω) is horizontally periodic, $\vec{\alpha}$ meets any nonhorizontal saddle connection. Thus, since $\vec{\alpha}$ and $\vec{b}$ are coherent, any nonhorizontal saddle connection comprising $\vec{b}$ must be a $(0, \pi)$-saddle connection. □

4.2 Flat grafting

For the remainder of this section, when we refer to a curve on (X, ω), we mean its geodesic representative.

Given a pair of coherent multicurves $(\vec{\alpha}, \vec{\beta})$, our strategy to realize them as multicylinders is to find a translation surface (X, ω) on which $\vec{\alpha}$ is a horizontal multicylinder, and then deform (X, ω) to make each curve of $\vec{\beta}$ into a cylinder while ensuring $\vec{\alpha}$ remains horizontal. A deformation that accomplishes the second step is *horizontal grafting*, which was introduced in [5]. We direct the reader to that paper for a more thorough discussion of this procedure.

Construction 4.4 (horizontal grafting) Let (X, ω) be a translation surface and suppose that c is a geodesic curve on (X, ω) not isotopic to the core curve of any cylinder. Let $\{s_1, \dots, s_l\}$ denote the saddle connections appearing in c, counted without multiplicity. Cutting (X, ω) along c results in a translation surface with piecewise-geodesic boundary, two geodesic segments for each s_i.

Suppose for the moment that no s_i is horizontal. Then the *grafting* of (X, ω) along c (by some distance t) is obtained by gluing each copy of s_i in $X \setminus c$ to a parallelogram with sides $\langle t, 0\rangle$ and s_i, then gluing the horizontal sides of the parallelograms together according to how c runs along the s_i. See Figure 6.

If some s_i is horizontal, then there are two different choices for how to define the grafting of (X,ω) along c. In either case, the grafted surface is obtained as a limit of rotating (X,ω) (either counterclockwise or clockwise) by a small bit, grafting, and then rotating back. This can be thought of as "shearing" the surface along that saddle and identifying segments of s_i with the horizontal edges of nonhorizontal parallelograms.

Remark 4.5 When (X,ω) has a marking $\varphi\colon S\to X$, the grafted surface (X',ω') inherits a marking $\varphi'\colon S\to X'$ since the grafting procedure takes place entirely in a neighborhood of c. However, depending on the angle that c makes at each cone point, grafting may split apart cone points of (X,ω), changing which stratum it lives in while preserving the genus of the surface. Therefore, there is no canonical way to mark both the translation surface *and its zeros* that is consistent under grafting. This is one of the reasons that we have decided to focus on the realizability of curves on a closed surface as cylinders, as opposed to considering curves relative to zeros.

We record below a number of properties of grafting, proofs of which can be found in [5].

Lemma 4.6 *Let all notation be as above. Orient c arbitrarily, inducing an orientation on each s_i.*

1. *If c is a concatenation of I-saddle connections, where I is a proper closed subinterval of $[0,\pi]$, then there is a choice of horizontal grafting such that the grafted surface is a translation surface.*
2. *If c contains at least one nonhorizontal saddle connection, then there exists some distance t such that c is realized as a cylinder on the grafting of (X,ω) along c by t.*

Disjointness of each curve in $\vec{\beta}$ allows us to graft along each curve in $\vec{\beta}$ without interfering with the cylindrical property of the previously grafted curves in $\vec{\beta}$. This is because grafting is a local procedure.

Lemma 4.7 *Let b_1 be a core curve of a cylinder and b_2 a noncylinder curve on some translation surface (X,ω). If $\iota(b_1,b_2)=0$, then b_1 is realized as a cylinder with the same slope as before on any surface obtained by grafting along b_2.*

Proof By Lemma 4.1, b_1 and b_2 are in minimal position. Hence, since $\iota(b_1,b_2)=0$, their geodesic representatives do not intersect. Recall that when we graft horizontally along b_2, we cut along b_2 and glue in a parallelogram p_i adjacent to each saddle connection s_i on b_2. In particular, $(X,\omega)\setminus b_2$ is isometric to the complement of the p_i's in the grafted surface. Thus, b_1 remains a cylinder on any grafting of b_2. □

Because curves in $\vec{\beta}$ are disjoint, we can now take the translation surface which realizes $\vec{\alpha}$ as the horizontal foliation and graft horizontally along each curve of $\vec{\beta}$ until $(\vec{\alpha},\vec{\beta})$ are multicylinders. This is shown in the backwards direction of Theorem 4.8.

In what follows, a triple of oriented multicurves $(\vec{\gamma}_1,\vec{\gamma}_2,\vec{\gamma}_3)$ is coherent if every pair $(\vec{\gamma}_i,\vec{\gamma}_j)$ is coherent.

Theorem 4.8 *Let $(\vec{\alpha}, \vec{\beta})$ be a pair of multicurves on surface S which contains no separating curves. Then $(\vec{\alpha}, \vec{\beta})$ can be realized as a pair of multicylinders with $\vec{\alpha}$ a directional multicylinder on some translation surface if and only if there exists a filling pair $(\vec{\alpha}', \vec{\gamma})$ such that $\vec{\alpha} \subset \vec{\alpha}'$ and $(\vec{\alpha}', \vec{\gamma}, \vec{b})$ is coherent for each $\vec{b} \subset \vec{\beta}$.*

Proof Consider some translation surface which realizes $\vec{\alpha}$ as a directional multicylinder and $\vec{\beta}$ as an arbitrary multicylinder. Postcomposing with an element of $\mathsf{GL}_2(\mathbb{R})$ as necessary, we may assume that each curve of $\vec{\alpha} = \vec{a}_1 \cup \cdots \cup \vec{a}_m$ is horizontal and oriented in the $+x$-direction. Moreover, we can shear the surface enough so that the angle of the holonomy of each curve of $\vec{\beta}$ lies in the set $\left[0, \frac{\pi}{2}\right) \cup \left[\pi, \frac{3\pi}{2}\right)$. Call this new surface (X, ω).

Suppose the curves of $\vec{\beta}$ which are core curves of horizontal cylinders on (X, ω) are $\vec{h}_1, \dots, \vec{h}_k$. As in Lemma 2.9, we can choose a local period coordinate chart for the ambient stratum and cut out a nonzero $\mathbb{R}$-linear subspace V of $H^1(X, \mathrm{Zeros}(\omega); \mathbb{C})$ by stipulating that

$$\mathrm{Im}(\mathrm{hol}(\vec{a}_i)) = 0 \quad \text{and} \quad \mathrm{Im}(\mathrm{hol}(\vec{h}_j)) = 0$$

for all $i = 1, \dots, m$ and $j = 1, \dots, k$. Recall by Lemma 2.5 there is some relatively open subset $U \subset V$ containing (X, ω) in the stratum in which each element of U realizes $(\vec{\alpha}, \vec{\beta})$ as a pair of multicylinders. Furthermore, U can be chosen sufficiently small enough so that $\vec{\beta}$ still points in direction $\left[0, \frac{\pi}{2}\right) \cup \left[\pi, \frac{3\pi}{2}\right)$ on every surface. Since V is cut out by integral-linear equations, rational points $H^1(X, \mathrm{Zeros}(\omega); \mathbb{Q} \oplus i\mathbb{Q})$ are dense in V and so we can find some square-tiled surface

$$(X', \omega') \in H^1(X, \mathrm{Zeros}(\omega); \mathbb{Q} \oplus i\mathbb{Q}) \cap U.$$

Since square-tiled surfaces are vertically and horizontally periodic, we may consider the filling pair $(\vec{\alpha}', \vec{\gamma})$ where $\vec{\alpha}'$ are the horizontal core curves (oriented in the $+x$ direction) and $\vec{\gamma}$ are the vertical core curves (oriented in the $+y$ direction). In particular, $\vec{\alpha} \subset \vec{\alpha}'$, and each curve $\vec{b} \in \vec{\beta}$ is coherent with $\vec{\alpha}'$ and $\vec{\gamma}$. (Note that $\vec{\alpha}'$ contains any $\vec{h}_j$'s and some curves of $\vec{b}$ may still be disjoint from $\vec{\alpha}'$.)

Conversely, assume such a pair $(\vec{\alpha}', \vec{\gamma})$ exists. If some curves of $\vec{\beta}$ do not intersect the curves of $\vec{\alpha}'$, then we can repartition the pair

$$\vec{\alpha}'' := \vec{\alpha}' \cup \mathbb{S}_{\vec{\alpha}'}(\vec{\beta}) \text{ and } \vec{\beta}' := \vec{\beta} - \mathbb{S}_{\vec{\alpha}'}(\vec{\beta})$$

so that every curve in $\vec{\beta}'$ intersects $\vec{\alpha}''$. Note that though $(\vec{\alpha}'', \vec{\gamma})$ may not be coherent, eg, when $\vec{b}_{i_j}$ intersects $\vec{\gamma}$ in the opposite direction as $\vec{\alpha}'$ does, the pair is still coherently orientable, as we can simply flip the orientations on each $\vec{b}_{i_k}$.

By Construction 2.6, there is some square-tiled surface (X, ω) that realizes $\vec{\alpha}''$ and $\vec{\gamma}$ as its horizontal and vertical cylinders, respectively. Moreover, we may assume $\vec{\alpha}'$ points in the $+x$ direction and $\vec{\gamma}$ the $+y$ direction. Coherence with $(\vec{\alpha}', \vec{\gamma})$ combined with disjointness from $\vec{\alpha}'' - \vec{\alpha}'$ ensures that each curve $\vec{b} \subset \vec{\beta}'$ is coherent with $\vec{\alpha}''$ and $\vec{\gamma}$. Corollary 4.3 applied using $\vec{\alpha}''$ as the horizontal cylinders, followed by rotating the surface by $\pi/2$ and reapplying the corollary, implies that the geodesic representatives of $\vec{b}$ on (X,ω) are concatenations of either only $\left[0,\frac{\pi}{2}\right]$-, $\left[\frac{\pi}{2},\pi\right]$-, $\left[\pi,\frac{3\pi}{2}\right]$-, or $\left[\frac{3\pi}{2},2\pi\right]$-saddle connections (or a cylinder pointing in one of those intervals). Since these are proper subintervals of $[0,\pi]$, Lemma 4.6(1) grants us that grafting along $\vec{b}$ will yield a translation surface.

Suppose that $\vec{\beta}' = \vec{b}_1 \cup \cdots \cup \vec{b}_m$. We will first show $(\vec{\alpha}', \vec{b}_1)$ can be simultaneously realized as cylinders on some translation surface using horizontal grafting (see Construction 4.4). If this is already the case on (X,ω), we then proceed to $(\vec{\alpha}, \vec{b}_1 \cup \vec{b}_2)$. Because no curve in $\vec{\beta}'$ is disjoint from $\vec{\alpha}''$, all curves of $\vec{\beta}'$ have a nonhorizontal saddle connection. By Lemma 4.6(2), we can horizontally graft along $\vec{b}_1$ enough to realize it as a cylinder on some other translation surface (X_1,ω_1). Because horizontal grafting preserves the horizontal foliation, $\vec{\alpha}''$ will remain as the complete set of horizontal cylinders on (X_1,ω_1).

Next, we will realize $\vec{b}_2$ on (X_1,ω_1) and repeat the above procedure. Assume $\vec{b}_2$ is not the core of a cylinder, or else we can move onto $\vec{b}_3$. By Lemma 4.1, $\vec{b}_2$ is disjoint from $\vec{b}_1$ because $\iota(\vec{b}_1,\vec{b}_2) = 0$ and $\vec{b}_1$ and $\vec{b}_2$ are in minimal position. We note that the vertical foliation of the grafted surface is no longer $\vec{\gamma}$. Nonetheless, $\vec{b}_1$ and $\vec{b}_2$ did not intersect transversely on (X,ω) and (X,ω) is isometric to (X_1,ω_1) away from $\vec{b}_1$ and the inserted parallelograms. Hence, the saddle connections comprising $\vec{b}_2$ on (X_1,ω_1) all point in the same directions as they did on (X,ω): in particular, they all lie in a proper subinterval of $[0,\pi]$. Again, we can graft horizontally along $\vec{b}_2$ enough to obtain a translation surface (X_2,ω_2) that realizes $\vec{b}_2$ as the core of a cylinder. Moreover, by Lemma 4.7, $\vec{b}_1$ is still the core of a cylinder on (X_2,ω_2), and $\vec{\alpha}''$ remains as the cores of the complete set of horizontal cylinders.

We continue iterating this process for each noncylinder curve in $\vec{\beta}'$ to realize $\vec{\beta}$ as a multicylinder on some translation surface (X_N,ω_N) where $\vec{\alpha}$ is a directional multicylinder. □

5 Pairwise coherence

Recall that two oriented multicurves $\vec{\alpha}$ and $\vec{\beta}$ on S are said to be *pairwise coherent* if for any pair of curves $\vec{a_i} \subset \vec{\alpha}$ and $\vec{b_j} \subset \vec{\beta}$, the pair $(\vec{a_i}, \vec{b_j})$ is coherent. Lemma 2.2 implies that pairwise coherence is necessary for two multicurves to be simultaneously realizable as multicylinders.

In this section, we provide two examples of a pair of pairwise coherent, but not coherently orientable, multicurves $\vec{\alpha}$ and $\vec{\beta}$ on some S. Our first, Example 5.1, is an instance where $\vec{\alpha}$ and $\vec{\beta}$ can be simultaneously realized as a pair of multicylinders on some translation surface, whereas in Example 5.2, they cannot.

Example 5.1 Consider the multicurves $\vec{\alpha} = \vec{a}_1 \cup \vec{a}_2$ and $\vec{\beta} = \vec{b}_1 \cup \vec{b}_2$ on S_2 as shown to the right in Figure 7. The picture to the left is a translation surfaces which realizes $\vec{\alpha}$ and $\vec{\beta}$ as cylinders. However, by observation of all possible assignments of orientation to $\vec{a}_1$, $\vec{a}_2$, $\vec{b}_1$, and $\vec{b}_2$, we see there is no combination of orientations that makes $\vec{\alpha}$ and $\vec{\beta}$ coherent.

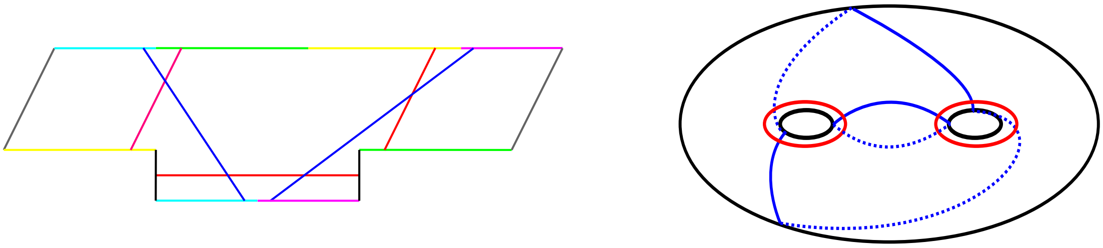
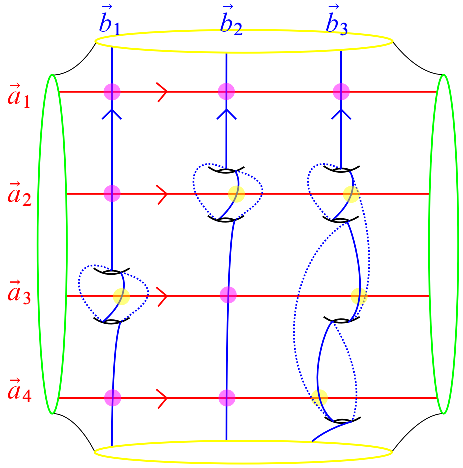


Figure 7: A set of multicurves which are not coherently orientable are still simultaneously realizable as cylinders on this translation surface. Left: polygonal representation. Right: topological representation.

Example 5.2 Next, we provide an example of two pairwise coherent multicurves that cannot be realized as a pair of multicylinders on any translation surface. Let multicurves $\vec{\alpha} = \vec{a}_1 \cup \vec{a}_2 \cup \vec{a}_3 \cup \vec{a}_4$ and $\vec{\beta} = \vec{b}_1 \cup \vec{b}_2 \cup \vec{b}_3$ on S_3 be as in Figure 8. These multicurves are pairwise coherent but are not coherently orientable.

Suppose by contradiction there is some translation surface (X, ω) that realizes $\vec{\alpha}$ and $\vec{\beta}$ as cylinders. Thus, there exists an assignment of angles, $\{\angle a_1, \angle a_2, \angle a_3, \angle a_4, \angle b_1, \angle b_2, \angle b_3\}$ relative to the positive x-axis recording the angles of the cylinders. Furthermore, the direction in which $\vec{a}_i$ contacts $\vec{b}_j$ agrees with the algebraic intersection numbers on S_3 shown in Figure 8. As we shall see, it is impossible to find any such collection of angles $\{\angle a_1, \angle a_2, \angle a_3, \angle a_4, \angle b_1, \angle b_2, \angle b_3\}$ that agrees with the prescribed intersection numbers.

When assigning angles to curves, without loss of generality, we can choose any one curve to begin with and give it any angle. So, let us set $\angle b_1 = 0$. Then $\angle a_1 \in (\pi, 2\pi)$ and $\angle a_2 \in (\pi, 2\pi)$ because $\hat{\imath}(a_1, b_1) > 0$ and $\hat{\imath}(a_2, b_1) > 0$. We note that $\angle a_1$ cannot equal $\angle a_2$, for if it did then $\vec{b}_2$ could not have



	$\vec{b}_1$	$\vec{b}_2$	$\vec{b}_3$
$\vec{a}_1$	$+$	$+$	$+$
$\vec{a}_2$	$+$	$-$	$-$
$\vec{a}_3$	$-$	$+$	$-$
$\vec{a}_4$	$+$	$+$	$-$

Figure 8: In the table to the right, the intersection pattern between the two multicurves is given by $\hat{\imath}(\vec{a}_i, \vec{b}_j)$, where $+$ indicates $\hat{\imath}(\vec{a}_i, \vec{b}_j) > 0$ and $-$ indicates $\hat{\imath}(\vec{a}_i, \vec{b}_j) < 0$.

constant slope on (X,ω) because it is not coherent with the foliation in the direction of $\angle a_1 = \angle a_2$. Thus, we will break up the rest of this proof into two cases, $\angle a_1 > \angle a_2$ and $\angle a_1 < \angle a_2$, each with two subcases; see Figure 9. See Figure 8 for illustrations.

(1) We begin with the assumption that $\angle a_1 > \angle a_2$.

In this case, $\hat{\imath}(\vec{a}_1,\vec{b}_2) > 0$ tells us that $\angle b_2 \in \bigl((\angle a_1, 2\pi) \cup [0, \angle a_1 - \pi)\bigr)$ and $\hat{\imath}(\vec{a}_2,\vec{b}_2) < 0$ means that $\angle b_2 \in (\angle a_2 - \pi, \angle a_2)$. Together, we have that

$$\angle b_2 \in \bigl((\angle a_1, 2\pi) \cup [0, \angle a_1 - \pi)\bigr) \cap (\angle a_2 - \pi, \angle a_2),$$

which simplifies to

$$\angle b_2 \in (\angle a_2 - \pi, \angle a_1 - \pi).$$

Since $\vec{b}_3$ has the same algebraic intersections with $\vec{a}_1$ and $\vec{a}_2$, we get the same conclusion for $\angle b_3$.

We can rule out the case that $\angle b_2 = \angle b_3$ as above, as this would imply that $\vec{a}_3$ would not be coherent with the foliation in the direction of $\angle b_2$ and $\angle b_3$. Since $\angle b_2$ and $\angle b_3$ lie in the same interval, we have the following two subcases to examine:

(a) Suppose first that $\angle b_3 > \angle b_2$. Then, because $\hat{\imath}(\vec{a}_3,\vec{b}_2) > 0$ and $\hat{\imath}(\vec{a}_3,\vec{b}_3) < 0$, we have that

$$\angle a_3 \in \bigl((\angle b_2 + \pi, 2\pi) \cup (0, \angle b_2)\bigr) \cap (\angle b_3, \angle b_3 + \pi),$$

which simplifies to

$$\angle a_3 \in (\angle b_2 + \pi, \angle b_3 + \pi) \subset (\pi, 2\pi).$$

However, this implies that $\hat{\imath}(\vec{a}_3,\vec{b}_1) > 0$, whereas the table in Figure 8 states that $\hat{\imath}(\vec{a}_3,\vec{b}_1) < 0$, resulting in a contradiction.

(b) Now, suppose $\angle b_3 < \angle b_2$. We have that $\hat{\imath}(\vec{a}_4,\vec{b}_2) > 0$ and $\hat{\imath}(\vec{a}_4,\vec{b}_3) < 0$. Therefore,

$$\angle a_4 \in \bigl((\angle b_2 + \pi, 2\pi) \cup (0, \angle b_2)\bigr) \cap (\angle b_3, \angle b_3 + \pi),$$

which simplifies to

$$\angle a_4 \in (\angle b_3, \angle b_2) \subset (0, \pi).$$

However, we have arrived at a contradiction again since this implies that $\hat{\imath}(\vec{a}_4,\vec{b}_1) < 0$, whereas Figure 8 states that $\hat{\imath}(\vec{a}_4,\vec{b}_1) > 0$.

(2) For the second case, we assume $\angle a_1 < \angle a_2$. Because $\hat{\imath}(\vec{a}_1,\vec{b}_2) > 0$ and $\hat{\imath}(\vec{a}_2,\vec{b}_2) < 0$, we see that

$$\angle b_2 \in \bigl((\angle a_1, 2\pi) \cup [0, \angle a_1 - \pi)\bigr) \cap (\angle a_2 - \pi, \angle a_2),$$

which simplifies to

$$\angle b_2 \in (\angle a_1, \angle a_2).$$

By identical reasoning using $\angle b_3$ in place of $\angle b_2$, we have that $\angle b_3$ lies in the same interval. We again split into two cases:

(a) If $\angle b_3 > \angle b_2$, then because $\hat{\imath}(\vec{a}_4,\vec{b}_2) > 0$ and $\hat{\imath}(\vec{a}_4,\vec{b}_3) < 0$, we have that

$$\angle a_4 \in (\angle b_2 - \pi, \angle b_2) \cap \bigl((\angle b_3, 2\pi) \cup (0, \angle b_3 - \pi)\bigr),$$

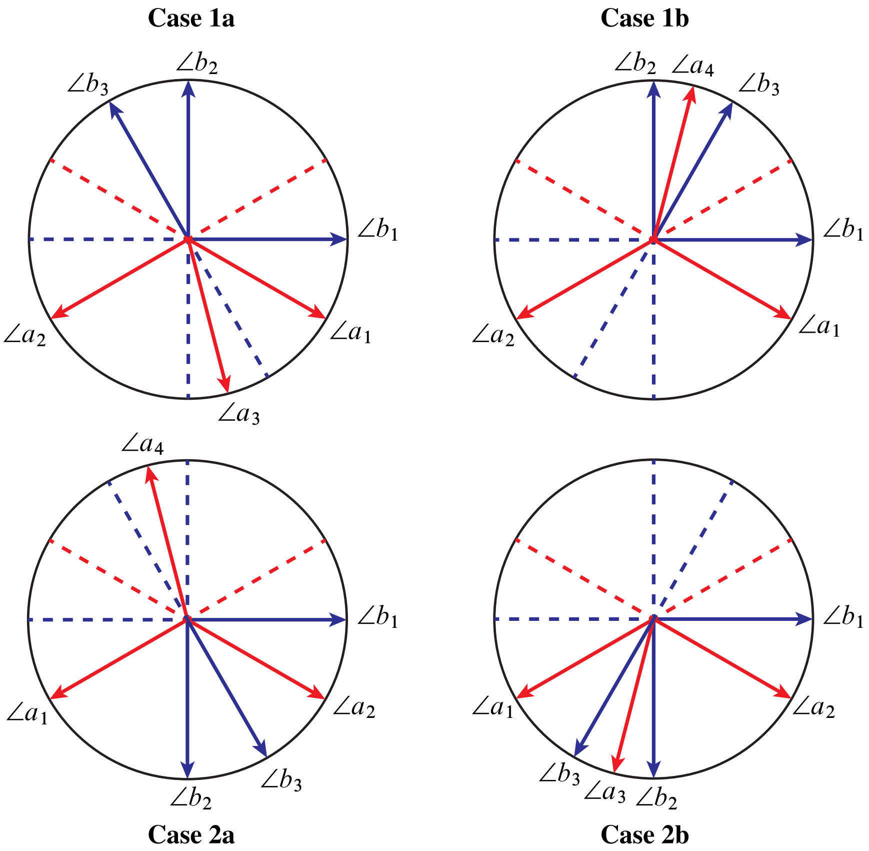


Figure 9: The ending states of each of the four cases in Example 5.2.

which simplifies to

$$\angle a_4 \in (\angle b_2 - \pi, \angle b_3 - \pi).$$

This implies that $\angle a_4 \in (0, \pi)$ and hence $\hat{\imath}(\vec{a}_4, \vec{b}_1) < 0$, but this contradicts Figure 8 where $\hat{\imath}(\vec{a}_4, \vec{b}_1) > 0$.

(b) If $\angle b_3 < \angle b_2$, then because $\hat{\imath}(\vec{a}_3, \vec{b}_2) > 0$ and $\hat{\imath}(\vec{a}_3, \vec{b}_3) < 0$, we have that

$$\angle a_3 \in (\angle b_2 - \pi, \angle b_2) \cap \big((\angle b_3, 2\pi) \cup (0, \angle b_3 - \pi)\big),$$

which simplifies to

$$\angle a_3 \in (\angle b_3, \angle b_2).$$

This result implies that $\angle a_3 \in (\pi, 2\pi)$ and hence $\hat{\imath}(\vec{a}_3, \vec{b}_1) > 0$, contradicting Figure 8 where $\hat{\imath}(\vec{a}_3, \vec{b}_1) < 0$.

We conclude no assignment of angles to $\vec{\alpha}$ and $\vec{\beta}$ on (X, ω) is compatible with the specified algebraic intersection numbers. Therefore, $\vec{\alpha}$ and $\vec{\beta}$ cannot be realized as a pair of multicylinders on any translation surface.

We observe that all of the curves in the example above are homologically independent, so they will satisfy ($\star$). Moreover, our argument used only the intersection numbers of these curves, not their precise configuration on the surface, so this obstruction can be detected at the level of homology.

References

[1] **T Aougab**, **W Menasco**, **M Nieland**, *Origamis associated to minimally intersecting filling pairs*, Pacific J. Math. 317 (2022) 1–20 MR

[2] **A Calderon**, **N Salter**, *Higher spin mapping class groups and strata of abelian differentials over Teichmüller space*, Adv. Math. 389 (2021) art. id. 107926 MR

[3] **H Chang**, **X Jin**, **W W Menasco**, *Origami edge-paths in the curve graph*, Topology Appl. 298 (2021) art. id. 107730 MR

[4] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR

[5] **S-W Fu**, *Flat grafting deformations of quadratic differentials on surfaces*, Geom. Dedicata 214 (2021) 119–138 MR

[6] **A Hatcher**, **W Thurston**, *A presentation for the mapping class group of a closed orientable surface*, Topology 19 (1980) 221–237 MR

[7] **L Jeffreys**, *Single-cylinder square-tiled surfaces and the ubiquity of ratio-optimising pseudo-Anosovs*, Trans. Amer. Math. Soc. 374 (2021) 5739–5781 MR

[8] **R C Penner**, *A construction of pseudo-Anosov homeomorphisms*, Trans. Amer. Math. Soc. 310 (1988) 179–197 MR

[9] **K Rafi**, *A characterization of short curves of a Teichmüller geodesic*, Geom. Topol. 9 (2005) 179–202 MR

[10] **N Salter**, *Higher spin mapping class groups in algebraic and flat geometry*, expository notes (2020) Available at `https://nsalter.science.nd.edu/expository-notes/cuernavacalectures.pdf`

[11] **A Wright**, *Cylinder deformations in orbit closures of translation surfaces*, Geom. Topol. 19 (2015) 413–438 MR

[12] **A Zorich**, *Flat surfaces*, from "Frontiers in number theory, physics, and geometry, I" (P E Cartier, B Julia, P Moussa, P Vanhove, editors), Springer (2006) 437–583 MR

[13] **A Zorich**, *Explicit Jenkins–Strebel representatives of all strata of abelian and quadratic differentials*, J. Mod. Dyn. 2 (2008) 139–185 MR

JA: *Department of Mathematics, Cornell University*
Ithaca, NY, United States
JB: *Swarthmore, PA, United States*
AC: *Department of Mathematics, University of Chicago*
Chicago, IL, United States
JL: *Prairie Village, KS, United States*
TS: *Department of Mathematics, Statistics, and Computer Science, University of Illinois Chicago*
Chicago, IL, United States

`ja742@cornell.edu`, `jmcibarkdoll@gmail.com`, `aaroncalderon@uchicago.edu`,
`jenav.lor@gmail.com`, `tsands3@uic.edu`

Received: 22 April 2024 Revised: 31 July 2024

Geometry & Topology Publications, an imprint of **mathematical sciences publishers** msp

Algebraic & Geometric Topology **25**:8 (2025) 5059–5111
DOI: 10.2140/agt.2025.25.5059
Published: 20 November 2025

Involutive Khovanov homology and equivariant knots

TAKETO SANO

For strongly invertible knots, we define an involutive version of Khovanov homology, and from it derive a pair of integer-valued invariants $(\underline{s}, \overline{s})$, which is an equivariant version of Rasmussen's s-invariant. Using these invariants, we reprove that the infinite family of knots J_n introduced by Hayden each admits exotic pairs of slice disks. Our construction is intended to give a Khovanov-theoretic analogue of the formalism given by Dai, Mallick and Stoffregen in involutive knot Floer theory.

57K18

1 Introduction

A *strongly invertible knot* is a knot K in S^3 equipped with an involution τ of S^3 that reverses the orientation of K. While strongly invertible knots have been studied for decades (see Sakuma [26]), recent developments in knot theory and low-dimensional topology gave rise to new directions in its research and its applications, as in Watson [33], Snape [30], Hayden [12], Hayden and Sundberg [13], Alfieri and Boyle [2], Boyle and Issa [7], Lipshitz and Sarkar [21] and Dai, Mallick and Stoffregen [10].

Dai, Mallick and Stoffregen [10] use *involutive knot Floer homology* to define integer-valued invariants $\underline{V}^{\tau}_0, \overline{V}^{\tau}_0, \underline{V}^{\iota\tau}_0, \overline{V}^{\iota\tau}_0$ of strongly invertible knots, and show that there is an infinite family of knots J_n, each admitting exotic pairs of slice disks. Previously, the result for the special case $J_0 = 17nh_{73}$ (also known as the *positron knot*) was proved by Hayden and Sundberg [13] using Khovanov homology, by distinguishing the cobordism maps induced from the two slice disks. In this paper, we adapt the formalism of involutive knot Floer homology to the Khovanov side, and recover the general result systematically.

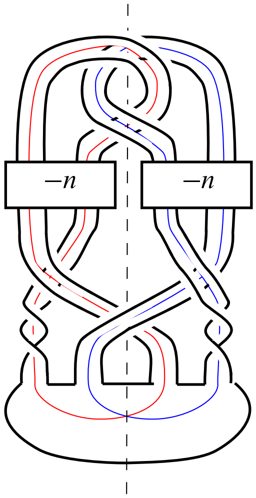


Figure 1: The knot J_n and the two slice disks D_n, D_n' obtained by compressing along the two colored circles.

For strongly invertible knots, we define an involutive version of Khovanov homology, called the *involutive Khovanov homology*, and from it derive a pair of integer-valued invariants $(\underline{s}, \overline{s})$, which is an equivariant version of Rasmussen's s*-invariant* [25] for ordinary knots.

Theorem 1 *For each $n \geq 0$, the strongly invertible knot J_n of Figure 1 has*

$$\underline{s}(J_n) = 0 < 2 \leq \overline{s}(J_n).$$

As we shall see later, Theorem 1 implies that each J_n never admits a *simple isotopy-equivariant* slice disk. In particular, the two slice symmetric disks D_n, D_n' of J_n depicted in Figure 1 are not smoothly isotopic rel J_n. Combining Theorem 1 with the fact that D_n and D_n' are topologically isotopic, which is proved by Hayden [12] using the result of Conway and Powell [9], we may conclude that these disks form an exotic pair of slice disks of J_n. This argument is completely analogous to the proof of [10, Theorem 7.11].

The definition of involutive Khovanov homology follows the formalism of involutive knot Floer homology; see Hendricks and Manolescu [14], Zemke [36], Alfieri, Kang and Stipsicz [3] and Dai, Mallick and Stoffregen [10]. Given a strongly invertible knot diagram (D, τ), there is an induced involution τ on the Khovanov complex $\mathrm{CKh}(D)$ over $\mathbb{F}_2$. Using this we define:

Definition 1.1 The *involutive Khovanov complex* of (D, τ) is defined by

$$\mathrm{CKhI}(D, \tau) = \mathrm{Cone}\big(\mathrm{CKh}(D) \xrightarrow{Q(1+\tau)} Q\mathrm{CKh}(D)\big),$$

where Q is a formal variable of $Q^2 = 0$. The homology of $\mathrm{CKhI}(D, \tau)$ is denoted by $\mathrm{KhI}(D, \tau)$ and is called the *involutive Khovanov homology* of (D, τ).

Theorem 2 *The isomorphism class of* $\mathrm{KhI}(D,\tau)$ *is an invariant of the strongly invertible knot* (K,τ).

Hereafter, we make τ implicit and omit it from CKhI and KhI. We note that Definition 1.1 is also valid for the deformed versions of Khovanov homology obtained by replacing the defining Frobenius algebra, and Theorem 2 still holds. There are also reduced versions, denoted by $\mathrm{CKhI}_r(D)$ and $\mathrm{KhI}_r(D)$, which are also invariants of strongly invertible knots.

Our equivariant invariant $(\underline{s},\overline{s})$ is defined using the Bar-Natan's deformation, given by the Frobenius algebra $A = R[X]/(X(X+H))$ over the ring $R = \mathbb{F}_2[H]$ with $\deg(H) = -2$. Let us first recall the definition of the $\mathbb{F}_2$-*Rasmussen invariant* $s = s^{\mathbb{F}_2}$, as characterized by Kotelskiy, Watson and Zibrowius in [18, Proposition 3.8]. For any knot K, it is known that the reduced Bar-Natan homology $\mathrm{BN}_r(K)$ has a single $\mathbb{F}_2[H]$-tower in homological grading 0. Then $s(K)$ is defined as the quantum grading of its generator:

$$\mathrm{BN}_r(K) \cong h^0 q^{s(K)} \mathbb{F}_2[H] \oplus (\mathrm{Tor}).$$

For a strongly invertible knot K, it is proved that the *reduced involutive Bar-Natan homology* $\mathrm{BNI}_r(K)$ has two $\mathbb{F}_2[H]$-towers, one in homological grading 0 and another in homological grading 1. Thus we may define $\underline{s}(K), \overline{s}(K)$ as the quantum gradings of their generators:

$$\mathrm{BNI}_r(K) \cong h^0 q^{\underline{s}(K)} \mathbb{F}_2[H] \oplus h^1 q^{\overline{s}(K)} \mathbb{F}_2[H] \oplus (\mathrm{Tor}).$$

The pair $(\underline{s}(K), \overline{s}(K))$ is called the *equivariant Rasmussen invariant* of K.

Theorem 3 *The equivariant Rasmussen invariant* $(\underline{s}(K), \overline{s}(K))$ *is an invariant of the strongly invertible knot* K, *satisfying*

$$\underline{s}(K) \le s(K) \le \overline{s}(K).$$

As is true for the ordinary s-invariant, our equivariant s-invariant is directly computable using computers. In particular, we have computed the invariants for the three strongly invertible knots given by Hayden and Sundberg in [13], which are proved therein to admit pairs of nonsmoothly isotopic slice disks.

Proposition 1.2 *The three strongly invertible slice knots* $K = m(9_{46}), 15n_{103488}, 17nh_{73}$ *have*

$$\underline{s}(K) = 0 < 2 = \overline{s}(K).$$

Table 1 shows the computed BNI_r for $J_0 = 17nh_{73}$, from which we can see that there are indeed two $\mathbb{F}_2[H]$ summands, one in bigrading $(0,0)$ and another in $(1,2)$, and the remaining summands are copies of $\mathbb{F}_2[H]/(H) = \mathbb{F}_2$. The computed BNI_r for the other two knots will be given in Section 5. Combined with Corollary 1.11, we recover that these knots admit pairs of nonsmoothly isotopic slice disks.

We further study properties of the equivariant invariant $(\underline{s},\overline{s})$, that are analogous to that of the ordinary s.

Proposition 1.3 *For the mirror* K^* *of* K, *we have*

$$\underline{s}(K^*) = -\overline{s}(K).$$

	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
18	.	.	.	.	.	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
16	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	.	.	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}^2$	$\mathbb{F}$	.	.
12	.	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}^2$	$\mathbb{F}$	.	.	.
10	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
8	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}^3$	$\mathbb{F}^3$	$\mathbb{F}$	.	.	.	.	.
6	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
4	.	.	.	$\mathbb{F}$	$\mathbb{F}^2$	$\mathbb{F}$	.	.	.	.	.	.	.	.
2	.	.	$\mathbb{F}$	$\mathbb{F}[H] \oplus \mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.	.	.
0	.	.	$\mathbb{F}[H]$	.	.	.	.	.	.	.	.	.	.	.
-2	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.	.	.	.	.	.

Table 1: $\mathrm{BNI}_r(J_0)$.

Proposition 1.4 *For another strongly invertible knot K',*

$$\underline{s}(K)+\underline{s}(K') \le \underline{s}(K \# K') \le \underline{s}(K)+\overline{s}(K') \le \overline{s}(K \# K') \le \overline{s}(K)+\overline{s}(K').$$

Here # denotes the equivariant connected sum.

Proposition 1.5 *Let K^+, K^- be strongly invertible knots such that K^- is obtained by applying an "equivariant negative crossing change" to K^+* (see Definition 4.15). *Then*

$$\underline{s}(K^-) \le \underline{s}(K^+) \le \underline{s}(K^-)+2a,$$

where $a=1$ if the move is performed on-axis, and $a=2$ if performed off-axis. The same holds for $\overline{s}$.

Proposition 1.6 *Let K^+, K^- be strongly invertible knots such that K^- is obtained by applying a "4-strand equivariant generalized negative crossing change" to K^+* (see Definition 4.24). *Then*

$$\underline{s}(K^-) \le \underline{s}(K^+) \le \underline{s}(K^-)+4a,$$

where $a=1$ if the move is performed on-axis, and $a=2$ if performed off-axis. The same holds for $\overline{s}$.

Proposition 1.7 *The positive (p,q)-torus knot $T_{p,q}$ has*

$$\underline{s}(T_{p,q})=\overline{s}(T_{p,q})=(p-1)(q-1)$$

with respect to the unique inverting involution.

The lower bound for the 4-genus

$$|s(K)| \le 2g_4(K)$$

is one of the significant properties of s, that led to the reproof of the *Milnor conjecture* [24]. This is an implication of its behavior under cobordisms. In order to prove an analogous result for the equivariant invariant, we impose the following condition on cobordisms.

Definition 1.8 Let L, L' be two strongly invertible links sharing the same involution τ of S^3. A *simple equivariant cobordism* between L and L' is an oriented cobordism S in $S^3 \times I$ between L, L' satisfying $(\tau \times \mathrm{id})(S) = S$. A *simple isotopy-equivariant cobordism* between L, L' is defined similarly, except that $(\tau \times \mathrm{id})(S)$ is only required to be isotopic to S rel boundary.

Proposition 1.9 *Suppose there is a connected, simple isotopy-equivariant cobordism S between two strongly invertible knots K, K'. Then,*

$$|\underline{s}(K) - \underline{s}(K')| \leq -\chi(S) \quad \textit{and} \quad |\overline{s}(K) - \overline{s}(K')| \leq -\chi(S).$$

In particular, $\underline{s}$ and $\overline{s}$ are invariant under simple isotopy-equivariant concordances.

We may naturally define the *simple equivariant genus* $\widetilde{\mathrm{sg}}_4(K)$ and the *simple isotopy-equivariant genus* $\widetilde{\mathrm{sig}}_4(K)$ of a strongly invertible knot K. Obviously, we have inequalities

$$\begin{tikzcd}
 & \widetilde{g}_4(K) \arrow[dr, phantom, sloped, "\leq"] & \\
g_4(K) \arrow[ur, phantom, sloped, "\leq"] \arrow[dr, phantom, sloped, "\leq"] & & \widetilde{\mathrm{sg}}_4(K) \leq \widetilde{g}_3(K) \\
 & \widetilde{\mathrm{sig}}_4(K) \arrow[ur, phantom, sloped, "\leq"] &
\end{tikzcd}$$

where g_4 is the ordinary 4-genus, while $\widetilde{g}_3$ and $\widetilde{g}_4$ are the equivariant 3- and 4-genus, respectively. Proposition 1.9 implies:

Corollary 1.10 *Both $|\underline{s}(K)|$ and $|\overline{s}(K)|$ bound $2\,\widetilde{\mathrm{sig}}_4(K)$ from below.*

Corollary 1.11 *If either $|\underline{s}(K)|$ or $|\overline{s}(K)|$ is greater than $2g_4(K)$, then no slice surfaces S of K realizing $g(S) = g_4(K)$ are simple isotopy-equivariant. In particular, S and $(\tau \times \mathrm{id})(S)$ are not smoothly isotopic rel K.*

Now Theorem 1 and the implication that the two slice disks D_n, D'_n of Figure 1 are not smoothly isotopic rel J_n follows from Theorem 3, Propositions 1.2 and 1.6 and Corollary 1.11, for we have

$$0 = \underline{s}(J_0) \leq \underline{s}(J_1) \leq \cdots \leq \underline{s}(J_n) \leq s(J_n) = 0 \quad \text{and} \quad 2 = \overline{s}(J_0) \leq \overline{s}(J_1) \leq \cdots \leq \overline{s}(J_n).$$

We finally remark that the restriction to *simple* equivariant cobordisms is due to the way the cobordism maps on Khovanov homology are defined. However, this restriction may lead to a deeper understanding of the *equivariant concordance group* $\widetilde{\mathscr{C}}$, defined by Sakuma in [26, Section 4]. It is known that the Smith conjecture fails in higher dimensions, in particular, there is an involution $\overline{\tau}$ on S^4 whose fixed-point set is a knotted sphere; see Gordon [11]. This implies that nonsimple equivariant cobordisms do exist. We may define the *simple equivariant concordance group* $\widetilde{\mathrm{s}\mathscr{C}}$, whose elements are simple equivariant concordance classes of directed strongly invertible knots, with the operation given by the equivariant connected sum. We question whether the surjective homomorphism $\widetilde{\mathrm{s}\mathscr{C}} \to \widetilde{\mathscr{C}}$ is injective or not, and also the existence of knots such that $\widetilde{g}_4(K) < \widetilde{\mathrm{sg}}_4(K)$.

Organization This paper is organized as follows. In Section 2, we define the *involutive Khovanov complex* CKhI for *involutive link diagrams*, and prove its invariance up to chain homotopy under the *involutive Reidemeister moves*. The reduced version is also introduced therein. In Section 3, we define the *equivariant Rasmussen invariant* $(\underline{s}, \overline{s})$ for *strongly invertible links*, not in the way presented above, but instead using the divisibility of the *equivariant Lee classes*. In Section 4, the above-stated properties of $(\underline{s}, \overline{s})$ are proved, including the proof that the two definitions of $(\underline{s}, \overline{s})$ coincide. In Section 5, the proof of Theorem 1 will be restated, followed by an observation of the result. In Section 6, we briefly state that an analogous construction is possible for 2-periodic links by considering a modified involution. In the appendix, we give a list of KhI and BNI for prime knots with up to 7 crossings, obtained by direct computations.

Acknowledgements The author thanks Masaki Taniguchi, Kouki Sato and Makoto Sakuma for helpful discussions, and the referee for helpful comments. The author was supported by JSPS KAKENHI grant 23K12982, RIKEN iTHEMS Program and academist crowdfunding.

2 Involutive Khovanov homology

Throughout this paper, we work in the smooth category and assume all objects and maps to be smooth. We assume that the reader is familiar with the construction of Khovanov homology [17], in particular Bar-Natan's reformulation given in [5]. Let R be a commutative ring with unity of char $R = 2$, and A a Frobenius algebra of the form $A = R[X]/(X(X+h))$ with $h \in R$, determined by $\varepsilon(1) = 0$, $\varepsilon(X) = 1$. For a link diagram D, let $\mathrm{CKh}(D; R, h)$ denote the Khovanov chain complex of D obtained from the above Frobenius algebra A, and $\mathrm{Kh}(D; R, h)$ its homology. Typically we consider the following cases:

$$(R, h) = (\mathbb{F}_2, 0),\ (\mathbb{F}_2, 1),\ (\mathbb{F}_2[H], H),$$

each corresponding to the original Khovanov homology [17], the filtered and the bigraded Bar-Natan homology [5] over $\mathbb{F}_2$. For the third case, we assume that H has $\deg(H) = -2$, so that the chain complex admits a bigrading. We usually make (R, h) implicit and omit it from the notations.

2.1 Definition

Hereafter, we assume that any involution τ on S^3 has fixed-point set S^1. (From the resolution of the Smith conjecture [32; 6], the fixed-point set of an involution of S^3 with nonempty fixed-point set is necessarily an unknotted circle.) The fixed-point set $\mathrm{Fix}(\tau)$ is called the *axis* of τ. We follow [22] for the definitions of involutive links and their equivalence.

Definition 2.1 An *involutive link* (L, τ) in S^3 is an oriented link L equipped with an involution τ on S^3 such that $\tau(L) = L$, possibly altering the orientations on some of its components. Two involutive links are *equivalent* if they are isotopic through involutive links. An involutive link (L, τ) is *strongly invertible* if τ reverses the orientation of L, *2-periodic* if τ preserves the orientation of L. Strongly invertible links and 2-periodic links are together called *equivariant links*.

We will draw involutive link diagrams so that the axis is projected as a straight vertical line, and the diagram is symmetric with respect to the axis, as in Figure 1. For an involutive link diagram (D,τ), the induced involution τ on $\mathrm{CKh}(D)$ is defined as follows. For each state s for D, there is a unique state s' such that $D(s)$ and $D(s')$ are symmetric with respect to the axis $\mathrm{Fix}(\tau)$. For each standard generator x in state s, its image $\tau(x)$ is defined to be the generator in the state s' that corresponds to x under the symmetry. It is obvious that τ commutes with d and is an involution on $\mathrm{CKh}(D)$ with bidegree $(0,0)$.

Definition 2.2 The *involutive Khovanov complex* of (D,τ) is defined by

$$\mathrm{CKhI}(D,\tau)=\mathrm{Cone}\big(\mathrm{CKh}(D)\xrightarrow{Q(1+\tau)}Q\mathrm{CKh}(D)\big),$$

where Q is a formal variable of $Q^2=0$. The homology of $\mathrm{CKhI}(D,\tau)$ is denoted by $\mathrm{KhI}(D,\tau)$ and is called the *involutive Khovanov homology* of (D,τ).

Remark 2.3 Our $\mathrm{CKhI}(D,\tau)$ is different from the triply graded complex $\mathrm{CKh}_\tau(D)$ given in [22], where $\mathrm{CKh}_\tau(D)=\mathrm{CKh}(D;\mathbb{F}_2,0)$ as $\mathbb{F}_2$-modules, but the differential is defined as $\partial=d+1+\tau$.

Hereafter we make τ implicit and omit it from CKhI and KhI. When explicitly describing elements and maps, we often regard $\mathrm{CKhI}(D)$ as the direct sum of two copies of $\mathrm{CKh}(D)$,

$$\mathrm{CKhI}(D)=\mathrm{CKh}(D)\oplus\mathrm{CKh}(D)[1],$$

with the differential given by

$$\begin{pmatrix} d & \\ 1+\tau & d\end{pmatrix}.$$

As in the noninvolutive case, the differential preserves the quantum grading if $h=0$ or $\deg(h)=-2$; or is quantum-grading nondecreasing if $h\neq 0$ and $\deg(h)=0$. In the former case, we regard $\mathrm{CKhI}(D)$ as a bigraded complex, and in the latter as a filtered complex.

Formally, a chain complex over $\mathbb{F}_2$ equipped with an involution τ is called a *τ-complex*. For τ-complex C, let C_τ denote the complex

$$C_\tau=\mathrm{Cone}(1+\tau).$$

This correspondence is functorial in the following sense. For homotopic chain maps f,g (of any degree) with homotopy h, we write

$$f\simeq_h g$$

to mean

$$f+g=dh+hd.$$

A chain map f between τ-complexes C,C' is *homotopy τ-equivariant* if $\tau f\simeq f\tau$. A *τ-conjugate* of f is defined by

$$f^\tau=\tau f\tau.$$

Obviously f is homotopy τ-equivariant if and only if $f \simeq f^\tau$. Two homotopy τ-equivariant chain maps f, g with homotopies

$$\tau f \simeq_{h_f} f\tau \quad \text{and} \quad \tau g \simeq_{h_g} g\tau$$

are *coherently homotopic* with homotopy h if $f \simeq_h g$ and

$$\tau h + h\tau \simeq h_g + h_f.$$

The following lemmas are easy to verify and will be used throughout this paper.

Lemma 2.4 *A homotopy τ-equivariant chain map $f\colon C \to C'$ with homotopy $\tau f \simeq_{h_f} f\tau$ induces a chain map*

$$f_\tau\colon C_\tau \to C'_\tau \quad \text{given by} \quad f_\tau = \begin{pmatrix} f & \\ h_f & f \end{pmatrix}.$$

In particular, we have

$$1_\tau = 1.$$

For another homotopy τ-equivariant chain map $g\colon C' \to C''$ with homotopy $\tau g \simeq_{h_g} g\tau$, the composition gf is homotopy τ-equivariant with homotopy

$$h_{gf} = h_g f + g h_f,$$

and the above correspondence gives

$$(gf)_\tau = g_\tau f_\tau.$$

Lemma 2.5 *Suppose that $f, g\colon C \to C'$ are homotopy τ-equivariant chain maps with homotopies $\tau f \simeq_{h_f} f\tau$ and $\tau g \simeq_{h_g} g\tau$, and are coherently homotopic with homotopies h and k satisfying $f \simeq_h g$ and $\tau h + h\tau \simeq_k h_g + h_f$. Then the induced maps*

$$f_\tau, g_\tau\colon C_\tau \to C'_\tau$$

are homotopic with homotopy

$$\begin{pmatrix} h & \\ k & h \end{pmatrix}.$$

Lemma 2.6 *Let f be a homotopy τ-equivariant homotopy equivalence with a homotopy τ-equivariant homotopy inverse g and coherent homotopies*

$$gf \simeq 1 \quad \text{and} \quad fg \simeq 1.$$

Then the induced map

$$f_\tau\colon C_\tau \to C'_\tau$$

is a homotopy equivalence with homotopy inverse g_τ.

2.2 Invariance

Next we prove the invariance of KhI. In [22], Lobb and Watson introduced the eight *involutive Reidemeister moves* (see Figure 2) and proved that two involutive links are equivalent if and only if those diagrams are related by a sequence of involutive Reidemeister moves and equivariant planar isotopies. Theorem 2 follows from the following stronger proposition.

Proposition 2.7 *Let D, D' be two involutive link diagrams related by one of the involutive Reidemeister moves. Then there is a chain homotopy equivalence*

$$\rho\colon \mathrm{CKhI}(D) \to \mathrm{CKhI}(D')$$

such that the following diagram commutes:

$$\begin{tikzcd}
Q\mathrm{CKh}(D)[1] \arrow[r, hook] \arrow[d, "\rho"] & \mathrm{CKhI}(D) \arrow[r, two heads] \arrow[d, "\rho"] & \mathrm{CKh}(D) \arrow[d, "\rho"] \\
Q\mathrm{CKh}(D')[1] \arrow[r, hook] & \mathrm{CKhI}(D') \arrow[r, two heads] & \mathrm{CKh}(D')
\end{tikzcd}$$

Here the two vertical arrows on the left and the right are given by the composition of the standard chain homotopy equivalences given in [5] *that corresponds to some decomposition of the involutive move into a sequence of ordinary Reidemeister moves.*

In order to prove Proposition 2.7, instead of explicitly constructing chain homotopy equivalences and chain homotopies for each of the moves, we prove the existences of the desired maps in a uniform way.

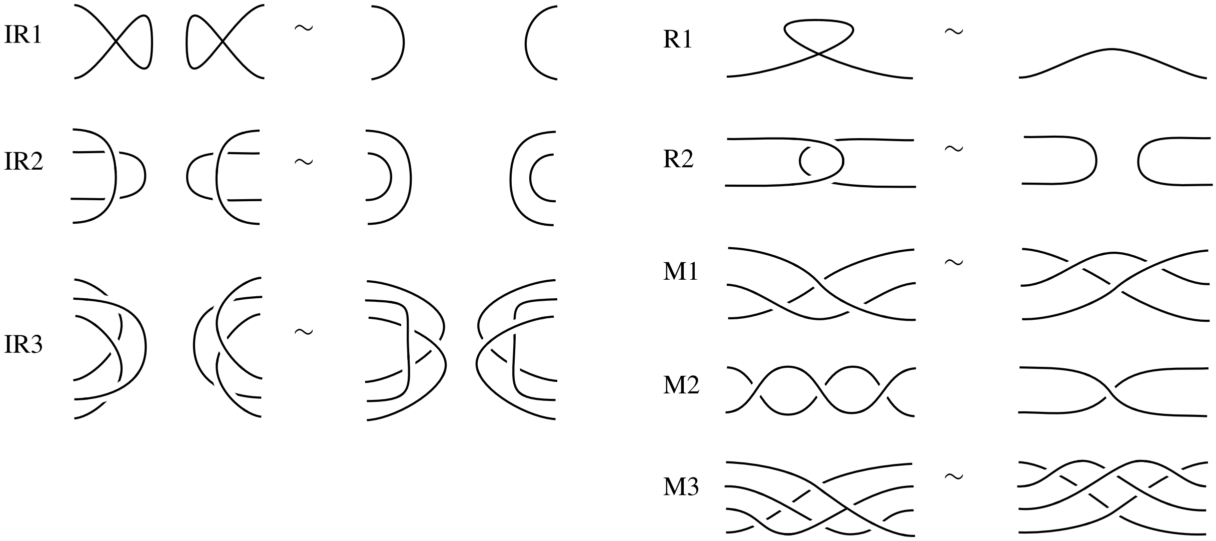


Figure 2: Involutive Reidemeister moves.

2.2.1 General strategy Any involutive Reidemeister move can be decomposed into a sequence of ordinary Reidemeister moves. By composing the corresponding maps given in [5] we get explicit homotopy equivalences F, G between $\mathrm{CKh}(D)$ and $\mathrm{CKh}(D')$, together with homotopies $GF \simeq_H I$ and $FG \simeq_{H'} I$. In general, these maps are not strictly τ-equivariant. Nevertheless, we can prove the following.

Claim 2.8 *F and G are homotopy τ-equivariant, ie there exists homotopies $\tau F \simeq_{h_F} F\tau$ and $\tau G \simeq_{h_G} G\tau$.*

Claim 2.9 *H and H' are coherent homotopies, ie there exists homotopies $\tau H + H\tau \simeq h_G F + G h_F$ and $\tau H' + H'\tau \simeq h_F G + F h_G$.*

Then the invariance follows from Lemma 2.6. Thus the proof is reduced to proving Claims 2.8 and 2.9 for each of the moves. Recall from [5, Definition 8.5] that a tangle diagram T is Kh-*simple* if any degree 0 automorphism (up to homotopy) of $C(T)$ is homotopic to $\pm I$. Here $C(T)$ is the *formal Khovanov complex* introduced in [5], which is a complex in the additive closure of the category $\mathrm{Cob}^3(\partial T)$. Here we strengthen the condition as follows.

Definition 2.10 A chain complex C (in any additive category) is *simple* if any degree 0 automorphism (up to homotopy) of C is homotopic to $\pm I$, and any degree $n \neq 0$ self-chain map $C \to C$ is null-homotopic. A tangle diagram T is Kh-*simple* if the complex $C(T)$ is simple.

A complex C being simple is equivalent to the condition that the homology of the End-complex $\mathrm{End}(C) = \mathrm{Hom}(C, C)$ has only $\pm I$ as the degree 0 multiplicative units (with respect to the composition) and its homology is supported only on degree 0. One can easily prove that [5, Lemmas 8.6–8.9] also hold for our stronger definition, namely:

Lemma 2.11 *Simplicity is preserved under chain-homotopy equivalences.*

Lemma 2.12 *Parings are* Kh-*simple. Here, a pairing is a tangle diagram that has no crossings and no closed components.*

Lemma 2.13 *A tangle diagram T is* Kh-*simple if and only if TX is* Kh-*simple. Here TX is a tangle diagram obtained by adding one extra crossing X somewhere along the boundary of T.*

The following lemmas will also be useful. Here C and C' are complexes in any additive category.

Lemma 2.14 *Suppose there are two degree 0 chain homotopy equivalences $F, F' \colon C \to C'$ with homotopy inverses G, G' respectively. If C is simple, then F and F', as well as G and G', are homotopic up to sign.*

Proof The map GF' is an automorphism on C with a homotopy inverse $G'F$. Since C is simple, we have $GF' \simeq \pm I$. Assuming that $GF' \simeq I$, we get $F \simeq FGF' \simeq F'$ and $G \simeq GF'G' \simeq G'$. □

Lemma 2.15 *Suppose there are degree* 0 *chain maps* $F, F'\colon C \to C'$ *and* $G, G'\colon C' \to C$, *together with homotopies* $F \simeq_{h_F} F'$, $G \simeq_{h_G} G'$, $GF \simeq_H I$ *and* $G'F' \simeq_{H'} I$. *If* C *is simple, then* $H - H'$ *and* $h_G F + G' h_F$ *are homotopic.*

Proof On one hand, we have

$$GF - G'F' = d(H - H') + (H - H')d.$$

On the other hand, we have

$$\begin{aligned} GF - G'F' &= (G - G')F + G'(F - F') \\ &= (dh_G + h_G d)F + G'(dh_F + h_F d) \\ &= d(h_G F + G' h_F) + (h_G F + G' h_F)d. \end{aligned}$$

Thus $(H - H') - (h_G F + G' h_F)$ is a degree -1 chain map, which is necessarily null-homotopic since C is simple. Therefore $H - H' \simeq h_G F + G' h_F$. □

2.2.2 IR1–IR3

Proof of Proposition 2.7, cases IR1–IR3 We define maps F, G by the compositions

$$F = F_2 F_1 \quad \text{and} \quad G = G_1 G_2,$$

where F_1, F_2 are the standard maps corresponding to the Reidemeister moves performed on the left and the right part of the diagram respectively (see Figure 3), and G_1, G_2 are those homotopy inverses. The conjugate maps are given by

$$F^\tau = F_2^\tau F_1^\tau \quad \text{and} \quad G^\tau = G_1^\tau G_2^\tau.$$

Since F_1 and F_2 (resp. G_1 and G_2) are commutative, we may write

$$F^\tau = F_1^\tau F_2^\tau \quad \text{and} \quad G^\tau = G_2^\tau G_1^\tau.$$

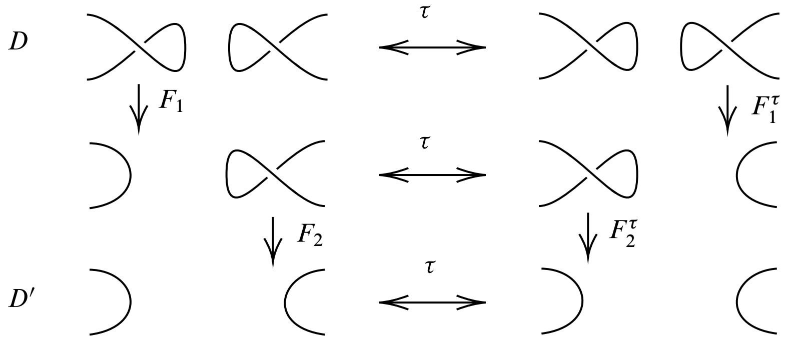


Figure 3: F and F^τ.

Since the tangle parts appearing in the move are Kh-simple, using Lemmas 2.14 and 2.15 we obtain homotopies with the desired properties such as $F_1 \simeq F_2^\tau$ and $F_2 \simeq F_1^\tau$. Thus we obtain the desired properties stated in Claims 2.8 and 2.9, such as $F \simeq F^\tau$. Here we are implicitly extending the maps defined locally for tangle diagrams to maps defined globally for the link diagrams, using the planar algebra structures explained in [5, Section 5]. □

2.2.3 R1, R2 and M1–M3

Proof of Proposition 2.7, cases R1, R2 and M1–M3 Observe that each of these moves occurs locally on a disk that intersects the axis, and the tangles appearing in the move are Kh-simple. Thus Claims 2.8 and 2.9 immediately follow from Lemmas 2.14 and 2.15. □

Remark 2.16 By using the explicit description of the maps, for some of the moves we may take the desired homotopies in a much simpler form. For R1 and R2, the homotopy equivalences F, G and the homotopies H, H' are strictly τ-equivariant. For R3, we may take $\tau F \simeq_{h_F} F\tau$ and $\tau G \simeq_{h_G} G\tau$ so that $Gh_F = 0$, $Fh_G = 0$, and $\tau H + H\tau = h_G F$, $\tau H' + H'\tau = h_F G$.

2.3 Reduced version

Next, we define a reduced version of the involutive Khovanov homology. Let us first recall the definition of the reduced complex in the noninvolutive setting. For a pointed link diagram D, the *reduced Khovanov complex*[1] $\mathrm{CKh}_r(D)$ is defined as the subcomplex of $\mathrm{CKh}(D)$ generated by the standard generators each labeled X on the pointed circle. The *coreduced complex* $\mathrm{CKh}'_r(D)$ is defined as the quotient $\mathrm{CKh}(D)/\mathrm{CKh}_r(D)$.

Definition 2.17 A *pointed involutive link* (L, τ) is an involutive link equipped with a basepoint on $\mathrm{Fix}(\tau) \cap L$.

Note that 2-periodic links cannot be pointed, since $\mathrm{Fix}(\tau) \cap L = \varnothing$. For a pointed involutive link diagram D, it is obvious that the (co)reduced complexes are invariant under the involution τ. Thus we may define:

Definition 2.18 Let (D, τ) be a pointed involutive link diagram. The *reduced involutive Khovanov complex* is defined as

$$\mathrm{CKhI}_r(D, \tau) = \mathrm{Cone}(\mathrm{CKh}_r(D) \xrightarrow{Q(1+\tau)} Q\mathrm{CKh}_r(D)).$$

Similarly, the *coreduced involutive Khovanov complex* is defined as

$$\mathrm{CKhI}'_r(D, \tau) = \mathrm{Cone}(\mathrm{CKh}'_r(D) \xrightarrow{Q(1+\tau)} Q\mathrm{CKh}'_r(D)).$$

---

[1] The conventional notation for the reduced Khovanov complex is $\widetilde{\mathrm{CKh}}$. Here we changed the notation to avoid putting too much decorations on the letters.

Again, we usually omit τ from the notations of CKhI_r. In the noninvolutive setting, the reduced and the coreduced complexes can be interchanged by the following automorphism σ.

Definition 2.19 The Frobenius algebra automorphism σ on A is defined by

$$1 \mapsto 1, \quad X \mapsto X + h.$$

Its induced automorphism on $\mathrm{CKh}(D)$ is also denoted by σ.

Proposition 2.20 *The automorphism σ is an involution that commutes with τ.*

Thus σ induces an involution on $\mathrm{CKhI}(D)$, which is again denoted σ. Analogous to the noninvolutive case, we have:

Proposition 2.21 *There are isomorphisms*

$$\mathrm{CKhI}'_r(D) \cong \sigma(\mathrm{CKhI}_r(D)) \quad \textit{and} \quad \mathrm{CKhI}_r(D) \cong \mathrm{CKhI}(D)/\sigma(\mathrm{CKhI}_r(D)).$$

Proof Immediate from [28, Propositions 3.12 and 3.14]. □

Proposition 2.22 *There is a short exact sequence*

$$\mathrm{CKhI}_r(D) \hookrightarrow \mathrm{CKhI}(D) \twoheadrightarrow \mathrm{CKhI}'_r(D).$$

Proof The short exact sequence

$$\mathrm{CKh}_r(D) \hookrightarrow \mathrm{CKh}(D) \twoheadrightarrow \mathrm{CKh}'_r(D)$$

is τ-equivariant, and hence we obtain maps between exact sequences

$$\begin{tikzcd}
\mathrm{CKhI}_r \arrow[r, hook] \arrow[d, equal] & \mathrm{CKhI} \arrow[r, two heads] \arrow[d, equal] & \mathrm{CKhI}/\mathrm{CKhI}_r \arrow[d, dashed] \\
\mathrm{Cone}(\mathrm{CKh}_r \to Q\mathrm{CKh}_r) \arrow[r, hook] & \mathrm{Cone}(\mathrm{CKh} \to Q\mathrm{CKh}) \arrow[r, two heads] & \mathrm{Cone}(\mathrm{CKh}'_r \to Q\mathrm{CKh}'_r)
\end{tikzcd}$$

The right dashed arrow is an isomorphism from the five lemma. □

Proposition 2.23 *The short exact sequence of Proposition 2.22 splits.*

Before proceeding to the proof, let us review the corresponding results in the noninvolutive setting. Shumakovitch [29] proved that the $\mathbb{F}_2$-Khovanov homology splits (ie when $(R,h) = (\mathbb{F}_2, 0)$), and lately Wigderson [35] extended this result to the $\mathbb{F}_2$-bigraded Bar-Natan homology (ie $(R,h) = (\mathbb{F}_2[H], H)$). Here we briefly review Wigderson's construction, which works whenever $\operatorname{char} R = 2$. First, as an R-module, the coreduced complex $\mathrm{CKh}'_r(D)$ can be identified with the submodule of $\mathrm{CKh}(D)$ generated by the standard generators each labeled 1 on the pointed circle. Then $\mathrm{CKh}(D) = \mathrm{CKh}'_r(D) \oplus \mathrm{CKh}_r(D)$ as R-modules, and the differential d of $\mathrm{CKh}(D)$ can be described as

$$d = \begin{pmatrix} d_1 & \\ f & d_X \end{pmatrix},$$

where d_X and d_1 are differentials of $\mathrm{CKh}_r(D)$ and $\mathrm{CKh}'_r(D)$, respectively, and

$$f\colon \mathrm{CKh}'_r(D) \to \mathrm{CKh}_r(D)$$

is the map given by restricting d to $\mathrm{CKh}'_r(D)$ and then projecting onto $\mathrm{CKh}_r(D)$. It follows from $d^2 = 0$ that f is a chain map, and thus $\mathrm{CKh}(D)$ may be regarded as the cone of f. Next, a null-homotopy κ of f is constructed as follows. For each $i \geq 0$, define

$$\kappa_i\colon \mathrm{CKh}'_r(D) \to \mathrm{CKh}_r(D), \quad x = \underline{1} \otimes \cdots \mapsto \sum_{\textcircled{0}^X, \dots, \textcircled{i}^X} \underline{X} \otimes \cdots.$$

Here, the underline indicates the label for the pointed circle, and the sum runs over all choices of $i + 1$ circles C_j labeled X in x. Inside the summation, the label on the pointed circle is changed from 1 to X while the labels of C_j are changed from X to 1. Then define

$$\kappa = \sum_{i \geq 0} h^i \kappa_i\colon \mathrm{CKh}'_r(D) \to \mathrm{CKh}_r(D).$$

It is proved in [35] that κ gives a null-homotopy of f, and hence $1 + \kappa$ gives a section of the quotient map $\mathrm{CKh}(D) \to \mathrm{CKh}'_r(D)$.

Proof of Proposition 2.23 Suppose D is a pointed involutive link diagram. It is obvious from the construction of κ that it commutes with τ, and hence

$$\begin{pmatrix} 1 + \kappa & \\ & 1 + \kappa \end{pmatrix}$$

gives a section of the quotient map $\mathrm{CKhI}(D) \to \mathrm{CKhI}'_r(D)$. □

Finally, we prove the invariance of the (co)reduced involutive homologies. In order to prove an analogue of Proposition 2.7, we need extra consideration for the moves that involve the basepoint, which are R1 and M1 with the basepoint placed on the horizontal strand. In fact, one can check that the corresponding maps do not restrict to the reduced complexes

$$\begin{tikzcd}
\mathrm{CKh}_r(D) \arrow[r, hook] \arrow[d, dotted, "\times"] & \mathrm{CKh}(D) \arrow[d, "\rho"] \\
\mathrm{CKh}_r(D') \arrow[r, hook] & \mathrm{CKh}(D)
\end{tikzcd}$$

Thus we restrict the diagrams and the moves that are allowed for the reduced complexes.

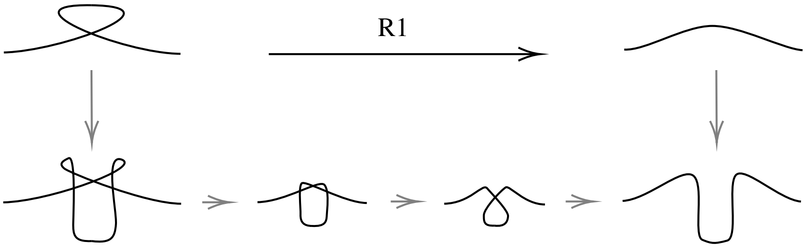


Figure 4: Modifying the R1-move.

Definition 2.24 A diagram of a pointed involutive link diagram is *normal* if the basepoint of the link is placed at the bottommost on the axis, and the horizontal strand containing the basepoint is directed rightwards.

Obviously, any pointed involutive link possesses a normal diagram. Moreover,

Proposition 2.25 *Suppose D, D' are normal pointed involutive link diagrams that represent the same pointed involutive link. Then there is a sequence of involutive Reidemeister moves whose intermediate diagrams are also normal.*

Proof Take any sequence of involutive Reidemeister moves between D and D'.

$$D = D_0 \to D_1 \to \cdots \to D_N = D'.$$

We modify this sequence, by increasing the number of moves if necessary, so that the intermediate diagrams are all normal.

Step 1 For each move $D_i \to D_{i+1}$, we may transform the diagrams by pulling down the horizontal strands that contain the basepoints so that they are placed at the bottommost on the axes (which can be realized by sequences of involutive Reidemeister moves). Moreover, we can show that the two transformed diagrams D'_i, D'_{i+1} can be related by a sequence of involutive Reidemeister moves that fix the basepoints:

$$\begin{tikzcd}
D_i \arrow[rrr, "\text{move}"] \arrow[d, "\text{pull}"'] & & & D_{i+1} \arrow[d, "\text{pull}"] \\
D'_i \arrow[r, "\text{move}"] & \cdots \arrow[rr, "\text{move}"] & & D'_{i+1}
\end{tikzcd}$$

The claim is clear when the original move occurs above the pointed strand, or when the moves are off-axis which are IR1–IR3. We must consider the case where the move occurs below the pointed strand, or contains the pointed strand itself. Figure 4 depicts the modification for the R1 move that contains the pointed strand. The claim for the remaining moves R2 and M1–M3 can be checked similarly.

Step 2 The remaining work is to undo the changes in the direction of the pointed strand due to the R1 moves. Note that applying an R1 move to the bottom strand is equivalent to half-twisting the upper parts and then applying an overall half-rotation with respect to the axis. Since the diagram is involutive, a half-rotation has the effect of only changing the orientations on the components of the link. We modify each R1 move by only twisting the upper parts (which can be represented by a sequence of involutive moves) while keeping the bottom strand fixed. Since the bottom strands of D and D' are both pointed rightwards, the R1 moves in total must be applied an even number of times. Thus the effects of skipping the overall half-rotations cancel, and we obtain a desired sequence of involutive Reidemeister moves between D and D'. □

Proposition 2.26 *Let D and D' be normal diagrams related by an involutive Reidemeister move. The chain homotopy equivalence ρ and the corresponding chain homotopies given in Proposition 2.7 restrict to*

$$\rho\colon \mathrm{CKhI}_r(D) \to \mathrm{CKhI}_r(D')$$

and the following diagram commutes:

$$\begin{tikzcd}
Q\mathrm{CKh}_r(D)[1] \arrow[r, hook] \arrow[d, "\rho"] & \mathrm{CKhI}_r(D) \arrow[r, two heads] \arrow[d, "\rho"] & \mathrm{CKh}_r(D) \arrow[d, "\rho"] \\
Q\mathrm{CKh}_r(D')[1] \arrow[r, hook] & \mathrm{CKhI}_r(D') \arrow[r, two heads] & \mathrm{CKh}_r(D')
\end{tikzcd}$$

The same statement holds for the coreduced counterparts.

Proof Since the basepoints are fixed by the move, the chain maps and chain homotopies of Proposition 2.7 restrict to the (co)reduced complexes. □

We conclude that for a pointed involutive link L with normal diagram D, the chain homotopy equivalence classes of the (co)reduced complexes $\mathrm{CKhI}_r(D)$, $\mathrm{CKhI}'_r(D)$ are invariants of L. Those homologies are denoted by $\mathrm{KhI}_r(L)$ and $\mathrm{KhI}'_r(L)$ and called the *(co)reduced involutive Khovanov homologies* of L.

2.4 Mirrors

Next, we study the behavior of the involutive complexes under mirrors. The arguments are straightforward extensions of [28, Section 3.5.2] to the involutive setting. Consider the standard perfect pairing on A

$$\langle\cdot,\cdot\rangle\colon A \otimes A \to R$$

given by $\langle x, y\rangle = \varepsilon(xy)$. The associated duality isomorphism $\mathsf{D}\colon A \to A^*$ such that $\langle x, y\rangle = \mathsf{D}(x)(y)$ is given by

$$\mathsf{D}(1) = X^*, \quad \mathsf{D}(X) = 1^* + hX^*,$$

where $\{1^*, X^*\}$ is the dual basis for A^* to the basis $\{1, X\}$ for A. For convenience, put $Y = X + h$, and consider another basis $\{1, Y\}$ for A and its dual basis $\{1^\dagger, Y^\dagger\}$ for A^*. Then we have

$$\mathsf{D}(1) = Y^\dagger, \quad \mathsf{D}(X) = 1^\dagger,$$

and

$$\langle X, X\rangle = h, \quad \langle X, Y\rangle = \langle Y, X\rangle = 0, \quad \langle Y, Y\rangle = h.$$

Note that when $h = 0$, the duality isomorphism D coincides with the ordinary self-dual isomorphism.

For a link diagram D, the duality isomorphism D induces a chain isomorphism

$$\mathsf{D}\colon \mathrm{CKh}(D^*) \xrightarrow{\cong} \mathrm{CKh}(D)^*,$$

where D^* denotes the mirror of D, and $\mathrm{CKh}(D)^*$ denotes the dual complex of $\mathrm{CKh}(D)$ with bigrading $(\mathrm{CKh}(D)^*)^{i,j} = \mathrm{CKh}^{-i,-j}(D)^*$. This gives a perfect pairing

$$\langle \cdot, \cdot \rangle\colon \mathrm{CKh}(D) \otimes \mathrm{CKh}(D^*) \to R.$$

Now suppose (D, τ) is an involutive link diagram.

Lemma 2.27 $$\mathsf{D}\tau = \tau^*\mathsf{D}\colon \mathrm{CKh}(D^*) \to \mathrm{CKh}(D)^*.$$

Proof By considering an $1X$-labeled generator x for D^* and an $1Y$-labeled generator y for D, we easily see that $\mathsf{D}(\tau x)(y) = \mathsf{D}(x)(\tau y)$ holds. □

Lemma 2.28 $$\mathrm{CKhI}(D)^*[1] \cong \mathrm{Cone}\big(\mathrm{CKh}(D)^* \xrightarrow{1+\tau^*} \mathrm{CKh}(D)^*\big).$$

Proof Obvious. □

Proposition 2.29 *There is an isomorphism*

$$\mathsf{D}\colon \mathrm{CKhI}(D^*) \xrightarrow{\cong} \mathrm{CKhI}(D)^*[1].$$

Proof The above results give isomorphisms,

$$\begin{aligned}\mathrm{CKhI}(D^*) &= \mathrm{Cone}\big(\mathrm{CKh}(D^*) \xrightarrow{1+\tau} \mathrm{CKh}(D^*)\big)\\ &\cong \mathrm{Cone}\big(\mathrm{CKh}(D)^* \xrightarrow{1+\tau^*} \mathrm{CKh}(D)^*\big)\\ &\cong \mathrm{CKhI}(D)^*[1].\end{aligned}$$ □

Next we see that the duality isomorphism also respects the (co)reduced complexes. By Proposition 2.21, the coreduced complex $\mathrm{CKhI}'_r(D)$ may be identified with the subcomplex $\sigma(\mathrm{CKhI}_r(D))$, and the reduced complex $\mathrm{CKhI}_r(D)$ with the quotient $\mathrm{CKhI}(D)/\sigma(\mathrm{CKhI}_r(D))$. Thus there is a short exact sequence in the reversed direction,

$$\mathrm{CKhI}'_r(D) \xrightarrow{i'} \mathrm{CKhI}(D) \xrightarrow{p'} \mathrm{CKhI}_r(D).$$

Proposition 2.30 *The isomorphism* D *of Proposition 2.29 induces isomorphisms on the* (co)*reduced complexes*, *such that the following diagram commutes*

$$\begin{tikzcd}
\mathrm{CKhI}_r(D^*) \arrow[r, "i"] \arrow[d, dashed, "\mathsf{D}"] & \mathrm{CKhI}(D^*) \arrow[r, "p"] \arrow[d, "\mathsf{D}"] & \mathrm{CKhI}'_r(D^*) \arrow[d, dashed, "\mathsf{D}"] \\
\mathrm{CKhI}_r(D)^* \arrow[r, "(p')^*"] & \mathrm{CKhI}(D)^* \arrow[r, "(i')^*"] & \mathrm{CKhI}'_r(D)^*
\end{tikzcd}$$

Proof Combine Proposition 2.29 with [28, Proposition 3.36]. □

Proposition 2.31 *There are perfect pairings*

$$\langle\cdot,\cdot\rangle\colon \mathrm{CKhI}(D)\otimes\mathrm{CKhI}(D^*)\to R \quad \textit{and} \quad \langle\cdot,\cdot\rangle_r\colon \mathrm{CKhI}_r(D)\otimes\mathrm{CKhI}_r(D^*)\to R$$

such that the following diagram commutes:

$$\begin{tikzcd}
\mathrm{CKhI}_r(D)\otimes\mathrm{CKhI}_r(D^*) \arrow[r, "\langle\cdot,\cdot\rangle_r"] \arrow[d, "i\otimes i"] & R \arrow[d, "h\cdot"] \\
\mathrm{CKhI}(D)\otimes\mathrm{CKhI}(D^*) \arrow[r, "\langle\cdot,\cdot\rangle"] & R
\end{tikzcd}$$

Proof Combine Proposition 2.29 with [28, Propositions 3.33, 3.37]. □

3 Equivariant Rasmussen invariants

The focus of this section is strongly invertible knots and links. We mainly consider the case where $h \neq 0$, typically

$$(R,h) = (\mathbb{F}_2, 1),\ (\mathbb{F}_2[H], H),\ (\mathbb{F}_2[H^{\pm}], H).$$

3.1 Equivariant Lee classes

Recall that in the noninvolutive setting, if $h \in R$ is invertible, then for a link diagram D the homology $\mathrm{Kh}(D)$ is generated by the *Lee classes* $\alpha(D,o)$ of D, each corresponding to an orientation o on D; see [19; 31]. Here we recall the construction.

Algorithm 3.1 Given a link diagram D, the *ab-coloring* on its Seifert circles is defined as follows: separate $\mathbb{R}^2$ into regions by the Seifert circles of D, and color the regions in the checkerboard fashion, with the unbounded region colored white. For each Seifert circle, let it inherit the orientation from D, and assign to it a if it sees a black region to the left with respect to the orientation, or b otherwise; see Figure 5.

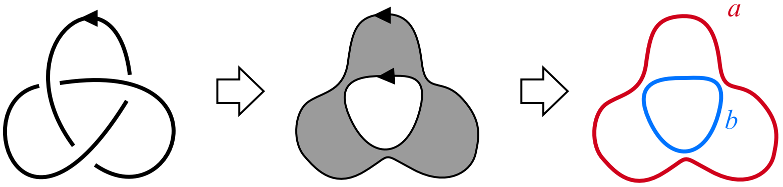


Figure 5: The ab-coloring on the Seifert circles of K.

Definition 3.2 Let D be a link diagram. There is a unique state s of D where the resolved diagram $D(s)$ gives the Seifert circles of D. With the ab-coloring on the Seifert circles, define an element $\alpha(D)$ in $\mathrm{CKh}(D)$ for that state by labeling each circle by X if it is colored a, and $Y = X + h$ if it is colored b. Similarly, for any orientation o on D, we define an element $\alpha(D, o)$ by the same procedure after reorienting D by o. These elements $\alpha(D, o)$ are in fact cycles, and are called the *Lee cycles* of D. The homology classes are called the *Lee classes* of D.

That $\alpha(D, o)$ is indeed a cycle can be seen from the fact that each crossing of D connects two differently colored strands in the resolved diagram, and merging the strands results in $XY = 0$. Note that the automorphism σ on $\mathrm{CKh}(D)$ interchanges $\alpha(D, o)$ and $\alpha(D, \bar{o})$, where $\bar{o}$ is the reversed orientation of o. We will frequently consider such pairs, so we write $\beta(D, o)$ for $\alpha(D, \bar{o})$. In particular when o is the given orientation of D, we write $\alpha(D)$ for $\alpha(D, o)$ and $\beta(D)$ for $\beta(D, o) = \alpha(D, \bar{o})$. The homological gradings of $\alpha(D, o)$ are all even, and in particular $\alpha(D)$, $\beta(D)$ has homological grading 0.

Proposition 3.3 *For a strongly invertible link diagram D, the Lee cycles are invariant under τ.*

Proof We only prove the case when o is the given orientation of D, since the other cases can be proved by reorienting D by o. From the definition of τ, it is obvious that the orientation-preserving state s is preserved by τ. Since $D(s)$, the diagram obtained by resolving D according to s, is symmetric with respect to the axis, we see that each circle C in $D(s)$ is either disjoint from the axis or intersects the axis exactly twice. For the first case, it is obvious that τ preserves the label on C. For the second case, there is a unique circle C' that is symmetric with C. Since D is strongly invertible, C and C' are oriented oppositely with respect to the reflection about the axis, and hence oriented equally as circles in $\mathbb{R}^2$. Thus from Algorithm 3.1 the two circles are colored the same. This shows that $\alpha(D)$ is invariant under τ. □

Hereafter we assume that D is a strongly invertible link diagram. For each orientation o on D, there are two elements $\alpha(D, o)$, $Q\alpha(D, o)$ in $\mathrm{CKhI}(D)$, which are denoted hereafter by $\underline{\alpha}(D, o)$ and $\overline{\alpha}(D, o)$. Proposition 3.3 implies that these are cycles in $\mathrm{CKhI}(D)$. By abuse of notation, the homology classes of $\underline{\alpha}(D, o)$, $\overline{\alpha}(D, o)$ are also denoted by the same symbols.

Definition 3.4 The cycles $\underline{\alpha}(D, o)$, $\overline{\alpha}(D, o)$ in $\mathrm{CKhI}(D)$ are called the *equivariant Lee cycles* of D, and those homology classes the *equivariant Lee classes* of D.

Note that $\underline{\alpha}(D,o)$ has even homological grading whereas $\overline{\alpha}(D,o)$ has odd. In particular, $\underline{\alpha}(D), \underline{\beta}(D)$ have homological grading 0, and $\overline{\alpha}(D), \overline{\beta}(D)$ have 1.

Example 3.5 For the left-handed trefoil diagram of Figure 5, we have

$$\underline{\alpha}(D) = X \otimes Y \in \mathrm{CKhI}^0(D) \quad \text{and} \quad \overline{\alpha}(D) = Q(X \otimes Y) \in \mathrm{CKhI}^1(D),$$

where X corresponds to the outer circle, and Y to the inner circle. The other cycles $\underline{\beta}(D), \overline{\beta}(D)$ are obtained by swapping X and Y.

The following proposition states the relation between the ordinary Lee cycles $\{\alpha(D,o)\}$ and the equivariant Lee cycles $\{\underline{\alpha}(D,o), \overline{\alpha}(D,o)\}$.

Proposition 3.6 *In the short exact sequence*

$$\mathrm{CKh}(D)[1] \xhookrightarrow{Q} \mathrm{CKhI}(D) \overset{q}{\twoheadrightarrow} \mathrm{CKh}(D)$$

the ordinary and the equivariant Lee cycles correspond as

$$\alpha(D,o) \overset{Q}{\longmapsto} \overline{\alpha}(D,o) \quad \textit{and} \quad \underline{\alpha}(D,o) \overset{q}{\longmapsto} \alpha(D,o).$$

It is well known in the noninvolutive setting that the $\mathbb{Q}$*-Lee homology* is freely generated by the Lee classes [19, Theorem 4.2]. More generally, whenever $h \in R$ is invertible, then the corresponding homology is freely generated by the Lee classes [31, Theorem 4.2], [27, Proposition 2.9]. Analogous statement also hold in the involutive setting.

Proposition 3.7 *If $h \in R$ is invertible, the involutive Khovanov homology* $\mathrm{KhI}(D)$ *is freely generated by the* $2^{|D|+1}$ *equivariant Lee classes.*

Proof The proof is completely similar to the proof for the noninvolutive case given in [34], also explained in detail in [20]. To explain briefly, first note that when h is invertible, we may take $\{X, Y\}$ as a basis for A. An *admissible coloring* of D is a coloring with a or b on the edges of D such that each crossing admits a resolution that determines the colors of the two arc segments accordingly; see Figure 6. Now $\mathrm{CKh}(D)$ can be decomposed into subcomplexes, each corresponding to an admissible coloring of D, generated by the XY-labeled generators that match the coloring. If there is a crossing such that the four incident edges are colored the same (as in the first two pictures of Figure 6), then the generators can be canceled in pairs, resulting in a trivial complex. By contracting all such subcomplexes, we will be left with subcomplexes each corresponding to an admissible coloring such that the four incident edges at each crossing have two different colors (as in the third picture of Figure 6), which in turn corresponds one-to-one to an orientation o of D. Such subcomplex is generated by the single Lee cycle $\alpha(D,o)$, and thus $\mathrm{CKh}(D)$ is chain homotopy equivalent to a complex generated by the Lee cycles with trivial differential.

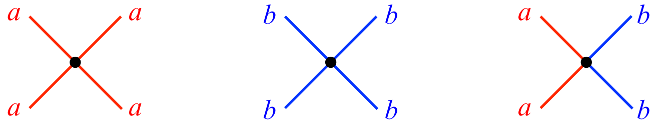


Figure 6: Local picture of an admissibly colored diagram.

This method also works in our case by performing the same cancellations in both $\mathrm{CKh}(D)$ and $Q\mathrm{CKh}(D)$. The only concern is that there are arrows

$$Q(1+\tau)\colon \mathrm{CKh}(D) \to Q\mathrm{CKh}(D)$$

in the differential of $\mathrm{CKhI}(D)$, but it will not affect the cancellation process, since the arrows are only running from the $\mathrm{CKh}(D)$ to $Q\mathrm{CKh}(D)$ and hence will not produce new arrows between the remaining canceling pairs. $\square$

In [27, Proposition 2.13] we showed in the noninvolutive setting that the behaviors of the Lee classes under the Reidemeister moves can be described explicitly. The same formula also holds in the involutive setting. Hereafter, $w(D)$ denotes the writhe of D, and $r(D)$ denotes the number of Seifert circles of D. The difference function δf of a unary function f is defined as $\delta f(x,y) = f(y) - f(x)$.

Proposition 3.8 *Suppose $h \in R$ is invertible. Let D, D' be strongly invertible link diagrams related by an involutive Reidemeister move. Under the isomorphism*

$$\rho\colon \mathrm{KhI}(D) \to \mathrm{KhI}(D')$$

given in Proposition 2.7, the equivariant Lee classes modulo torsions correspond as

$$\begin{aligned}
\underline{\alpha}(D) &\overset{\rho}{\longmapsto} h^j \underline{\alpha}(D'), & \overline{\alpha}(D) &\overset{\rho}{\longmapsto} h^j \overline{\alpha}(D'),\\
\underline{\beta}(D) &\overset{\rho}{\longmapsto} h^j \underline{\beta}(D'), & \overline{\beta}(D) &\overset{\rho}{\longmapsto} h^j \overline{\beta}(D'),
\end{aligned}$$

where

$$j = \frac{\delta w(D,D') - \delta r(D,D')}{2}.$$

Similar statements also hold for the other Lee classes, after appropriately reorienting D and D' with respect to the choice of the orientation o on D.

Proof From Proposition 3.7, the long exact sequence induced from the short exact sequence of Proposition 3.6 splits, and from Proposition 2.7 the isomorphism ρ fits into the following commutative diagram

$$\begin{tikzcd}
0 \arrow[r] & \mathrm{Kh}(D)[1] \arrow[r] \arrow[d, "\rho"] & \mathrm{KhI}(D) \arrow[r] \arrow[d, dashed, "\rho"] & \mathrm{Kh}(D) \arrow[r] \arrow[d, "\rho"] & 0 \\
0 \arrow[r] & \mathrm{Kh}(D')[1] \arrow[r] & \mathrm{KhI}(D') \arrow[r] & \mathrm{Kh}(D') \arrow[r] & 0
\end{tikzcd}$$

From [27, Proposition 2.13] and the fact that ρ preserves the homological grading, it follows that

$$\begin{tikzcd}
\alpha(D) \arrow[r, mapsto] \arrow[d, mapsto, "\rho"] & \overline{\alpha}(D) \arrow[d, mapsto, dashed, "\rho"] & \underline{\alpha}(D) \arrow[r, mapsto] \arrow[d, mapsto, dashed, "\rho"] & \alpha(D) \arrow[d, mapsto, "\rho"] \\
h^j\alpha(D') \arrow[r, mapsto] & h^j\overline{\alpha}(D') & h^j\underline{\alpha}(D') \arrow[r, mapsto] & h^j\alpha(D')
\end{tikzcd}$$

The proof for the β-classes is similar. □

Next, we consider the reduced setting. Hereafter, whenever the (co)reduced complexes are considered, it is implicitly assumed that the diagram is pointed and normal. Let $O(D)$ be the set of all orientations on D, and $O^+(D)$ be the subset of $O(D)$ consisting of orientations o whose orientation on the based component coincides with that of D.

First, we review the noninvolutive setting, which is extensively studied in [28]. For each $o \in O^+(D)$, the cycle $\alpha(D, o)$ lies in the reduced complex $\mathrm{CKh}_r(D) \subset \mathrm{CKh}(D)$. This will be denoted by $\alpha_r(D, o)$ for the sake of distinction. The counterpart $\beta(D, o)$ lies in $\sigma\mathrm{CKh}'_r(D) \subset \mathrm{CKh}(D)$, which is isomorphic to the coreduced complex $\mathrm{CKh}'_r(D)$. Let $\beta_r(D, o)$ denote the corresponding cycle in $\mathrm{CKh}'_r(D)$. $\beta_r(D, o)$ can be described by simply replacing the label of $\beta(D, o)$ on the pointed circle from $Y = \sigma(X)$ to 1.

Now we return to the involutive setting. The *(co)reduced equivariant Lee cycles* $\underline{\alpha}_r(D, o)$, $\overline{\alpha}_r(D, o) \in \mathrm{CKhI}_r(D)$ and $\underline{\beta}_r(D, o)$, $\overline{\beta}_r(D, o) \in \mathrm{CKhI}'_r(D)$ are defined in the same way. The following propositions can be easily verified.

Proposition 3.9 *Under the maps in the short exact sequence of Proposition 2.22, for each* $o \in O^+(D)$, *the Lee cycles in the unreduced, reduced, and coreduced complexes correspond as*

$$\underline{\alpha}(D, o) \xmapsto{i} \underline{\alpha}_r(D, o), \quad \underline{\beta}(D, o) \xmapsto{p} h\underline{\beta}_r(D, o),$$
$$\overline{\alpha}(D, o) \xmapsto{i} \overline{\alpha}_r(D, o), \quad \overline{\beta}(D, o) \xmapsto{p} h\overline{\beta}_r(D, o).$$

Proposition 3.10 *If* $h \in R$ *is invertible, then the reduced homology* $\mathrm{KhI}_r(D)$ *is freely generated by the* $2^{|D|}$ *reduced equivariant Lee classes* $\{\underline{\alpha}_r(D, o), \overline{\alpha}_r(D, o)\}_{o \in O^+(D)}$. *Similarly, the coreduced homology* $\mathrm{KhI}'_r(D)$ *is freely generated by the* $2^{|D|}$ *coreduced equivariant Lee classes* $\{\underline{\beta}_r(D, o), \overline{\beta}_r(D, o)\}_{o \in O^+(D)}$.

Proposition 3.11 *Suppose* $h \in R$ *is invertible. Let* D, D' *be two strongly invertible link diagrams (pointed and normal) related by an involutive Reidemeister move. Under the isomorphism given in Proposition 2.26*

$$\rho\colon \mathrm{KhI}_r(D) \to \mathrm{KhI}_r(D'),$$

the reduced equivariant Lee classes correspond as

$$\underline{\alpha}_r(D) \xmapsto{\rho} h^j\underline{\alpha}_r(D'), \quad \overline{\alpha}_r(D) \xmapsto{\rho} h^j\overline{\alpha}_r(D'), \qquad \text{where } j = \frac{\delta w(D, D') - \delta r(D, D')}{2}.$$

Similar statements hold for the coreduced counterparts.

Finally, we state the correspondence of the unreduced and the (co)reduced Lee cycles under the splitting of Proposition 3.12. First, we consider the noninvolutive setting.

Proposition 3.12 *Under the splitting of* [35]

$$\mathrm{CKh}(D) \cong \mathrm{CKh}'_r(D) \oplus \mathrm{CKh}_r(D),$$

for each $o \in O^+(D)$, the unreduced and the (co)reduced Lee cycles correspond as

$$\alpha(D,o) \mapsto (0, \alpha_r(D,o)), \quad \beta(D,o) \mapsto (h\beta_r(D,o), \alpha_r(D,o)).$$

Proof Here we only give a sketch. The isomorphism

$$\mathrm{CKh}(D) = \mathrm{Cone}(f) \xrightarrow{\cong} \mathrm{CKh}'_r(D) \oplus \mathrm{CKh}_r(D)$$

is given by

$$\begin{pmatrix} 1 & \\ \kappa & 1 \end{pmatrix}$$

using the null-homotopy κ described in Section 2.3. From this description, it is obvious that α maps to $(0,\alpha)^T$. To see that β maps to $(h\beta_r, \alpha)^T$, put

$$\beta = \underline{Y} \otimes x.$$

Here the underline indicates the label corresponding to the pointed circle. With the vector notation β is represented as $(h\underline{1} \otimes x, \underline{X} \otimes x)^T$. Note that $\underline{1} \otimes x$ is exactly β_r, so it remains to prove that

$$h\kappa(\underline{1} \otimes x) + \underline{X} \otimes x = \alpha_r.$$

If we define $\kappa_{-1} \colon \mathrm{CKh}'_r(D) \to \mathrm{CKh}_r(D)$ by

$$\kappa_{-1}(\underline{1} \otimes \cdots) = \underline{X} \otimes \cdots \quad \text{and} \quad \bar{\kappa} = \sum_{i \geq -1} h^{i+1}\kappa_i \colon \mathrm{CKh}'_r(D) \to \mathrm{CKh}_r(D),$$

then the aimed equation can be written as

$$\bar{\kappa}(\beta_r) = \alpha_r.$$

This is a purely algebraic problem and can be proved by the induction on the number of the Seifert circles. □

Proposition 3.13 *Under the splitting given in Proposition 2.23*

$$\mathrm{CKhI}(D) \cong \mathrm{CKhI}'_r(D) \oplus \mathrm{CKhI}_r(D),$$

the equivariant Lee cycles correspond as

$$\begin{aligned} \underline{\alpha}(D,o) &\mapsto (0, \underline{\alpha}_r(D,o)), & \underline{\beta}(D,o) &\mapsto (h\underline{\beta}_r(D,o), \underline{\alpha}_r(D,o)), \\ \bar{\alpha}(D,o) &\mapsto (0, \bar{\alpha}_r(D,o)), & \bar{\beta}(D,o) &\mapsto (h\bar{\beta}_r(D,o), \bar{\alpha}_r(D,o)), \end{aligned}$$

for each $o \in O^+(D)$.

Proof Immediate from Proposition 3.12. □

3.2 Divisibility of equivariant Lee classes

Hereafter we assume that R is a PID and h is prime (hence nonzero and noninvertible), typically $(R, h) = (\mathbb{F}_2[H], H)$.

Definition 3.14 Let M be a finitely generated free R-module. The *h-divisibility* of an element z in M is defined by

$$d_h(z) = \max\{k \geq 0 \mid z \in h^k M\}.$$

Divisibilities can be compared by homomorphisms. Suppose M, N are finitely generated free R-modules, $f : M \to N$ is a homomorphism, and $z \in M$, $w \in N$ are elements such that $f(z) = h^j w$ for some $j \in \mathbb{Z}$. Then we have

$$d_h(z) \leq j + d_h(w).$$

In particular when f is an isomorphism, the equality holds. Divisibilities can also be compared after inverting h. Namely, if z, w are elements in M such that $z \otimes 1 = h^j (w \otimes 1)$ in $h^{-1}M = M \otimes_R (h^{-1}R)$, then we have

$$d_h(z) = j + d_h(w).$$

See [28, Lemmas 4.2–4.7] for details.

Definition 3.15 For each orientation o on D, define nonnegative integers $\underline{d}_h(D, o)$ and $\overline{d}_h(D, o)$ by the *h-divisibility* (modulo torsion) of the equivariant Lee classes $\underline{\alpha}(D, o), \overline{\alpha}(D, o) \in \mathrm{KhI}(D)/\mathrm{Tor}$ respectively, ie

$$\underline{d}_h(D, o) = \max\{k \geq 0 \mid \underline{\alpha}(D, o) \in (h^k)(\mathrm{KhI}(D)/\mathrm{Tor})\},$$
$$\overline{d}_h(D, o) = \max\{k \geq 0 \mid \overline{\alpha}(D, o) \in (h^k)(\mathrm{KhI}(D)/\mathrm{Tor})\}.$$

Example 3.16 Consider the left-handed trefoil diagram of Figure 5. In Example 3.5 we had

$$\underline{\alpha}(D) = X \otimes Y \in \mathrm{CKhI}^0(D).$$

Now, consider the element x of homological grading -1 depicted in the left side of Figure 7. Note that x is τ-invariant and hence $(d + 1 + \tau)x = dx = X \otimes X$. Now

$$\underline{\alpha}(D) \sim X \otimes Y + X \otimes X = h(X \otimes 1),$$

and we have $\underline{d}_h(D) \geq 1$. In fact in this case the equality holds. Similarly we have $\overline{d}_h(D) = 1$.

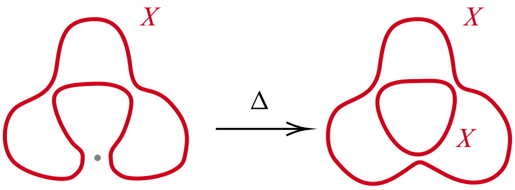


Figure 7: Elements x and dx.

In [27; 28] we defined a similar quantity in the noninvertible setting, namely for each orientation o on D,

$$d_h(D,o) = \max\{k \geq 0 \mid \alpha(D,o) \in (h^k)(\mathrm{Kh}(D)/\operatorname{Tor})\}.$$

Proposition 3.17 $$\underline{d}_h(D,o) \leq d_h(D,o) \leq \overline{d}_h(D,o).$$

Proof Immediate from Proposition 3.6. □

Proposition 3.18 $$\underline{d}_h(\bigcirc) = \overline{d}_h(\bigcirc) = 0.$$

Proposition 3.19 *If D is a positive diagram, then $\underline{d}_h(D) = \overline{d}_h(D) = 0$.*

Proof Take a sequence of involutive diagrams

$$D \to D_1 \to \cdots \to D_N$$

by symmetrically resolving positive crossings, so that the final diagram D_N is a disjoint union of symmetric circles. This induces a sequence of quotient maps

$$\mathrm{CKhI}(D) \to \mathrm{CKhI}(D_1) \to \cdots \to \mathrm{CKhI}(D_N)$$

and gives

$$0 \leq \underline{d}_h(D) \leq \cdots \leq \underline{d}_h(D_N) = 0,$$

and similarly $\overline{d}_h(D) = 0$. □

The following lemmas will be used in the coming sections, each stating that $\underline{d}_h(D,o), \overline{d}_h(D,o)$ can be described in several ways. Here, the h-divisibilities are considered modulo torsions.

Lemma 3.20 $$d_h(\underline{\alpha}(D,o)) = d_h(\underline{\beta}(D,o)), \quad d_h(\overline{\alpha}(D,o)) = d_h(\overline{\beta}(D,o)).$$

Proof Consider the automorphism σ. □

Lemma 3.21 $$\begin{aligned} d_h(\underline{\alpha}_r(D,o)) &= d_h(\underline{\beta}_r(D,o)) = \underline{d}_h(D,o), \\ d_h(\overline{\alpha}_r(D,o)) &= d_h(\overline{\beta}_r(D,o)) = \overline{d}_h(D,o). \end{aligned}$$

Proof Immediate from Proposition 3.13. □

Lemma 3.22 $$d_h(\underline{\alpha}(D,o) + \underline{\beta}(D,o)) = \underline{d}_h(D,o) + 1, \quad d_h(\overline{\alpha}(D,o) + \overline{\beta}(D,o)) = \overline{d}_h(D,o) + 1.$$

Proof Again immediate from Proposition 3.13. □

3.3 Definition of the equivariant invariants

Proposition 3.8 and Proposition 3.6 justifies the following definition.

Definition 3.23 For a strongly invertible link L with diagram D, define

$$\underline{s}_h(L) = 2\underline{d}_h(D) + w(D) - r(D) + 1 \quad \text{and} \quad \overline{s}_h(L) = 2\overline{d}_h(D) + w(D) - r(D) + 1.$$

The pair $(\underline{s}_h(L), \overline{s}_h(L))$ is called the *equivariant Rasmussen* invariant of L.

In [27; 28], for a (noninvolutive) link L with diagram D, the invariant $s_h(L)$ is defined as

$$s_h(L) = 2d_h(D) + w(D) - r(D) + 1.$$

Proposition 3.24
$$\underline{s}_h(L) \le s_h(L) \le \overline{s}_h(L).$$

Proof Immediate from Proposition 3.17. □

Corollary 3.25 *For a strongly invertible knot K, we have*

$$\underline{s}_h(K) \le s(K) \le \overline{s}_h(K),$$

where $s(K)$ is the $\mathbb{F}_2$-Rasmussen invariant of K.

Proof Proof that $s_h(K)$ coincides with $s(K)$ when $\operatorname{char}(R) = 2$ is in [28, Theorem 2, Proposition 4.36]. □

Proposition 3.26
$$\underline{s}_h(\bigcirc) = \overline{s}_h(\bigcirc) = 0.$$

Proposition 3.27 *If K is positive with positive diagram D, then*

$$\underline{s}_h(K) = \overline{s}_h(K) = n(D) - r(D) + 1,$$

where $n(D)$ is the number of crossings of D.

Corollary 3.28 *For the positive (p,q)-torus knot $T_{p,q}$,*

$$\underline{s}_h(T_{p,q}) = \overline{s}_h(T_{p,q}) = (p-1)(q-1)$$

with respect to the unique inverting involution τ of $T_{p,q}$.

Example 3.29 For the left-handed trefoil 3_1, we have $\underline{s}_h(K) = \overline{s}_h(K) = -2$ from the computations in Example 3.16. For the right-handed trefoil $m(3_1)$, we have $\underline{s}_h(m(3_1)) = \overline{s}_h(m(3_1)) = 2$ from Corollary 3.28.

4 Properties of the equivariant invariants

In this section properties of $(\underline{s}_h, \overline{s}_h)$ stated in Section 1 will be proved in more generality. Throughout, we assume that R is a PID and h is prime.

4.1 Mirror formula

Proposition 4.1 *Consider a strongly invertible link diagram D and its mirror D^*. With the perfect pairing of Proposition 2.31, for any $o \in O(D)$ and $o' \in O(D^*)$, we have*

$$\langle \underline{\alpha}(D,o), \overline{\alpha}(D^*,o') \rangle = \langle \overline{\alpha}(D,o), \underline{\alpha}(D^*,o') \rangle = \begin{cases} h^{r(D,o)} & \text{if } o' = o^*, \\ 0 & \text{otherwise,} \end{cases}$$

where $r(D,o)$ denotes the number of Seifert circles of D reoriented by o. Similarly for the reduced versions, we have

$$\langle \underline{\alpha}_r(D,o), \overline{\alpha}_r(D^*,o') \rangle_r = \langle \overline{\alpha}_r(D,o), \underline{\alpha}_r(D^*,o') \rangle_r = \begin{cases} h^{r(D,o)-1} & \text{if } o' = o^*, \\ 0 & \text{otherwise.} \end{cases}$$

Proof Obvious from the observation that the Seifert circles of D and D^* are identical, together with $\langle X, X \rangle = \langle Y, Y \rangle = h$ and $\langle X, Y \rangle = 0$. □

Proposition 4.2 *For a strongly invertible knot diagram D which is also pointed and normal, we have*

$$\underline{d}_h(D) + \overline{d}_h(D^*) = \overline{d}_h(D) + \underline{d}_h(D^*) = r(D) - 1.$$

Proof The formula is proved using the reduced Lee classes. Take a generator z of $\mathrm{KhI}^0_r(D) \cong R$ and put

$$\underline{\alpha}_r(D) = a h^d z,$$

where $d = \underline{d}_h(D)$ and $h \nmid a$. Similarly take a generator w of $\mathrm{KhI}^1_r(D^*) \cong R[1]$ and put

$$\overline{\alpha}_r(D^*) = b h^{d'} w,$$

where $d = \overline{d}_h(D^*)$ and $h \nmid b$. The perfect pairing of Proposition 2.31 induces a perfect pairing

$$\langle \cdot, \cdot \rangle_r : (\mathrm{KhI}_r(D)/\operatorname{Tor}) \otimes (\mathrm{KhI}_r(D^*)/\operatorname{Tor}) \to R,$$

and from Proposition 4.1 we have

$$\langle \underline{\alpha}_r(D), \overline{\alpha}_r(D^*) \rangle_r = ab h^{d+d'} \langle z, w \rangle_r = h^{r(D)-1}.$$

Now $\langle z, w \rangle_r$ must be a unit of R, and since h is assumed to be prime, we must have that $a, b \in R$ are both units and

$$\underline{d}_h(D) + \overline{d}_h(D^*) = r(D) - 1.$$

The other equation follows from a similar argument. □

Proposition 4.3 *For a strongly invertible knot K,*

$$\underline{s}_h(K^*) = -\overline{s}_h(K).$$

Proof Immediate from Proposition 4.2. □

Example 4.4 Compare Example 3.29.

The proof of Proposition 4.2 also shows that the following two elements

$$\underline{\xi}_r(D) = h^{-\underline{d}_h(D)} \underline{\alpha}_r(D) \quad \text{and} \quad \overline{\xi}_r(D) = h^{-\overline{d}_h(D)} \overline{\alpha}_r(D)$$

form a basis of $\mathrm{KhI}_r(D)/\operatorname{Tor} \cong R \oplus R[1]$. Similarly,

$$\underline{\zeta}_r(D) = h^{-\underline{d}_h(D)} \underline{\beta}_r(D) \quad \text{and} \quad \overline{\zeta}_r(D) = h^{-\overline{d}_h(D)} \overline{\beta}_r(D)$$

form a basis of $\mathrm{KhI}'_r(D)/\operatorname{Tor} \cong R \oplus R[1]$. Under the identification of Proposition 3.13, the elements corresponding to $\underline{\xi}_r(D), \overline{\xi}_r(D)$ are

$$\underline{\xi}(D) = h^{-\underline{d}_h(D)} \underline{\alpha}(D) \quad \text{and} \quad \overline{\xi}(D) = h^{-\overline{d}_h(D)} \overline{\alpha}(D),$$

and the elements corresponding to $\underline{\zeta}_r(D), \overline{\zeta}_r(D)$ are

$$\underline{\zeta}(D) = h^{-\underline{d}_h(D)-1}(\underline{\alpha}(D) + \underline{\beta}(D)) \quad \text{and} \quad \overline{\zeta}(D) = h^{-\overline{d}_h(D)-1}(\overline{\alpha}(D) + \overline{\beta}(D)).$$

Thus the four elements $\underline{\zeta}(D)$, $\underline{\xi}(D)$, $\overline{\zeta}(D)$ and $\overline{\xi}(D)$ form a basis of $\mathrm{KhI}(D)/\operatorname{Tor} \cong R^2 \oplus R[1]^2$. Propositions 3.8 and 3.11 imply that all of these classes are invariant under the Reidemeister moves.

In particular when R is graded and $\deg(h) = -2$, we see that

$$\underline{s}_h(K) = \mathrm{gr}_q(\underline{\xi}_r(K)) \quad \text{and} \quad \overline{s}_h(K) = \mathrm{gr}_q(\overline{\xi}_r(K)).$$

Thus when $(R, h) = (\mathbb{F}_2[H], H)$ the definition of $(\underline{s}, \overline{s})$ for strongly invertible knots given in Section 1 coincides with Definition 3.23. We summarize:

Proposition 4.5

$$\begin{aligned} \mathrm{KhI}(K) &= R\langle \underline{\zeta}(K), \underline{\xi}(K), \overline{\zeta}(K), \overline{\xi}(K) \rangle \oplus (\mathrm{Tor}), \\ \mathrm{KhI}_r(K) &= R\langle \underline{\xi}_r(K), \overline{\xi}_r(K) \rangle \oplus (\mathrm{Tor}), \\ \mathrm{KhI}'_r(K) &= R\langle \underline{\zeta}_r(K), \overline{\zeta}_r(K) \rangle \oplus (\mathrm{Tor}). \end{aligned}$$

Example 4.6 For the simplest example $D = \bigcirc$, we have

$$\begin{aligned} \underline{\alpha}(D) &= X, & \underline{\beta}(D) &= Y, \\ \overline{\alpha}(D) &= QX, & \overline{\beta}(D) &= QY, \end{aligned}$$

and $\underline{d}_h(D) = \bar{d}_h(D) = 0$, so

$$\begin{aligned} \underline{\xi}(D) &= X, & \underline{\zeta}(D) &= 1, \\ \bar{\xi}(D) &= QX, & \bar{\zeta}(D) &= Q1, \end{aligned}$$

and we obtain

$$\mathrm{KhI}(D) = R\langle 1, X, Q1, QX\rangle = A \oplus QA,$$

as expected.

4.2 Connected sum formula

Arguments in this section are inspired by [15], where the connected sum formula for the $\underline{d}$-, $\bar{d}$-invariants in involutive Heegaard Floer homology is proved.

For strongly invertible links L, L', the (equivariant) *disjoint union* $L \sqcup L'$ and the (equivariant) *connected sum* $L \#_b L'$ along an equivariant band b are defined in the obvious ways so that the resulting links are also strongly invertible. Note that different choices of b will in general give nonequivalent links, but here we make the choice implicit and omit b from the notation.[2] The corresponding operations for strongly invertible link diagrams D, D' are also defined. When we write $D \sqcup D'$, it is assumed that D and D' are disjoint as diagrams. When we write $D \# D'$, it is assumed that the band is untwisted and no crossings are produced by the surgery.

Proposition 4.7 *There is a canonical isomorphism*

$$\mathrm{CKhI}(D \sqcup D') \cong \mathrm{Cone}(\,\mathrm{CKh}(D) \otimes \mathrm{CKh}(D') \xrightarrow{Q(1\otimes 1+\tau\otimes\tau)} Q(\mathrm{CKh}(D) \otimes \mathrm{CKh}(D'))\,).$$

Under this identification, we have

$$\underline{\alpha}(D \sqcup D') = \alpha(D) \otimes \alpha(D'), \quad \bar{\alpha}(D \sqcup D') = Q(\alpha(D) \otimes \alpha(D')).$$

Proof Obvious from the canonical isomorphism

$$\mathrm{CKh}(D \sqcup D') \cong \mathrm{CKh}(D) \otimes \mathrm{CKh}(D'),$$

which holds for any (R, h); see [17, Section 7.4]. □

Proposition 4.8 *There are chain maps*

$$\begin{tikzcd} \mathrm{CKhI}(D \sqcup D') \arrow[r, shift left, "m"] & \mathrm{CKhI}(D \# D') \arrow[l, shift left, "\Delta"] \end{tikzcd}$$

corresponding to the band surgery from $D \sqcup D'$ *to* $D \# D'$ *and its reverse.*

Proof The corresponding maps in the noninvolutive setting are τ-invariant. □

---

[2] For *directed* strongly invertible knots K, K', there is a canonical choice of the band b from K to K', and the equivariant connected sum is defined without ambiguity. See [26].

Proposition 4.9 *For strongly invertible links L, L', we have*

$$\underline{s}_h(L \# L') - 1 \le \underline{s}_h(L \sqcup L') \le \underline{s}_h(L \# L') + 1,$$

and similarly for $\overline{s}_h$.

Proof Under the maps of Proposition 4.8, we have

$$m(\alpha(D) \otimes \alpha(D')) = h\alpha(D \# D') \quad \text{and} \quad \Delta\alpha(D \# D') = \alpha(D) \otimes \alpha(D'),$$

hence

$$\underline{d}_h(D \# D') \le \underline{d}_h(D \sqcup D') \le \underline{d}_h(D \# D') + 1.$$

This gives the desired inequality. □

Lemma 4.10 *Suppose C, C' are τ-complexes over $\mathbb{F}_2$. Put*

$$C_\tau = \operatorname{Cone}(C \xrightarrow{1+\tau} C), \quad C'_\tau = \operatorname{Cone}(C' \xrightarrow{1+\tau} C') \quad \text{and} \quad C^{\otimes}_\tau = \operatorname{Cone}(C \otimes C' \xrightarrow{1\otimes 1+\tau\otimes\tau} C \otimes C').$$

Let $z \in C$ and $z' \in C'$ be τ-invariant cycles. In the following, $\sim$ denotes homologous.

(1) *Let $x, y \in C$ and $x', y' \in C'$ be elements such that*

$$\begin{pmatrix} z \\ 0 \end{pmatrix} \sim \begin{pmatrix} x \\ y \end{pmatrix}, \quad \begin{pmatrix} z' \\ 0 \end{pmatrix} \sim \begin{pmatrix} x' \\ y' \end{pmatrix}$$

in C_τ and in C'_τ respectively. Then in $C^{\otimes}_\tau$,

$$\begin{pmatrix} z \otimes z' \\ 0 \end{pmatrix} \sim \begin{pmatrix} x \otimes x' \\ x \otimes y' + y \otimes \tau x \end{pmatrix}.$$

(2) *Let $x, y \in C$ and $x', y' \in C'$ be elements such that*

$$\begin{pmatrix} z \\ 0 \end{pmatrix} \sim \begin{pmatrix} x \\ y \end{pmatrix}, \quad \begin{pmatrix} 0 \\ z' \end{pmatrix} \sim \begin{pmatrix} x' \\ y' \end{pmatrix}$$

in C_τ and in C'_τ respectively. Then in $C^{\otimes}_\tau$,

$$\begin{pmatrix} 0 \\ z \otimes z' \end{pmatrix} \sim \begin{pmatrix} x \otimes x' \\ y \otimes x' + \tau x \otimes y' \end{pmatrix}.$$

Proof (1) Put

$$\begin{pmatrix} z \\ 0 \end{pmatrix} - \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} d & \\ 1+\tau & d \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} z' \\ 0 \end{pmatrix} - \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} d & \\ 1+\tau & d \end{pmatrix} \begin{pmatrix} a' \\ b' \end{pmatrix}.$$

Then the boundary of

$$\begin{pmatrix} x \otimes a' + a \otimes x' + a \otimes da' \\ x \otimes b' + b \otimes \tau x' + (1+\tau) a \otimes \tau a' \end{pmatrix}$$

gives the desired relation.

(2) Put

$$\begin{pmatrix} z \\ 0 \end{pmatrix} - \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} d & \\ 1+\tau & d \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ z' \end{pmatrix} - \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} d & \\ 1+\tau & d \end{pmatrix} \begin{pmatrix} a' \\ b' \end{pmatrix}.$$

Then the boundary of

$$\begin{pmatrix} x \otimes a' \\ x \otimes b' + \tau a \otimes y' + (1 \otimes 1 + \tau \otimes \tau)(a \otimes a') + da \otimes b' + b \otimes da' \end{pmatrix}$$

gives the desired relation. □

Proposition 4.11 *For strongly invertible knots K, K',*

$$\underline{s}_h(K) + \underline{s}_h(K') \le \underline{s}_h(K \# K') \le \underline{s}_h(K) + \overline{s}_h(K') \le \overline{s}_h(K \# K') \le \overline{s}_h(K) + \overline{s}_h(K').$$

Proof We will prove the first and the third inequalities, which is sufficient to prove Proposition 4.11 from the mirror formula Proposition 4.3. Take $z = \alpha(D) + \beta(D)$ in $\mathrm{CKh}(D)$ and $z' = \alpha(D') + \beta(D')$ in $\mathrm{CKh}(D')$. Observe that under the chain map m of Proposition 4.8, we have

$$\begin{aligned} \begin{pmatrix} z \otimes z' \\ 0 \end{pmatrix} &\overset{m}{\longmapsto} h \begin{pmatrix} \alpha(D \# D') + \beta(D \# D') \\ 0 \end{pmatrix} = h(\underline{\alpha}(D \# D') + \underline{\beta}(D \# D')), \\ \begin{pmatrix} 0 \\ z \otimes z' \end{pmatrix} &\overset{m}{\longmapsto} h \begin{pmatrix} 0 \\ \alpha(D \# D') + \beta(D \# D') \end{pmatrix} = h(\overline{\alpha}(D \# D') + \overline{\beta}(D \# D')). \end{aligned}$$

From Lemma 3.22, there are elements $x, y \in \mathrm{CKh}(D)$ such that,

$$\begin{pmatrix} z \\ 0 \end{pmatrix} \sim h^{\underline{d}_h(D)+1} \begin{pmatrix} x \\ y \end{pmatrix}$$

in $\mathrm{CKhI}(D)$, modulo torsion in homology. By inverting h, we may assume that they are strictly homologous in $h^{-1}\mathrm{CKhI}(D)$. Similarly, there are elements $x', y' \in \mathrm{CKh}(D')$ such that

$$\begin{pmatrix} z' \\ 0 \end{pmatrix} \sim h^{\underline{d}_h(D')+1} \begin{pmatrix} x' \\ y' \end{pmatrix}$$

in $h^{-1}\mathrm{CKhI}(D')$. From Lemma 4.10 (1), we have

$$\begin{pmatrix} z \otimes z' \\ 0 \end{pmatrix} \sim h^{\underline{d}_h(D)+\underline{d}_h(D')+2} \begin{pmatrix} x \otimes x' \\ x \otimes y' + y \otimes \tau x \end{pmatrix}$$

in $h^{-1}\mathrm{CKhI}(D \sqcup D')$. Under the map m of Proposition 4.8, the left-hand side maps to

$$h(\underline{\alpha}(D \# D') + \underline{\beta}(D \# D'))$$

in $h^{-1}\mathrm{CKhI}(D \# D')$. Its homology class in $h^{-1}\mathrm{KhI}(D \# D')$ is the $h^{\underline{d}_h(D \# D')+2}$ multiple of the class $\underline{\zeta}(D \# D')$ of Proposition 4.5. The homology class of

$$m \begin{pmatrix} x \otimes x' \\ x \otimes y' + y \otimes \tau x \end{pmatrix}$$

is also some h^e multiple of $\underline{\zeta}(D \# D')$ for some $e \geq 0$ (because x, y, x' and y' were taken before inverting h). Thus it follows that

$$\underline{d}_h(D) + \underline{d}_h(D') \leq \underline{d}_h(D \# D').$$

This implies the first inequality. Similarly for the third inequality, there are elements $x'', y'' \in \mathrm{CKh}(D')$ such that

$$\begin{pmatrix} 0 \\ z' \end{pmatrix} \sim h^{\bar{d}_h(D')+1} \begin{pmatrix} x'' \\ y'' \end{pmatrix}$$

in $h^{-1}\mathrm{CKhI}(D')$. From Lemma 4.10 (2), we have

$$\begin{pmatrix} 0 \\ z \otimes z' \end{pmatrix} \sim h^{\underline{d}_h(D)+\bar{d}_h(D')+2} \begin{pmatrix} x \otimes x'' \\ y \otimes x'' + \tau x \otimes y'' \end{pmatrix}.$$

By a similar argument, we obtain

$$\underline{d}_h(D) + \bar{d}_h(D') \leq \bar{d}_h(D \# D')$$

which implies the third inequality. □

4.3 Behavior under crossing changes

Proposition 4.12 *Let D^+ be a link diagram with at least one positive crossing, and D^- be a diagram obtained from D^+ by applying a negative crossing change to one of the positive crossings of D^+. There are homological grading preserving chain maps*

$$\begin{tikzcd} \mathrm{CKh}(D^+) \arrow[r, shift left, "\Phi^-"] & \mathrm{CKh}(D^-) \arrow[l, shift left, "\Phi^+"] \end{tikzcd}$$

such that the Lee cycles correspond as

$$\alpha(D^+) \overset{\Phi^-}{\longmapsto} \alpha(D^-), \quad \alpha(D^-) \overset{\Phi^+}{\longmapsto} h\alpha(D^+).$$

Proof Let x be the positive crossing of D^+ on which the crossing change is performed. Let D_0, D_1 be the 0-, 1-resolved diagram of D^+ at x respectively. Then $\mathrm{CKh}(D^+)$ may be described as a cone of the surgery map

$$\mathrm{CKh}(D_0) \xrightarrow{e} \mathrm{CKh}(D_1).$$

Similarly, by 1-resolving D^- at the corresponding crossing, we see that $\mathrm{CKh}(D^-)$ can be described as a cone of

$$\mathrm{CKh}(D_1) \xrightarrow{e'} \mathrm{CKh}(D_0).$$

The setup is depicted in Figure 8 where the red arcs indicate the corresponding surgery maps. Note that $\alpha(D^+)$ and $\alpha(D^-)$ are identical, and that they both belong to $\mathrm{CKh}(D_0)$.

Now we define chain maps

$$\begin{tikzcd} \mathrm{CKh}(D^+) \arrow[r, shift left, "\Phi^-"] & \mathrm{CKh}(D^-) \arrow[l, shift left, "\Phi^+"] \end{tikzcd}$$

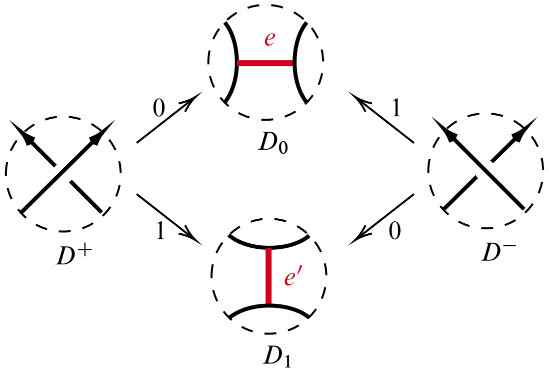


Figure 8: Diagrams $D^{\pm}$ and their resolutions.

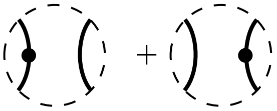
so that they fit into the following commutative diagram

$$\begin{tikzcd}
0 \arrow[r] \arrow[d, dashed, leftrightarrow, "0"] & \mathrm{CKh}(D_0) \arrow[r, "e"] \arrow[d, dashed, shift right, "\Phi^-"'] & \mathrm{CKh}(D_1) \arrow[d, dashed, leftrightarrow, "0"] \\
\mathrm{CKh}(D_1) \arrow[r, "e'"] & \mathrm{CKh}(D_0) \arrow[r] \arrow[u, dashed, shift right, "\Phi^+"'] & 0
\end{tikzcd}$$

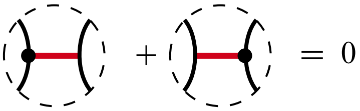
giving the desired chain maps between the complexes. First, define $\Phi^- = \mathrm{id}_{D_0}$, which is obviously a chain map satisfying

$$\alpha(D^+) \overset{\Phi^-}{\longmapsto} \alpha(D^-).$$

Next, we define Φ^+ as



where each dot represents the multiplication by X on the circle it is drawn on. To verify that Φ^+ is a chain map, it suffices to show that $e\Phi^+ = 0 = \Phi^+ e'$. The first equation can be described pictorially as



which obviously holds, since dots can move freely within their connected components. The second equation can be proved similarly. Finally, we see that

$$\alpha(D^-) \overset{\Phi^+}{\longmapsto} h\alpha(D^+)$$

from the local description of $\alpha(D^+)$ together with $X^2 = hX$ and $XY = 0$. □

Remark 4.13 The crossing change maps $\Phi^{\pm}$ of Proposition 4.12 partially appear in [4, Figures 3,4]. Φ^+ also appears in [16, Section 3] in the form of a morphism in the category Cob^3.

Proposition 4.14 *Let L^+, L^- be links such that L^- is obtained by applying a negative crossing change to L^+. Then*

$$s_h(L^-) \le s_h(L^+) \le s_h(L^-) + 2.$$

Proof Let D^+, D^- be diagrams of L^+, L^- respectively, such that D^- is obtained by applying a negative crossing change to a single crossing of D^+. From Proposition 4.12, we have

$$d_h(D^+) \le d_h(D^-) \le d_h(D^+) + 1.$$

Thus,

$$\begin{aligned} s_h(L^-) &= 2d_h(D^-) + w(D^-) - r(D^-) + 1 \\ &\le 2(d_h(D^+) + 1) + (w(D^+) - 2) - r(D^+) + 1 = s_h(L) \\ &\le 2d_h(D^-) + (w(D^-) + 2) - r(D^-) + 1 = s_h(L^-) + 2. \end{aligned}$$ □

Definition 4.15 An *equivariant negative crossing change* on a strongly invertible link L is an operation that is either a single negative crossing change on a crossing lying on the axis of L, or two negative crossing changes on crossings x and $\tau(x)$ lying off the axis.

Proposition 4.16 *Let L^+, L^- be strongly invertible links such that L^- is obtained by applying an equivariant negative crossing change to L^+. Then*

$$\underline{s}_h(L^-) \le \underline{s}_h(L^+) \le \underline{s}_h(L^-) + 2a,$$

where $a = 1$ if the move is performed on-axis, and $a = 2$ if performed off-axis. The same holds for $\overline{s}_h$.

Proof If the move is performed on-axis, the maps $\Phi^\pm$ are strictly τ-equivariant and hence induce maps between the involutive complexes. If the move is performed off-axis, then we may define equivariant crossing change maps

$$\Phi^\pm = \Phi_2^\pm \Phi_1^\pm,$$

where $\Phi_1^\pm$ and $\Phi_2^\pm$ are the nonequivariant crossing change maps corresponding to the off-axis moves. By an argument similar to the proof of Proposition 2.7 for moves IR1–IR3, we see that $\Phi^\pm$ is strictly τ-equivariant and hence induce maps between the involutive complexes. □

4.4 Behavior under generalized crossing changes

Next, we extend the results in the previous section to *generalized crossing changes*, introduced by Cochran and Tweedy in [8].

Definition 4.17 For $n \ge 1$, a $2n$*-strand generalized negative* (*resp. positive*) *crossing change* on a link L is a modification of L by adding a *positive* (resp. *negative*) full twist on $2n$ parallel strands of L, where n strands are oriented one way and the other are oriented the other; see Figure 9

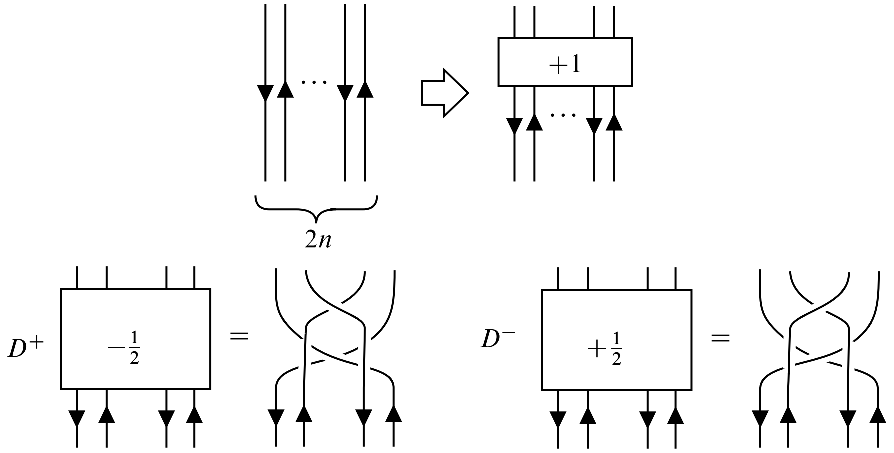


Figure 9: Top: a $2n$-strand generalized negative crossing change. Bottom: diagrams D^+ and D^-.

Note that the case $n=1$ gives the ordinary negative (resp. positive) crossing change. Hereafter we only consider the case $n=2$.

Proposition 4.18 *Let D^+, D^- be diagrams that differ locally as in Figure 9. Then there are homological grading preserving chain maps*

$$\mathrm{CKh}(D^+) \underset{\Phi^+}{\overset{\Phi^-}{\rightleftarrows}} \mathrm{CKh}(D^-)$$

such that the Lee classes modulo torsions correspond as

$$\alpha(D^+) \overset{\Phi^-}{\longmapsto} \alpha(D^-), \quad \alpha(D^-) \overset{\Phi^+}{\longmapsto} h^2\alpha(D^+).$$

The idea of the proof is similar to that of Proposition 4.12 but we need some preparations. First recall from [5, Section 4] the definition of *strong deformation retracts*.

Definition 4.19 A chain map $r : C \to C'$ between chain complexes (in any additive category) is a *strong deformation retract*[3] if there is a chain map $i : C' \to C$ (called the *inclusion*) and a homotopy h on C, satisfying

(i) $ri = 1$,

(ii) $ir - 1 = dh + hd$,

(iii) $hi = 0$,

(iv) $rh = 0$, and

(v) $h^2 = 0$.

The following lemma is a generalization of [5, Lemma 4.5].

---

[3] Conditions (iii)–(v) are called the *side conditions*. In [5, Definition 4.3] only conditions (i)–(iii) are imposed, but we can easily check that the remaining two conditions also hold for the R1 and the R2 maps.

Lemma 4.20 *Suppose X, Y, Z, W are chain complexes (in any additive category), and there are maps f, g, k, l (not necessarily chain maps) such that the square of complexes*

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "k"'] & Y \arrow[d, "g"] \\
Z \arrow[r, "l"] & W
\end{tikzcd}$$

forms a chain complex C. Furthermore suppose Y has a strong deformation retract $r\colon Y \to Y'$ with inclusion i and homotopy h. Then the square of complexes

$$\begin{tikzcd}
X \arrow[r, "rf"] \arrow[d, "k"'] \arrow[dr, "ghf"] & Y' \arrow[d, "gi"] \\
Z \arrow[r, "l"] & W
\end{tikzcd}$$

forms a chain complex C' and is a strong deformation retract of C.

Proof The differentials of C and C' may be expressed by matrices

$$D = \begin{pmatrix} d_X & & & \\ f & d_Y & & \\ k & & d_Z & \\ & g & l & d_W \end{pmatrix}, \quad D' = \begin{pmatrix} d_X & & & \\ rf & d_{Y'} & & \\ k & & d_Z & \\ ghf & gi & l & d_W \end{pmatrix},$$

and one can see that $D^2 = 0$ implies $(D')^2 = 0$. Furthermore, one can check that

$$R = \begin{pmatrix} 1 & & & \\ & r & & \\ & & 1 & \\ & gh & & 1 \end{pmatrix}, \quad I = \begin{pmatrix} 1 & & & \\ hf & i & & \\ & & 1 & \\ & & & 1 \end{pmatrix}, \quad H = \begin{pmatrix} 0 & & & \\ & h & & \\ & & 0 & \\ & & & 0 \end{pmatrix},$$

gives a strong deformation retract $R\colon C \to C'$ with inclusion I and homotopy H. □

Lemma 4.21 *Let D^+, D^- be diagrams that differ locally as in Figure 9. The complex $\mathrm{CKh}(D^+)$ strongly deformation retracts onto the complex E described in Figure 10. Similarly, the complex $\mathrm{CKh}(D^-)$ strongly deformation retracts onto the complex E^- described in Figure 10. (Descriptions for some arrows are omitted, since we will not use them.)*

Proof By considering the 0 and 1 resolutions of the bottom left and the bottom right crossings of D^+, we see that the complex $\mathrm{CKh}(D^+)$ can be expressed as a square of complexes

$$\begin{tikzcd}
\mathrm{CKh}(D^+_{00}) \arrow[r, "e"] \arrow[d, "e'"] & \mathrm{CKh}(D^+_{10}) \arrow[d, "e'"] \\
\mathrm{CKh}(D^+_{01}) \arrow[r, "e"] & \mathrm{CKh}(D^+_{11})
\end{tikzcd}$$

where e and e' denote the surgery map corresponding to the bottom left and the bottom right crossings, respectively. Observe that the resolved diagrams D^+_{00}, D^+_{10}, D^+_{01} can be simplified by performing R2

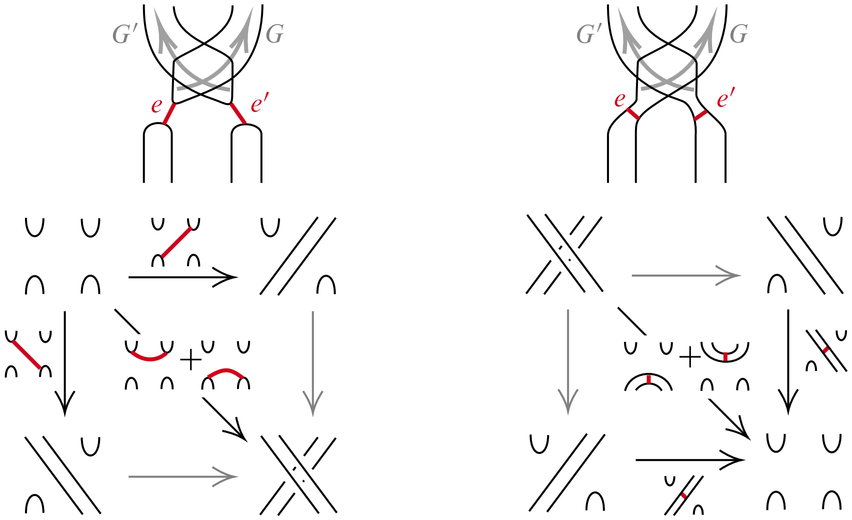


Figure 10: Simplifying $\mathrm{CKh}(D^+)$ and $\mathrm{CKh}(D^-)$. Top: D^+, left, and D^-, right. Bottom: E^+, left, and E^-, right.

moves; see Figure 10. Since the map G for the R2-move defined in [5, Section 4.2] is a strong deformation retract, we may apply Lemma 4.20 repeatedly and obtain a strong deformation retract E^+ of $\mathrm{CKh}(D^+)$ of the form

$$\begin{tikzcd}
E^+_{00} \arrow[r, "G'eF"] \arrow[d, "e'"] \arrow[dr, "e'H'eF"] & E^+_{10} \arrow[d, "e'F'"] \\
E^+_{01} \arrow[r, "eF"] & E^+_{11}
\end{tikzcd}$$

Here, G and G' are the maps for the R2-moves, F, F' and H, H' are the corresponding inclusions and homotopies. By unraveling the explicit maps, one can check that the maps are given as in Figure 10. Similarly $\mathrm{CKh}(D^-)$ can be expressed as a square of complexes

$$\begin{tikzcd}
\mathrm{CKh}(D^-_{00}) \arrow[r, "e"] \arrow[d, "e'"] & \mathrm{CKh}(D^-_{10}) \arrow[d, "e'"] \\
\mathrm{CKh}(D^-_{01}) \arrow[r, "e"] & \mathrm{CKh}(D^-_{11})
\end{tikzcd}$$

and the resolved diagrams D^-_{10}, D^-_{01}, D^-_{11} can be simplified by performing R2 moves. Thus we obtain a strong deformation retract E^- of $\mathrm{CKh}(D^-)$ of the form

$$\begin{tikzcd}
E^-_{00} \arrow[r, "Ge"] \arrow[d, "G'e'"] \arrow[dr, "GeH'e'"] & E^-_{10} \arrow[d, "e'"] \\
E^-_{01} \arrow[r, "GeF'"] & E^-_{11}
\end{tikzcd}$$

and again one can check that the maps are given as in Figure 10. □

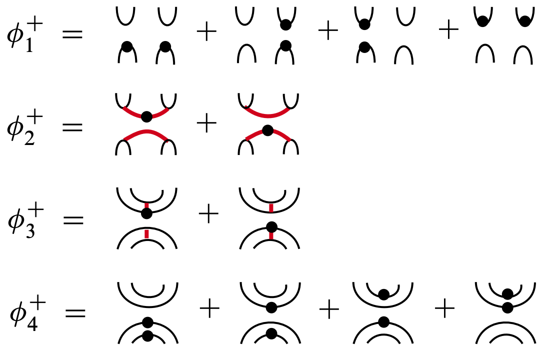
Proof of Proposition 4.18 First we define chain maps

$$\begin{tikzcd} E^+ \arrow[r, shift left, "\phi^-"] & E^- \arrow[l, shift left, "\phi^+"] \end{tikzcd}$$

between the two simplified complexes E^+, E^- of Lemma 4.21. For ϕ^-, since E^+_{00} and E^-_{11} are identical, we may define $\phi^- = \mathrm{id}$ on E^+_{00} and 0 elsewhere. Obviously this is a chain map, since we have $\phi^- d^+ = 0 = d^- \phi^-$. Next, ϕ^+ is defined by the sum of the following four maps $\phi^+_1, \phi^+_2, \phi^+_3, \phi^+_4$:

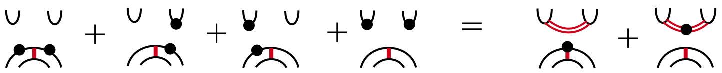
$$\begin{tikzcd}
 & & \underline{E^+_{00}} \arrow[r] \arrow[dr] \arrow[drr] & E^+_{10} \arrow[dr] & \\
 & & & E^+_{01} \arrow[r] & E^+_{11} \\
E^-_{00} \arrow[r] \arrow[dr] \arrow[drr] \arrow[uurr, dashed, "\phi^+_3"] \arrow[urrrr, dashed, "\phi^+_4"] & E^-_{10} \arrow[dr] & & & \\
 & E^-_{01} \arrow[r] & \underline{E^-_{11}} \arrow[uuu, dashed, "\phi^+_1"] \arrow[uurr, dashed, "\phi^+_2"'] & &
\end{tikzcd}$$

with explicit descriptions:



Note that E^+_{00} and E^-_{11} only have homological grading 0 (within the displayed area), whereas E^+_{11} has homological grading in range $[0, 4]$ and E^-_{00} has homological grading in range $[-4, 0]$. One can see that each ϕ^+_i has domain and codomain in the homological grading 0 parts.

To show that $\phi^+ = \sum_i \phi^+_i$ actually defines a chain map, it suffices to verify that all the diagrams displayed in Figure 11 commute. Here each number in the parenthesis indicates the homological grading. This can be checked by the explicit descriptions of the maps ϕ^+_i and the differentials of E^+, E^- given in Figure 10. For example, the commutativity of the third diagram is equivalent to



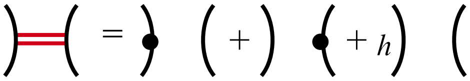
$$\begin{tikzcd}
E^-_{10}(-1) \arrow[r, "d"] \arrow[d] & E^-_{11}(0) \arrow[d, "\phi_1^+"] \\
0 \arrow[r] & E^+_{00}(0)
\end{tikzcd}
\quad
\begin{tikzcd}
E^-_{01}(-1) \arrow[r, "d"] \arrow[d] & E^-_{11}(0) \arrow[d, "\phi_1^+"] \\
0 \arrow[r] & E^+_{00}(0)
\end{tikzcd}
\quad
\begin{tikzcd}
E^-_{00}(-1) \arrow[r, "d"] \arrow[d, "d"] & E^-_{11}(0) \arrow[d, "\phi_1^+"] \\
E^-_{00}(0) \arrow[r, "\phi_3^+"] & E^+_{00}(0)
\end{tikzcd}
\quad
\begin{tikzcd}
E^-_{00}(-1) \arrow[r, "d"] \arrow[d, "d"] & E^-_{11}(0) \arrow[d, "\phi_2^+"] \\
E^-_{00}(0) \arrow[r, "\phi_4^+"] & E^+_{11}(0)
\end{tikzcd}$$

$$\begin{tikzcd}
E^-_{00}(0) \arrow[r, "\phi_3^+"] \arrow[d] & E^+_{00}(0) \arrow[d, "d"] \\
0 \arrow[r] & E^+_{10}(1)
\end{tikzcd}
\quad
\begin{tikzcd}
E^-_{00}(0) \arrow[r, "\phi_3^+"] \arrow[d] & E^+_{00}(0) \arrow[d, "d"] \\
0 \arrow[r] & E^+_{01}(1)
\end{tikzcd}
\quad
\begin{tikzcd}
E^-_{00}(0) \arrow[r, "\phi_3^+"] \arrow[d, "\phi_4^+"] & E^+_{00}(0) \arrow[d, "d"] \\
E^+_{11}(0) \arrow[r, "d"] & E^+_{11}(1)
\end{tikzcd}
\quad
\begin{tikzcd}
E^-_{11}(0) \arrow[r, "\phi_1^+"] \arrow[d] & E^+_{00}(0) \arrow[d, "d"] \\
0 \arrow[r] & E^+_{10}(1)
\end{tikzcd}$$

$$\begin{tikzcd}
E^-_{11}(0) \arrow[r, "\phi_1^+"] \arrow[d] & E^+_{00}(0) \arrow[d, "d"] \\
0 \arrow[r] & E^+_{01}(1)
\end{tikzcd}
\quad
\begin{tikzcd}
E^-_{11}(0) \arrow[r, "\phi_1^+"] \arrow[d, "\phi_2^+"] & E^+_{00}(0) \arrow[d, "d"] \\
E^+_{11}(0) \arrow[r, "d"] & E^+_{11}(1)
\end{tikzcd}$$

Figure 11

Here, a doubled arc in the right hand side represents a handle attachment (or a saddle move applied twice), and the equation can be checked using the *neck cutting relation*:



Verifications are left to the reader. Now the desired maps $\Phi^\pm$ between $\mathrm{CKh}(D^+)$ and $\mathrm{CKh}(D^-)$ are defined as

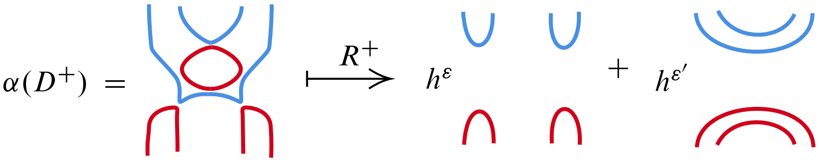
$$\mathrm{CKh}(D^+) \underset{I^+}{\overset{R^+}{\rightleftarrows}} E^+ \underset{\phi^+}{\overset{\phi^-}{\rightleftarrows}} E^- \underset{R^-}{\overset{I^-}{\rightleftarrows}} \mathrm{CKh}(D^-).$$

where $R^\pm$ and $I^\pm$ are the retractions and inclusions respectively.

Finally, the correspondence between the Lee classes can be checked by comparing those images in E^+ and E^- under the retractions $R^\pm$. The retractions are given by the following matrices

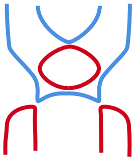
$$R^+ = \begin{pmatrix} G & & & \\ G'eH & G' & & \\ & & G & \\ e'H'eH & e'H' & eH & 1 \end{pmatrix}, \quad R^- = \begin{pmatrix} 1 & & & \\ & G & & \\ & & G' & \\ & & GeH' & G \end{pmatrix}.$$

The image of $\alpha(D^+) \in \mathrm{CKh}(D^+_{00})$ can be directly computed as

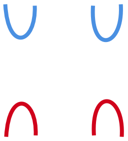


Here $\varepsilon, \varepsilon' \in \{0, 1\}$ are determined by how the arcs are connected outside the displayed area. Similarly, the image of $\alpha(D^-) \in \mathrm{CKh}(D^-_{11})$ is

$$\alpha(D^-) =$$



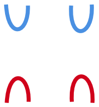
$$\xmapsto{R^-} h^{\varepsilon}$$



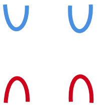
First,

$$\phi^- R^+(\alpha(D^+)) = R^-(\alpha(D^-))$$

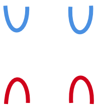
is obvious by definition. Next, from the explicit description of ϕ_1^+ and ϕ_2^+,



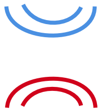
$$\xmapsto{f_1^+} h^2$$



and



$$\xmapsto{f_2^+} h^{(1-\varepsilon)+\varepsilon'+1}$$



Thus

$$\phi^+ R^-(\alpha(D^-)) = h^2 R^+(\alpha(D^+)),$$

and the proof is complete. □

Proposition 4.22 *Let L^+, L^- be links such that L^- is obtained from L^+ by applying a 4-strand generalized negative crossing change on L^+. Then*

$$s_h(L^-) \le s_h(L^+) \le s_h(L^-) + 4.$$

Proof A generalized negative crossing change can be realized by first applying Reidemeister moves on the four parallel strands and then modifying the negative half-twist part to a positive half-twist. Thus the result follows from Proposition 4.18 by an argument similar to the proof of Proposition 4.14. □

Remark 4.23 In [23, Theorem 1.11], the lower bound

$$s^{\mathbb{Q}}(L^-) \le s^{\mathbb{Q}}(L^+)$$

is given for the $\mathbb{Q}$-Rasmussen invariant (with no restrictions on the number of strands). It is obtained by realizing the move as a connected genus-0 cobordism from L^+ to L^- in $\overline{\mathbb{C}P^2} \setminus (B^4 \sqcup B^4)$. We expect that our map Φ^+ gives (a part of) the combinatorial description of this geometric map.

Now we prove the result for the equivariant case.

Definition 4.24 An *equivariant generalized negative crossing change* on a strongly invertible link L is an operation that is either a generalized negative crossing change on a set of strands lying on the axis, or two generalized negative crossing changes on two sets of strands that correspond by τ, lying off the axis.

Proposition 4.25 *Let L^+, L^- be strongly invertible links such that L^- is obtained by applying an equivariant* 4*-strand generalized negative crossing change to L^+. Then*

$$\underline{s}_h(L^-) \leq \underline{s}_h(L^+) \leq \underline{s}_h(L^-) + 4a,$$

where $a = 1$ if the move is performed on-axis, and $a = 2$ if performed off-axis. The same holds for $\overline{s}_h$.

Proof Let D^+, D^- be diagrams of L^+, L^- respectively such that D^+ and D^- are locally related by an equivariant generalized crossing change. First suppose the generalized negative crossing change is performed on-axis. Recall the definition of $\Phi^\pm$

$$\begin{tikzcd}
\mathrm{CKh}(D^+) \arrow[r, shift left, "R^+"] & E^+ \arrow[l, shift left, "I^+"] \arrow[r, shift left, "\phi^-"] & E^- \arrow[l, shift left, "\phi^+"] \arrow[r, shift left, "I^-"] & \mathrm{CKh}(D^-). \arrow[l, shift left, "R^-"]
\end{tikzcd}$$

The middle maps $\phi^\pm$ are strictly τ-equivariant. Moreover, since the tangle diagrams appearing in the moves are Kh-simple, from Lemma 2.14 it follows that the chain homotopy equivalences $R^\pm$ and $I^\pm$ are homotopy τ-equivariant. Thus the composite maps $\Phi^\pm$ are homotopy τ-equivariant and hence induce maps between the involutive complexes.

The proof of the case when the move is performed off-axis is similar to the proof of Proposition 4.16. □

4.5 Behavior under cobordisms

First recall that in the noninvolutive setting, given an oriented cobordism S between links L, L' in $\mathbb{R}^3$, there is a map between the corresponding complexes

$$\phi_S\colon \mathrm{CKh}(D^+) \to \mathrm{CKh}(D'),$$

where D, D' are diagrams of L, L' under a fixed projection $\mathbb{R}^3 \to \mathbb{R}^2$. The map ϕ_S is obtained by decomposing S into elementary cobordism and composing the corresponding maps between the intermediate complexes. It is proved in [5, Theorem 4] that ϕ_S is invariant up to chain homotopy under isotopies of S rel boundary. We prove an analogous statement in the involutive setting.

Definition 4.26 Let L, L' be two involutive links sharing the same involution τ of S^3. A *simple equivariant cobordism* between L and L' is an oriented cobordism S in $S^3 \times I$ between L, L' satisfying $(\tau \times \mathrm{id})(S) = S$. A *simple isotopy-equivariant cobordism* between L, L' is defined similarly, expect that $(\tau \times \mathrm{id})(S)$ is only required to be isotopic to S rel boundary.

Hereafter we simply write τS for $(\tau \times \mathrm{id})(S)$.

Lemma 4.27 *Let τ be an involution on S^3 and S be a cobordism between links L, L'. Then the following diagram commutes up to homotopy*

$$\begin{tikzcd}
\mathrm{CKh}(D) \arrow[r, "\phi_S"] \arrow[d, "\tau"] & \mathrm{CKh}(D') \arrow[d, "\tau"] \\
\mathrm{CKh}(\tau D) \arrow[r, "\phi_{\tau S}"] & \mathrm{CKh}(\tau D')
\end{tikzcd}$$

Proof First, assume that S is an elementary cobordism. For the Reidemeister moves R1, R2 and the three Morse moves M1–M3, one can check from the explicit descriptions that the square strictly commutes. For the R3 move, although it does not commute strictly, it follows from Kh-simplicity that $(\phi_S)^\tau = \tau\phi_S\tau$ and $\phi_{\tau S}$ are chain homotopic, so the square commutes up to homotopy.

For a general cobordism S, decompose it into elementary cobordisms as $S = S_1 \cup S_2 \cup \cdots \cup S_N$. Then $\tau S_1 \cup \tau S_2 \cup \cdots \cup \tau S_N$ gives an elementary cobordism decomposition of τS, and we obtain a homotopy commutative diagram

$$\begin{tikzcd}
C(D) \arrow[r, "\phi_{S_1}"] \arrow[d, "\tau"] & C(D_1) \arrow[r] \arrow[d, "\tau"] & \cdots \arrow[r] & C(D_{N-1}) \arrow[r, "\phi_{S_N}"] \arrow[d, "\tau"] & C(D') \arrow[d, "\tau"] \\
C(\tau D) \arrow[r, "\phi_{\tau S_1}"] & C(\tau D_1) \arrow[r] & \cdots \arrow[r] & C(\tau D_{N-1}) \arrow[r, "\phi_{\tau S_N}"] & C(\tau D')
\end{tikzcd}$$
□

Proposition 4.28 *Let S be a simple isotopy-equivariant cobordism between involutive links L and L'. Then there is a cobordism map*

$$\phi_S \colon \mathrm{CKhI}(D) \to \mathrm{CKhI}(D')$$

such that the following diagram commutes:

$$\begin{tikzcd}
Q\mathrm{CKh}(D)[1] \arrow[r, hook] \arrow[d, "\phi_S"] & \mathrm{CKhI}(D) \arrow[r, two heads] \arrow[d, "\phi_S"] & \mathrm{CKh}(D) \arrow[d, "\phi_S"] \\
Q\mathrm{CKh}(D')[1] \arrow[r, hook] & \mathrm{CKhI}(D') \arrow[r, two heads] & \mathrm{CKh}(D')
\end{tikzcd}$$

Proof By definition S and τS are isotopic rel boundary, so from [5, Theorem 4] there is a homotopy $\phi_S \simeq \phi_{\tau S}$. Together with Lemma 4.27, we have

$$\phi_S \simeq \phi_{\tau S} \simeq (\phi_S)^\tau.$$

Thus the result follows from Lemma 2.4. □

Now we restrict to strongly invertible links, and prove the behavior of the equivariant Lee classes under simple isotopy-equivariant cobordisms.

Proposition 4.29 *Suppose S is a simple isotopy-equivariant cobordism between strongly invertible links L, L', such that every component of S has boundary in L. Then under the map ϕ_S of Proposition 4.28, the equivariant Lee classes modulo torsions for the given orientations correspond as*

$$\underline{\alpha}(D) \overset{\phi_S}{\longmapsto} h^j \underline{\alpha}(D'), \quad \underline{\beta}(D) \overset{\phi_S}{\longmapsto} h^j \underline{\beta}(D'),$$
$$\overline{\alpha}(D) \overset{\phi_S}{\longmapsto} h^j \overline{\alpha}(D'), \quad \overline{\beta}(D) \overset{\phi_S}{\longmapsto} h^j \overline{\beta}(D'),$$

where

$$j = \frac{\delta w(D, D') - \delta r(D, D') - \chi(S)}{2}.$$

Proof By an argument similar to the proof of Proposition 3.8, the assertions are immediate from [27, Proposition 3.4]. □

Proposition 4.30 *Under the assumption of Proposition 4.29, we have*

$$\underline{s}_h(L) \le \underline{s}_h(L') - \chi(S) \quad \text{and} \quad \overline{s}_h(L) \le \overline{s}_h(L') - \chi(S).$$

Moreover if every component of S has boundary in both L and L', then we have

$$|\underline{s}_h(L) - \underline{s}_h(L')| \le -\chi(S) \quad \text{and} \quad |\overline{s}_h(L) - \overline{s}_h(L')| \le -\chi(S).$$

In particular, both $\underline{s}$ and $\overline{s}$ are invariant under simple isotopy-equivariant link concordances.

Proof Immediate from Proposition 4.29. □

Corollary 4.31 *For a strongly invertible knot K, both $|\underline{s}_h(K)|$ and $|\overline{s}_h(K)|$ bound $2\,\widetilde{\mathrm{sig}}_4(K)$ from below.*

Corollary 4.32 *If either $|\underline{s}_h(K)|$ or $|\overline{s}_h(K)|$ is greater than $2g_4(K)$, then no slice surfaces S of K realizing $g(S) = g_4(K)$ are simple isotopy-equivariant.*

5 The main theorem

In this section specialize to $(R, h) = (\mathbb{F}_2[H], H)$. The invariants $\underline{s}_H, \overline{s}_H$ will be denoted by $\underline{s}, \overline{s}$.

Proposition 1.2 *The three strongly invertible slice knots $K = m(9_{46})$, $15n_{103488}$, $17nh_{73}$ have*

$$\underline{s}(K) = 0 < 2 = \overline{s}(K).$$

Proof Proved by direct computations of BNI_r for the three knots. The result for $J_0 = 17nh_{73}$ is given in Table 1. The results for $m(9_{46})$ and $15n_{103488}$ are given in Table 2. □

	0	1	2	3	4	5	6	7
12	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
10	.	.	.	.	.	.	.	.
8	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
6	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
4	.	.	.	.	.	.	.	.
2	.	$\mathbb{F}[H]$	$\mathbb{F}$	.	.	.	.	.
0	$\mathbb{F}[H]$	.	.	.	.	.	.	.

	0	1	2	3	4	5	6	7	8	9	10	11
18	.	.	.	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
16	.	.	.	.	.	.	.	.	.	.	.	.
14	.	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
12	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}^2$	$\mathbb{F}$	.	.	.
10	.	.	.	.	.	.	.	.	.	.	.	.
8	.	.	.	.	$\mathbb{F}$	$\mathbb{F}^2$	$\mathbb{F}$	.	.	.	.	.
6	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.
2	.	$\mathbb{F}[H]$	$\mathbb{F}$	.	.	.	.	.	.	.	.	.
0	$\mathbb{F}[H]$	.	.	.	.	.	.	.	.	.	.	.

Table 2: $\mathrm{BNI}_r(m(9_{46}))$, left, and $\mathrm{BNI}_r(15n_{103488})$, right.

Theorem 1 *The strongly invertible knot J_n of Figure 1 has*

$$\underline{s}(J_n) = 0 < 2 \leq \overline{s}(J_n).$$

Proof Immediate from Theorem 3 and Propositions 1.2 and 1.6. □

To give some intuition for the inequality $\underline{s} < \overline{s}$, we give a human readable proof of a weaker inequality

$$\underline{s}(K) = 0 < 2 \leq \overline{s}(K)$$

specifically for $K = m(9_{46})$, inspired by the arguments given in [1]. Figure 12, left, depicts a diagram D of $m(9_{46})$ in the standard form of the pretzel knot $P(-3,3,-3)$. Given $w(D) = 3$, $r(D) = 8$, it suffices to show that

$$\underline{d}_h(D) = 2 < 3 \leq \overline{d}_h(D).$$

In Figure 12, the central diagram depicts the Lee cycle $\alpha(D)$ of D and the diagram on the right depicts another τ-invariant cycle z, which is obtained from $\alpha(D)$ by altering the labels of the two center circles to 1. By an argument similar to Example 3.16, we may find some τ-invariant element $x \in \mathrm{CKh}^{-1}(D)$ that gives

$$\alpha(D) \sim h^2 z,$$

where $\sim$ denotes homologous. Using $x \in \mathrm{CKhI}^{-1}(D)$ and $Qx \in \mathrm{CKhI}^{0}(D)$, we obtain

$$\underline{\alpha}(D) \sim h^2 z, \quad \overline{\alpha}(D) \sim h^2 Qz,$$

and hence

$$\underline{d}_h(D), \overline{d}_h(D) \geq 2.$$

Here, it is necessary that x is τ-invariant, otherwise $Q(1+\tau)x$ will produce some nonzero term in $Q\mathrm{CKh}(D)$.

Since K is slice, we have $s_h(D) = 0$ and hence $d_h(D) = 2$. Thus from Proposition 3.17 we obtain $\underline{d}_h(D) = 2$. It remains to show that the cycle $Qz \in \mathrm{CKhI}(D)$ is at least once more h-divisible. This is

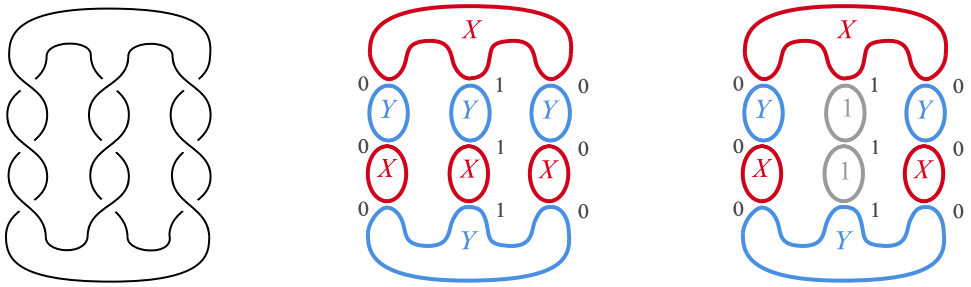


Figure 12: $m(9_{46})$.

equivalent to Qz being null-homologous *modulo* h in $\mathrm{CKhI}(D)$. Thus it suffices to set $h = 0$ and prove that there exists elements $a, b \in \mathrm{CKh}(D)$ such that

$$\begin{tikzcd} & b \arrow[r, mapsto, "d"] \arrow[d, mapsto, "Q(1+\tau)"] & 0 \\ Qa \arrow[r, mapsto, "d"] & Qz & \end{tikzcd} \qquad d(Qa) + Q(1+\tau)b = Qz.$$

Here it is necessary that $a \in \mathrm{CKh}(D)$ is *not* τ-invariant, otherwise we would also have $z \sim 0$ modulo h in $\mathrm{CKhI}(D)$ and $\underline{d}_h(D) \geq 3$. We leave it to the reader to find explicit elements a, b satisfying the above relation. Hence we conclude that $\overline{d}_h(D) \geq 3$.

This observation demonstrates the following facts:

- A cycle $z \in \mathrm{CKh}(D)$ must be symmetric (ie τ-invariant) to be a cycle in $\mathrm{CKhI}(D)$, whereas $Qz \in \mathrm{CKhI}(D)$ is always a cycle.
- A nontrivial boundary $dx \in \mathrm{CKh}(D)$ must be symmetric to be a boundary in $\mathrm{CKhI}(D)$, whereas $Q(dx) \in \mathrm{CKhI}(D)$ is always a boundary.
- A nonsymmetric cycle $z \in \mathrm{CKh}(D)$ gives a nontrivial boundary $Q(1+\tau)z$ in $Q\mathrm{CKh}(D) \subset \mathrm{CKhI}(D)$.

Therefore there are "less cycles" in $\mathrm{CKh}(D) \subset \mathrm{CKhI}(D)$ and "more boundaries" in $Q\mathrm{CKh}(D) \subset \mathrm{CKhI}(D)$, which allow $\underline{s} < \overline{s}$ to happen. A more simplified proof of Proposition 1.2 using reduction techniques is to be presented in a future paper.

6 On 2-periodic links

Analogous constructions for 2-periodic links are possible by using $\sigma\tau$ instead of τ, where σ is the involution given in Definition 2.19. Since σ is induced from a Frobenius algebra isomorphism, it commutes with any map coming from Bar-Natan's category $\mathrm{Cob}^3_{/l}(B)$, and many of the arguments in Sections 2 and 3 run in parallel for 2-periodic knots and links. The exceptions are (i) equivariant connected sums are not defined and (ii) the reduced complexes cannot be defined for 2-periodic links. Here we only state some of the basic definitions and results.

Definition 6.1 Given an involutive link (D, τ), define

$$\mathrm{CKhI}(D, \sigma\tau) = \mathrm{Cone}(\, \mathrm{CKh}(D) \xrightarrow{Q(1+\sigma\tau)} Q\mathrm{CKh}(D)\,).$$

Proposition 6.2 *The chain homotopy type of* $\mathrm{CKhI}(D, \sigma\tau)$ *is an invariant of the involutive link.*

Hereafter we assume that (D, τ) is 2-periodic.

Proposition 6.3 *The Lee cycles $\alpha(D, o)$ in $\mathrm{CKh}(D)$ are invariant under $\sigma\tau$.*

Proof Similar to the proof of Proposition 3.3, except that no Seifert circles intersect $\mathrm{Fix}(\tau)$, and that each symmetric pair of Seifert circles are oriented oppositely. □

Definition 6.4 The cycles $\underline{\alpha}(D, o) = \alpha(D, o)$ and $\overline{\alpha}(D, o) = Q\alpha(D, o)$ in $\mathrm{CKhI}(D, \sigma\tau)$ are called the *equivariant Lee cycles* of D.

Proposition 6.5 *If $h \in R$ is invertible, the homology $\mathrm{KhI}(D, \sigma\tau)$ is freely generated by the $2^{|D|+1}$ equivariant Lee classes.*

Hereafter we additionally assume that R is a PID and h is prime.

Proposition 6.6 *Let $\underline{d}(D)$ and $\overline{d}(D)$ be the h-divisibility (modulo torsion) of the equivariant Lee classes $\underline{\alpha}(D), \overline{\alpha}(D)$ in $\mathrm{KhI}(D, \sigma\tau)$. Then the quantities*

$$\underline{s}_h(L) = 2\underline{d}_h(D) + w(D) - r(D) + 1,$$
$$\overline{s}_h(L) = 2\overline{d}_h(D) + w(D) - r(D) + 1,$$

are invariants of the corresponding 2-periodic link L, satisfying

$$\underline{s}_h(L) \leq s_h(L) \leq \overline{s}_h(L).$$

In particular when K is a 2-periodic knot, then

$$\underline{s}_h(K) \leq s(K) \leq \overline{s}_h(K)$$

where $s(K)$ is the $\mathbb{F}_2$-Rasmussen invariant.

Proposition 6.7 *Let S be a simple isotopy-equivariant cobordism between 2-periodic links L and L'. Then we have*

$$\underline{s}_h(L) \leq \underline{s}_h(L') - \chi(S) \quad \text{and} \quad \overline{s}_h(L) \leq \overline{s}_h(L') - \chi(S).$$

Moreover if every component of S has boundary in both L and L', then we have

$$|\underline{s}_h(L) - \underline{s}_h(L')| \leq -\chi(S) \quad \text{and} \quad |\overline{s}_h(L) - \overline{s}_h(L')| \leq -\chi(S).$$

In particular, both $\underline{s}$ and $\overline{s}$ are invariant under simple isotopy-equivariant link concordances.

Corollary 6.8 *For a 2-periodic knot K, both $|\underline{s}_h(K)|$ and $|\overline{s}_h(K)|$ bound $2\,\widetilde{\mathrm{sig}}_4(K)$ from below.*

Appendix Computations for small prime knots

Here, the reduced involutive Khovanov homologies for prime knots with up to 7 crossings are given, computed by a program[4] developed by the author, with the diagrams given in [22, Section 6.5] as the input data.

Before describing the results, we review some of the basic facts on strongly invertible knots. See [26, Section 3] for the references.

Proposition A.9 (1) *Every torus knot admits exactly one inverting involution.*

(2) *An invertible hyperbolic knot admits exactly one or two inverting involutions*; *two when it has* (*cyclic or free*) *period* 2, *or one otherwise.*

(3) *For a fully amphichiral hyperbolic knot* K,

  (a) *if* K *admits a unique inverting involution* τ, *then* $(K,\tau) \cong (K,\tau)^*$,

  (b) *if* K *admits two inverting involutions* τ, τ', *then* $(K,\tau) \cong (K,\tau')^*$.

Thus for a torus knot (T) and a fully amphichiral hyperbolic knot (HA), it suffices to perform the computation for only one of its involutions. For a nonamphichiral hyperbolic knot (HN), there are possibly two nonequivalent involutions. In fact, all nonamphichiral hyperbolic knots with up to 7 crossings have exactly two nonequivalent involutions, and are distinguished by Sakuma's η-polynomial; see [26, Appendix].

The following list shows the computation results for strongly invertible prime knots with up to 7 crossings. Each item displays the name of the strongly invertible knot K, together with its type: (T), (HA) or (HN). For a (HN) type knot, the two distinct strongly invertible knots are distinguished by suffixes a, b, such as 5_{2a} and 5_{2b}. Its reduced involutive Khovanov homology $(R,h) = (\mathbb{F}_2, 0)$ is displayed on the left and its reduced involutive Bar-Natan homology $(R,h) = (\mathbb{F}_2[H], H)$ on the right. Note that both theories are bigraded, and the bigrading of each summand can be extracted from its computed generator. For the sake of readability, $\mathbb{F}_2$ is simply written as $\mathbb{F}$ and also $\mathbb{F}_2[H]/(H)$ is replaced with $\mathbb{F}$. From the latter table, one can read off the values of the equivariant Rasmussen invariants $(\underline{s}, \overline{s})$, however all knots in this list have $\underline{s} = \overline{s} = s^{\mathbb{F}_2}$. One can observe that BNI_r is generally not a direct sum of two shifted copies of BN_r, as can be seen in 7_{4b} and 7_{7b}. It is also notable that these two have order two H-torsions in BNI_r.

For the purpose of distinguishing nonequivalent involutions on the same knot, we see that the only successful ones are 7_4 and 7_7, so our invariants are not as strong as Sakuma's η-polynomial and Lobb–Watson's triply graded invariant CKh_τ; see [22]. Some additional structures, such as filtrations, might be given on KhI to further strengthen the invariant.

---

[4] The program is available at https://github.com/taketo1024/yui.

- 3_1 (T)

	0	1	2	3	4
8	.	.	.	$\mathbb{F}$	$\mathbb{F}$
6	.	.	$\mathbb{F}$	$\mathbb{F}$	.
4	.	.	.	.	.
2	$\mathbb{F}$	$\mathbb{F}$	.	.	.

	0	1	2	3	4
8	.	.	.	$\mathbb{F}$	$\mathbb{F}$
6	.	.	.	.	.
4	.	.	.	.	.
2	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.

- 4_1 (HA)

	-2	-1	0	1	2	3
4	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
2	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
0	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
-2	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
-4	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.

	-2	-1	0	1	2	3
4	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
2	.	.	.	.	.	.
0	.	.	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.
-2	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
-4	.	.	.	.	.	.

- 5_1 (T)

	0	1	2	3	4	5	6
14	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
12	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
10	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
8	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
6	.	.	.	.	.	.	.
4	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.

	0	1	2	3	4	5	6
14	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
12	.	.	.	.	.	.	.
10	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
8	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.
4	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.	.

- 5_{2a}, 5_{2b} (HN)

	0	1	2	3	4	5	6
12	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
10	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
8	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
6	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.
4	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
2	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.

	0	1	2	3	4	5	6
12	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
10	.	.	.	.	.	.	.
8	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
6	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
4	.	.	.	.	.	.	.
2	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.	.

- $6_{1a}, 6_{1b}$ (HN)

	-2	-1	0	1	2	3	4	5
8	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
6	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
4	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
2	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.
0	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
-2	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
-4	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.

	-2	-1	0	1	2	3	4	5
8	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
6	.	.	.	.	.	.	.	.
4	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
2	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
0	.	.	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.
-2	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
-4	.	.	.	.	.	.	.	.

- $6_{2a}, 6_{2b}$ (HN)

	-2	-1	0	1	2	3	4	5
10	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
8	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.
6	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.
4	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.
2	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
0	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
-2	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.

	-2	-1	0	1	2	3	4	5
10	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
8	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
6	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
4	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
2	.	.	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.
0	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
-2	.	.	.	.	.	.	.	.

- 6_3 (HA)

	-3	-2	-1	0	1	2	3	4
6	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
4	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.
2	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.
0	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.
-2	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
-4	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.
-6	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.

	-3	-2	-1	0	1	2	3	4
6	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
4	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
2	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
0	.	.	.	$\mathbb{F}[H]\oplus\mathbb{F}$	$\mathbb{F}[H]\oplus\mathbb{F}$	.	.	.
-2	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
-4	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
-6	.	.	.	.	.	.	.	.

- 7_1 (T)

	0	1	2	3	4	5	6	7	8
20	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
18	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
16	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
14	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
12	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
10	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
6	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.

	0	1	2	3	4	5	6	7	8
20	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
18	.	.	.	.	.	.	.	.	.
16	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
14	.	.	.	.	.	.	.	.	.
12	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
10	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
6	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.	.	.	.

- $7_{2a}, 7_{2b}$ (HN)

	0	1	2	3	4	5	6	7	8
16	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
14	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
12	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
10	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.
8	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
6	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.
4	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.
2	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.

	0	1	2	3	4	5	6	7	8
16	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
14	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
10	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
8	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
6	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.
2	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.	.	.	.

- $7_{3a}, 7_{3b}$ (HN)

18	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
16	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
14	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.
12	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.
10	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
8	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.
6	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.
4	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8

18	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
16	.	.	.	.	.	.	.	.	.
14	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.
12	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
10	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
8	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.
4	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8

- 7_{4a} (HN)

16	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
14	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
12	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.
10	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.
8	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
6	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.	.	.
4	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.	.
2	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8

16	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
14	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.
10	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
8	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
6	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.
2	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8

- 7_{4b} (HN)

16	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
14	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
12	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
10	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.
8	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
6	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.
4	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.
2	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8

16	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
14	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
10	.	.	.	.	$\mathbb{F}[H]/(H^2)$	$\mathbb{F}$	.	.	.
8	.	.	.	$\mathbb{F}$	.	.	.	.	.
6	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.
2	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8

- $7_{5a}, 7_{5b}$ (HN)

18	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
16	.	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.
14	.	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.
12	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.
10	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.	.
8	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.	.	.
6	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.
4	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8

18	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
16	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
14	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.
12	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.
10	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
8	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.
4	$\mathbb{F}[H]$	$\mathbb{F}[H]$	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8

- $7_{6a}, 7_{6b}$ (HN)

2	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
0	.	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.
-2	.	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.
-4	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.
-6	.	.	.	$\mathbb{F}^4$	$\mathbb{F}^4$	.	.	.	.
-8	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.	.	.
-10	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.	.
-12	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
	-5	-4	-3	-2	-1	0	1	2	3

2	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
0	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
-2	.	.	.	.	.	$\mathbb{F}[H]\oplus\mathbb{F}$	$\mathbb{F}[H]\oplus\mathbb{F}$	.	.
-4	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.
-6	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
-8	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
-10	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.
-12	.	.	.	.	.	.	.	.	.
	-5	-4	-3	-2	-1	0	1	2	3

- 7_{7a} (HN)

6	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
4	.	.	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.
2	.	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.
0	.	.	.	.	$\mathbb{F}^4$	$\mathbb{F}^4$	.	.	.
-2	.	.	.	$\mathbb{F}^4$	$\mathbb{F}^4$	.	.	.	.
-4	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.	.	.
-6	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.	.
-8	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
	-4	-3	-2	-1	0	1	2	3	4

6	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
4	.	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.
2	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
0	.	.	.	.	$\mathbb{F}[H]\oplus\mathbb{F}^2$	$\mathbb{F}[H]\oplus\mathbb{F}^2$	.	.	.
-2	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
-4	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.
-6	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.
-8	.	.	.	.	.	.	.	.	.
	-4	-3	-2	-1	0	1	2	3	4

- 7_{7b} (HN)

6	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
4	.	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.
2	.	.	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.
0	.	.	.	.	$\mathbb{F}^3$	$\mathbb{F}^3$	.	.	.
-2	.	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.
-4	.	.	$\mathbb{F}^2$	$\mathbb{F}^2$	.	.	.	.	.
-6	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.
-8	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.	.	.	.
	-4	-3	-2	-1	0	1	2	3	4

6	.	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$
4	.	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.
2	.	.	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.
0	.	.	.	.	$\mathbb{F}[H]\oplus\mathbb{F}$	$\mathbb{F}[H]\oplus\mathbb{F}$	.	.	.
-2	.	.	.	$\mathbb{F}$	$\mathbb{F}$	.	.	.	.
-4	.	.	$\mathbb{F}[H]/(H^2)$	$\mathbb{F}$	.	.	.	.	.
-6	.	$\mathbb{F}$	.	.	.	.	.	.	.
-8	.	.	.	.	.	.	.	.	.
	-4	-3	-2	-1	0	1	2	3	4

References

[1] **T Abe**, *State cycles which represent the canonical class of Lee's homology of a knot*, Topology Appl. 159 (2012) 1146–1158 MR

[2] **A Alfieri**, **K Boyle**, *Strongly invertible knots, invariant surfaces, and the Atiyah–Singer signature theorem*, Michigan Math. J. 74 (2024) 845–861 MR

[3] **A Alfieri**, **S Kang**, **A I Stipsicz**, *Connected Floer homology of covering involutions*, Math. Ann. 377 (2020) 1427–1452 MR

[4] **A Alishahi**, *Unknotting number and Khovanov homology*, Pacific J. Math. 301 (2019) 15–29 MR

[5] **D Bar-Natan**, *Khovanov's homology for tangles and cobordisms*, Geom. Topol. 9 (2005) 1443–1499 MR

[6] **H Bass**, **J W Morgan** (editors), *The Smith conjecture*, Pure Appl. Math. 112, Academic, Orlando, FL (1984) MR

[7] **K Boyle**, **A Issa**, *Equivariant 4-genera of strongly invertible and periodic knots*, J. Topol. 15 (2022) 1635–1674 MR

[8] **T D Cochran**, **E Tweedy**, *Positive links*, Algebr. Geom. Topol. 14 (2014) 2259–2298 MR

[9] **A Conway**, **M Powell**, *Characterisation of homotopy ribbon discs*, Adv. Math. 391 (2021) art. id. 107960 MR

[10] **I Dai**, **A Mallick**, **M Stoffregen**, *Equivariant knots and knot Floer homology*, J. Topol. 16 (2023) 1167–1236 MR

[11] **C M Gordon**, *On the higher-dimensional Smith conjecture*, Proc. London Math. Soc. 29 (1974) 98–110 MR

[12] **K Hayden**, *Corks, covers, and complex curves*, preprint (2021) arXiv 2107.06856

[13] **K Hayden**, **I Sundberg**, *Khovanov homology and exotic surfaces in the 4-ball*, J. Reine Angew. Math. 809 (2024) 217–246 MR

[14] **K Hendricks**, **C Manolescu**, *Involutive Heegaard Floer homology*, Duke Math. J. 166 (2017) 1211–1299 MR

[15] **K Hendricks**, **C Manolescu**, **I Zemke**, *A connected sum formula for involutive Heegaard Floer homology*, Selecta Math. 24 (2018) 1183–1245 MR

[16] **N Ito**, **J Yoshida**, *A cobordism realizing crossing change on $\mathfrak{sl}_2$ tangle homology and a categorified Vassiliev skein relation*, Topology Appl. 296 (2021) art. id. 107646 MR

[17] **M Khovanov**, *A categorification of the Jones polynomial*, Duke Math. J. 101 (2000) 359–426 MR

[18] **A Kotelskiy**, **L Watson**, **C Zibrowius**, *Immersed curves in Khovanov homology*, preprint (2019) arXiv 1910.14584

[19] **E S Lee**, *An endomorphism of the Khovanov invariant*, Adv. Math. 197 (2005) 554–586 MR

[20] **L Lewark**, *The Rasmussen invariant of arborescent and of mutant links*, Master's thesis, ETH Zürich (2009) Available at `http://www.math.jussieu.fr/~lewark/Master-Lukas-Lewark.pdf`

[21] **R Lipshitz**, **S Sarkar**, *Khovanov homology of strongly invertible knots and their quotients*, from "Frontiers in geometry and topology" (P M N Feehan, L L Ng, P S Ozsváth, editors), Proc. Sympos. Pure Math. 109, Amer. Math. Soc., Providence, RI (2024) 157–182 MR

[22] **A Lobb**, **L Watson**, *A refinement of Khovanov homology*, Geom. Topol. 25 (2021) 1861–1917 MR

[23] **C Manolescu**, **M Marengon**, **S Sarkar**, **M Willis**, *A generalization of Rasmussen's invariant, with applications to surfaces in some four-manifolds*, Duke Math. J. 172 (2023) 231–311 MR

[24] **J Milnor**, *Singular points of complex hypersurfaces*, Annals of Mathematics Studies 61, Princeton Univ. Press (1968) MR

[25] **J Rasmussen**, *Khovanov homology and the slice genus*, Invent. Math. 182 (2010) 419–447 MR

[26] **M Sakuma**, *On strongly invertible knots*, from "Algebraic and topological theories" (M Nagata, S Araki, A Hattori, editors), Kinokuniya, Tokyo (1986) 176–196 MR

[27] **T Sano**, *A description of Rasmussen's invariant from the divisibility of Lee's canonical class*, J. Knot Theory Ramifications 29 (2020) art,id. 2050037 MR

[28] **T Sano**, **K Sato**, *A family of slice-torus invariants from the divisibility of Lee classes*, Topology Appl. 357 (2024) art. id. 109059 MR

[29] **A N Shumakovitch**, *Torsion of Khovanov homology*, Fund. Math. 225 (2014) 343–364 MR

[30] **M Snape**, *Homological invariants of strongly invertible knots*, PhD thesis, University of Glasgow (2018) Available at `http://theses.gla.ac.uk/id/eprint/39015`

[31] **P Turner**, *Khovanov homology and diagonalizable Frobenius algebras*, J. Knot Theory Ramifications 29 (2020) 1950095, 10 MR

[32] **F Waldhausen**, *Über Involutionen der 3-Sphäre*, Topology 8 (1969) 81–91 MR

[33] **L Watson**, *Khovanov homology and the symmetry group of a knot*, Adv. Math. 313 (2017) 915–946 MR

[34] **S Wehrli**, *A spanning tree model for Khovanov homology*, J. Knot Theory Ramifications 17 (2008) 1561–1574 MR

[35] **Y Wigderson**, *The Bar-Natan theory splits*, J. Knot Theory Ramifications 25 (2016) 1650014, 19 MR

[36] **I Zemke**, *Connected sums and involutive knot Floer homology*, Proc. Lond. Math. Soc. 119 (2019) 214–265 MR

Interdisciplinary Theoretical and Mathematical Sciences Program, RIKEN
Wako, Japan

`taketo.sano@riken.jp`

Received: 4 May 2024 Revised: 22 November 2024

msp

Algebraic & Geometric Topology **25**:8 (2025) 5113–5136
DOI: 10.2140/agt.2025.25.5113
Published: 20 November 2025

A diagrammatic computation of abelian link invariants

DAVID CIMASONI
LIVIO FERRETTI
JESSICA LIU

We show how the multivariable signature and Alexander polynomial of a colored link can be computed from a single symmetric matrix naturally defined from a colored link diagram. In the case of a single variable, it coincides with the matrix introduced by Kashaev (2021), which was recently proven to compute the Levine–Tristram signature and the Alexander polynomial of oriented links (see Liu, 2023 and Cimasoni and Ferretti, 2024). As a corollary, we obtain a multivariable extension of Kauffman's (1983) determinantal model of the Alexander polynomial, recovering a result of Zibrowius (2017).

57K10

1 Introduction

As its title suggests, the aim of the present article is to give a way of computing several classical link invariants directly from a diagram. Before specifying these invariants, let us mention that this story is best told in the context of *colored links*, that we now recall.

Given an integer $\mu > 0$, a *μ-colored link* is an oriented link $L \subset S^3$ each of whose components is endowed with a *color* in $\{1, \dots, \mu\}$ in such a way that all these colors are used. Two colored links are isotopic if they are related by an ambient isotopy which respects the orientation and color of all components. Clearly, a 1-colored link is just an oriented link, while a μ-component μ-colored link is an oriented ordered link. A μ-colored link can be described by an oriented link diagram D with colored components, an object which we will refer to as a *μ-colored diagram* (see Figure 1 for an example). As usual, a crossing v of D is naturally endowed with a sign that we denote by $\operatorname{sgn} v = \pm 1$; see Figure 1. Finally, a crossing will be called *monochromatic* if the two corresponding strands are of the same color, and *bichromatic* otherwise.

The invariants we are interested in computing are the classical abelian invariants of a μ-colored link L, namely its multivariable *signature* and *nullity* $\sigma_L, \eta_L \colon (S^1 \setminus \{1\})^\mu \to \mathbb{Z}$, and its multivariable *Alexander polynomial* Δ_L in the normalized form given by the *Conway function* ∇_L. In the case $\mu = 1$, these are the well-known *Levine–Tristram signature* and *nullity* and *Alexander–Conway polynomial*, without doubt among the most studied of link invariants. We refer to Section 2 for the definition of these classical objects.

As our main result will show, these invariants can all be computed from a single symmetric matrix $\tau_D(x)$, whose coefficients are functions of formal variables $x = \{x_j, x_{jk} \mid 1 \leq j, k \leq \mu\}$ indexed by (unordered pairs of) colors. We now give its definition.

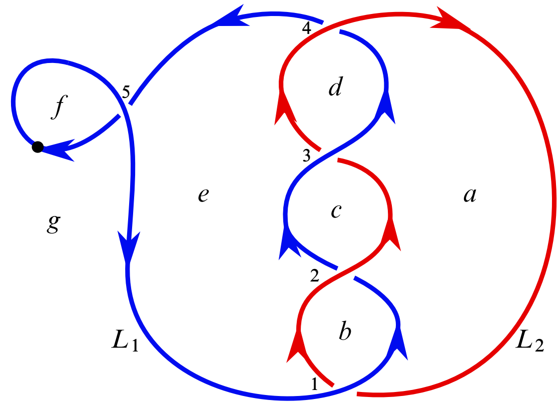


Figure 1: A 2-colored diagram D for a 2-colored link $L = L_1 \cup L_2$. The crossings are labeled 1 through 5 and the regions are labeled a through g. Crossings 1 to 4 are positive and bichromatic, while crossing 5 is negative and monochromatic. The marked point on L_1 will serve a further purpose.

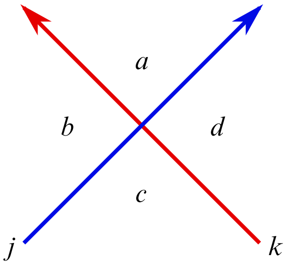
Definition 1.1 Given a μ-colored diagram D, let $\tau_D(x)$ be the symmetric matrix with rows and columns indexed by the regions of D defined by

$$\tau_D(x) = \sum_v \frac{\operatorname{sgn} v}{\sqrt{1-x_j^2}\sqrt{1-x_k^2}} \tau_v(x),$$

where the sum is over all crossings of D, the indices $j, k \in \{1, \dots, \mu\}$ are the (possibly identical) colors of the two strands crossing at v, and the only nonvanishing coefficients of the matrix $\tau_v(x)$ are given in Figure 2.

Also, we shall denote by $\tilde{\tau}_D(x)$ the matrix obtained by removing the two rows and columns corresponding to two adjacent regions of D determined by a marked point on D. We will assume without loss of generality that this point is on a strand of color 1.

Note that if the regions a, b, c, d around a crossing v are not all distinct, then one should add the corresponding rows and columns of $\tau_v(x)$. This happens in the following example.



	a	b	c	d
a	x_{jk}	x_j	1	x_k
b	x_j	$2x_jx_k - x_{jk}$	x_k	1
c	1	x_k	x_{jk}	x_j
d	x_k	1	x_j	$2x_jx_k - x_{jk}$

Figure 2: A crossing v together with the corresponding 4×4 minor of $\tau_v(x)$. The incoming left strand is of color j, the incoming right strand of color k, and the four adjacent regions are a, b, c, and d.

Example 1.2 Consider the 2-colored diagram D illustrated in Figure 1. Ordering the regions alphabetically, the corresponding matrix is given by

$$\tau_D(x)=\frac{1}{\sqrt{1-x_1^2}\sqrt{1-x_2^2}}\begin{bmatrix} 4(2x_1x_2-x_{12}) & 2x_2 & 2x_1 & 2x_2 & 4 & 0 & 2x_1\\ 2x_2 & 2x_{12} & 1 & 0 & 2x_1 & 0 & 1\\ 2x_1 & 1 & 2x_{12} & 1 & 2x_2 & 0 & 0\\ 2x_2 & 0 & 1 & 2x_{12} & 2x_1 & 0 & 1\\ 4 & 2x_1 & 2x_2 & 2x_1 & 4(2x_1x_2-x_{12}) & 0 & 2x_2\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 2x_1 & 1 & 0 & 1 & 2x_2 & 0 & 2x_{12} \end{bmatrix}$$
$$-\frac{1}{1-x_1^2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 2x_1^2-x_{11} & 1 & 2x_1\\ 0 & 0 & 0 & 0 & 1 & 2x_1^2-x_{11} & 2x_1\\ 0 & 0 & 0 & 0 & 2x_1 & 2x_1 & 2x_{11}+2 \end{bmatrix},$$

where the first summand contains the contributions from the four (positive) bichromatic crossings and the second summand is the contribution from the (negative) monochromatic crossing.

Here is our main result.

Theorem 1.3 *Let D be an arbitrary μ-colored diagram for a μ-colored link L.*

(i) *For any $\omega=(\omega_1,\dots,\omega_\mu)\in(S^1\setminus\{1\})^\mu$, the signature and nullity of L are given by*

$$\sigma_L(\omega)=\tfrac{1}{2}(\operatorname{sign}\tilde\tau_D(\omega)-w_m(D))\quad\text{and}\quad\eta_L(\omega)=\tfrac{1}{2}\operatorname{null}\tilde\tau_D(\omega),$$

where $w_m(D)$ is the sum of the signs of all monochromatic crossings of D, and $\tau_D(\omega)$ stands for the evaluation of $\tau_D(x)$ at

$$x_j=\operatorname{Re}(\omega_j^{1/2}),\quad x_{jk}=\operatorname{Re}(\omega_j^{1/2}\omega_k^{1/2}).$$

(ii) *If D is connected, then the Conway function of L satisfies*

$$\nabla_L^2(t_1,\dots,t_\mu)=\frac{1}{(t_1-t_1^{-1})^2}\Big(\prod_v(-\operatorname{sgn}v)\big(\tfrac{1}{2}(t_j-t_j^{-1})\big)\big(\tfrac{1}{2}(t_k-t_k^{-1})\big)\Big)\cdot\det\tilde\tau_D(t^2),$$

where the product is over all crossings of D, the indices j,k are the (possibly identical) colors of the two strands crossing at v, and $\tau_D(t^2)$ stands for the evaluation of $\tau_D(x)$ at

$$x_j=\tfrac{1}{2}(t_j+t_j^{-1}),\quad x_{jk}=\tfrac{1}{2}(t_jt_k+t_j^{-1}t_k^{-1}).$$

Several remarks are in order.

Remarks (i) We need to fix one square root of each coordinate $\omega_j \in S^1 \setminus \{1\}$ of ω. Our choice is to take $\omega_j = e^{i\theta_j}$ with $\theta_j \in (0, 2\pi)$, and $\omega_j^{1/2} = e^{i\theta_j/2}$. In other words, $\omega_j^{1/2}$ denotes the unique square root such that $\operatorname{Im}(\omega_j^{1/2})$ lies in $(0, 1]$. In particular, we have $(\bar{\omega}_j)^{1/2} = -(\omega_j^{1/2})$, and $\sqrt{1-x_j^2} = \operatorname{Im}(\omega_j^{1/2})$. Note that $x_j^2 \neq 1$, so $\tau_D(\omega)$ is a well-defined symmetric *real* matrix.

(ii) When defining $\tau_D(t^2)$ in the second point of Theorem 1.3, we need to explain how we evaluate $\sqrt{1-x_j^2}$, since there is a sign ambiguity: we set $\sqrt{1-x_j^2} = \frac{1}{2i}(t_j^{-1} - t_j)$. (See also Note 3.2.)

(iii) In both points of Theorem 1.3, the evaluations of the formal variables satisfy $x_{jj} = 2x_j^2 - 1$ for all j. Therefore, if a crossing v is monochromatic, then the matrix $\tau_v(x)$ can be written in a simple form which only depends on the single variable x_j.

(iv) In particular, if $\mu = 1$, then $\tau_D(x)$ depends on a single variable. This matrix was first introduced by Kashaev in [13]; see discussion below.

(v) In principle, it would be enough to only define the matrix $\tau_D(t^2)$ to state and prove Theorem 1.3. The reason we chose to introduce the matrix $\tau_D(x)$ is merely historical: in doing so, we explicitly present our results as an extension of Kashaev's work.

(vi) In [13], Kashaev studied the effect of Reidemeister moves on the matrix $\tau_D(x)$ in the single-variable case, proving that a certain *modified S-equivalence* class of the matrix is a link invariant. We expect a similar result to hold in the multivariable setting.

(vii) As it will become apparent from the proof of the second point, our construction naturally produces the square of the Conway function, so we cannot hope to recover its sign. (See also Note 1.6.)

Example 1.4 Consider once again the 2-colored link illustrated in Figure 1. From the corresponding matrix $\tau_D(x)$ given in Example 1.2, we compute

$$\tilde{\tau}_D(t^2) = \frac{-4}{(t_1-t_1^{-1})(t_2-t_2^{-1})} \times \begin{bmatrix} 2(t_1t_2^{-1}+t_1^{-1}t_2) & t_2+t_2^{-1} & t_1+t_1^{-1} & t_2+t_2^{-1} & 4 \\ t_2+t_2^{-1} & t_1t_2+t_1^{-1}t_2^{-1} & 1 & 0 & t_1+t_1^{-1} \\ t_1+t_1^{-1} & 1 & t_1t_2+t_1^{-1}t_2^{-1} & 1 & t_2+t_2^{-1} \\ t_2+t_2^{-1} & 0 & 1 & t_1t_2+t_1^{-1}t_2^{-1} & t_1+t_1^{-1} \\ 4 & t_1+t_1^{-1} & t_2+t_2^{-1} & t_1+t_1^{-1} & 2(t_1t_2^{-1}+t_1^{-1}t_2)-\frac{t_2-t_2^{-1}}{t_1-t_1^{-1}} \end{bmatrix},$$

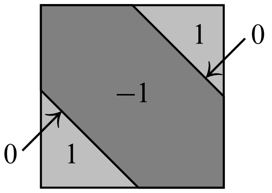
and obtain

$$\det \tilde{\tau}_D(t^2) = -\left(\frac{-4}{(t_1-t_1^{-1})(t_2-t_2^{-1})}\right)^5 (t_1-t_1^{-1})(t_2-t_2^{-1})(t_1t_2+t_1^{-1}t_2^{-1})^2.$$

To compute $\operatorname{sign} \tilde{\tau}_D(\omega)$, we evaluate $\tilde{\tau}_D(t^2)$ at $t_i = \omega_i^{1/2}$, and recall that the signature can change value only when the nullity changes value. Since $\operatorname{Im}(\omega_j^{1/2}) \in (0, 1]$, the computation of the determinant immediately implies that the nullity of $\tilde{\tau}_D(\omega)$ vanishes on the complement of $\Sigma := \{(\omega_1, \omega_2) \in (S^1 \setminus \{1\})^2 \mid \omega_1\omega_2 = -1\}$, so its signature is constant on the connected components of this complement. Then, an easy but tedious computation of minors shows that the nullity of $\tilde{\tau}_D(\omega)$ is constant equal to 2 on Σ, so the signature is also constant along each component of Σ. In particular, Theorem 1.3 implies that $\eta_L(\omega) = 1$ if $\omega_1\omega_2 = -1$, and $\eta_L(\omega) = 0$ otherwise. As for the signature, the values of $\operatorname{sign}\tilde{\tau}_D(\omega)$ can now be computed by picking one point in each of those connected components. For example, taking $\omega_1 = \omega_2 = -1$ we obtain

$$\tilde{\tau}_D(\omega) = \begin{bmatrix} 4 & 0 & 0 & 0 & 4 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 0 \\ 4 & 0 & 0 & 0 & 3 \end{bmatrix},$$

from which we compute $\operatorname{sign}\tilde{\tau}_D(\omega) = -3$ and hence, by Theorem 1.3, $\sigma_L(\omega) = -1$. The other values of σ_L can be computed in a similar way. The result can be summarized in the formula $\sigma_L(\omega_1,\omega_2) = -\operatorname{sign}\big(\operatorname{Re}((1-\omega_1)(1-\omega_2))\big)$, and is plotted below, where the domain $(S^1 \setminus \{1\})^2$ is shown as a square:



Finally, we also get $\pm\nabla_L(t_1,t_2) = t_1t_2 + t_1^{-1}t_2^{-1}$.

The first point of this theorem provides a practical new way of computing multivariable signatures, but it also yields much simpler proofs of known properties of this invariant. For example, consider the following situation: let L be the $(\mu-1)$-colored link obtained from a μ-colored link L' by identifying the colors $\mu-1$ and μ; then, for all $(\omega_1,\dots,\omega_{\mu-1}) \in (S^1\setminus\{1\})^{\mu-1}$, we have the equality

$$\sigma_L(\omega_1,\dots,\omega_{\mu-1}) = \sigma_{L'}(\omega_1,\dots,\omega_{\mu-1},\omega_{\mu-1}) - \sum \operatorname{lk}(K_{\mu-1},K_\mu),$$

the sum being over all components $K_{\mu-1} \subset L$ of color $\mu-1$ and all components $K_\mu \subset L$ of color μ. The original proof of this fact is rather tedious; see [4, Proposition 2.5]. It is an amusing exercise to check that this fact immediately follows from Theorem 1.3. In particular, given a μ-component μ-colored link $L' = K_1 \cup \dots \cup K_\mu$, the underlying 1-colored link L satisfies the equality

$$\xi(L) := \sigma_L(-1) + \sum_{i<j} \operatorname{lk}(K_i,K_j) = \sigma_{L'}(-1,\dots,-1).$$

This, together with the straightforward [4, Proposition 2.8], yields a one line proof of the main result of [17]: the integer $\xi(L)$ does not depend on the orientation of the components of L.

The second point of this theorem implies a corollary that we now present. Given a connected colored diagram D, let K_D be the matrix whose rows are indexed by the crossings of D, whose columns are

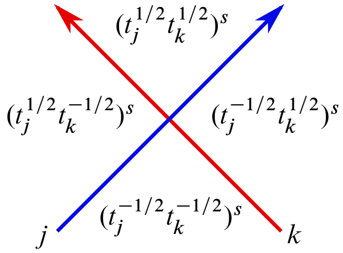

$(t_j^{1/2}t_k^{1/2})^s$
$(t_j^{1/2}t_k^{-1/2})^s$
$(t_j^{-1/2}t_k^{1/2})^s$
$(t_j^{-1/2}t_k^{-1/2})^s$
j
k

Figure 3: The labels in the definition of K_D around a vertex v with $s = \operatorname{sgn} v$.

indexed by the regions of D, and whose coefficients are defined by the label of the corners in Figure 3. (If a region abuts a corner from two sides, then the corresponding labels should be added.) Finally, let $\widetilde{K}_D$ denote the square matrix obtained from K_D by removing two columns corresponding to two adjacent regions (separated by a strand of color 1).

Corollary 1.5 *If D is a connected diagram for a colored link L, then*

$$\det \widetilde{K}_D = \pm(t_1 - t_1^{-1})\nabla_L(t_1, \dots, t_\mu).$$

Note 1.6 In fact, what we get from Corollary 1.5 is the symmetrized Alexander polynomial: since the sign of the determinant depends on the order of the rows and columns of $\widetilde{K}_D$, ie on the numbering of the regions and crossings of D, we cannot determine the sign of the Conway function.

Example 1.7 Consider one last time the 2-colored link illustrated in Figure 1. The matrix K_D equals

$$\begin{bmatrix} t_1^{-1/2}t_2^{1/2} & t_1^{1/2}t_2^{1/2} & 0 & 0 & t_1^{1/2}t_2^{-1/2} & 0 & t_1^{-1/2}t_2^{-1/2} \\ t_1^{1/2}t_2^{-1/2} & t_1^{-1/2}t_2^{-1/2} & t_1^{1/2}t_2^{1/2} & 0 & t_1^{-1/2}t_2^{1/2} & 0 & 0 \\ t_1^{-1/2}t_2^{1/2} & 0 & t_1^{-1/2}t_2^{-1/2} & t_1^{1/2}t_2^{1/2} & t_1^{1/2}t_2^{-1/2} & 0 & 0 \\ t_1^{1/2}t_2^{-1/2} & 0 & 0 & t_1^{-1/2}t_2^{-1/2} & t_1^{-1/2}t_2^{1/2} & 0 & t_1^{1/2}t_2^{1/2} \\ 0 & 0 & 0 & 0 & 1 & 1 & t_1 + t_1^{-1} \end{bmatrix},$$

from which we compute once again $\pm\nabla_L(t_1, t_2) = t_1 t_2 + t_1^{-1} t_2^{-1}$.

Let us now put our results in the context of the preexisting literature.

In 2018, Kashaev [13] defined the matrix $\tau_D(x)$ in the case $\mu = 1$, and conjectured Theorem 1.3 in this special case. Recently, Cimasoni and Ferretti [3] provided a proof of the second part of this conjecture by establishing a connection with Kauffman's determinantal model of the Alexander polynomial [15]; they also proved the first part of the Kashaev conjecture in a very restrictive case and indirect way. Immediately afterwards, Liu [16] gave a complete proof of the conjecture. Joining our efforts, we now extend Liu's approach to the general multivariable case in Theorem 1.3 and in its proof.

As for Corollary 1.5, in the case $\mu = 1$ it is nothing but Kauffman's aforementioned model for the Alexander polynomial [15]. Interestingly, Kauffman did state a multivariable version of his model (a detailed proof was only given many years later by Sato [18]), but it is different from our model.

However, Zibrowius [20] gave a state sum model for the multivariable Conway function which uses the same labels as the ones of Figure 3 up to a sign; using an extension of Kauffman's *clock theorem* [15], this state sum model can be turned into a determinantal model which coincides with the one of Corollary 1.5. This latter fact can be found in Zibrowius's PhD thesis [19, Chapter I.4] (but not in [20]). Therefore, and even though our proof is completely different, Corollary 1.5 is not a new result in the strict sense.

Let us finally mention that Friedl–Kausik–Quintanilha [11] recently provided an algorithm for the computation of generalized Seifert matrices (see Section 2.1) for colored links given as closures of colored braids. Since such matrices can be used to define σ_L, η_L and ∇_L, this method yields an algorithmic computation of these invariants. However, the remarkable feature of Theorem 1.3 remains: a new way of computing these invariants from a single symmetric matrix obtained directly from a colored diagram.

This paper is organized as follows. In Section 2, we recall the necessary background on generalized Seifert matrices (Section 2.1), multivariate signatures of colored links (Section 2.2), and the Conway function (Section 2.3). Section 3 contains the proof of our results, namely the first and second points of Theorem 1.3 in Sections 3.1 and 3.2, respectively, and of Corollary 1.5 in Section 3.3. A slightly informal last Section 3.4 contains results on the Alexander module.

Acknowledgments

We thank Claudius Zibrowius for informing us of his work on the Kauffman model, and the referee for very valuable suggestions. David Cimasoni and Livio Ferretti are supported by the Swiss NSF grant 200021-212085. Jessica Liu is partially supported by NSERC CGS-D.

2 Background

The aim of this section is to briefly recall the necessary background for our work: we start in Section 2.1 with the definition of C-complexes and generalized Seifert forms, then move on to multivariate signatures in Section 2.2, before dealing with the Conway function in Section 2.3.

2.1 Generalized Seifert surfaces and matrices

Seifert surfaces and matrices are well-known tools in the construction and study of (single-variable) abelian link invariants, such as the Levine–Tristram signature and the Alexander polynomial. Less well known is the fact that multivariate invariants can be defined and studied via generalized Seifert surfaces, known as C-complexes. We now introduce these objects, following [2; 8].

To do so, we will use the notation $L = L_1 \cup \cdots \cup L_\mu$ for a μ-colored link, where L_i is the sublink of L consisting of all the components of color i.

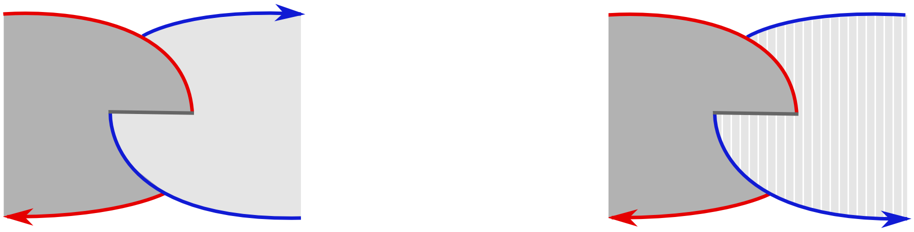


Figure 4: A positive clasp intersection (left), and a negative one (right).

Definition 2.1 A *C-complex* for a μ-colored link $L = L_1 \cup \cdots \cup L_\mu$ is a union $S = S_1 \cup \cdots \cup S_\mu$ of surfaces embedded in S^3 satisfying the following conditions:

(i) For all i, the surface S_i is a (possibly disconnected) Seifert surface for L_i.

(ii) For all $i \neq j$, the surfaces S_i and S_j are either disjoint or intersect in a finite number of *clasps*; see Figure 4.

(iii) For all i, j, k pairwise distinct, the intersection $S_i \cap S_j \cap S_k$ is empty.

A C-complex for a 1-colored link L is nothing but a (possibly disconnected) Seifert surface for the oriented link L. The existence of a C-complex for any given colored link is easy to establish [2]. On the other hand, the corresponding notion of S-equivalence is more difficult to prove; see [10] for the recently corrected statement.

These C-complexes allow us to define *generalized Seifert forms* as follows. For any choice of signs $\varepsilon = (\varepsilon_1, \dots, \varepsilon_\mu) \in \{\pm 1\}^\mu$, let

$$\alpha^\varepsilon \colon H_1(S) \times H_1(S) \to \mathbb{Z}$$

be the bilinear form given by $\alpha^\varepsilon(x, y) = \operatorname{lk}(x^\varepsilon, y)$, where x^ε denotes a well-chosen representative of the homology class $x \in H_1(S)$ pushed off S_i in the ε_i-normal direction (see [4] for a more formal definition). We denote by A^ε the corresponding *generalized Seifert matrices*, defined with respect to a fixed basis of $H_1(S)$. One easily checks the equality

$$A^{-\varepsilon} = (A^\varepsilon)^{\mathrm{T}} \tag{1}$$

for all $\varepsilon \in \{\pm 1\}^\mu$. The two generalized Seifert matrices A^-, A^+ of a 1-colored link L coincide with the usual Seifert matrix A of the oriented link L and its transposed matrix A^{T}.

The general principle regarding these matrices is the following: what Seifert matrices can do in one variable for oriented links, generalized Seifert surfaces can do in μ-variables for μ-colored links. In Sections 2.2 and 2.3 we illustrate this principle with two examples of invariants.

2.2 Signatures and nullities of colored links

Fix a C-complex S for a μ-colored link L and a basis of $H_1(S)$. Consider an element $\omega = (\omega_1, \dots, \omega_\mu)$ of $\mathbb{T}^\mu_* := (S^1 \setminus \{1\})^\mu$, and set

$$H(\omega) := \sum_{\varepsilon \in \{\pm 1\}^\mu} \Big(\prod_{i=1}^{\mu} (1 - \overline{\omega}_i^{\varepsilon_i}) \Big) A^\varepsilon. \tag{2}$$

Using (1), one easily checks that $H(\omega)$ is a Hermitian matrix and hence admits a well-defined signature $\operatorname{sign} H(\omega) \in \mathbb{Z}$ and nullity $\operatorname{null} H(\omega) \in \mathbb{Z}_{\geq 0}$.

Definition 2.2 [4] The *signature* and *nullity* of the μ-colored link L are functions

$$\sigma_L, \eta_L \colon \mathbb{T}^{\mu}_{*} \to \mathbb{Z}$$

defined by $\sigma_L(\omega) := \operatorname{sign} H(\omega)$ and $\eta_L(\omega) := \operatorname{null} H(\omega)$, respectively.

The fact that these functions are well-defined invariants, ie do not depend on the choice of the C-complex S for L, relies on the aforementioned notion of S-equivalence [4; 10]. In the case $\mu = 1$, the functions $\sigma_L, \eta_L \colon S^1 \setminus \{1\} \to \mathbb{Z}$ are the signature and nullity of the Hermitian matrix $(1-\omega)A + (1-\overline{\omega})A^{\mathrm{T}}$: they coincide with the Levine–Tristram signature and nullity of the oriented link L. We refer to the recent survey [5] for background on this classical invariant.

In a nutshell, all the remarkable properties of the Levine–Tristram signature extend to the multivariable setting. For example, the function σ_L is constant on the connected components of the complement in $\mathbb{T}^{\mu}_{*}$ of the zeros of the multivariable Alexander polynomial $\Delta_L(t_1, \dots, t_\mu)$ [4] (see Section 2.3 below). Also, if $(\omega_1, \dots, \omega_\mu) \in \mathbb{T}^{\mu}_{*}$ is not the root of any Laurent polynomial $p(t_1, \dots, t_\mu)$ with $p(1, \dots, 1) = \pm 1$, then $\sigma_L(\omega_1, \dots, \omega_\mu)$ and $\eta_L(\omega_1, \dots, \omega_\mu)$ are invariant under *topological concordance* of colored links [6].

2.3 The Conway function of a colored link

The one-variable Alexander polynomial $\Delta_L(t)$ of an oriented link L can be generalized to a μ-variable polynomial invariant $\Delta_L(t_1, \dots, t_\mu)$ of a μ-colored link L, a fact known to Alexander himself [1]. To do so, consider the exterior $X_L := S^3 \setminus \nu(L)$ of L and the surjective group homomorphism

$$\pi_1(X_L) \to \mathbb{Z}^{\mu}, \quad [\gamma] \mapsto \bigl(\operatorname{lk}(\gamma, L_1), \dots, \operatorname{lk}(\gamma, L_\mu)\bigr).$$

This defines a regular $\mathbb{Z}^{\mu}$-cover $\widehat{X}_L$ of X_L whose homology groups are hence equipped with the structure of a module over the group ring $\mathbb{Z}[\mathbb{Z}^{\mu}] = \mathbb{Z}[t_1^{\pm 1}, \dots, t_\mu^{\pm 1}]$. In particular, the module $\mathscr{A}_L := H_1(\widehat{X}_L)$ is called the (multivariable) *Alexander module* of L (see Section 3.4), and a greatest common divisor of the elements of its first elementary ideal is the *Alexander polynomial* of L.

This Laurent polynomial in μ-variables is only well defined up to multiplication by units of the ring $\mathbb{Z}[t_1^{\pm 1}, \dots, t_\mu^{\pm 1}]$, ie up to a sign and powers of the variables. This later indeterminacy can be easily overcome by harnessing the symmetry of Δ and requiring it to satisfy $\Delta_L(t_1^{-1}, \dots, t_\mu^{-1}) = \pm \Delta_L(t_1, \dots, t_\mu)$, but the sign issue is a nontrivial one.

The solution was suggested by Conway in his landmark paper [7]. He claimed the existence of a well-defined rational function ∇_L satisfying

$$\nabla_L(t_1, \dots, t_\mu) \doteq \begin{cases} (1/(t_1 - t_1^{-1}))\Delta_L(t_1^2) & \text{if } \mu = 1; \\ \Delta(t_1^2, \dots, t_\mu^2) & \text{if } \mu > 1, \end{cases} \tag{3}$$

where $\doteq$ stands for the equality up to multiplication by $\pm t_1^{\nu_1}\cdots t_\mu^{\nu_\mu}$ with $\nu_1,\dots,\nu_\mu\in\mathbb{Z}$. The first explicit construction of this *Conway function* was given by Hartley [12] using free differential calculus, but we will make use of the following geometric construction [2]. Given any *connected* C-complex $S=S_1\cup\cdots\cup S_\mu$ for a L, consider the matrix

$$A_S := \sum_{\varepsilon\in\{\pm1\}^\mu}\Big(\prod_{i=1}^{\mu}\varepsilon_i t_i^{\varepsilon_i}\Big)A^\varepsilon. \tag{4}$$

Then, the Conway function of L is given by

$$\nabla_L(t_1,\dots,t_\mu)=(\operatorname{sgn} S)\Big(\prod_{i=1}^{\mu}(t_i-t_i^{-1})^{\chi(S\setminus S_i)-1}\Big)\det(-A_S), \tag{5}$$

where $\operatorname{sgn} S$ denotes the product of the signs of the clasps of S (recall Figure 4). Note that in the case $\mu=1$, (3) and (5) lead to the formula $\Delta_L(t)=\det(t^{-1/2}A-t^{1/2}A^{\mathrm T})$, the classical definition of the Alexander–Conway polynomial of the oriented link L [14].

This geometric construction of the Conway function yields straightforward proofs of the various properties of this invariant. In particular, it yields a "geometric explanation" of the local relations that can be used to compute it from a link diagram; see [2] for more details.

3 Proofs of the main results

We will now provide proofs of our results. More precisely, we start in Section 3.1 with the demonstration of the first part of Theorem 1.3 on signatures and nullities. Section 3.2 deals with the second part on the Conway function, while Section 3.3 contains the proof of Corollary 1.5 on the multivariable Kauffman model. Finally, Section 3.4 consists in a slightly informal discussion on the Alexander module.

3.1 Signatures and nullities

We discuss how to compute the multivariable signature, proving part (i) of Theorem 1.3 which we now restate for convenience.

Proposition 3.1 *Let D be a diagram for a μ-colored link L. For any $\omega=(\omega_1,\dots,\omega_\mu)\in(S^1\setminus\{1\})^\mu$, the signature and nullity of L are given by*

$$\sigma_L(\omega)=\tfrac12(\operatorname{sign}\tilde\tau_D(\omega)-w_m(D))\quad\textit{and}\quad \eta_L(\omega)=\tfrac12\operatorname{null}\tilde\tau_D(\omega),$$

where $w_m(D)$ is the sum of the signs of all monochromatic crossings of D, and $\tau_D(\omega)$ stands for the evaluation of $\tau_D(x)$ at $x_j=\operatorname{Re}(\omega_j^{1/2})$ and $x_{jk}=\operatorname{Re}(\omega_j^{1/2}\omega_k^{1/2})$.

Proof Fix an arbitrary μ-colored link L, and let rL denote the μ-colored link L with reverse orientation but same coloring as L. Let $L\,\#_1\,rL$ denote a connected sum of L and rL along two components of color 1. Unlike for knots, the isotopy type of the connected sum of (colored) links is not well defined. However, any two such connected sums have the same signature and nullity, as these invariants behave

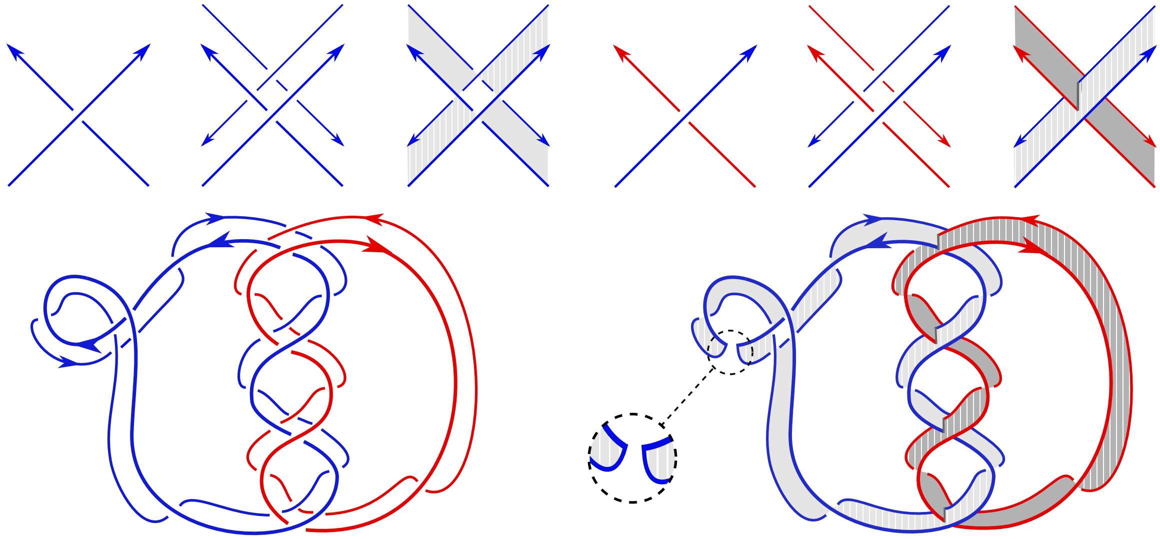


Figure 5: First row: the construction of a C-complex S for $L \#_1 rL$ near a monochromatic crossing (left three images) and a bichromatic crossing (right three images). Second row, left: the diagram for the disjoint sum of L and rL obtained from D of Figure 1. Second row, right: the corresponding C-complex S for $L \#_1 rL$.

additively under this ill-defined operation; this follows from Propositions 2.12 and 2.5 of [4]. Since the signature and nullity are unchanged by reversing the orientation (see [4, Corollary 2.9]), the relations

$$\sigma_{L\#_1 rL}(\omega) = 2\sigma_L(\omega) \quad \text{and} \quad \eta_{L\#_1 rL}(\omega) = 2\eta_L(\omega) \tag{6}$$

hold for any such connected sum.

The idea of the proof is to use a diagram D for L to construct a C-complex S for $L \#_1 rL$ whose first homology has a basis given by classes of loops corresponding to the regions and the monochromatic crossings of D — minus the two regions near the connected sum. By taking generalized Seifert matrices with respect to this basis, we show that the matrix $H(\omega)$ used in Definition 2.2 is congruent to a block-diagonal matrix of the form $\tilde{\tau}_D(\omega) \oplus Z$ with $\sigma(Z) = -w_m(D)$ and null $Z = 0$. Combined with (6), this completes the proof.

We now give the details. To construct a C-complex for $L \#_1 rL$ from D, we use the following procedure (see Figure 5 (top row) for the construction near crossings, and Figure 5 (bottom row) for an example).

(i) At each crossing of D, draw a copy of the corresponding crossing for rL "a bit above and behind" the crossing of D.

(ii) Connect the remaining strands of rL to each other following along the edges of D, possibly creating an additional crossing along each edge (with rL passing under L). This yields a diagram for the disjoint sum of L and rL.

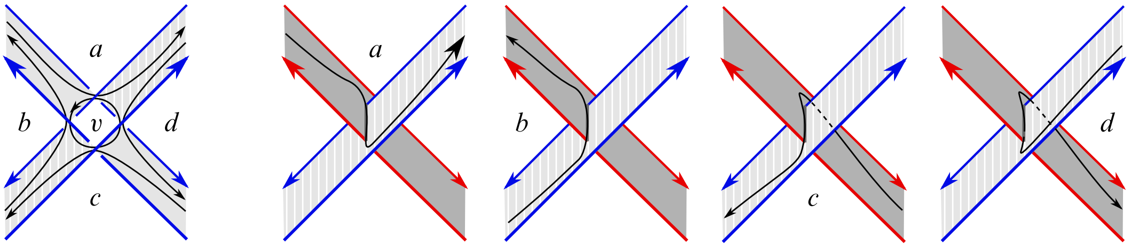


Figure 6: The five cycles of S near a monochromatic crossing (left, one image) and the four cycles of S near a bichromatic crossing (right, four images). The labels a, b, c, d, v for the regions are used for the local linking matrices in Figure 8.

(iii) Create a clasp intersection near each bichromatic crossing of D and apply the usual Seifert algorithm near each monochromatic crossing of D. This yields a C-complex for the disjoint sum of L and rL.

(iv) Finally, pick a point on a strand of color 1 in D and cut the corresponding surface at that place. The result is a C-complex S for (some version of) $L \#_1 rL$.

Note that S deformation retracts onto a graph defined as follows: take the 4-regular graph underlying the diagram D, add a loop at each vertex corresponding to a monochromatic vertex, and remove the edge along which the connected sum was performed. As a consequence, a natural basis for $H_1(S)$ is given by classes of cycles corresponding to the regions and monochromatic crossings of D — minus the two regions adjacent to where the connected sum happens. We use the convention that the cycles representing our basis of $H_1(S)$ are oriented counterclockwise in the plane of D where S is drawn, and denote by the same letter a region or monochromatic crossing and its corresponding cycle of $H_1(S)$. Using this explicit basis, we now study the local contribution to generalized Seifert matrices near crossings of D.

For the remainder of the proof, we adopt the following convention: we say a crossing v has color $(j,k)=(j_v,k_v)$ if its incoming left strand is color j, and incoming right strand has color k, as in Figure 2. If $j=k$, we may simply say it has color j.

As illustrated in Figure 6, there are five homology classes in $H_1(S)$ coming into play near a monochromatic crossing of D, and four near a bichromatic crossing. To compute the Seifert forms locally, we need to choose a convention for drawing the pushouts so that no contribution to the linking numbers comes from the crossings that occur along the edges of D: this is illustrated in Figure 7.

The local contribution to the matrices A^ε near different types of crossings of D are given in Figure 8. Since the C-complex near a negative bichromatic crossing is the mirror image of the C-complex near a positive bichromatic crossing, the contribution near a negative bichromatic crossing can be obtained by changing the sign of the contribution near a positive crossing. Moreover, by the symmetry (1), the contribution for opposite choices of ε can be obtained by transposition. In conclusion, the local linking of all possible cases can be computed from those in Figure 8. For the sake of clarity, let us mention that in the case of monochromatic crossings as well, one could compute the contribution for negative crossings

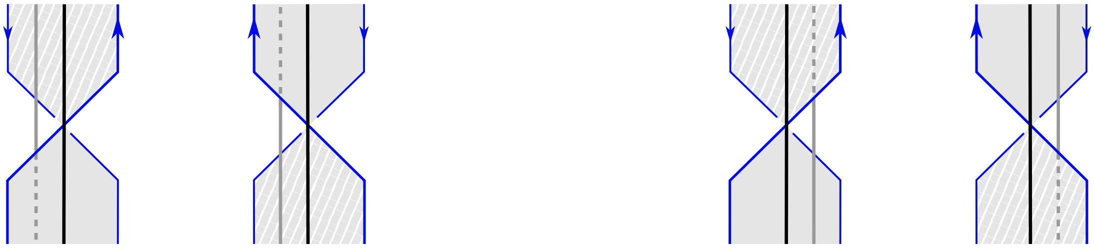


Figure 7: Convention for drawing the pushouts in the positive (left) and negative (right) directions near a crossing of $L\#_1 rL$ occurring along an edge of D: if the pushout appears behind the surface (dotted gray) then it is drawn between the original curve (black) and the diagram for L (thick blue); if the pushout appears in front of the surface (solid gray) then it is drawn between the curve and the diagram for rL (thin blue).

from the one for positive crossings: the linking numbers involving the curves b and d (or their pushout) is unchanged, while the others change in a controlled way. However, since the precise relation is less immediately evident, we prefer to include both matrices in Figure 8.

We now write $H(\omega)$ as a sum over crossings of D. For a crossing v, let A_v^ε denote the square matrix (of size equal to the first Betti number of S) given by the contribution to A^ε from the linking near v; in other words, A_v^ε is zero everywhere except in the 5×5 or 4×4 minor corresponding to the homology classes coming into play near v, where its values are given by the local contributions to linking numbers given by the matrices from Figure 8. We have

$$H(\omega)=\sum_{\varepsilon\in\{\pm1\}^\mu}\Bigl(\prod_{i=1}^{\mu}(1-\overline{\omega}_i^{\varepsilon_i})\Bigr)A^\varepsilon=\sum_v\sum_{\varepsilon\in\{\pm1\}^\mu}\Bigl(\prod_{i=1}^{\mu}(1-\overline{\omega}_i^{\varepsilon_i})\Bigr)A_v^\varepsilon=:\sum_v H_v, \tag{7}$$

where $H_v=\sum_\varepsilon\bigl(\prod_i(1-\overline{\omega}_i^{\varepsilon_i})\bigr)A_v^\varepsilon$, and the sums indexed by v always refer to the sum over all crossings of D. Note that A_v^ε is entirely specified by ε_{j_v} and ε_{k_v}. Thus, we use $A_v^{(\alpha,\beta)}$ to denote A_v^ε for any ε with $\varepsilon_{j_v}=\alpha$ and $\varepsilon_{k_v}=\beta$. If v is monochromatic with $\varepsilon_{j_v}=\varepsilon_{k_v}=\alpha$, we simply write $A_v^{(\alpha)}$.

lk	a	b	c	d	v
$a^{(1)}$	$-\frac{1}{2}$	$-\frac{1}{2}$	0	0	1
$b^{(1)}$	0	0	0	0	0
$c^{(1)}$	0	$-\frac{1}{2}$	$-\frac{1}{2}$	0	1
$d^{(1)}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0	-1
$v^{(1)}$	0	1	0	0	-1

lk	a	b	c	d	v
$a^{(1)}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	0	0
$b^{(1)}$	0	0	0	0	0
$c^{(1)}$	0	$-\frac{1}{2}$	$\frac{1}{2}$	0	0
$d^{(1)}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0	-1
$v^{(1)}$	-1	1	-1	0	1

lk	a	b	c	d
$a^{(1,1)}$	$-\frac{1}{2}$	$-\frac{1}{2}$	0	0
$b^{(1,1)}$	0	0	0	0
$c^{(1,1)}$	0	$\frac{1}{2}$	$-\frac{1}{2}$	0
$d^{(1,1)}$	$\frac{1}{2}$	-1	$\frac{1}{2}$	0

lk	a	b	c	d
$a^{(1,-1)}$	0	0	0	0
$b^{(1,-1)}$	$-\frac{1}{2}$	$\frac{1}{2}$	0	0
$c^{(1,-1)}$	1	$-\frac{1}{2}$	0	$-\frac{1}{2}$
$d^{(1,-1)}$	$-\frac{1}{2}$	0	0	$\frac{1}{2}$

Figure 8: The local contributions to linking numbers near crossings of D, where the curves are labeled as in Figure 6. First two tables: contributions near a positive (first) and negative (second) monochromatic crossing of color j, where $x^{(1)}$ stands for x^ε with $\varepsilon_j=1$. Last two tables: contributions near a positive bichromatic crossing of colors (j,k), with x^ε denoted by $x^{(\varepsilon_j,\varepsilon_k)}$.

When v is monochromatic of color $j = j_v$, the matrix H_v can be rewritten as

$$\begin{aligned} H_v &= \sum_{(\varepsilon_1,\dots,\hat{\varepsilon}_j,\dots,\varepsilon_\mu)\in\{\pm1\}^{\mu-1}} \Bigg(\prod_{\substack{i=1\\ i\neq j}}^{\mu}(1-\bar{\omega}_i^{\varepsilon_i})\Bigg)((1-\bar{\omega}_j)A_v^{(1)}+(1-\omega_j)A_v^{(-1)}) \\ &= \Bigg(\prod_{\substack{i=1\\ i\neq j}}^{\mu}(1-\bar{\omega}_i)(1-\omega_i)\Bigg)((1-\bar{\omega}_j)A_v^{(1)}+(1-\omega_j)A_v^{(-1)}), \end{aligned}$$

where the second equality uses the relation $(1-\bar{\omega}_i)+(1-\omega_i) = (1-\bar{\omega}_i)(1-\omega_i)$. Hence, using the notation $s_i := (1-\omega_i)$, the matrix H_v for a monochromatic crossing v of color j is given by

$$H_v = \Bigg(\prod_{\substack{i=1\\ i\neq j}}^{\mu}|s_i|^2\Bigg)(\bar{s}_j A_v^{(1)} + s_j A_v^{(-1)}) =: \Bigg(\prod_{\substack{i=1\\ i\neq j}}^{\mu}|s_i|^2\Bigg)A_v, \tag{8}$$

while for a bichromatic crossing of colors (j,k), it is given by

$$H_v = \Bigg(\prod_{\substack{i=1\\ i\neq j,k}}^{\mu}|s_i|^2\Bigg)(\bar{s}_j\bar{s}_k A_v^{(1,1)} + \bar{s}_j s_k A_v^{(1,-1)} + s_j\bar{s}_k A_v^{(-1,1)} + s_j s_k A_v^{(-1,-1)}) =: \Bigg(\prod_{\substack{i=1\\ i\neq j,k}}^{\mu}|s_i|^2\Bigg)A_v, \tag{9}$$

where we use the notation

$$A_v := \begin{cases} \bar{s}_j A_v^{(1)} + s_j A_v^{(-1)} & \text{if } v \text{ is monochromatic of color } j; \\ \bar{s}_j\bar{s}_k A_v^{(1,1)} + \bar{s}_j s_k A_v^{(1,-1)} + s_j\bar{s}_k A_v^{(-1,1)} + s_j s_k A_v^{(-1,-1)} & \text{if } v \text{ is bichromatic of colors } j,k. \end{cases}$$

Plugging (8) and (9) into (7), we get

$$H(\omega) = \sum_v H_v = \sum_v \Bigg(\prod_{i\neq j_v,k_v}^{\mu}|s_i|^2\Bigg)A_v. \tag{10}$$

Writing the Hermitian matrix $H(\omega)$ as a block matrix of the form

$$H(\omega) = \begin{pmatrix} X & Y \\ Y^* & Z \end{pmatrix},$$

where the first line and column correspond to regions and the second line and column correspond to monochromatic crossings, we see that Z is a diagonal matrix with coefficient corresponding to the monochromatic crossing v given by $-\big(\prod_{i=1}^{\mu}|s_i|^2\big)\operatorname{sgn} v$. In particular, Z is invertible (and hence has nullity $\operatorname{null} Z = 0$), while its signature is equal to $\sigma(Z) = -w_m(D)$. Furthermore, $H(\omega)$ is congruent to the block diagonal matrix

$$MH(\omega)M^* = \begin{pmatrix} X - YZ^{-1}Y^* & 0 \\ 0 & Z^{-1} \end{pmatrix} \tag{11}$$

via $M = \begin{pmatrix} I & -YZ^{-1} \\ 0 & Z^{-1} \end{pmatrix}$. Since $\sigma(Z^{-1}) = \sigma(Z) = -w_m(D)$ and $\operatorname{null} Z^{-1} = 0$, it remains to show that the matrix $X - YZ^{-1}Y^*$ coincides with $\tilde{\tau}_D(\omega)$ up to transformations that do not affect the signature and nullity.

$$\begin{array}{c|cccc}
 & a & b & c & d \\
\hline
a & \frac{\omega_j+\bar\omega_j}{2} & -\frac{1+\bar\omega_j}{2} & 1 & -\frac{1+\omega_j}{2} \\
b & -\frac{1+\omega_j}{2} & 1 & -\frac{1+\omega_j}{2} & \omega_j \\
c & 1 & -\frac{1+\bar\omega_j}{2} & \frac{\omega_j+\bar\omega_j}{2} & -\frac{1+\omega_j}{2} \\
d & -\frac{1+\bar\omega_j}{2} & \bar\omega_j & -\frac{1+\bar\omega_j}{2} & 1
\end{array}
\qquad
\begin{array}{c|cccc}
 & a & b & c & d \\
\hline
a & -\frac{\bar s_j\bar s_k+s_js_k}{2} & \frac{\bar s_k(\omega_j-\bar\omega_j)}{2} & s_j\bar s_k & \frac{s_j(\bar\omega_k-\omega_k)}{2} \\
b & \frac{s_k(\bar\omega_j-\omega_j)}{2} & \frac{s_j\bar s_k+\bar s_js_k}{2} & \frac{s_j(\bar\omega_k-\omega_k)}{2} & -s_js_k \\
c & \bar s_js_k & \frac{\bar s_j(\omega_k-\bar\omega_k)}{2} & -\frac{\bar s_j\bar s_k+s_js_k}{2} & \frac{s_k(\bar\omega_j-\omega_j)}{2} \\
d & \frac{\bar s_j(\omega_k-\bar\omega_k)}{2} & -\bar s_j\bar s_k & \frac{\bar s_k(\omega_j-\bar\omega_j)}{2} & \frac{s_j\bar s_k+\bar s_js_k}{2}
\end{array}$$

Figure 9: The local contribution to $X-YZ^{-1}Y^*$ for a positive monochromatic and bichromatic crossing v. If v is a negative crossing, the matrix is the negative of the corresponding matrix for a positive crossing. Left: v positive monochromatic crossing of color j. Right: v positive bichromatic crossing of colors (j,k).

To determine $X-YZ^{-1}Y^*$, let us fix two regions a and b. Note that

$$(YZ^{-1}Y^*)_{a,b}=\sum_v Y_{a,v}Z^{-1}_{v,v}\overline{Y}_{b,v},$$

and that it is only possible for both $Y_{a,v}$ and $Y_{b,v}$ to be nonzero if the regions a and b are both adjacent to the crossing v. Therefore, the matrix $YZ^{-1}Y^*$ is a sum over crossings, where the contribution at each crossing is a matrix that is zero everywhere except in the 4×4 minor corresponding to the four adjacent regions of the crossing. The same is then true for $X-YZ^{-1}Y^*$. For a bichromatic crossing v, there is no column of Y that corresponds to v, so this 4×4 minor is nothing but the nonzero 4×4 minor of A_v. For a monochromatic crossing v, the 4×4 minor of $X-YZ^{-1}Y^*$ is obtained by performing the corresponding matrix operations to the 5×5 minor of A_v. The computation is similar to the single variable case; see [16] for more details. The explicit values for these local contributions to $X-YZ^{-1}Y^*$ are given in Figure 9. Since the minor for a negative crossing turns out to be the negative of the minor for a positive one, we only provide these minors in the case of positive crossings.

We need to perform one last change of basis, which we now describe. If the sublink L_i winds around the region a a total of α_i times, then multiply the basis element corresponding to a with $\prod_{i=1}^{\mu}(-\omega_i^{-1/2})^{\alpha_i}$. This change of basis alters the matrices in Figure 9 in the following way. If v is a monochromatic crossing of color j, then L_j winds around b once more than it does around a and c, and it winds around d once fewer. Thus the rows corresponding to b and d are multiplied by $-\omega_j^{-1/2}$ and $-\omega_j^{1/2}$ respectively, and the columns are multiplied by the conjugates $-\omega_j^{1/2}$ and $-\omega_j^{-1/2}$. If v is a bichromatic crossing of colors (j,k), then L_j winds once more around a and b than it does around c and d, and L_k winds once more around b and c than it does around a and d. Thus we multiply the row for a by $-\omega_j^{-1/2}$, the row for c by $-\omega_k^{-1/2}$, and the row for b by $\omega_j^{-1/2}\omega_k^{-1/2}$, and we multiply the corresponding columns by the conjugates.

Remarkably, the sum of local contributions to $X-YZ^{-1}Y^*$ from Figure 9 can now be written in terms of the single matrix $\tau_v(\omega)$, the evaluation of $\tau_v(x)$ at $x_j=\mathrm{Re}(\omega_j^{1/2})$ and $x_{jk}=\mathrm{Re}(\omega_j^{1/2}\omega_k^{1/2})$, in both the monochromatic and bichromatic cases. Indeed, for a monochromatic crossing v, it coincides with $(\operatorname{sgn} v)\tau_v(\omega)$, while for a bichromatic crossing, it yields

$$4 \operatorname{sgn} v \sqrt{1-x_{j_v}^2}\sqrt{1-x_{k_v}^2}\,\tau_v(\omega).$$

The result of the matrix $X - YZ^{-1}Y^*$ after this change of basis thus gives

$$\sum_{v \text{ monochr}} \Bigl(\prod_{i \neq j_v} |s_i|^2\Bigr) \operatorname{sgn} v\, \tau_v(\omega) + \sum_{v \text{ bichr}} \Bigl(\prod_{i \neq j_v, k_v} |s_i|^2\Bigr) 4 \operatorname{sgn} v \sqrt{1-x_{j_v}^2}\sqrt{1-x_{k_v}^2}\,\tau_v(\omega) \tag{12}$$
$$= \frac{1}{4}\Bigl(\prod_{i=1}^{\mu} |s_i|^2\Bigr) \sum_v \frac{\operatorname{sgn} v}{\sqrt{1-x_{j_v}^2}\sqrt{1-x_{k_v}^2}}\,\tau_v(\omega) = \frac{1}{4}\Bigl(\prod_{i=1}^{\mu} |s_i|^2\Bigr)\tilde{\tau}_D(\omega).$$

The positive constant $\frac{1}{4}\prod_{i=1}^{\mu}|s_i|^2$ affects neither the signature nor the nullity, so we have

$$2\sigma_L(\omega) = \sigma(X - YZ^{-1}Y^*) + \sigma(Z^{-1}) = \sigma(\tilde{\tau}_D(\omega)) - w_m(D),$$
$$2\eta_L(\omega) = \eta(X - YZ^{-1}Y^*) + \eta(Z^{-1}) = \eta(\tilde{\tau}_D(\omega)). \qquad \square$$

Note 3.2 In the manipulations of matrices throughout this proof, we never used the fact that the ω_j are complex numbers (except, of course, when computing signatures): the only property needed is that $\omega_j\bar{\omega}_j = 1$. Therefore, everything works equally well if we consider the $\omega_j^{1/2}$ as *formal variables*, and set

$$\bar{\omega}_j^{1/2} := \omega_j^{-1/2}, \quad x_j = \operatorname{Re}(\omega_j^{1/2}) := \tfrac{1}{2}(\omega_j^{1/2} + \omega_j^{-1/2}), \quad \sqrt{1-x_j^2} = \operatorname{Im}(\omega_j^{1/2}) := \tfrac{1}{2i}(\omega_j^{1/2} - \omega_j^{-1/2}).$$

3.2 The Conway function

We now discuss how to compute the Conway function of a colored link from the matrix $\tilde{\tau}_D(x)$ and prove the second point of Theorem 1.3, which we now restate for convenience.

Proposition 3.3 *If D is a connected μ-colored diagram for a μ-colored link L, we have the equality*

$$\nabla_L^2(t_1, \dots, t_\mu) = \frac{1}{(t_1 - t_1^{-1})^2}\Bigl(\prod_v (-\operatorname{sgn} v)\bigl(\tfrac{1}{2}(t_j - t_j^{-1})\bigr)\bigl(\tfrac{1}{2}(t_k - t_k^{-1})\bigr)\Bigr) \cdot \det \tilde{\tau}_D(t^2),$$

where the product is over all crossings of D, the indices j, k are the (possibly identical) colors of the two strands crossing at v, and $\tau_D(t^2)$ stands for the evaluation of $\tau_D(x)$ at

$$x_j = \tfrac{1}{2}(t_j + t_j^{-1}), \quad x_{jk} = \tfrac{1}{2}(t_j t_k + t_j^{-1}t_k^{-1}).$$

Our starting point is (5), which expresses the Conway function ∇_L in terms of the matrix A_S associated to a connected C-complex S for L; recall (4). If $H(\omega)$ denotes the matrix defined in (2), which is used for computing the multivariable signature, and we consider $\omega_i^{-1/2} =: t_i$ as a *formal variable*, we have

$$H(t^{-2}) := H(t_1^{-2}, \dots, t_\mu^{-2}) = \sum_{\varepsilon \in \{\pm 1\}^\mu} \Bigl(\prod_{i=1}^{\mu}(1 - t_i^{2\varepsilon_i})\Bigr) A^\varepsilon = \sum_{\varepsilon \in \{\pm 1\}^\mu} \Bigl(\prod_{i=1}^{\mu} \varepsilon_i t_i^{\varepsilon_i} \varepsilon_i (t_i^{-\varepsilon_i} - t_i^{\varepsilon_i})\Bigr) A^\varepsilon \tag{13}$$
$$= \Bigl(\prod_{i=1}^{\mu} -(t_i - t_i^{-1})\Bigr) \sum_{\varepsilon \in \{\pm 1\}^\mu} \Bigl(\prod_{i=1}^{\mu} \varepsilon_i t_i^{\varepsilon_i}\Bigr) A^\varepsilon = (-1)^\mu \Bigl(\prod_{i=1}^{\mu}(t_i - t_i^{-1})\Bigr) A_S.$$

Hence, the Conway function can in fact be computed from the matrix H.

In order to prove Proposition 3.3, we adopt the same strategy as in the computation of the signature: starting from a *connected* diagram D, we use the C-complex S for $L \#_1 rL$ constructed in the previous section to compute the Conway function of $L \#_1 rL$, and conclude by applying well-known formulas relating the Conway function of a connected sum to the Conway functions of the summands. By construction, requiring D to be connected precisely means that S is connected.

Let $D = D_1 \cup \cdots \cup D_\mu$ be a connected, μ-colored diagram of a μ-colored link L and let $S = S_1 \cup \cdots \cup S_\mu$ be the C-complex for $L \#_1 rL$ defined in the proof of Proposition 3.1. As before, the notation $\#_1$ stands for the connected sum performed along a component of color 1. Let n_m and n_b denote the number of monochromatic and bichromatic crossings of D, respectively, and $n = n_m + n_b$ be the total number of crossings. Similarly, let $n_{m,i}$ and $n_{b,i}$ denote respectively the number of monochromatic and bichromatic crossings of D *without* the color i. Finally, recall that $\operatorname{sgn} S$ denotes the product of the signs of the clasps of S (see Figure 4).

Lemma 3.4 *With the notations above, the C-complex S satisfies:*

(i) $\operatorname{sgn} S = \prod_{v \text{ bichr}} \operatorname{sgn} v$, *where the product is taken over all bichromatic crossings.*

(ii) *Its first Betti number is equal to $b_1(S) = n + n_m$ and is even.*

(iii) $\chi(S \setminus S_1) = -n_{b,1} - 2n_{m,1}$ *and* $\chi(S \setminus S_i) = 1 - n_{b,i} - 2n_{m,i}$ *for all* $i \neq 1$.

Proof The first equality is clear by construction, since S has one clasp for each bichromatic crossing of D, and the sign of the clasp is equal to the sign of the corresponding crossing. To check the second point, let r denote the number of regions of D. By construction, we have $b_1(S) = (r-2) + n_m$, while an Euler characteristic computation yields the equality $r-2 = n$. Since n_b is always even, it follows that $b_1(S) = n + n_m = n_b + 2n_m$ is also even. As for the third point, one just needs to notice that $S \setminus S_i$ deformation retracts onto a graph Γ_i constructed from (the planar projection of) the diagram $D \setminus D_i$ by adding one loop to each monochromatic crossing and, if $i \neq 1$, by removing one edge of color 1 (which corresponds to performing the connected sum). The number of vertices of Γ_i minus the number of its edges yields the result. □

To shorten our formulas, we use the notation $s_b := \prod_{v \text{ bichr}} \operatorname{sgn} v$ and $s_m := \prod_{v \text{ monochr}} \operatorname{sgn} v$, where the product is taken respectively over all bichromatic and monochromatic crossings of D. We are now ready to prove Proposition 3.3.

Proof of Proposition 3.3 In what follows, we always evaluate H at $\omega_i = t_i^{-2}$, considered as formal variables, and rely on Note 3.2 to use the computations from the proof of Proposition 3.1 in this formal setting. Recall that, by (11) and the ensuing discussion, there exists a matrix M' such that

$$M' H(\omega) M'^* = \tilde{\tau}'_D(\omega) \oplus Z^{-1},$$

where the matrix

$$Z = \left(\prod_{i=1}^{\mu} (1-\omega_i)(1-\omega_i^{-1}) \right) \operatorname{diag}(-\operatorname{sgn} v)$$

is indexed by the monochromatic crossings of D, and $\det M' = \det M'^{*} = \det Z^{-1}$ — in the notations from the proof of Proposition 3.1, $\tilde{\tau}'_D(\omega)$ is the matrix $X - YZ^{-1}Y^{*}$ after the final change of basis, as in (12). Therefore,

$$\det H(\omega) = \det Z \det \tilde{\tau}'_D(\omega) = (-1)^{n_m} s_m \biggl(\prod_{i=1}^{\mu} (1-\omega_i)(1-\omega_i^{-1}) \biggr)^{n_m} \det \tilde{\tau}'_D(\omega).$$

Evaluating at $\omega_i = t_i^{-2}$, and using the equality $(1-t_i^2)(1-t_i^{-2}) = -(t_i - t_i^{-1})^2$, we obtain

$$\det H(t^{-2}) = (-1)^{n_m + \mu n_m} s_m \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{2n_m} \biggr) \det \tilde{\tau}'_D(t^2). \tag{14}$$

Furthermore, since $\tilde{\tau}'_D$ is a matrix of size n and (12) yields

$$\tilde{\tau}'_D(\omega) = \frac{1}{4} \biggl(\prod_{i=1}^{\mu} (1-\omega_i)(1-\omega_i^{-1}) \biggr) \tilde{\tau}_D(\omega),$$

we have

$$\det \tilde{\tau}'_D(t^2) = (-1)^{\mu n} \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{2n} \biggr) \det\bigl(\tfrac{1}{4} \tilde{\tau}_D(t^2)\bigr). \tag{15}$$

Putting everything together, and writing t for $(t_1, \dots, t_\mu)$, we obtain

$$\begin{aligned}
&\nabla_L^2(t) \\
&\quad = \nabla_L(t) \nabla_{rL}(t) = \frac{1}{t_1 - t_1^{-1}} \nabla_{L \#_1 rL}(t) \\
&\quad \overset{(5)}{=} \frac{1}{t_1 - t_1^{-1}} (-1)^{b_1(S)} (\operatorname{sgn} S) \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{\chi(S \setminus S_i) - 1} \biggr) \det A_S \\
&\quad \overset{(13)}{=} \frac{s_b}{t_1 - t_1^{-1}} \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{\chi(S \setminus S_i) - 1} \biggr) (-1)^{\mu b_1(S)} \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{-b_1(S)} \biggr) \det H(t^{-2}) \\
&\quad \overset{(14)}{=} \frac{s_b}{t_1 - t_1^{-1}} \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{\chi(S \setminus S_i) - 1 - b_1(S)} \biggr) (-1)^{n_m + \mu n_m} s_m \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{2n_m} \biggr) \det \tilde{\tau}'_D(t^2) \\
&\quad \overset{(15)}{=} \frac{(-1)^{n_m + \mu n_m} s_b s_m}{t_1 - t_1^{-1}} \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{\chi(S \setminus S_i) - 1 - b_1(S) + 2n_m} \biggr) (-1)^{\mu n} \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{2n} \biggr) \det\bigl(\tfrac{1}{4} \tilde{\tau}_D(t^2)\bigr) \\
&\quad = \frac{(-1)^{n_m}}{(t_1 - t_1^{-1})^2} \cdot \frac{s_b s_m}{4^n} \biggl(\prod_{i=1}^{\mu} (t_i - t_i^{-1})^{-n_{b,i} - 2n_{m,i} + n + n_m} \biggr) \det \tilde{\tau}_D(t^2),
\end{aligned}$$

where in the first line we used Corollary 2 and Proposition 5 of [2]. The following equalities derive, as indicated, from (5), (13), (14) and (15), together with the first point of Lemma 3.4 in the third line (with (13)), the second point in the fourth line (with (14)), and the second and third points in the last line.

To conclude, we note that the exponent $-n_{b,i} - 2n_{m,i} + n + n_m$ appearing in the last line is equal to the number of bichromatic crossings involving a strand of color i plus twice the number of monochromatic crossings of color i. Therefore,

$$\frac{1}{4^n}(s_b s_m)\prod_{i=1}^{\mu}(t_i-t_i^{-1})^{-n_{b,i}-2n_{m,i}+n+n_m}=\prod_v(\operatorname{sgn} v)\bigl(\tfrac{1}{2}(t_j-t_j^{-1})\bigr)\bigl(\tfrac{1}{2}(t_k-t_k^{-1})\bigr),$$

where the product on the right-hand side is over all crossings of D and the indices j,k are the two (possibly identical) colors of strands crossing at v. The proposition now follows from observing that $(-1)^{n_m}=(-1)^n$ since $n=n_m+n_b$ and n_b is even. □

3.3 The multivariate Kauffman model

Having finished the proof of Theorem 1.3, we now turn our attention to Corollary 1.5.

Starting from a *connected* diagram D of a μ-colored link, let K_D (or simply K) be the matrix defined by the labels in Figure 3, and $\widetilde{K}$ be the square matrix obtained from K by removing two columns corresponding to two adjacent regions of D separated by a strand of color 1. Corollary 1.5 is a direct consequence of Proposition 3.3 together with the following lemma.

Lemma 3.5 *Let $S=(S_{v,v})$ be the diagonal matrix indexed by the crossings of D with coefficients*

$$S_{v,v}=\frac{-4\operatorname{sgn} v}{(t_j-t_j^{-1})(t_k-t_k^{-1})},$$

where j and k are the colors of the two strands meeting at v. Then, we have

$$\tau_D(t^2)=K^{\mathrm{T}}SK.$$

We start by proving Corollary 1.5, before addressing the proof of Lemma 3.5.

Proof of Corollary 1.5 By Lemma 3.5, we have $\tilde{\tau}_D(t^2)=\widetilde{K}^{\mathrm{T}}S\widetilde{K}$, yielding the equality

$$(\det\widetilde{K})^2=\det S^{-1}\det\tilde{\tau}_D(t^2).$$

Since

$$\det S^{-1}=\prod_v(-\operatorname{sgn} v)\bigl(\tfrac{1}{2}(t_j-t_j^{-1})\bigr)\bigl(\tfrac{1}{2}(t_k-t_k^{-1})\bigr),$$

Proposition 3.3 implies that $(\det\widetilde{K})^2=(t_1-t_1^{-1})^2\nabla_L^2(t_1,\dots,t_\mu)$, and Corollary 1.5 follows. □

Proof of Lemma 3.5 Recall that $K=(K_{v,f})$ is a matrix with rows indexed by the crossings of D and columns indexed by the regions of D. Let us fix two regions f and g of D. By definition, the corresponding coefficient of $K^{\mathrm{T}}SK$ is

$$(K^{\mathrm{T}}SK)_{f,g}=\sum_v\frac{-4\operatorname{sgn} v}{(t_j-t_j^{-1})(t_k-t_k^{-1})}K_{v,f}K_{v,g}, \tag{16}$$

while the corresponding coefficient of $\tau_D(t^2)$ is

$$(\tau_D(t^2))_{f,g}=\sum_v\frac{-4\operatorname{sgn} v}{(t_j-t_j^{-1})(t_k-t_k^{-1})}(\tau_v(t^2))_{f,g}. \tag{17}$$

In both cases, the sum is over all crossings of D (and j,k denote the colors of the strands crossing at v), but the only nonzero contributions come from the crossings adjacent to both f and g.

Comparing the labels of Figure 2 evaluated at $x_j = \frac{1}{2}(t_j + t_j^{-1})$ and $x_{jk} = \frac{1}{2}(t_j t_k + t_j^{-1} t_k^{-1})$ with the labels of Figure 3, one notices an interesting relation. To state it precisely, let $\mathbb{Q}(t)$ denote the field of fractions of $\mathbb{Z}[t_1^{\pm 1}, \dots, t_\mu^{\pm 1}]$, and let $\varphi\colon \mathbb{Q}(t) \to \mathbb{Q}(t)$ be the involution induced by $t_i \mapsto t_i^{-1}$ for all i. We claim that the following equality holds:

$$(\tau_v(t^2))_{f,g} = \tfrac{1}{2}(K_{v,f} K_{v,g} + \varphi(K_{v,f} K_{v,g})). \tag{18}$$

The proof of this claim is divided into three cases, depending on the relative positions of the regions f and g. Let us first assume that f and g are two different regions of the same checkerboard color. In such a case, we have $K_{v,f} K_{v,g} = 1 = (\tau_v(t^2))_{f,g}$ for each crossing v incident to both f and g, so (18) holds. Let us now assume that f and g are regions with different checkerboard colors, meeting at a crossing v with strands of colors j and k. If f and g are adjacent to the strand of color j (resp. k), we get $K_{v,f} K_{v,g} = t_k^{\pm 1}$ (resp. $t_j^{\pm 1}$). Since the coefficient of $\tau_v(t^2)$ is $x_k = \frac{1}{2}(t_k + t_k^{-1})$ (resp. $x_j = \frac{1}{2}(t_j + t_j^{-1})$), (18) holds in this case as well. Finally, let us assume that $f = g$. For a crossing v incident to f, we get either $K_{v,f}^2 = (t_j t_k)^{\pm 1}$ or $K_{v,f}^2 = (t_j^{-1} t_k)^{\pm 1}$, depending on the position of f around v. Similarly, the corresponding coefficient of $\tau_v(t^2)$ is either $x_{jk} = \frac{1}{2}(t_j t_k + t_j^{-1} t_k^{-1})$ or $2x_j x_k - x_{jk} = \frac{1}{2}(t_j t_k^{-1} + t_j^{-1} t_k)$, respectively. This concludes the proof of (18).

The equations (16), (17) and (18) immediately imply the equality

$$\tau_D(t^2) = \tfrac{1}{2}(K^{\mathrm{T}} S K + \varphi(K^{\mathrm{T}} S K)),$$

where φ is applied to matrices coefficientwise. To conclude the proof of Lemma 3.5, it remains to check that $\varphi(K^{\mathrm{T}} S K) = K^{\mathrm{T}} S K$. This fact being surprisingly technical, we make it the object of a final separate lemma. □

Lemma 3.6 *Let $\mathbb{Q}(t)$ denote the field of fractions of $\mathbb{Z}[t_1^{\pm 1}, \dots, t_\mu^{\pm 1}]$, and let $\varphi\colon \mathbb{Q}(t) \to \mathbb{Q}(t)$ be the involution induced by $t_i \mapsto t_i^{-1}$ for all i. Then, we have the equality $\varphi(K^{\mathrm{T}} S K) = K^{\mathrm{T}} S K$.*

Proof We have to show that, for any two regions f and g of D, the coefficient

$$(K^{\mathrm{T}} S K)_{f,g} = \sum_v \frac{-4 \operatorname{sgn} v}{(t_j - t_j^{-1})(t_k - t_k^{-1})} K_{v,f} K_{v,g}$$

is invariant under φ, where the sum is taken over all crossings adjacent to both f and g. We will consider several cases, according to the relative positions of f and g with respect to the crossings.

First of all, if f and g are two regions of the same checkerboard color, each common crossing v contributes a term $-4 \operatorname{sgn} v / ((t_j - t_j^{-1})(t_k - t_k^{-1}))$ to the coefficient $(K^{\mathrm{T}} S K)_{a,b}$. Since all these terms are invariant under φ, this case is checked.

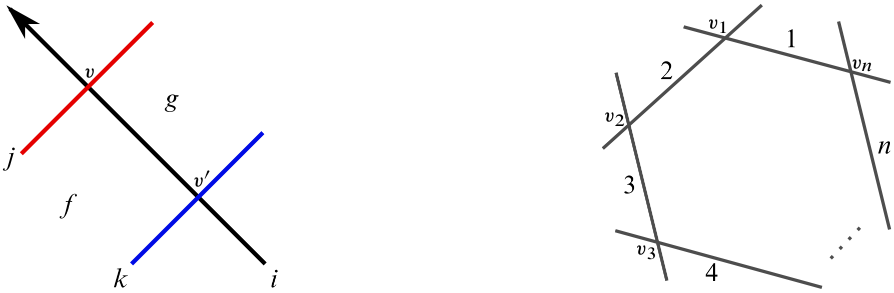


Figure 10: The conventions in the proof of Lemma 3.6.

Next, suppose that f and g have different checkerboard colors. Each edge of D adjacent to both regions gives two contributions to the sum, one for each crossing adjacent to the edge. So, let us consider a common edge with endpoints v and v', and suppose without loss of generality that the colors and orientations of the strands are as in Figure 10 (left). The two contributions sum up to

$$\frac{-4st_j^{-s}}{(t_i-t_i^{-1})(t_j-t_j^{-1})}+\frac{-4s't_k^{s'}}{(t_i-t_i^{-1})(t_k-t_k^{-1})},$$

where $s=\operatorname{sgn} v$ and $s'=\operatorname{sgn} v'$. Proving that the term displayed above is invariant under φ is clearly equivalent to showing that

$$G:=st_j^{-s}(t_k-t_k^{-1})+s't_k^{s'}(t_j-t_j^{-1})$$

satisfies $\varphi(G)=-G$. Expanding the products and denoting by χ the characteristic function, one checks that G is equal to

$$(t_jt_k-t_j^{-1}t_k^{-1})(\chi_{s'=1}-\chi_{s=-1})+(t_jt_k^{-1}-t_j^{-1}t_k)(\chi_{s=-1}-\chi_{s'=-1}),$$

which is clearly antisymmetric, thus finishing this case.

Finally, let us consider the diagonal coefficient corresponding to a region f. Suppose that, when moving around the boundary of f counterclockwise, one encounters n crossings $v_1,\dots,v_n$ of respective signs $s_1,\dots,s_n$. (It can happen that f abuts the same crossing from two sides, but since the corresponding labels are added, our computations remain valid in this case.) Let us also number the edges of the boundary from 1 to n as in Figure 10 (right). To each edge, we assign a sign $\varepsilon_i\in\{\pm1\}$, where $\varepsilon_i=1$ if the edge i is oriented coherently with the counterclockwise orientation of the boundary of f, and $\varepsilon_i=-1$ otherwise. Without loss of generality, we can assume that all the edges have different colors, that we also denote by $1,\dots,n$; in the general case, if two colors coincide, one simply needs to identify the corresponding variables in the following computations, a transformation which does not affect the symmetry.

With these notations and the help of Figure 3, one computes

$$(K^{\mathrm{T}}SK)_{f,f}=\sum_{i=1}^{n}\frac{-4s_i(t_i^{\varepsilon_{i+1}}t_{i+1}^{-\varepsilon_i})^{s_i}}{(t_i-t_i^{-1})(t_{i+1}-t_{i+1}^{-1})}.$$

As before, proving that this coefficient is invariant under φ is equivalent to showing that

$$G := \prod_{j=1}^{n}(t_j - t_j^{-1}) \sum_{i=1}^{n} \frac{s_i (t_i^{\varepsilon_{i+1}} t_{i+1}^{-\varepsilon_i})^{s_i}}{(t_i - t_i^{-1})(t_{i+1} - t_{i+1}^{-1})}$$

satisfies $\varphi(G) = (-1)^n G$. Expanding as a sum of monomials, we obtain

$$\begin{aligned} G &= \sum_{i=1}^{n} \Big(\prod_{j \neq i, i+1} (t_j - t_j^{-1}) \Big) s_i (t_i^{\varepsilon_{i+1}} t_{i+1}^{-\varepsilon_i})^{s_i} = \sum_{i=1}^{n} \sum_{\alpha \in \{\pm 1\}^{n-2}} \Big(\prod_{j \neq i, i+1} \alpha_j t_j^{\alpha_j} \Big) s_i (t_i^{\varepsilon_{i+1}} t_{i+1}^{-\varepsilon_i})^{s_i} \\ &= \sum_{\beta \in \{\pm 1\}^n} c_\beta t_1^{\beta_1} \cdots t_n^{\beta_n} \end{aligned}$$

for some coefficient c_β. To compute these coefficients explicitly, let us define for each $\beta \in \{\pm 1\}^n$ the (possibly empty) set of indices $I_\beta = \{i \in \{1, \dots, n\} \mid \beta_i = \varepsilon_{i+1} s_i,\ \beta_{i+1} = -\varepsilon_i s_i\}$. We then have

$$c_\beta = \sum_{i \in I_\beta} \beta_1 \cdots \beta_{i-1} s_i \beta_{i+2} \cdots \beta_n = \sum_{i \in I_\beta} s_i \beta_i \beta_{i+1} (\beta_1 \cdots \beta_n) = \beta_1 \cdots \beta_n \sum_{i \in I_\beta} s_i \beta_i \beta_{i+1} = \beta_1 \cdots \beta_n d_\beta,$$

with $d_\beta = \sum_{i \in I_\beta} s_i \beta_i \beta_{i+1}$. The desired equality $\varphi(G) = (-1)^n G$ is equivalent to $c_{-\beta} = (-1)^n c_\beta$ for all $\beta \in \{\pm 1\}^n$, which in turns is equivalent to $d_{-\beta} = d_\beta$.

Therefore, we are left with the proof of the equality $d_{-\beta} = d_\beta$ for all $\beta \in \{\pm 1\}^n$. Given any such β and any index $i \in \{1, \dots, n\}$, define $\tilde\beta$ by $\tilde\beta_i = -\beta_i$ and $\tilde\beta_j = \beta_j$ for $j \neq i$. A straightforward but slightly cumbersome computation yields

$$d_\beta - d_{\tilde\beta} = \begin{cases} -\varepsilon_i \beta_i & \text{if } (\beta_{i-1}, \beta_{i+1}) = (-\varepsilon_i s_{i-1}, -\varepsilon_i s_i), \\ \varepsilon_i \beta_i & \text{if } (\beta_{i-1}, \beta_{i+1}) = (\varepsilon_i s_{i-1}, \varepsilon_i s_i), \\ 0 & \text{otherwise.} \end{cases}$$

This expression is invariant if we replace β by $-\beta$. It thus follows that, for any two $\beta, \beta' \in \{\pm 1\}^n$, we have $d_\beta - d_{\beta'} = d_{-\beta} - d_{-\beta'}$. To prove the equality $d_{-\beta} = d_\beta$ for all β, we therefore only need to check it for a single β. Taking $\beta = (1, \dots, 1)$, we get

$$d_\beta - d_{-\beta} = \sum_{i : (\varepsilon_i, \varepsilon_{i+1}) = (-s_i, s_i)} s_i - \sum_{i : (\varepsilon_i, \varepsilon_{i+1}) = (s_i, -s_i)} s_i = \sum_{i : \varepsilon_i \neq \varepsilon_{i+1}} \varepsilon_{i+1} = 0,$$

since there is an even number of crossings at which ε_i changes sign, going from 1 to -1 in exactly half of the cases and from -1 to 1 in the others. □

3.4 The Alexander module

We conclude this article with a slightly informal discussion on yet another abelian link invariant, namely the Alexander module (recall its definition from Section 2.3).

The question we address is whether it is possible to obtain a presentation of (the square of) the Alexander module $\mathscr{A}_L$ of a μ-colored link L from the matrix $\tilde\tau_D(x)$ associated to a colored diagram D for L. As we will see, the answer is positive in the case $\mu = 1$, but not in general.

First, recall that $\mathscr{A}_L$ does not admit a square presentation matrix over $\Lambda = \mathbb{Z}[t_1^{\pm 1}, \dots, t_\mu^{\pm 1}]$ if $\Delta_L \neq 0$ and $\mu \geq 4$ [9]. For this reason alone, there is no hope of answering the above question positively in general. However, the module $\mathscr{A}_L$ does admit a square presentation matrix over the localized ring

$$\Lambda_S := \mathbb{Z}[t_1^{\pm 1}, \dots, t_\mu^{\pm 1}, (t_1 - 1)^{-1}, \dots, (t_\mu - 1)^{-1}].$$

By Corollary 3.6 of [4], the generalized Seifert matrices can be used to compute such a square presentation matrix. Therefore, it is natural to hope that the strategy developed in this work could be applied to this invariant as well. However, since the change of variables $2x_j = t_j^{1/2} + t_j^{-1/2}$ makes use of fractional powers of the variables, we will need to work over the slightly larger ring

$$\Lambda'_S := \mathbb{Z}[t_1^{\pm 1/2}, \dots, t_\mu^{\pm 1/2}, (t_1 - 1)^{-1}, \dots, (t_\mu - 1)^{-1}].$$

In the case $\mu = 1$, this program can be carried out, yielding the following result. Let D be a connected diagram for an oriented link L. As one easily checks, the coefficients of the matrix $\tau_D(x)$ are polynomials in $2x =: y$. Let $\mathscr{M}_D$ denote the $\mathbb{Z}[y]$-module presented by the matrix $\tilde{\tau}_D(y)$. Then, we have an isomorphism of Λ'_S-modules

$$\mathscr{M}_D \otimes_{\mathbb{Z}[y]} \Lambda'_S \simeq \mathscr{A}_L^{\oplus 2} \otimes_\Lambda \Lambda'_S,$$

where $\Lambda'_S = \mathbb{Z}[t^{\pm 1/2}, (t-1)^{-1}]$ is a $\mathbb{Z}[y]$-module via the ring homomorphism $\mathbb{Z}[y] \to \Lambda'_S$ mapping y to $t^{1/2} + t^{-1/2}$, and a Λ-module via the natural inclusion $\Lambda \to \Lambda'_S$. Less formally, one can say that the matrix $\tilde{\tau}_D(y)$ is a presentation matrix of $\mathscr{A}_L^{\oplus 2}$ over Λ'_S via the substitution $y = t^{1/2} + t^{-1/2}$. If $L = K$ is a knot, then the multiplication by $(t-1)$ is invertible in $\mathscr{A}_K$. As a consequence, the matrix $\tilde{\tau}_D(y)$ presents $\mathscr{A}_K^{\oplus 2}$ over the ring $\mathbb{Z}[t^{\pm 1/2}]$.

Unfortunately, these results do not carry over to the case $\mu > 1$. Indeed, let D be a μ-colored diagram for a μ-colored link L, and assume that each pair of colors meet in D. Then, using Corollary 3.6 of [4], it is possible to prove that the matrix $\tilde{\tau}_D(x)$ presents the Alexander module of $L \#_1 rL$ over the ring

$$\mathbb{Z}\left[\tfrac{1}{2}, t_1^{\pm 1/2}, \dots, t_\mu^{\pm 1/2}, (t_1 - 1)^{-1}, \dots, (t_\mu - 1)^{-1}\right],$$

under the substitutions $x_j = \frac{1}{2}(t_j + t_j^{-1})$ and $x_{jk} = \frac{1}{2}(t_j t_k + t_j^{-1} t_k^{-1})$. However, the isomorphism

$$\mathscr{A}_{L \#_1 L'} \simeq \mathscr{A}_L \oplus \mathscr{A}_{L'}$$

that we used in the case $\mu = 1$ is no longer valid in general for $\mu > 1$. In other words, the additivity under connected sum enjoyed by the other abelian invariants considered in this work does not extend to the Alexander module in general.

References

[1] **J W Alexander**, *Topological invariants of knots and links*, Trans. Amer. Math. Soc. 30 (1928) 275–306 MR

[2] **D Cimasoni**, *A geometric construction of the Conway potential function*, Comment. Math. Helv. 79 (2004) 124–146 MR

[3] **D Cimasoni**, **L Ferretti**, *On the Kashaev signature conjecture*, Fund. Math. 266 (2024) 275–287 MR

[4] **D Cimasoni**, **V Florens**, *Generalized Seifert surfaces and signatures of colored links*, Trans. Amer. Math. Soc. 360 (2008) 1223–1264 MR

[5] **A Conway**, *The Levine–Tristram signature: a survey*, from "2019–20 MATRIX annals" (D R Wood, J de Gier, C E Praeger, T Tao, editors), MATRIX Book Ser. 4, Springer, Cham (2021) 31–56 MR

[6] **A Conway**, **M Nagel**, **E Toffoli**, *Multivariable signatures, genus bounds, and* 0.5*-solvable cobordisms*, Michigan Math. J. 69 (2020) 381–427 MR

[7] **J H Conway**, *An enumeration of knots and links, and some of their algebraic properties*, from "Computational Problems in Abstract Algebra" (J Leech, editor), Pergamon (1970) 329–358 MR

[8] **D Cooper**, *The universal abelian cover of a link*, from "Low-dimensional topology" (R Brown, T L Thickstun, editors), London Math. Soc. Lecture Note Ser. 48, Cambridge Univ. Press (1982) 51–66 MR

[9] **R H Crowell**, **D Strauss**, *On the elementary ideals of link modules*, Trans. Amer. Math. Soc. 142 (1969) 93–109 MR

[10] **C W Davis**, **T Martin**, **C Otto**, *Moves relating C-complexes: a correction to Cimasoni's "A geometric construction of the Conway potential function"*, Topology Appl. 302 (2021) art. id. 107799 MR

[11] **S Friedl**, **C Kausik**, **J P Quintanilha**, *An algorithm to calculate generalized Seifert matrices*, J. Knot Theory Ramifications 31 (2022) art. id. 2250068 MR

[12] **R Hartley**, *The Conway potential function for links*, Comment. Math. Helv. 58 (1983) 365–378 MR

[13] **R Kashaev**, *On symmetric matrices associated with oriented link diagrams*, from "Topology and geometry — a collection of essays dedicated to Vladimir G Turaev" (A Papadopoulos, editor), IRMA Lect. Math. Theor. Phys. 33, Eur. Math. Soc., Zürich (2021) 131–145 MR

[14] **L H Kauffman**, *The Conway polynomial*, Topology 20 (1981) 101–108 MR

[15] **L H Kauffman**, *Formal knot theory*, Mathematical Notes 30, Princeton Univ. Press (1983) MR

[16] **J Liu**, *A proof of the Kashaev signature conjecture*, preprint (2023) arXiv 2311.01923

[17] **K Murasugi**, *On the signature of links*, Topology 9 (1970) 283–298 MR

[18] **M Sato**, *On the Conway potential function introduced by Kauffman*, preprint (2011) arXiv 1103.2449

[19] **C B Zibrowius**, *On a Heegaard Floer theory for tangles*, PhD thesis, University of Cambridge (2017) Available at `https://doi.org/10.17863/CAM.8706`

[20] **C B Zibrowius**, *Kauffman states and Heegaard diagrams for tangles*, Algebr. Geom. Topol. 19 (2019) 2233–2282 MR

Section de mathématiques, Université de Genève
Genève, Switzerland

Section de mathématiques, Université de Genève
Genève, Switzerland

Department of Mathematics, University of Toronto
Toronto, ON, Canada

`david.cimasoni@unige.ch`, `livio.ferretti@unige.ch`, `jessliu@math.toronto.edu`

Received: 27 May 2024 Revised: 4 October 2024

Algebraic & Geometric Topology **25**:8 (2025) 5137–5152
DOI: 10.2140/agt.2025.25.5137
Published: 20 November 2025

Equivariant double-slice genus, stabilization, and equivariant stabilization

MALCOLM GABBARD

We define the equivariant double-slice genus and equivariant superslice genus of a strongly invertible knot. We prove lower bounds for both the equivariant double-slice genus and the equivariant superslice genus. Using these bounds we find a family of knots which are double-slice and equivariantly slice, but have equivariant double-slice genus at least n. Using this result, we construct unknotted symmetric 2-spheres which do not bound symmetric 3-balls. Additionally, using double-slice and superslice genera we find effective lower bounds for 1-handle stabilization distance and identify a possible method for using equivariant double-slice and superslice genera to bound the symmetric 1-handle stabilization distance for symmetric surfaces.

57K10, 57K45; 57K40

1 Introduction

Given a knot $K \subset S^3$, its *double-slice genus* $g_{\mathrm{ds}}(K)$ was first defined by Livingston–Meier in [14] as the minimal genus of an unknotted surface $\Sigma \subset S^4$ such that Σ intersects an equatorial S^3 transversely with intersection K. In 2021, Chen [5] was able to construct a lower bound for the double-slice genus using Casson–Gordon invariants which allowed him to prove that there are slice knots with arbitrarily large double-slice genus. A later result by Orson–Powell in [17] using a new lower bound for double-slice genus from signatures refined this result, finding slice knots with double-slice genus exactly n for all n.

The discussion of equivariant genus of symmetric knots originates from Sakuma's work [19] on the equivariant concordance group of strongly invertible knots. A *strongly invertible knot* (K,τ) is a knot $K \subset S^3$ and an involution $\tau\colon S^3 \to S^3$ satisfying $\tau(K) = K$ and $\mathrm{fix}(\tau) = S^1$ intersecting K in two points. The equivariant concordance group of strongly invertible knots has received considerable attention in recent years, including in [3; 4; 7; 8], providing many obstructions to strongly invertible knots being equivariantly slice, ie bounding a symmetric disk in B^4. A more general analysis was recently initiated by Boyle–Issa in [3] where they define for a strongly invertible knot (K,τ) its *equivariant* 4*-genus* $\tilde{g}_4(K,\tau)$ to be the minimal genus of a properly embedded surface $\Sigma \subset B^4$ with $\partial\Sigma = K$, such that for some extension $\bar{\tau}\colon B^4 \to B^4$ of τ, $\bar{\tau}(\Sigma) = \Sigma$. Using knot Floer homology, Dai–Mallick–Stoffregen [6] were able to find slice knots which, viewed as strongly invertible knots, have arbitrarily large equivariant 4-genus.

Combining these ideas, we define the *equivariant double-slice genus* $\tilde{g}_{\mathrm{ds}}(K,\tau)$ of a strongly invertible knot (K,τ) to be the minimal genus of an equivariantly unknotted surface $\Sigma \subset S^4$ of which K appears

---

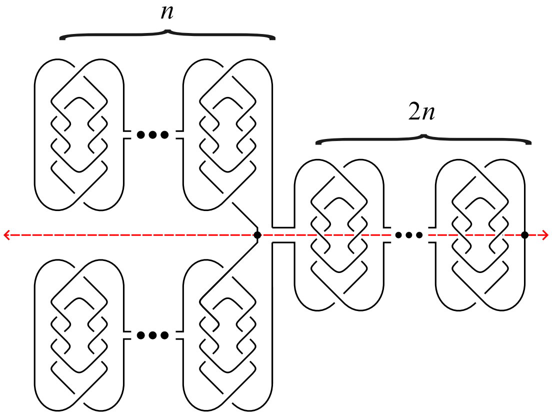


Figure 1: The knot (K_n, τ).

as the cross-section. In other words, it is the minimal genus of a surface Σ intersecting S^3 transversely with intersection K and bounding a handlebody H such that $\bar{\tau}(H) = H$ for some orientation preserving extension of τ to S^4. In order to differentiate the equivariant double-slice genus from the equivariant 4-genus and the double-slice genus, we prove the following lower bound:

Theorem 1.1 *Let (K, τ) be a strongly invertible knot and let K_0 and K_1 be the knots formed from the union of an arc of K with, respectively, the half-axes h_0 and h_1. Then*

$$\min\{g_{\mathrm{ds}}(K_0), g_{\mathrm{ds}}(K_1)\} \leq \tilde{g}_{\mathrm{ds}}(K, \tau).$$

This lower bound allows, to some extent, questions about the equivariant double-slice genus of a strongly invertible knot to be answered in terms of the nonequivariant double-slice genus of these other knots K_0 and K_1, allowing us to make use of the well-studied bounds for double-slice knots. Using this lower bound, we are able to prove that the equivariant double-slice genus of a knot (K, τ) does not depend on a combination of its double-slice genus and equivariant 4-genus.

Theorem 1.2 *The knot (K_n, τ) depicted in Figure 1 satisfies the following*:

(1) *K_n is double-slice.*

(2) *(K_n, τ) is equivariantly slice.*

(3) *$\tilde{g}_{\mathrm{ds}}(K_n, \tau) \geq n$.*

1.1 Equivariantly superslice knots

Similar to double-slice genus, there is a notion of *superslice* genus $g_{\mathrm{ss}}(K)$ defined by Chen in [5] as the minimal genus of a surface $\Sigma \subset B^4$ with boundary K such that the double of Σ along K in S^4 is unknotted. We extend this discussion to the equivariant setting by defining the *equivariant superslice genus* $\tilde{g}_{\mathrm{ss}}(K, \tau)$ of a knot to be the minimal genus of a surface $\Sigma \subset B^4$ such that $\bar{\tau}(\Sigma) = \Sigma$ and the double of Σ is equivariantly unknotted. Using similar techniques to the double-slice setting we are able to prove the following:

Theorem 1.3 *Let (K,τ) be a strongly invertible knot and let K_0 and K_1 be the knots formed from the union of an arc of K with, respectively, the half-axes h_0 and h_1. Then*

$$\min\{g_{\mathrm{ss}}(K_0), g_{\mathrm{ss}}(K_1)\} \le \tilde{g}_{\mathrm{ss}}(K,\tau).$$

1.2 Symmetric 2-knots

Similar to strongly invertible knots, one could ask what are the properties of 2-knots invariant under some symmetry of S^4. Using Theorem 1.3, we are able to construct interesting examples of symmetric 2-knots that distinguish them significantly from both nonequivariant 2-knots and strongly invertible knots. Namely, we are able to prove the following result:

Theorem 1.4 *There exists a symmetric 2-sphere in $(S^4,\bar{\tau})$ which bounds a 3-ball but bounds no equivariant 3-ball.*

1.3 Stabilization distance of disks rel boundary

The internal stabilization distance of embedded surfaces in 4-manifolds is a well-studied area [1; 2; 9; 10; 11; 12; 16; 20], which generally asks the following: given two embedded surfaces Σ_1 and Σ_2, how many internal stabilizations (additions of 2-dimensional 1-handles to Σ_1 and Σ_2) are necessary before Σ_1 and Σ_2 are isotopic? In recent work of Miller and Powell [16], this question was extended to internal stabilizations of properly embedded surfaces rel boundary. In particular, the authors define the 1*-handle stabilization distance* $d_1(\Sigma_1,\Sigma_2)$ between smoothly and properly embedded genus g surfaces $\Sigma_1,\Sigma_2 \subset B^4$ with common boundary K to be the minimal $n \subset \mathbb{N}$ such that Σ_1 and Σ_2 become ambiently isotopic rel boundary after each has been stabilized at most n times.

Using basic properties of double-slice and superslice genus, we are able to prove the following lower bound for the 1-handle stabilization distance of surfaces satisfying certain conditions:

Theorem 1.5 *Let Σ_1 and Σ_2 be properly embedded genus h surfaces with boundary $K \subset S^3$ such that $\Sigma_1 \cup_K \Sigma_2 \subset (B^4,\Sigma_1) \cup_{(S^3,K)} (B^4,\Sigma_2)$ is unknotted. Then $d_1(\Sigma_1,\Sigma_2) \ge g_{\mathrm{ss}}(K)-h$.*

By letting K be double-slice this immediately yields the following corollary:

Corollary 1.6 *Let K be double-slice with $g_{\mathrm{ss}}(K)=n$, then K admits slice disks D_1 and D_2 such that $d_1(D_1,D_2) \ge n$.*

From here, we define a symmetric notion of 1-handle stabilization distance $\tilde{d}_1^{\tau}(\Sigma_1,\Sigma_2)$ for certain classes of symmetric surfaces and prove the following symmetric analog of Theorem 1.5.

Theorem 1.7 *Let $\Sigma_1,\Sigma_2 \subset B^4$ be properly embedded genus h surfaces with boundary K which are both $\bar{\tau}$-invariant. If $\Sigma_1 \cup_K -\Sigma_2 \subset (B^4,\Sigma_1) \cup_{(S^3,K)} (B^4,\Sigma_2)$ is equivariantly unknotted, then $\tilde{d}_1^{\tau}(\Sigma_1,\Sigma_2) \ge \frac{1}{2}\tilde{g}_{\mathrm{ss}}(K,\tau)-h$.*

1.4 Organization

Section 2 provides an overview of notation and conventions, as well as necessary background on double-slice genus and equivariant 4-genus. Section 3 defines the equivariant double-slice genus and covers some properties of involutions on handlebodies arising in this setting. In Section 4, we bound the equivariant double-slice genus, proving Theorem 1.1 as well as Theorem 1.2. In Section 5, we define equivariant superslice genus and prove analogous results to Section 4, as well as results about equivariantly knotted 2-spheres. In Section 6, we discuss stabilization of surfaces rel boundary and introduce a notion of equivariant stabilization.

Acknowledgments

Many thanks to Dave Auckly for countless helpful conversations and Evan Scott for insights which helped find a mistake in an earlier version of the paper. The author was partially supported by NSF grant DMS-1952755.

2 Background

In this section, we recall the necessary background pertaining to double-slice classical knots and equivariantly slice strongly invertible knots.

2.1 Conventions and classical 4-genus

A knot K will refer to a classical knot, namely the oriented image of a smooth embedding of S^1 into S^3. From a given knot K we have the following related knots:

- rK is the *reverse* of a knot K, that is, K with the opposite orientation.
- mK is the *mirror* of a knot K, which is the image of K under a reflection of S^3.
- $-K$ is the *inverse* of a knot K, and is the reverse of the mirror.

Recall that the 4*-genus* $g_4(K)$ of a knot K is the minimal genus of a smooth properly embedded surface $\Sigma \subset B^4$ with $\partial\Sigma = K$. If $g_4(K) = 0$, we say that K is *slice*. One well-known fact we will use consistently is that given a knot K, $K \# -K$ is slice.

2.2 Double-slice genus

The *double-slice genus* of a knot K, denoted by $g_{\mathrm{ds}}(K)$, was defined by Livingston and Meier [14] as the minimal genus of an unknotted smooth surface $\Sigma \subset S^4$ such that the intersection of Σ with an equatorial S^3 is K. By unknotted, we mean that the surface Σ bounds a smoothly embedded handlebody in S^4. If the double-slice genus of K is 0, ie K is the cross-section of an unknotted sphere, we say that K is *double-slice*. One important fact we know from Sumners [26] is that given any knot K, $K \# -K$ is double-slice.

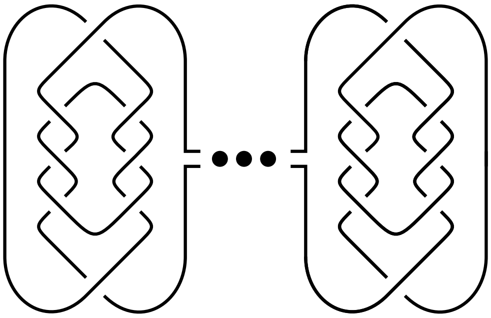


Figure 2: The knot L_n: n copies of 8_{20} summed together.

There are many obstructions to knots being double-slice, and a good survey is provided in [14]. The primary bound for double-slice genus we will use is Theorem 1.1 of [17] from Orson and Powell, which states:

Theorem 2.1 (Orson and Powell) *Let K be a knot in S^3 and $\sigma_\omega(K)$ be its signature function. Then*

$$g_{\mathrm{ds}}(K) \geq \max \sigma_\omega(K).$$

One important application of this theorem is the following example coming from Theorem 1.2 of [17] which we will refer to in future constructions:

Example 2.2 Let L be the knot denoted by 8_{20} in Rolfsen's table [18] and let $L_n = \#^n L$ as depicted in Figure 2. In [17] it is shown, using the additivity of the knot signature function, that $g_{\mathrm{ds}}(L_n) = n$, despite the fact that L_n is slice.

2.3 Equivariant 4-genus

To define the equivariant 4-genus, we first recall that a *strongly invertible knot* is a pair (K, τ) consisting of a classical knot K and an involution τ acting on S^3 such that $\operatorname{fix}(\tau) = S^1$ and $\operatorname{fix}(\tau) \cap K$ is two points. For a given strongly invertible knot (K, τ), we refer to the fixed point set as the *axis* and refer to the two intervals of the axis separated by the intersection with K as *half-axes*.

As it will often be useful to work with a specified half-axis, we recall that a *directed strongly invertible knot* is a triple (K, τ, h) consisting of a strongly invertible knot (K, τ) and an oriented half-axis denoted by h. Given a directed strongly invertible knot (K, τ, h), its *antipode* is the knot $a(K, \tau, h) = (K, \tau, h')$, where h' is the other choice of half-axis; an example is shown in Figure 3.

Given a strongly invertible knot (K, τ), Boyle and Issa define in [3] a notion of equivariant 4-genus $\tilde{g}_4(K, \tau)$, which we recall here.

Definition 2.3 Given a strongly invertible knot (K, τ) in S^3, an *equivariant surface* for (K, τ) is a connected, smoothly properly embedded surface $F \subset B^4$ with $\partial F = K \subset \partial B^4$ such that $\bar{\tau}(F) = F$, for a smooth extension $\bar{\tau}\colon B^4 \to B^4$ of τ.

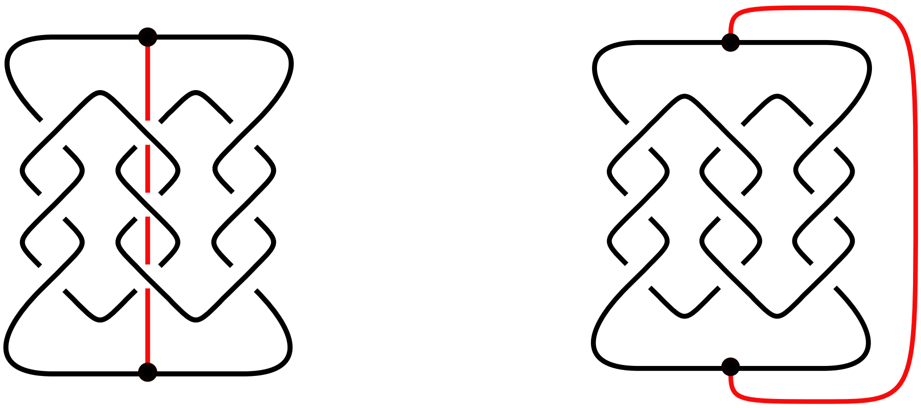


Figure 3: A strongly invertible knot $(9_{46}, \tau, h)$ and its antipode.

From here we define the *equivariant* 4-*genus* as in [3]:

Definition 2.4 The *equivariant* 4-*genus* of a strongly invertible knot (K, τ) is the minimal genus of an orientable equivariant surface for (K, τ), denoted by $\tilde{g}_4(K, \tau)$. When clear from context, we may instead write $\tilde{g}_4(K)$.

Given a strongly invertible knot (K, τ), we have that $(K \# -K, \tau)$ is equivariantly slice, where the connected sum is an equivariant connect sum banding together two directed strongly invertible knots with a symmetric band consistent with the orientations on the half-axis. For a more explicit description of the equivariant connect sum, see [19]. Another common operation we will use to construct strongly invertible knots is the *equivariant double* of a knot K, denoted by $D(K)$.

The double of a knot K is a strongly invertible knot $(D(K), \tau)$ where $D(K) = K \# rK$ and τ is the involution taking K to rK and vice versa, as shown in Figure 4. Of importance to our genus bounds is the following immediate result:

Proposition 2.5 *Given a knot K with* 4*-genus $g_4(K) = n$, its double $(D(K), \tau)$ has equivariant* 4*-genus $\tilde{g}_4(D(K)) \leq 2n$.*

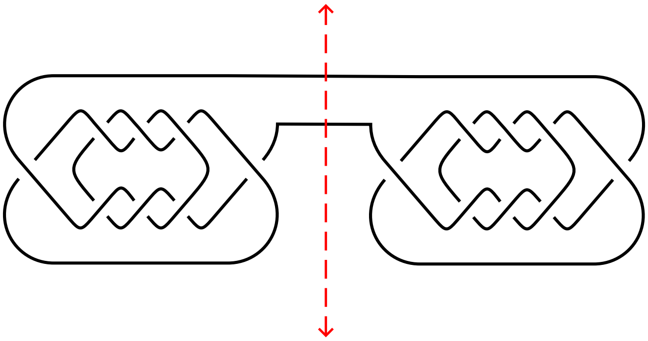


Figure 4: The equivariant double of 8_{20}.

3 Equivariant double-slice genus

Having defined both double-slice genus and equivariant 4-genus, we can now define the logical combination of the two, which we call the *equivariant double-slice genus*. In order to do this, we first discuss extensions of τ to S^4.

We divide the possible extension of τ into two categories: orientation preserving and orientation reversing. By Smith [23], the fixed point set of these extensions will either be S^2 or S^1, depending on if it is orientation preserving or reversing, respectively. We will refer to an orientation preserving extension of τ to S^4 as $\bar{\tau}$, which will be the primary focus of this paper. We discuss orientation reversing extensions briefly at the end of this section.

With this in mind, we now define an *equivariant slicing surface* and an *equivariant slicing handlebody*:

Definition 3.1 Given a directed strongly invertible knot $(K,\tau,h)\subset S^3$ and a smooth extension $\bar{\tau}$ of τ to S^4, an *equivariant slicing handlebody* of (K,τ,h) is a handlebody $H\subset S^4$ such that $\bar{\tau}(H)=H$ and H intersects the standard S^3 containing K transversely with $h\subset H$ and $\partial H\cap S^3=K$. We call ∂H an *equivariant slicing surface* and say that (K,τ,h) *divides* ∂H.

Remark 3.2 Since $H\cap S^3$ contains an arc h of the fixed point set, $\bar{\tau}$ preserves the tangent vector to h at a point of h. Since it setwise fixes the hemispheres of S^4, it also fixes the normal vector to S^3 at that point. Furthermore, the fixed point set is two-dimensional, so we conclude that $\bar{\tau}|_H$ is orientation reversing.

Remark 3.3 Given the equivariant connect sum of two directed strongly invertible knots, as defined in the previous section, we get a natural equivariant connect sum of equivariant slicing surfaces which we will make use of in future constructions.

To see that an equivariant slicing surface exists, we take an equivariant Seifert surface $F\subset S^3$ for (K,τ,h), guaranteed to exist by [3], and consider $F\times[-1,1]\subset S^4$, as in the nonequivariant setting described in [14]. $F\times[-1,1]$ is then an equivariant slicing handlebody for (K,τ,h). We say a surface is *equivariantly unknotted* if it bounds an equivariant handlebody, but do not require the possibly stronger condition that it is equivariant isotopic to some standard embedding. With this in mind, we define the *equivariant double-slice genus*:

Definition 3.4 The *equivariant double-slice genus* of a directed strongly invertible knot (K,τ,h), which we denote $\tilde{g}_{\mathrm{ds}}(K,\tau,h)$, is the minimal genus of an equivariant slicing handlebody for (K,τ,h). If $\tilde{g}_{\mathrm{ds}}(K,\tau,h)=0$, we say (K,τ,h) is *equivariantly double-slice*.

It is clear that the equivariant 4-genus of a directed strongly invertible knot is equal to that of its antipode, as any symmetric surface one bounds is also bounded by the other. It is less clear, and currently unknown to the author, if the same is true for the equivariant double-slice genus.

Open Question A *Given a directed strongly invertible knot and its antipode, does*

$$\tilde{g}_{\mathrm{ds}}(K,\tau,h)=\tilde{g}_{\mathrm{ds}}(a(K,\tau,h))?$$

Because this is unknown, we define the *equivariant double-slice genus* of a nondirected strongly invertible knot (K,τ) to be the minimum of $\tilde{g}_{\mathrm{ds}}(K,\tau,h)$ and $\tilde{g}_{\mathrm{ds}}(a(K,\tau,h))$.

The basic lower bounds $\tilde{g}_{\mathrm{ds}}(K,\tau)\geq g_{\mathrm{ds}}(K)$ and $\tilde{g}_{\mathrm{ds}}(K,\tau)\geq 2\tilde{g}_4(K,\tau)$ for the equivariant double-slice genus do not allow us to differentiate the equivariant double-slice genus from the double-slice genus and the equivariant 4-genus at the same time, as can be done with Theorem 1.1.

3.1 Involutions of handlebodies

Here we provide a brief discussion of the involutions of handlebodies and, more specifically, equivariant slicing handlebodies. We start by recalling a general result about involutions on handlebodies coming from the work of Kalliongis and McCullough in [13]. In their work, they discuss multiple decompositions of actions on handlebodies into simpler actions on smaller parts. We will make use of the first decomposition they discuss, called the vertical-horizontal decomposition. They define an involution on a bundle $\Sigma\times I$ to be *vertical* if it is of the form $1_\Sigma\times r$ and *horizontal* if it is of the form $\sigma\times I$ for some involution σ of Σ; see Figure 5.

Theorem 3.5 [13, Theorem 5.1] *Let τ be an orientation reversing involution of a handlebody H, and suppose that some component of* $\operatorname{fix}(h)$ *is* 2*-dimensional. We have a decomposition $H=H_0\cup\bigl(\bigcup_{j=1}^{r}H_j\bigr)$, where each piece is h-invariant, such that:*

1. *H_0 is an I-bundle over a connected* 2*-manifold, and the restriction of h to H_0 is horizontal. This action may be chosen to be a product action if and only if no component of* $\operatorname{fix}(h)$ *is a point or a Möbius band.*
2. *Each H_j is an I-bundle over a surface of negative Euler characteristic, which is a deformation retract of a component of* $\operatorname{fix}(h)$*, and the restriction of h to H_j is vertical.*
3. *$\{H_1,\dots,H_r\}$ are pairwise disjoint, and for $1\leq i\leq r$ each $H_0\cap H_i$ is a single* 2*-disk.*

In the case of an equivariant slicing handlebody, we get added restrictions on the vertical-horizontal decomposition coming from our knowledge of $\bar{\tau}$.

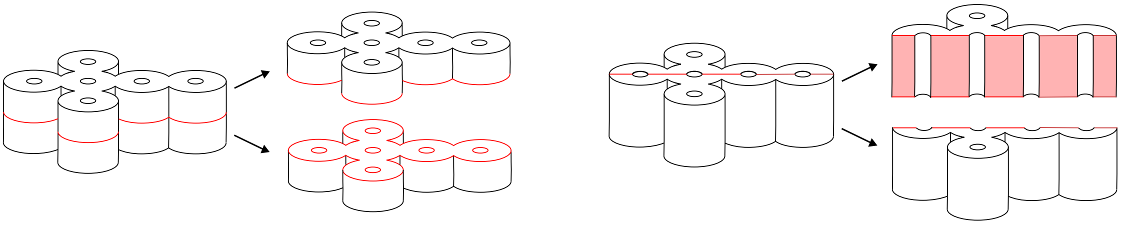


Figure 5: Example of vertical H_i components (left) and horizontal H_0 components (right) with fixed point set in red.

Lemma 3.6 *The fixed point set of an equivariant slicing handlebody for a directed strongly invertible knot* (K,τ,h) *is the disjoint union of some number of planar surfaces.*

Proof Let H be an equivariant slicing handlebody for (K,τ,h). By Remark 3.2, τ must be orientation reversing on H with 2-dimensional fixed point set. So, by [13], the fixed point set of $\tau|_H$ is a collection of surfaces. However, the fixed point set of $\tau|_H$ is a subset of the total fixed point set, a sphere. This means the surfaces comprising the fixed point set are limited to planar surfaces. □

With this restriction on the fixed point set, we can now rewrite Theorem 3.5 for equivariant slicing handlebodies to give us better restrictions on the H_i.

Proposition 3.7 *Let* H *be an equivariant slicing handlebody for a directed strongly invertible knot* (K,τ,h). *Then* H *has a decomposition* $H = H_0 \cup \bigl(\bigcup_{j=1}^{r} H_j\bigr)$, *where each piece is* τ*-invariant, such that:*

(1) H_0 *is an* I*-bundle over a connected planar surface, and the restriction of* τ *to* H_0 *is horizontal and a product action.*

(2) *Each* H_j *is an* I*-bundle over a planar surface which is not a disk or annulus, which is a deformation retract of a component of* $\mathrm{fix}(\tau)$, *and the restriction of* τ *to* H_j *is vertical.*

(3) $\{H_1,\dots,H_r\}$ *are pairwise disjoint, and for* $1 \le i \le r$ *each* $H_0 \cap H_i$ *is a single* 2*-disk.*

Proof We begin by verifying that the assumptions of Theorem 3.5 hold for an arbitrary equivariant slicing handlebody H, ie that $\tau|_H$ is orientation reversing and that $\mathrm{fix}(\tau|_H)$ contains a 2-dimensional component. The fact that $\tau|_H$ is orientation reversing was discussed in Remark 3.2. The fact that $\mathrm{fix}(\tau|_H)$ contains a 2-dimensional component comes from the fact that it contains a half-axis of $\mathrm{fix}(\tau) \subset S^3$. Since this subset is 1-dimensional, and since $\tau|_H$ is orientation reversing, it must then be a subset of a 2-dimensional subset of $\mathrm{fix}(\tau|_H)$. Thus, we know that H has a vertical horizontal decomposition as in Theorem 3.5.

We now look at the changes to (1), namely that H_0 is an I-bundle over a planar surface and that the restriction of h to H_0 is horizontal and a product action. The fact that H_0 is an I-bundle over a planar surface, as opposed to an arbitrary 2-manifold as in Theorem 3.5, is a direct result of Lemma 3.6. Similarly, by Lemma 3.6 no component of $\mathrm{fix}(\tau)$ is a point or Möbius band. So by (1) of Theorem 3.5, we have that the restriction of τ to H_0 is horizontal and is a product action.

For (2) the only change made from Theorem 3.5 is noting that the only surfaces that can appear are, by Lemma 3.6, planar surfaces. By (2) of Theorem 3.5, each must have a negative Euler characteristic and thus is not a disk or annulus.

Since there were no changes to (3), this completes the proof. □

Considering orientation preserving extensions instead of orientation reversing extensions changes the problem dramatically. The action then switches the hemispheres of S^4 and would be orientation preserving on the equivariant slicing handlebody for (K,τ). Our main results, which we now present, do not apply in this setting, as the proofs rely heavily on the behavior of the action on the handlebodies we discussed.

4 Bounds on equivariant double-slice genus

We prove Theorems 1.1 and 1.2 and discuss relevant examples. We first prove a version of Theorem 1.1 for directed strongly invertible knots, from which Theorem 1.1 follows immediately.

Theorem 4.1 *Let (K, τ, h) be a directed strongly invertible knot and K_0 be the union of h with an arc of K ending on the two fixed points. Then $g_{\mathrm{ds}}(K_0) \le \tilde{g}_{\mathrm{ds}}(K, \tau, h)$.*

Proof Let H be a minimal equivariant slicing handlebody for (K, τ, h). We will start by creating a decomposition of H which allows us to construct a useful slicing handlebody for K_0.

We start by decomposing H into $H = H_0 \cup \bigl(\bigcup_{j=1}^{r} H_j\bigr)$ as described in Proposition 3.7. Using this decomposition, we will show that the fixed point set $\bar{\tau}|_H$, which by abuse of notation we will refer to as simply τ, separates H into two identical components, H^1 and $H^2 = \tau(H^1)$. Smith proved that the fixed point set of an involution on a sphere is a $\mathbb{Z}_2$-Čech homology sphere [21; 22; 24; 25]. This means an involution on a circle has zero or two fixed points. It follows by extending the involution on a planar surface Σ to an involution on S^2 that a 1-dimensional fixed point set of an involution on Σ is either S^1 or a collection of properly embedded arcs, either of which separates the surface into identical components.

Thus, since τ acts trivially on fibers of H_0, the fixed point set of τ separates H_0 into two identical components, H_0^1 and H_0^2, satisfying $g(H_0) \ge g(H_0^i)$ as shown in Figure 5. For the H_j, the fixed point set is a planar surface separating the H_j into identical components, H_j^1 and H_j^2, with $H_j^1 = \tau(H_j^2)$ and $g(H_j) = g(H_j^i)$ for $i = 1, 2$ as shown in Figure 5. Note that all the attaching regions in both H_0 and the H_j are disks containing an arc of the fixed point set which are identified. Since both fixed point sets separate, the fixed point set of $H_0 \cup H_j$ separates. Iterating, we get that the fixed point set of H separates it into two handlebodies, H^1 and H^2, with decompositions given by connect sums of the H_j^1 and H_j^2, respectively. Thus, we have

$$g(H) = g(H_0) + \sum_{j=1}^{r} g(H_j) \ge g(H_0^i) + \sum_{j=1}^{r} g(H_j^i) = g(H^i).$$

By construction, the H^i are splitting handlebodies for the two "halves" of K corresponding to the two different arcs. □

This allows us to partially reduce questions about the equivariant double-slice genus of a knot K to the nonequivariant double-slice genus of a different knot K_0. With this, we are ready to prove Theorem 1.2.

Proof of Theorem 1.2 We will show the knot K_n pictured in Figure 1 satisfies the desired properties. To see that K_n is double-slice, we note that $K_n = (8_{20}\#{-8_{20}})^{2n}$. Since $8_{20}\#{-8_{20}}$ is double-slice, so is the $2n$ connect sum of it with itself. Now we show that it is equivariantly slice. First, note that 8_{20} is the pretzel knot $(3, -3, 2)$ and so, by Sakuma [19], is equivariantly slice. Additionally note that since 8_{20} is slice, $D(8_{20})$ is equivariantly slice. Since K_n can be seen as an equivariant connect sum of $2n$ copies of 8_{20} and n copies of $D(8_{20})$, which we just said are equivariantly slice, we know that K_n is also equivariantly slice.

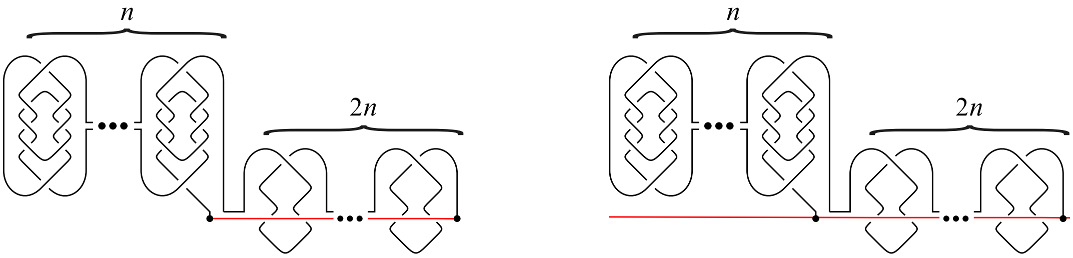


Figure 6: The decomposition of (K_n, τ) with both half-axes.

Lastly, we show that $\tilde{g}_{\mathrm{ds}}(K_n, \tau, h_n) \geq n$. To do this, we first see in Figure 6 that, regardless of which half-axis we pick, the knot we get from taking an arc of K_n union the half-axis is $\#^n \, 8_{20}$. Thus, by Theorem 1.1, we have that, for either choice of half-axis, $\tilde{g}_{\mathrm{ds}}(K_n, \tau, h_n) \geq g_{\mathrm{ds}}(\#^n \, 8_{20})$. As discussed in Example 2.2, using signature invariants and the work of Orson–Powell [17], we can see $g_{\mathrm{ds}}(\#^n \, 8_{20}) = n$. Thus, we get our last property, that $\tilde{g}_{\mathrm{ds}}(K_n, \tau) \geq n$. □

Example 4.2 For a slightly different construction, we will show that the hyperbolic knot (S, τ, h) depicted in Figure 7 (right) is double-slice, equivariantly slice, but not equivariantly double-slice. We will do this by showing that the knot (K, τ, h) depicted in Figure 7 (left) is double-slice and equivariantly slice, then identifying an invertible concordance from (S, τ, h) to (K, τ, h). Because the constructed concordance will be symmetric, this will show that (S, τ, h) is also double-slice and equivariantly slice. We will then apply Theorem 1.1 to see that it is not equivariantly double-slice.

Since K is the connect sum of the knot 8_{20} with itself, and 8_{20} is fully amphichiral (meaning $8_{20} = -8_{20}$), we have that $K = 8_{20} \# 8_{20}$ and this knot is therefore double-slice. Additionally, 8_{20} is slice, and therefore by Proposition 2.5, $K = D(8_{20})$ is equivariantly slice.

By adding symmetric grabbers to $8_9 \# 8_9$ to construct S as in Figure 7 (right), one gets a symmetric invertible concordance from (S, τ, h) to (K, τ, h). Therefore, (S, τ, h) is double-slice and equivariantly slice.

Lastly, to see that (S, τ, h) is not equivariantly double-slice, we apply Theorem 4.1. The knot we get by adding a half-axis and deleting an arc is 8_{20}, which has $g_{\mathrm{ds}}(8_{20}) = 1$. Therefore, $\tilde{g}_{\mathrm{ds}}(S, \tau, h) \geq 1$, meaning (S, τ, h) is not equivariantly double-slice.

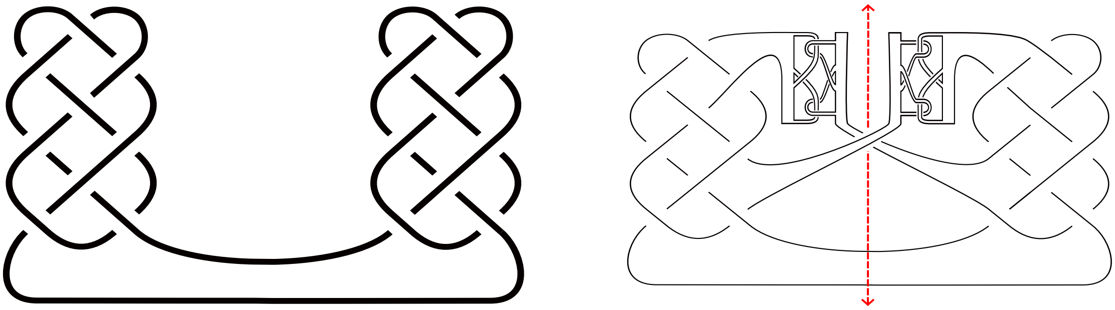


Figure 7: Left: the knot $8_9 \# 8_9$. Right: a knot S invertibly concordant to $8_9 \# 8_9$.

5 Equivariant superslice genus and equivariantly knotted 2-spheres

We define a notion of equivariant superslice genus and construct bounds similar to those for equivariant double-slice genus in Theorem 1.1. We will use this result in Section 5.2 to provide examples of unknotted symmetric 2-spheres which are symmetrically knotted.

5.1 Equivariant superslice genus

Similar to the double-slice genus, there is a concept of *superslice genus* first defined by Wenzhao Chen in [5]. The *superslice genus* of a knot K, denoted by $g_{\mathrm{ss}}(K)$, is the minimal genus of a slicing surface Σ for K such that Σ is symmetric about K. That is to say, $\Sigma \subset S^4$ is the double of a surface $F \subset B^4$ along its boundary $K \subset S^3$. We call Σ a *superslicing surface* for K, and the handlebody it bounds a *superslicing handlebody*. In the initial conception of superslice genus in [5], $g_{\mathrm{ss}}(K)$ is the genus of F as opposed to Σ, meaning it is exactly half of our conception. This change is made to keep the numbers consistent with those of the double-slice genus. If $g_{\mathrm{ss}}(K) = 0$, we say K is *superslice*.

While the double-slice genus is a clear lower bound for the superslice genus of a knot, in our work we will make use of a stronger lower bound proved in [5]:

Theorem 5.1 (Chen) *Given a knot K, let Σ be the two-fold branched cover of S^3 along K. Let n be the minimum number of generators for $H_1(\Sigma; \mathbb{Z})$. Then $n \leq g_{\mathrm{ss}}(K)$.*

In the same way we were able to define equivariant slicing handlebodies and surfaces, we now define *equivariant superslicing handlebodies* and *equivariant superslicing surfaces*.

Definition 5.2 Given a strongly invertible knot (K, τ, h), an *equivariant superslicing handlebody H* of (K, τ, h) is a superslicing handlebody which is also an equivariant slicing handlebody. We call ∂H an *equivariant superslicing surface* for (K, τ, h).

The construction used in Section 3 to guarantee the existence of an equivariant slicing handlebody also guarantees the existence of an equivariant superslicing handlebody. Thus, we can define the *equivariant superslice genus* of a directed strongly invertible knot (K, τ, h), denoted by $\tilde{g}_{\mathrm{ss}}(K, \tau, h)$.

Definition 5.3 The *equivariant superslice genus* of a strongly invertible knot (K, τ, h) is the minimal genus of an equivariant superslicing surface for (K, τ, h). If $\tilde{g}_{\mathrm{ss}}(K, \tau, h) = 0$, we say (K, τ) is *equivariantly superslice*.

As in the equivariant double-slice case, when considering a nondirected strongly invertible knot (K, τ), we define the *equivariant superslice genus* of (K, τ) to be the minimum of the equivariant superslice genus of the two directed strongly invertible knots obtainable from (K, τ).

Our primary bound for $\tilde{g}_{\mathrm{ss}}(K, \tau, h)$ follows exactly as in the equivariantly double-slice case:

Theorem 5.4 *Let* (K, τ, h) *be a directed strongly invertible knot and* K_0 *be the union of an arc of* K *with* h. *Then* $g_{\mathrm{ss}}(K_0) \leq \tilde{g}_{\mathrm{ss}}(K, \tau, h)$.

The proof of this is exactly the proof of Theorem 1.1, taking the equivariant slicing handlebody to be an equivariant superslicing handlebody.

Example 5.5 Using Theorem 5.4, one can check that the superslice knot J in Figure 8 (left) is not equivariantly superslice with the blue half-axis. When considered with the red axis, it is possible that $a(J, \tau, h)$ is equivariantly superslice, making it a possible example answering Open Question A in the negative.

5.2 Symmetrically knotted 2-spheres

For strongly invertible knots, Marumoto [15] proves that any two strongly invertible knots which are isotopic to the unknot are equivariantly isotopic to each other. We will show that this does not extend to equivariantly unknotted 2-spheres in (S^4, τ). We create symmetric 2-spheres that are unknotted in the classical sense (they bound a 3-ball) but do not bound any symmetric 3-ball, contrasting with the properties of strongly invertible knots in S^3.

Proof of Theorem 1.4 Consider the knot $(J_1, \tau) = (J, \tau, h) \# a(J, \tau, h)$ depicted in Figure 8 (right). We will construct a slicing 2-sphere S^2 for K such that $\tau(S^2) = S^2$ is not an equivariant slicing sphere, ie S^2 will be a symmetric 2-sphere bounding a 3-ball which bounds no invariant 3-ball.

First, we show that (J_1, τ) is not equivariantly double-slice (with either half-axis), which means any equivariant 2-sphere which it appears as a cross-section of cannot bound an equivariant 3-ball. This follows immediately from Theorem 4.1 as the decomposition gives a connect sum of 6_1 with itself, which is not double-slice.

Thus, to complete the proof, all we need is a symmetric 2-sphere for (J_1, τ) which we know is unknotted. To construct this, we first construct such a sphere for (J, τ). This sphere is the ribbon 2-sphere constructed by taking the double of the obvious ribbon disk for J. Since the ribbon disk is symmetric, so is the ribbon

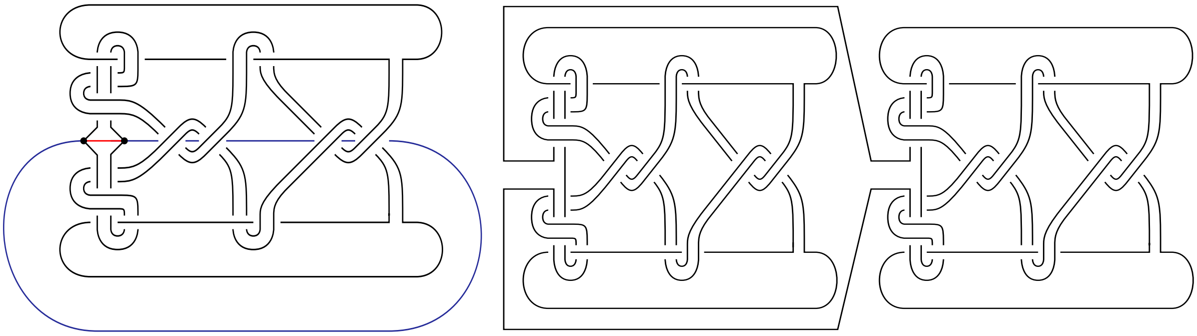


Figure 8: Left: superslice knot (J, τ, h) (blue) and $a(J, \tau, h)$ (red). Right: the knot $(J_1, \tau, h) \# a(J_1, \tau, h)$.

2-sphere. The fact that this 2-sphere is unknotted comes from elementary moves for ribbon surfaces: a homotopy of the core of the band in Figure 8 (left) to a trivial band between the two circles corresponds to an isotopy of the corresponding ribbon 2-sphere. Thus, (J, τ) appears as the cross-section of a symmetric unknotted 2-sphere which we call S.

Note that $a(J_1, \tau, h)$ also appears as a cross-section of the same 2-sphere S, as this 2-sphere does not depend on the choice of half-axis. Therefore, taking the equivariant connect sum of these two spheres, we get another symmetric sphere S^2 whose intersection with S^3 is $(J_1, \tau, h) \# a(J_1, \tau, h)$. Since this S^2 is the connect sum of two unknotted spheres, it itself is unknotted. Thus, this S^2 is a symmetric sphere bounding a 3-ball which does not bound any invariant 3-ball. □

6 Internal stabilization rel boundary

In Section 6.1, we describe how to use superslice genus to obstruct isotopy of surfaces rel boundary and bound their 1-handle stabilization distance. In Section 6.2, we define a notion of equivariant stabilization distance and extend the results of Section 6.1 to this new equivariant setting.

6.1 Internal stabilization bounds from double-slice genus

First, we must define some terms. Let $\Sigma_1, \Sigma_2 \subset B^4$ be two properly embedded surfaces with common boundary K. The 1*-handle stabilization distance* $d_1(\Sigma_1, \Sigma_2)$ from Σ_1 to Σ_2 is the minimum number of orientation preserving ambient 1-handles $\{h_i\}$ and $\{h'_i\}$ needed so that $\Sigma_1 \cup \{h_i\}$ is smoothly isotopic rel boundary to $\Sigma_2 \cup \{h'_i\}$, as defined in [16]. With this we are ready to prove Theorem 1.5.

Proof of Theorem 1.5 Let $d = d_1(\Sigma_1, \Sigma_2)$ and let $\Sigma'_1 = \Sigma_1 \cup \{h_i\}_{i=1}^d$ and $\Sigma'_2 = \Sigma_2 \cup \{h'_i\}_{i=1}^d$ be stabilized surfaces that are isotopic rel boundary. Since $\Sigma_1 \cup_K -\Sigma_2 \subset (B^4, \Sigma_1) \cup_{(S^3,K)} (B^4, \Sigma_2)$ is unknotted, the stabilized surface $\Sigma' = \Sigma'_1 \cup_K -\Sigma'_2 \subset (B^4, \Sigma'_1) \cup_{(S^3,K)} (B^4, \Sigma'_2)$ is also unknotted, as it is an unknotted handlebody $\Sigma_1 \cup_K -\Sigma_2$ union handles. Since Σ'_1 is isotopic rel boundary to Σ'_2, we can isotope Σ' in S^4 rel B^4 (the hemisphere containing Σ'_1) to get the double of (B^4, Σ'_1). This means that the double of (B^4, Σ'_1) is a superslicing surface for K and therefore $h + d \geq \frac{1}{2} g_{\mathrm{ss}}(K)$, ie $d \geq \frac{1}{2} g_{\mathrm{ss}}(K) - h$. We have $\frac{1}{2} g_{\mathrm{ss}}(K)$, not $g_{\mathrm{ss}}(K)$, because we want the genus of the surface bounded by the knot, not its double. Thus, $d_1(\Sigma_1, \Sigma_2) \geq \frac{1}{2} g_{\mathrm{ss}}(K) - h$. □

Letting K be double-slice, we are able to use this to obstruct certain slice disks of K from being isotopic rel boundary, from which Corollary 1.6 immediately follows.

In [14] a survey of double-slice knots with 12 or fewer crossings is conducted, finding 20 knots which are double-slice and providing an explicit description of the double-slicing via bands. Of these 20 knots, 17 have Alexander polynomial not equal to 1 and are therefore not superslice. Thus, for these 17 knots, the band diagrams given in [14] depict slice disks not isotopic rel boundary. This fact has been proven for the first of these 17 knots, 9_{46}, by others including Miller and Powell [16] and Sundberg and Swann [27].

Of the 17 knots with Alexander polynomial not equal to 1, the following 15 knots have $H_1(\Sigma, \mathbb{Z}) = \mathbb{Z}_n^2$ for some n, where Σ is the 2-fold branch cover of S^3 along one of the knots K:

$$9_{46}, 10_{99}, 10_{123}, 10_{155}, 11n_{74}, 12a_{427}, 12a_{1105}, 12n_{268}, \\ 12n_{397}, 12n_{414}, 12n_{605}, 12n_{636}, 12n_{706}, 12n_{817}, 12n_{838}.$$

Thus, if you take the connect sum $\#^m K$, where K is any of the 15 knots, you get that $H_1(\Sigma, \mathbb{Z}) = \mathbb{Z}_n^{2m}$. By Theorem 5.1, this means that $g_{\mathrm{ss}}(\#^m K) \geq 2m$. Combining this fact with Theorem 1.5, we get that any two slice disks arising from the same double-slicing of $\#^m K$ have stabilization distance at least m.

6.2 Equivariant stabilization

We now define an equivariant notion of 1-handle stabilization distance and reprove Theorem 1.5 in this equivariant setting.

Definition 6.1 Let ρ be a smooth $\mathbb{Z}_p$ action on B^4, $K \subset S^3$ be a ρ-invariant knot, and $\Sigma_1, \Sigma_2 \subset B^4$ be τ-invariant properly embedded surfaces with common boundary K. The *equivariant* 1*-handle stabilization distance*, denoted by $\tilde{d}_\rho(\Sigma_1, \Sigma_2)$, is the minimal number of orientation preserving ambient 1-handles $\{h_i\}$ and $\{h_i'\}$ needed so that $\rho(\Sigma_i \cup \{h_i\}) = \Sigma_i \cup \{h_i\}$ for $i \in \{1, 2\}$ and $\Sigma_1 \cup \{h_i\}$ is smoothly equivariantly isotopic rel boundary to $\Sigma_2 \cup \{h_i'\}$.

Restricting our attention to the action $\bar{\tau}$ discussed in Section 2.3 we can ask about equivariant stabilization distance bounds coming from equivariant double-slice and equivariant superslice genus. More specifically, we get the statement of Theorem 1.7, the proof of which follows exactly as in the nonequivariant setting. One interesting question about equivariant stabilization that Theorem 1.5 may be useful for is the following:

Open Question B *For every $n \in \mathbb{N}$, does there exist $\bar{\tau}$-invariant surfaces $\Sigma_1, \Sigma_2 \subset B^4$ which are isotopic but have $\tilde{d}_1^\tau(\Sigma_1, \Sigma_2) = n$?*

References

[1] **D Auckly**, **H J Kim**, **P Melvin**, **D Ruberman**, **H Schwartz**, *Isotopy of surfaces in* 4*-manifolds after a single stabilization*, Adv. Math. 341 (2019) 609–615 MR

[2] **R İ Baykur**, **N Sunukjian**, *Knotted surfaces in* 4*-manifolds and stabilizations*, J. Topol. 9 (2016) 215–231 MR

[3] **K Boyle**, **A Issa**, *Equivariant* 4*-genera of strongly invertible and periodic knots*, J. Topol. 15 (2022) 1635–1674 MR

[4] **J C Cha**, **K H Ko**, *On equivariant slice knots*, Proc. Amer. Math. Soc. 127 (1999) 2175–2182 MR

[5] **W Chen**, *A lower bound for the double slice genus*, Trans. Amer. Math. Soc. 374 (2021) 2541–2558 MR

[6] **I Dai**, **A Mallick**, **M Stoffregen**, *Equivariant knots and knot Floer homology*, J. Topol. 16 (2023) 1167–1236 MR

[7] **A Di Prisa**, *Equivariant algebraic concordance of strongly invertible knots*, J. Topol. 17 (2024) art. id. e70006 MR

[8] **A Di Prisa**, **G Framba**, *A new invariant of equivariant concordance and results on* 2*-bridge knots*, Algebr. Geom. Topol. 25 (2025) 1117–1132 MR

[9] **G Guth**, *For exotic surfaces with boundary, one stabilization is not enough*, preprint (2022) arXiv 2207.11847

[10] **K Hayden**, *An atomic approach to Wall-type stabilization problems*, preprint (2023) arXiv 2302.10127

[11] **K Hayden**, **S Kang**, **A Mukherjee**, *One stabilization is not enough for closed knotted surfaces*, preprint (2023) arXiv 2304.01504

[12] **A Juhász**, **I Zemke**, *Stabilization distance bounds from link Floer homology*, J. Topol. 17 (2024) art. id. e12338 MR

[13] **J Kalliongis**, **D McCullough**, *Orientation-reversing involutions on handlebodies*, Trans. Amer. Math. Soc. 348 (1996) 1739–1755 MR

[14] **C Livingston**, **J Meier**, *Doubly slice knots with low crossing number*, New York J. Math. 21 (2015) 1007–1026 MR

[15] **Y Marumoto**, *Relations between some conjectures in knot theory*, Math. Sem. Notes Kobe Univ. 5 (1977) 377–388 MR

[16] **A N Miller**, **M Powell**, *Stabilization distance between surfaces*, Enseign. Math. 65 (2019) 397–440 MR

[17] **P Orson**, **M Powell**, *A lower bound for the doubly slice genus from signatures*, New York J. Math. 27 (2021) 379–392 MR

[18] **D Rolfsen**, *Knots and links*, Mathematics Lecture Series 7, Publish or Perish, Houston, TX (1990) MR

[19] **M Sakuma**, *On strongly invertible knots*, from "Algebraic and topological theories" (M Nagata, S Araki, A Hattori, N Iwahori, editors), Kinokuniya, Tokyo (1986) 176–196 MR

[20] **O Singh**, *Distances between surfaces in* 4*-manifolds*, J. Topol. 13 (2020) 1034–1057 MR

[21] **P A Smith**, *Transformations of finite period*, Ann. of Math. 39 (1938) 127–164 MR

[22] **P A Smith**, *Transformations of finite period, II*, Ann. of Math. 40 (1939) 690–711 MR

[23] **P A Smith**, *Fixed-point theorems for periodic transformations*, Amer. J. Math. 63 (1941) 1–8 MR

[24] **P A Smith**, *Transformations of finite period, III: Newman's theorem*, Ann. of Math. 42 (1941) 446–458 MR

[25] **P A Smith**, *Transformations of finite period, IV: Dimensional parity*, Ann. of Math. 46 (1945) 357–364 MR

[26] **D W Sumners**, *Invertible knot cobordisms*, Comment. Math. Helv. 46 (1971) 240–256 MR

[27] **I Sundberg**, **J Swann**, *Relative Khovanov–Jacobsson classes*, Algebr. Geom. Topol. 22 (2022) 3983–4008 MR

Kansas State University
Manhattan, KS, United States

`malcolmga@ksu.edu`

Received: 27 May 2024 Revised: 10 October 2024

Algebraic & Geometric Topology **25**:8 (2025) 5153–5174
DOI: 10.2140/agt.2025.25.5153
Published: 20 November 2025

Coarse and bi-Lipschitz embeddability of subspaces of the Gromov–Hausdorff space into Hilbert spaces

NICOLÒ ZAVA

We discuss the embeddability of subspaces of the Gromov–Hausdorff space, which consists of isometry classes of compact metric spaces endowed with the Gromov–Hausdorff distance, into Hilbert spaces. These embeddings are particularly valuable for applications to topological data analysis. We prove that its subspace consisting of metric spaces with at most n points has asymptotic dimension $n(n-1)/2$. Thus, there exists a coarse embedding of that space into a Hilbert space. On the contrary, if the number of points is not bounded, then the subspace cannot be coarsely embedded into any uniformly convex Banach space and so, in particular, into any Hilbert space. Furthermore, we prove that, even if we restrict to finite metric spaces whose diameter is bounded by some constant, the subspace still cannot be bi-Lipschitz embedded into any finite-dimensional Hilbert space. We obtain both nonembeddability results by finding obstructions to coarse and bi-Lipschitz embeddings in families of isometry classes of finite subsets of the real line endowed with the Euclidean–Hausdorff distance.

46B85, 51F30; 54B20

1 Introduction

The Gromov–Hausdorff distance d_{GH} measures how two metric spaces resemble each other. It was introduced by Edwards in [27], and then rediscovered and generalised by Gromov [31]. Until around 2000, the Gromov–Hausdorff distance had been mainly used by pure mathematicians who were interested in the induced topology. That direction is still of great interest, and, as an example, we mention the two recent papers [3; 4].

In addition to the intrinsic interest in it, a great impulse to study the quantitative aspects of the Gromov–Hausdorff distance came from its applications in topological data analysis, which is a fast-growing subject aiming to use topological techniques to analyse a wide range of real-world data (see, for example, [13; 33], and [30] for a growing dataset of real-world applications). The Gromov–Hausdorff distance provides a theoretical framework to directly compare point clouds by considering them as metric spaces. This approach proved to be useful in shape recognition and comparison [43; 46; 47], which arises, for example, in molecular biology, databases of objects, face recognition and matching of articulated objects.

Comparing two metric spaces using the Gromov–Hausdorff distance directly is computationally expensive. Even approximating it within a factor of 3 for trees with unit edge length is NP-hard ([2; 59]; see also [43], where the author discussed the connection between computing the Gromov–Hausdorff distance and a class of NP-hard problems). Therefore, creating efficiently computable invariants to approximate the Gromov–Hausdorff distance is of particular interest. Following [45], an *invariant* ψ associates to a metric space X an element $\psi(X)$ of another metric space $(\mathfrak{Y}, d_{\mathfrak{Y}})$ in such a way that, if X and Y are two isometric metric spaces, then $\psi(X) = \psi(Y)$. Furthermore, an invariant ψ is *stable* if there exists a function $\rho_+ \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that

$$d_{\mathfrak{Y}}(\psi(X), \psi(Y)) \leq \rho_+(d_{\mathrm{GH}}(X, Y)) \tag{1}$$

for every pair of metric spaces X and Y. Stability implies that small perturbations of the metric spaces have a limited effect on the associated invariants. Therefore, considering similarity recognition, we avoid false negatives, which are situations where two metric spaces are very similar in the Gromov–Hausdorff distance, but their invariants are far apart. Furthermore, stable invariants can be used to provide lower bounds to the Gromov–Hausdorff distance as shown in [45]. We refer to the latter paper for a wide range of stable invariants. Additional examples are hierarchical clustering [14; 15] and persistence diagrams induced by the Vietoris–Rips, the Dowker, and the Čech filtrations ([17; 18]; see also [25] for details and applications of persistent homology).

In contrast to false negatives, even though still undesirable, it is often acceptable when an invariant produces false positives, where two dissimilar spaces are mapped to close values. In this paper, we study when stable invariants are actually bound to lose information because of the unavoidable creation of false positives. We focus our study on those invariants taking values in a Hilbert space. Those are particularly relevant for the applications in machine learning pipelines since many algorithms expect either data in the form of Euclidean vectors or at least access to a so-called feature map into a Hilbert space. As formally stated in the sequel, we prove that the existence of a stable invariant avoiding false positives strongly depends on a bound on the cardinality of the metric spaces.

Our approach to the problem requires notions and techniques from coarse geometry. Intuitively, this field, also known as large-scale geometry, focuses on large-scale, global properties of spaces ignoring local features. We refer to [52; 58] for a wide introduction. A map $\psi \colon (X, d_X) \to (Y, d_Y)$ between two metric spaces is said to be a *coarse embedding* if there exist two maps $\rho_-, \rho_+ \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $\rho_- \to \infty$ and, for every $x, y \in X$,

$$\rho_-(d_X(x, y)) \leq d_Y(\psi(x), \psi(y)) \leq \rho_+(d_X(x, y)). \tag{2}$$

In the case of stable invariants, ie, satisfying (1), a lower bound as in (2) prevents false positives since the larger the Gromov–Hausdorff distance, the larger the distance between the two associated invariants.

Coarse embeddings have been introduced by Gromov and extensively studied in coarse geometry. A crucial application of this theory is due to Yu, who proved in [65] that those metric spaces that can be coarsely embedded into a Hilbert space satisfy the Novikov and the coarse Baum–Connes conjectures generalising results contained in [64]. This result motivated two research directions. On one hand, since an explicit coarse embedding can be hard to construct, a plethora of conditions ensuring its existence have been defined and investigated. We refer the interested reader to [52] for a discussion on the topic and to [63] for more examples. Among these properties, if a space has finite asymptotic dimension, then it can be coarsely embedded into a Hilbert space [34; 65]. Asymptotic dimension is a large-scale counterpart of Lebesgue's covering dimension introduced in [32] (see also [7]). On the other hand, examples of metric spaces that cannot be coarsely embedded were constructed, for example, in [22; 39]. Showing that one of those pathological examples can be coarsely embedded into a metric space X is a technique to prove that X itself cannot be coarsely embedded into any Hilbert space.

Those two strategies have been adopted to prove if metric spaces emerging in different fields can be coarsely embedded into Hilbert spaces. In topological data analysis, collections of persistence diagrams endowed with various metrics represent a prominent example. It was proved in [48] that the space of persistence diagrams of at most n points endowed with the Hausdorff distance has finite asymptotic dimension, and so it can be coarsely embedded into a Hilbert space. This result, despite being nonconstructive, motivated further research in that direction that eventually led to explicit bi-Lipschitz and coarse embeddings in [6] and [49], respectively. In the opposite direction, it was proved in [11; 48; 62] that spaces of persistence diagrams with various metrics cannot be coarsely embedded into any Hilbert space. We also refer to [55], where the authors showed the equivalence of this problem with the embeddability of Wasserstein spaces.

Another example can be found in [29]. Motivated by the interest in invariants in crystallography and pharmaceutics (see [26]), the authors used the previously described strategy to prove that spaces of periodic point sets equipped with the Euclidean bottleneck distance cannot be coarsely embedded into any uniformly convex Banach space. The constructions used in that paper are adapted from those developed in [63] to prove the analogous noncoarse embeddability results for families of finite subsets of metric spaces endowed with the Hausdorff distance.

In this paper, we prove the following results.

Theorem A *The space $\mathcal{GH}^{\leq n}$ of isometry classes of metric spaces with at most n points endowed with the Gromov–Hausdorff distance has asymptotic dimension $n(n-1)/2$, and so it can be coarsely embedded into a Hilbert space.*

Theorem B *The space $\mathcal{GH}^{<\omega}$ of isometry classes of finite metric spaces endowed with the Gromov–Hausdorff distance cannot be coarsely embedded into any uniformly convex Banach space, and so, in particular, into any Hilbert space.*

As an immediate consequence of Theorem B, the same result holds for the *Gromov–Hausdorff space $\mathcal{GH}$*, which is the metric space of isometry classes of compact metric spaces equipped with the Gromov–Hausdorff distance. In the paper, we prove a stronger version of Theorem B stating that already the much smaller subspace of $\mathcal{GH}^{<\omega}$ consisting of all isometry classes of finite subsets of the real line cannot be coarsely embedded into any uniformly convex Banach space (Theorem 4.2). Thanks to a recent result due to Majhi, Vitter and Wenk [42], Theorem 4.2, and therefore also Theorem B, will follow from the fact that an obstruction to coarse embeddability is found in the space of isometry classes of finite subsets of the real line endowed with the Euclidean–Hausdorff distance (a modification of the Gromov–Hausdorff distance for subsets of $\mathbb{R}^d$).

We conclude the paper focussing on the subspace $\mathcal{GH}^{<\omega}_{\leq R}$ of $\mathcal{GH}^{<\omega}$ whose elements have diameter bounded by a constant $R > 0$. Since the diameter of this subspace is bounded, the map collapsing the space into a point is trivially a coarse embedding. An immediate follow-up question is whether it can be bi-Lipschitz embedded, as in the case of persistence diagrams with at most n points. We provide a partial negative answer.

Theorem C *$\mathcal{GH}^{<\omega}_{\leq R}$ cannot be bi-Lipschitz embedded into any finite-dimensional Hilbert space.*

Inspired by [16], where the authors proved that certain spaces of persistence diagrams cannot be bi-Lipschitz embedded into any finite-dimensional Hilbert space, we compute the Assouad dimension ([5]; see also [10] for an earlier definition) of $\mathcal{GH}^{<\omega}_{\leq R}$ and show that it is infinite. This dimension notion was in fact introduced to provide such embeddability obstructions. More precisely, we show that already the subset consisting of all isometry classes of finite subsets of an interval has infinite Assouad dimension and cannot be bi-Lipschitz embedded into any finite-dimensional Hilbert space. Again, using the aforementioned Majhi, Vitter and Wenk's theorem, we deduce our claims from the analogous results for the space of isometry classes of finite subsets of an interval endowed with the Euclidean–Hausdorff distance (Theorem 5.1 and Proposition 5.3).

The paper is organised as follows. In Section 2 we provide the needed background regarding the Gromov– and the Euclidean–Hausdorff distances. In Section 3, the asymptotic dimension is introduced and Theorem A is proved. Theorem B is shown in Section 4, and, finally, we define the Assouad dimension and provide Theorem C in Section 5. We conclude the paper discussing a list of questions in Section 5.1.

Acknowledgements The author was supported by the FWF Grant, Project number I4245-N35. The author would like to thank Thomas Weighill for the helpful discussions around Theorem 3.10, and Takamitsu Yamauchi for bringing to my attention the fundamental reference [35]. Furthermore, the author is thankful for the detailed and helpful comments of the reviewer of this manuscript.

Notation We denote by $\mathbb{N}$, $\mathbb{Q}$ and $\mathbb{R}$ the set of natural numbers including 0, the set of rational numbers, and the set of real numbers, respectively. For $c \in \mathbb{R}$, we also write

$$\mathbb{R}_{\geq c} = \{x \in \mathbb{R} \mid x \geq c\} \quad \text{and} \quad \mathbb{R}_{>c} = \{x \in \mathbb{R} \mid x > c\}.$$

For a set X, we denote by $|X|$ its cardinality. Moreover, for $n \in \mathbb{N}$, we define the following subsets of the power set of X:

$$[X]^{=n} = \{A \subseteq X \mid |A| = n\}, \quad [X]^{\leq n} = \bigcup_{k \leq n} [X]^{=n}, \quad [X]^{<\omega} = \bigcup_{k \in \mathbb{N}} [X]^{=n}.$$

2 The Gromov–Hausdorff distance and the Euclidean–Hausdorff distance

We recall some basic notions, the definitions of the Gromov–Hausdorff and the Euclidean–Hausdorff distances and their relationships. We refer to [12; 44; 45; 54; 61] for comprehensive discussions on the Gromov–Hausdorff distance.

Definition 2.1 A pair (X, d) consisting of a set X and a map $d\colon X \times X \to \mathbb{R}$ is called a *network* [19]. A network (X, d) is a *metric space* (and d is a *metric*) if it satisfies

- (M1) for every $x, y \in X$, $d(x, y) \geq 0$ and $d(x, x) = 0$;
- (M2) for every $x, y \in X$, $d(x, y) = 0$ if and only if $x = y$;
- (M3) for every $x, y \in X$, $d(x, y) = d(y, x)$;
- (M4) for every $x, y, z \in X$, $d(x, y) \leq d(x, z) + d(z, y)$.

Let us recall that an *isometry* between two networks (X, d_X) and (Y, d_Y) is a bijective map $\psi\colon X \to Y$ such that, for every $x, x' \in X$, $d_Y(\psi(x), \psi(x')) = d_X(x, x')$. In that case, X and Y are said to be *isometric*.

For a subset A of a metric space (X, d), its *diameter* is

$$\operatorname{diam} A = \sup_{x,y \in A} d(x, y).$$

A *correspondence* $\mathcal{R}$ between two sets X and Y is a relation $\mathcal{R} \subseteq X \times Y$ such that every $x \in X$ is in relation with at least one element $y \in Y$ and vice versa. Then, for every metric space (Z, d), the *Hausdorff distance* is defined as follows: for every $X, Y \subseteq Z$,

$$d_{\mathrm{H}}(X, Y) = \inf_{\mathcal{R} \subseteq X \times Y \text{ correspondence}} \sup_{(x,y) \in \mathcal{R}} d(x, y).$$

Definition 2.2 Given two metric spaces X and Y, their *Gromov–Hausdorff distance* d_{GH} is

$$d_{\mathrm{GH}}(X, Y) = \inf_{Z \text{ metric space}} \inf\bigl\{d_{\mathrm{H}}(i_X(X), i_Y(Y)) \mid i_X\colon X \to Z \text{ and } i_Y\colon Y \to Z \text{ isometric embeddings}\bigr\}.$$

The reader may notice an abuse of notation in the previous definition since all possible metric spaces form a proper class. However, the infimum value can be achieved by investigating just a set of spaces. Indeed, it is enough to consider the disjoint union $X \sqcup Y$ endowed with pseudometrics (where the distance between distinct points may be zero) whose restrictions to the subsets X and Y coincide with the original metrics. We refer to [12] for the details.

If two metric spaces are isometric, their Gromov–Hausdorff distance is 0. The converse implication does not hold in general. However, if X and Y are compact and $d_{\mathrm{GH}}(X, Y) = 0$, then X and Y are isometric.

Denote by $\mathcal{GH}$ the set of all isometry classes of compact metric spaces endowed with d_{GH}, where the Gromov–Hausdorff distance between two isometry classes is the Gromov–Hausdorff distance between any pair of representatives. Since two compact metric spaces are isometric if and only if their Gromov–Hausdorff distance is 0 (see, for example, [12]), $\mathcal{GH}$ is a metric space, also called the *Gromov–Hausdorff space*. Furthermore, we consider the subspace $\mathcal{GH}^{<\omega}$ of $\mathcal{GH}$ consisting of isometry classes of finite metric space, which is dense in $\mathcal{GH}$. Actually, the subspace $\mathcal{GH}^{<\omega}_{\mathbb{Q}}$ of isometry classes of finite spaces endowed with metrics taking values in $\mathbb{Q}$ is dense in $\mathcal{GH}$ [54]. Therefore, $\mathcal{GH}$ is separable.

The Gromov–Hausdorff distance can be alternatively characterised using correspondences. If (X, d_X) and (Y, d_Y) are two networks, and $\mathcal{R} \subseteq X \times Y$ is a correspondence between them, the *distortion of* $\mathcal{R}$ is

$$\operatorname{dis}\mathcal{R} = \sup_{(x_1,y_1),(x_2,y_2)\in\mathcal{R}} |d_X(x_1,x_2) - d_Y(y_1,y_2)|.$$

Definition 2.3 [19] Let X and Y be two networks. Then their *network distance* is

$$d_{\mathcal{N}}(X,Y) = \frac{1}{2} \inf_{\mathcal{R} \subseteq X \times Y \text{ correspondence}} \operatorname{dis}\mathcal{R}.$$

It is known that, if X and Y are metric spaces, then $d_{\mathrm{GH}}(X,Y) = d_{\mathcal{N}}(X,Y)$ (see, for example, [12]). A further characterisation of the Gromov–Hausdorff distance can be found in [36].

In [66], a characterisation of the network distance for *quasimetric spaces* (ie, networks satisfying (M1), (M2) and (M4)) in the spirit of Definition 2.2 is provided.

The Gromov–Hausdorff distance is difficult to compute even in simple cases. For example, the distance between spheres of different dimensions endowed with their geodesic distance is not known in general [1; 41]. To approximate it, it is convenient to consider another related distance.

Definition 2.4 Let X and Y be two subsets of $\mathbb{R}^d$. Consider them as metric spaces. Then, their *Euclidean–Hausdorff distance* d_{EH} is defined as

$$d_{\mathrm{EH}}(X,Y) = \inf\bigl\{d_{\mathrm{H}}(i_X(X), i_Y(Y)) \mid i_X\colon X \to \mathbb{R}^d \text{ and } i_Y\colon Y \to \mathbb{R}^d \text{ isometric embeddings}\bigr\}.$$

Using the following folklore result (see, for example, [9, Chapter IV, Section 38]), d_{EH} can be conveniently characterised.

Theorem 2.5 *If $f\colon X \to Y$ is an isometry between two subsets of $\mathbb{R}^d$, then there exists an isometry $\tilde{f}\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $\tilde{f}|_X = f$.*

Corollary 2.6 (see, for example, [3, Corollary 4.3]) *If X and Y are two subsets of $\mathbb{R}^d$, then*

$$d_{\mathrm{EH}}(X,Y) = \inf_{f\in\operatorname{Isom}(\mathbb{R}^d)} d_{\mathrm{H}}(X, f(Y)),$$

where $\operatorname{Isom}(\mathbb{R}^d)$ denotes the group of isometries of $\mathbb{R}^d$.

Clearly, $d_{\mathrm{EH}}(X,Y)\geq d_{\mathrm{GH}}(X,Y)$ for every pair X and Y of subsets of an Euclidean space $\mathbb{R}^d$. Moreover, the inequality can be strict (for example, see [44]). A lower bound on the Gromov–Hausdorff distance depending on the Euclidean–Hausdorff distance was provided in [44].

Theorem 2.7 *For every pair of compact subsets X and Y of $\mathbb{R}^d$,*

$$d_{\mathrm{GH}}(X,Y)\leq d_{\mathrm{EH}}(X,Y)\leq c_d\sqrt{M\cdot d_{\mathrm{GH}}(X,Y)},$$

where $M=\max\{\operatorname{diam}X,\operatorname{diam}Y\}$ and c_d is a constant depending only on the dimension d.

However, if X and Y are finite subsets of $\mathbb{R}$, linear lower bounds to d_{GH} depending on d_{EH} can be proved.

Theorem 2.8 [42, Theorem 3.2] *For every pair X and Y of compact subsets of $\mathbb{R}$,*

$$\tfrac{4}{5}d_{\mathrm{EH}}(X,Y)\leq d_{\mathrm{GH}}(X,Y)\leq d_{\mathrm{EH}}(X,Y).$$

If we denote by $\mathcal{EH}_1$ ($\mathcal{EH}_1^{<\omega}$) the space of isometry classes of compact (finite, respectively) subsets of the real line endowed with the Euclidean–Hausdorff distance, the canonical inclusion of $\mathcal{EH}_1$ into $\mathcal{GH}$ and that of $\mathcal{EH}_1^{<\omega}$ into $\mathcal{GH}^{<\omega}$ are bi-Lipschitz according to Theorem 2.8. Let us recall that a map $\psi\colon X\to Y$ between metric spaces is a *bi-Lipschitz embedding* if there are two linear maps $\rho_-\colon x\mapsto a\cdot x$ and $\rho_+\colon x\mapsto b\cdot x$, where $a,b>0$, satisfying (2).

3 The space of metric spaces of at most *n* points is coarsely embeddable into a Hilbert space

Given two subsets Y, Z of a metric space (X,d) and a radius $r\geq 0$ we write

$$\operatorname{dist}_d(Y,Z)=\operatorname{dist}(Y,Z)=\inf\{d(y,z)\mid y\in Y,\ z\in Z\}\quad\text{and}\quad B_d(Y,r)=\bigcup_{y\in Y}B_d(y,r),$$

where $B_d(y,r)$ denotes the closed ball centred in y with radius r.

A family of subsets $\mathcal{U}$ of a metric space X is said to be

- *uniformly bounded* if there exists $R\geq 0$ such that $\operatorname{diam}U\leq R$ for every $U\in\mathcal{U}$ (if we need to specify R, we say it is *R-bounded*);
- *r-disjoint* for some $r>0$ if $\operatorname{dist}(U,V)>r$ for every $U,V\in\mathcal{U}$ with $U\neq V$.

Definition 3.1 [32] Let X be a metric space. The *asymptotic dimension* of X is at most $n\in\mathbb{N}$ (and we write $\operatorname{asdim}X\leq n$) if for every $r>0$ there exists a uniformly bounded cover $\mathcal{U}=\mathcal{U}_0\cup\cdots\cup\mathcal{U}_n$ of X such that $\mathcal{U}_i$ is r-disjoint for every $i=0,\dots,n$. Furthermore, $\operatorname{asdim}X=n$ if $\operatorname{asdim}X\leq n$ and $\operatorname{asdim}X\not\leq n$, and $\operatorname{asdim}X=\infty$ if $\operatorname{asdim}X\not\leq m$ for every $m\in\mathbb{N}$.

Example 3.2 (see [52]) For every $n\in\mathbb{N}$, $\operatorname{asdim}\mathbb{R}^n=\operatorname{asdim}(\mathbb{R}_{\geq 0})^n=n$ where $\mathbb{R}^n$ is equipped with any p-norm, $p\in[1,\infty]$, and $(\mathbb{R}_{\geq 0})^n$ with any of the inherited metrics.

Let us recall two basic properties of the asymptotic dimension that we are going to use later in this section.

Proposition 3.3 (see [7]) *Let (X, d) be a metric space and $Y \subseteq X$ be a subspace. Then*

(a) $\operatorname{asdim} Y \leq \operatorname{asdim} X$, *and*

(b) $\operatorname{asdim} Y = \operatorname{asdim} X$ *provided that Y is* ***large*** *in X, ie, there exists $r \geq 0$ such that $B_d(Y, r) = X$.*

The goal of this section is to prove Theorem A, which we obtain as a particular case of the more general Corollary 3.11.

Lemma 3.4 *For every $n \in \mathbb{N}$, $\operatorname{asdim} \mathcal{GH}^{\leq n} \geq n(n-1)/2$.*

Proof By [35, Theorem 4.1], $\mathcal{GH}^{\leq n}$ contains isometric copies of arbitrarily large balls of $\mathbb{R}^{n(n-1)/2}$ endowed with the supremum metric. Then, [48, Lemma 2.10] implies that $\operatorname{asdim} \mathcal{GH}^{\leq n} \geq n(n-1)/2$. □

In order to prove the opposite inequality, we need to show different steps. Let us start with recalling a known result.

Theorem 3.5 ([56]; see also [37, Theorem 4.6]) *For every $n \in \mathbb{N}$, $\operatorname{asdim}([X]^{\leq n}, d_{\mathrm{H}}) \leq n \operatorname{asdim} X$. In particular, $\operatorname{asdim}[\mathbb{R}]^{\leq n} \leq n$ and $\operatorname{asdim}[\mathbb{R}_{\geq 0}]^{\leq n} \leq n$.*

For every positive integer $n \in \mathbb{N}$, let us define $\mathcal{X}_n$ as the metric subspace of $([\mathbb{R}_{\geq 0}]^{\leq n+1}, d_{\mathrm{H}})$ whose elements contain the point 0. According to Theorem 3.5, $\operatorname{asdim} \mathcal{X}_n \leq n+1$ since the asymptotic dimension is monotone (Proposition 3.3(a)). In the sequel, we improve that bound showing that $\operatorname{asdim} \mathcal{X}_n \leq n$. The proof outline is similar to and inspired by that of [48, Theorem 3.2]. First, we need a classical preliminary result.

Given two families $\mathcal{U}$ and $\mathcal{V}$ of subsets of a metric space X and $r > 0$, we define a new family of subsets as

$$\mathcal{U} \cup_r \mathcal{V} = \{N_r(U, \mathcal{V}) \mid U \in \mathcal{U}\} \cup \{V \in \mathcal{V} \mid \forall U \in \mathcal{U},\ \operatorname{dist}(V, U) > r\},$$

where $N_r(U, \mathcal{V}) = U \cup \bigcup \{V \in \mathcal{V} \mid \operatorname{dist}(U, V) \leq r\}$.

Let us immediately note that $\bigcup (\mathcal{V} \cup_r \mathcal{U}) \supseteq \bigcup \mathcal{V} \cup \bigcup \mathcal{U}$.

Lemma 3.6 [7, Proposition 24] *Let $\mathcal{U}$ be an r-disjoint, R-bounded family of subsets of X with $R \geq r$. Let $\mathcal{V}$ be a $5R$-disjoint, uniformly bounded family of subsets of X. Then $\mathcal{V} \cup_r \mathcal{U}$ is r-disjoint and uniformly bounded.*

Lemma 3.7 *For every $n \in \mathbb{N}$, $\operatorname{asdim} \mathcal{X}_n \leq n$.*

Proof Let us prove the result by induction.

If $n = 1$, then, for every $r > 0$, the usual uniformly bounded cover used to prove that the real line has asymptotic dimension 1 (see [7; 52]) can be adapted. More precisely, for every $r > 0$, define, for $i = 0, 1$,

$$\mathcal{U}_i = \{\{\{0, x\} \mid x \in V_k^i\} \mid k \in \mathbb{N}\}, \quad \text{where } V_k^i = [(4k + 2i)r, (4k + 2i + 2)r] \subseteq \mathbb{R}_{\geq 0}$$

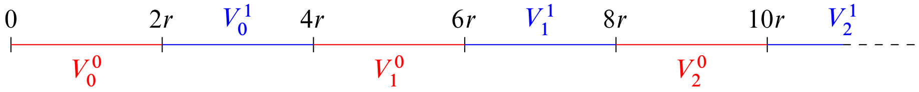


Figure 1: A representation of the uniformly bounded cover $\mathcal{V} = \mathcal{V}_0 \cup \mathcal{V}_1$ showing that $\operatorname{asdim} \mathbb{R}_{\geq 0} \leq 1$. In red, the elements of $\mathcal{V}_0 = \{V_k^0 \mid k \in \mathbb{N}\}$ and, in blue, the subsets contained in the family $\mathcal{V}_1 = \{V_k^1 \mid k \in \mathbb{N}\}$.

(see Figure 1). Then, each $\mathcal{U}_i$ is r-disjoint and $\mathcal{U}_0 \cup \mathcal{U}_1$ forms a uniformly bounded cover of $\mathcal{X}_1$. Hence, $\operatorname{asdim} \mathcal{X}_1 \leq 1$.

Suppose now that the assertion is true for some $n \in \mathbb{N}$. We want to show that $\operatorname{asdim} \mathcal{X}_{n+1} \leq n+1$. Fix a positive $r > 0$. Then we can write

$$\mathcal{X}_{n+1} = \widetilde{\mathcal{X}_{n+1}} \cup B_{d_{\mathrm{H}}}(\mathcal{X}_n, r),$$

where $\widetilde{\mathcal{X}_{n+1}}$ consists of subsets A of $\mathbb{R}_{\geq 0}$ such that, if $x \in A$ has $|x| \leq r$, then $x = 0$. It indeed represents a partition. In fact, every element $C \in \mathcal{X}_{n+1}$ can be decomposed $C = C_0 \cup C_1$ where $\max_{x \in C_0} |x| \leq r$ and $\min_{x \in C_1} |x| > r$. Then, either $C_0 = \{0\}$ and so $C \in \widetilde{\mathcal{X}_{n+1}}$, or $d_{\mathrm{H}}(C, C_1 \cup \{0\}) \leq r$ and $C_1 \cup \{0\} \in \mathcal{X}_n$.

Let us now consider $\widetilde{\mathcal{X}_{n+1}}$, and define the subspace $\mathcal{Y}_n = \{X \setminus \{0\} \mid X \in \widetilde{\mathcal{X}_{n+1}}\}$ of $[\mathbb{R}_{\geq 0}]^{\leq n+1}$. Monotonicity of the asymptotic dimension (Proposition 3.3(a)) implies that $\operatorname{asdim} \mathcal{Y}_n \leq n+1$. Hence, there exists a uniformly bounded cover $\mathcal{W} = \mathcal{W}_0 \cup \cdots \cup \mathcal{W}_{n+1}$ of $\mathcal{Y}_n$ such that $\mathcal{W}_i$ is r-disjoint for every $i = 0, \dots, n+1$. Let $R \geq r$ be an upper bound to the diameter of the elements in $\mathcal{W}$. For every $i = 0, \dots, n+1$, construct the family of subsets $\widetilde{\mathcal{W}_i} = \{W \cup \{0\} \mid W \in \mathcal{W}_i\}$ of $\widetilde{\mathcal{X}_{n+1}}$. Note that $\widetilde{\mathcal{W}} = \widetilde{\mathcal{W}_0} \cup \cdots \cup \widetilde{\mathcal{W}_{n+1}}$ is a cover of $\widetilde{\mathcal{X}_{n+1}}$. Furthermore, each element of $\widetilde{\mathcal{W}}$ has diameter bounded by R and it is easy to see that each of the families $\widetilde{\mathcal{W}_i}$, $i = 0, \dots, n+1$, is r-disjoint.

Since $\mathcal{X}_n$ is large in $B_{d_{\mathrm{H}}}(\mathcal{X}_n, r)$, $\operatorname{asdim} B_{d_{\mathrm{H}}}(\mathcal{X}_n, r) \leq n \leq n+1$ by the inductive hypothesis and Proposition 3.3(b). Hence, there exists a uniformly bounded cover $\mathcal{V} = \mathcal{V}_0 \cup \cdots \cup \mathcal{V}_{n+1}$ of $B_{d_{\mathrm{H}}}(\mathcal{X}_n, r)$ such that $\mathcal{V}_i$ is $5R$-disjoint for every $i = 0, \dots, n+1$.

Define, for every $i = 0, \dots, n+1$,

$$\mathcal{U}_i = \widetilde{\mathcal{W}_i} \cup_r \mathcal{V}_i.$$

Hence, $\mathcal{U} = \mathcal{U}_0 \cup \cdots \cup \mathcal{U}_{n+1}$ is a uniformly bounded cover of $\mathcal{X}_{n+1} = \bigcup \widetilde{\mathcal{W}} \cup \bigcup \mathcal{V}$ and $\mathcal{U}_i$ is r-disjoint for every $i = 0, \dots, n+1$ according to Lemma 3.6. Thus, $\operatorname{asdim} \mathcal{X}_{n+1} \leq n+1$. □

On the family of all isometry classes of finite networks, the network distance $d_{\mathcal{N}}$ is a *pseudometric*, ie, it satisfies the properties (M1), (M3) and (M4) stated in Definition 2.1 [19]. Consider the equivalence relation between finite networks given by $X \sim Y$ if $d_{\mathcal{N}}(X, Y) = 0$. We refer to [20] where this equivalence relation is completely characterised. The space obtained by quotienting the family of all isometry classes of finite networks under this equivalence relation is a metric space (it is a standard way to obtain a metric space out of a pseudometric space; see, for example, [12, Proposition 1.1.5]). We denote it by $\mathcal{N}^{<\omega}$. Let us recall that the distance between two elements in $\mathcal{N}^{<\omega}$ is the network distance between any two representatives. For the sake of simplicity, we will be directly working with representatives of objects in $\mathcal{N}^{<\omega}$ instead of their equivalence classes without explicit mention. Let us also consider the following metric subspaces of $\mathcal{N}^{<\omega}$:

- $\mathcal{S}^{<\omega}$ – the family of equivalence classes of finite *pseudo-semi-metric spaces*, ie, networks satisfying (M1) and (M3);
- $\mathcal{N}^{\leq n}$ – the family of equivalence classes of networks whose cardinality is at most n;
- $\mathcal{S}^{\leq n} = \mathcal{S}^{<\omega} \cap \mathcal{N}^{\leq n}$.

Define the map $\mathcal{D}\colon \mathcal{N}^{<\omega} \to [\mathbb{R}]^{<\omega}$ as follows: for every $(X, d) \in \mathcal{N}^{<\omega}$,

$$\mathcal{D}(X) = d(X \times X) \subseteq \mathbb{R}.$$

The subset $\mathcal{D}(X)$ is called the *distance set* of X. Note that $\mathcal{D}|_{\mathcal{S}^{\leq n}}\colon \mathcal{S}^{\leq n} \to \mathcal{X}_{n(n-1)/2}$. For the sake of simplicity, we denote by $\mathcal{D}$ also the restriction.

Proposition 3.8 [20, Proposition 4.3.4] *The map $\mathcal{D}\colon \mathcal{N}^{<\omega} \to ([\mathbb{R}]^{<\omega}, d_{\mathrm{H}})$ is well defined and* ***2-Lipschitz*** *(ie, $d_{\mathrm{H}}(\mathcal{D}(X), \mathcal{D}(Y)) \leq 2 d_{\mathcal{N}}(X, Y)$ for every $X, Y \in \mathcal{N}^{<\omega}$).*

Proof Let $(X, d_X), (Y, d_Y) \in \mathcal{N}^{<\omega}$ and $\mathcal{R}$ be a correspondence such that $R = \operatorname{dis}\mathcal{R} = 2d_{\mathcal{N}}(X, Y)$. For every $x_1, x_2 \in X$ pick $y_1, y_2 \in Y$ satisfying $(x_i, y_i) \in \mathcal{R}$, for $i = 1, 2$. Then, $|d_X(x_1, x_2) - d_Y(y_1, y_2)| \leq R$. Since the same argument can be carried out also for every pair of points $y_1, y_2 \in Y$, $d_{\mathrm{H}}(\mathcal{D}(X), \mathcal{D}(Y)) \leq R$. This inequality also implies that the map is well defined. □

Even if we restrict ourselves to consider only metric spaces that are finite subsets of the real line, the map $\mathcal{D}$ is not injective [8]. However, it has a further property that allows us to deduce an upper bound to the asymptotic dimension of $\mathcal{GH}^{\leq n}$.

Let us recall that, for a positive integer $k \in \mathbb{N}$, a map $\varphi\colon (X, d_X) \to (Y, d_Y)$ between metric spaces is *coarsely k-to-1* [50] if

- there exists a map $\rho_+\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that, for every $x, y \in X$,
  $$d_Y(\varphi(x), \varphi(y)) \leq \rho_+(d_X(x, y))$$
  (ie, φ is *bornologous*);
- for every $R \geq 0$, there exists $S \geq 0$ such that for every $y \in Y$ there are $x_1, \dots, x_k \in X$ satisfying
  $$\varphi^{-1}(B_{d_Y}(y, R)) \subseteq \bigcup_{i=1}^{k} B_{d_X}(x_i, S).$$

Coarsely k-to-1 maps play a very important role in providing bounds for the asymptotic dimension.

Theorem 3.9 ([23; 50]; see also [24]) *Let $\varphi\colon X \to Y$ be a coarsely k-to-1 surjective map between metric spaces. Then,*

$$\operatorname{asdim} X \leq \operatorname{asdim} Y \leq k(\operatorname{asdim} X + 1) - 1.$$

Theorem 3.10 *The map $\mathcal{D}\colon \mathcal{S}^{\leq n} \to \mathcal{D}(\mathcal{S}^{\leq n}) = \mathcal{X}_{n(n-1)/2}$ is coarsely k-to-1 for some suitable k.*

Proof Proposition 3.8 implies that the map is bornologous.

Let us fix $R \in \mathbb{R}_{\geq 0}$ and $D \in \mathcal{D}(\mathcal{S}^{\leq n}) = \mathcal{X}_{n(n-1)/2}$. In particular, $0 \in D$. Suppose that $(Y, d_Y) \in \mathcal{S}^{\leq n}$ satisfies $d_{\mathrm{H}}(D, \mathcal{D}(Y)) \leq R$. Then, there exists a function $f : \mathcal{D}(Y) \to D$ such that $|f(a) - a| \leq R$ for every $a \in \mathcal{D}(Y)$. Furthermore, without loss of generality, we can require that $f(0) = 0$ (note that $0 \in \mathcal{D}(Y)$).

Define a new object $X_Y \in \mathcal{S}^{\leq n}$ as $X_Y = (Y, f \circ d_Y)$. Then,

(a) $X_Y \in \mathcal{S}^{\leq n}$;

(b) $\mathcal{D}(X_Y) \subseteq D$;

(c) $d_{\mathcal{N}}(Y, X_Y) \leq \frac{1}{2} \operatorname{dis} \operatorname{id} \leq \frac{1}{2} R$.

Items (a) and (b) imply that

$$|\{X_Y \mid Y \in \mathcal{S}^{\leq n} : d_{\mathrm{H}}(\mathcal{D}(Y), D) \leq R\}| \leq |D|^{\frac{1}{2}n(n-1)} \leq \left(\tfrac{1}{2}n(n-1) + 1\right)^{\frac{1}{2}n(n-1)} =: k.$$

Hence, $\mathcal{D}$ is a coarsely k-to-1 map. □

Corollary 3.11 *For every $n \in \mathbb{N}$, $\operatorname{asdim} \mathcal{GH}^{\leq n} = \operatorname{asdim} \mathcal{S}^{\leq n} = n(n-1)/2$.*

Proof The claim is implied by

$$\tfrac{1}{2}n(n-1) \leq \operatorname{asdim} \mathcal{GH}^{\leq n} \leq \operatorname{asdim} \mathcal{S}^{\leq n} \leq \operatorname{asdim} \mathcal{D}(\mathcal{S}^{\leq n}) = \operatorname{asdim} \mathcal{X}_{\frac{1}{2}n(n-1)} \leq \tfrac{1}{2}n(n-1),$$

where the first inequality is stated in Lemma 3.4, the second one follows from Proposition 3.3(a), the third one from Theorem 3.9, and the last from Lemma 3.7. □

Thus, Theorem A follows since having finite asymptotic dimension implies the existence of a coarse embedding into a Hilbert space [34; 65].

Remark 3.12 Similarly, we can show that $\mathcal{D}\colon \mathcal{N}^{\leq n} \to [\mathbb{R}]^{n^2}$ is coarsely $(n^2)^{n^2}$-to-1 surjective map and so $\operatorname{asdim} \mathcal{N}^{\leq n} \leq \operatorname{asdim} [\mathbb{R}]^{\leq n^2} \leq n^2$ by Theorems 3.9 and 3.5. Hence,

$$\tfrac{1}{2}n(n-1) \leq \operatorname{asdim} \mathcal{N}^{\leq n} \leq n^2.$$

In particular, $\mathcal{N}^{\leq n}$ can be coarsely embedded into a Hilbert space.

4 Coarse nonembeddability into Hilbert spaces

Let us now recall some definitions and results coming from coarse geometry as they will be the crucial stepping stones to prove our main results.

Suppose that the map $\psi\colon (X,d_X)\to(Y,d_Y)$ between metric spaces is a coarse embedding. If two maps $\rho_-\colon\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ and $\rho_+\colon\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ are such that $\rho_-\to\infty$ and (2) is fulfilled, then we call ρ_- and ρ_+ *control functions*, and we say that ψ is a (ρ_-,ρ_+)*-coarse embedding*.

Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of metric spaces and X be another metric space. We say that a family of maps $\{i_k\colon X_k\to X\}_{k\in\mathbb{N}}$ is a *coarse embedding* if there exist two maps $\rho_-,\rho_+\colon\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ such that i_k is a (ρ_-,ρ_+)-coarse embedding for every $k\in\mathbb{N}$. We say that X *contains a coarse disjoint union of* $\{X_k\}_{k\in\mathbb{N}}$ if there exists a coarse embedding $\{i_k\colon X_k\to X\}$ such that

$$\operatorname{dist}(i_n(X_n),i_m(X_m))\xrightarrow{m+n\to\infty}\infty.$$

Let us recall that a Banach space $(A,\|\cdot\|)$ is *uniformly convex* if, for every $0<\varepsilon\leq 2$, there exists $\delta>0$ so that, for any two vectors $x,y\in A$ with $\|x\|=\|y\|=1$, the condition $\|x-y\|\geq\varepsilon$ implies that $\|(x+y)/2\|\leq 1-\delta$. Hilbert spaces are, in particular, uniformly convex Banach spaces.

Theorem 4.1 [39] *There exists a sequence $\{X_k\}_{k\in\mathbb{N}}$ of finite metric spaces such that, if a metric space X contains a coarse disjoint union of $\{X_k\}_{k\in\mathbb{N}}$, then X cannot be coarsely embedded into any uniformly convex Banach space.*

The goal of this section is to prove the following result.

Theorem 4.2 *$\mathcal{EH}_1^{<\omega}$ cannot be coarsely embedded into any uniformly convex Banach space.*

Hence, Theorem B immediately follows thanks to Theorem 2.8 since a composite of coarse embeddings is still a coarse embedding.

To prove Theorem 4.2, we intend to apply Theorem 4.1. Following the approach used in [63], let us first isometrically embed an arbitrary finite metric space into a more manageable space.

Lemma 4.3 (Kuratowski embedding) *For every finite metric space X, there are $m,n\in\mathbb{N}\setminus\{0\}$ such that X can be isometrically embedded into $([0,m]^n,d_m^n)$, where $d_m^n((x_i)_i,(y_i)_i)=\max_{i=1,\dots,n}|x_i-y_i|$ for every $(x_i)_i,(y_i)_i\in[0,m]^n$.*

Therefore, in order to apply Theorem 4.1, we show that $\mathcal{EH}_1^{<\omega}$ contains a coarse disjoint union of the family $\{[0,m]^n\mid m,n\in\mathbb{N}\setminus\{0\}\}$. We intend to define the coarse embeddings $\varphi_m^n\colon[0,m]^n\to[\mathbb{R}]^{<\omega}$ recursively. To do it, let us fix a bijection $T\colon(\mathbb{N}\setminus\{0\})^2\to\mathbb{N}\setminus\{0\}$ defined as follows: for every pair $(m,n)\in\mathbb{N}$,

$$T(m,n)=m+\sum_{i=2}^{m+n-1}(i-1)=m+\tfrac{1}{2}(m+n-2)(m+n-1).$$

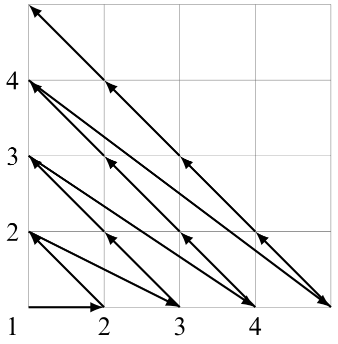


Figure 2: A representation of the map $T^{-1}\colon \mathbb{N}\setminus\{0\}\to(\mathbb{N}\setminus\{0\})^2$.

We represent in Figure 2 the sequence of the points $T(m,n)$. By construction, for every two pairs $(m,n),(m',n')\in\mathbb{N}\setminus\{0\}$, $m+n\le m'+n'$ provided that $T(m,n)\le T(m',n')$. For the sake of simplicity, for two pairs $(m,n),(m',n')\in(\mathbb{N}\setminus\{0\})^2$, let us write $(m,n)\preceq(m',n')$ if $T(m,n)\le T(m',n')$, and $(m,n)\prec(m',n')$ if $(m,n)\preceq(m',n')$ and $(m,n)\ne(m',n')$.

To construct the maps φ_m^n, we need various parameters. Let us define a sequence $\{a_i(m)\}_{i\in\mathbb{N}\setminus\{0\}}$ of positive real values depending on m as

$$a_i(m)=4m(i-1).$$

Furthermore, for every $m,n\in\mathbb{N}\setminus\{0\}$, we inductively construct

$$D(m,n)=\max\{4m(n+2),\max_{(m',n')\prec(m,n)}D(m',n')+m+2^{T(m,n)}\}.$$

Thus, $D(1,1)=12$ (according to the notation that $\max\varnothing=-\infty$), and $D(m,n)\ge 4m(n+2)=a_n(m)+12m$.

Let us define, for every $m,n\in\mathbb{N}\setminus\{0\}$, a map $\varphi_m^n\colon[0,m]^n\to[\mathbb{R}]^{=n+1}$ as follows: for every $(x_i)_i\in[0,m]^n$,

$$\varphi_m^n((x_i)_i)=\{a_i(m)+x_i\mid i=1,\dots,n\}\cup\{D(m,n)\}$$

(see Figure 3). This construction should be compared with that provided in [63]. The crucial difference is the last point $D(m,n)$. Its purpose is to disincentivise the action of isometries and conveniently increase the diameter of the image of φ_m^n (see Lemma 4.4).

Lemma 4.4 *The map* $\varphi_m^n\colon([0,m]^n,d_m^n)\to\mathcal{EH}_1^{<\omega}$ *is a coarse embedding whose control functions are independent of* m *and* n*. More precisely,* φ_m^n *is a* (ρ_-,ρ_+)*-coarse embedding, where* $\rho_-\colon x\mapsto x/2$ *and* $\rho_+=\mathrm{id}$*. Furthermore, for every* $(x_i)_i\in[0,m]^n$*, we have* $\operatorname{diam}\varphi_m^n((x_i)_i)\in[D(m,n)-m,\,D(m,n)]$.

Proof It is easy to see that

$$d_{\mathrm{EH}}\big(\varphi_m^n((x_i)_i),\varphi_m^n((y_i)_i)\big)\le d_{\mathrm{H}}\big(\varphi_m^n((x_i)_i),\varphi_m^n((y_i)_i)\big)=d_m^n((x_i)_i,(y_i)_i)$$

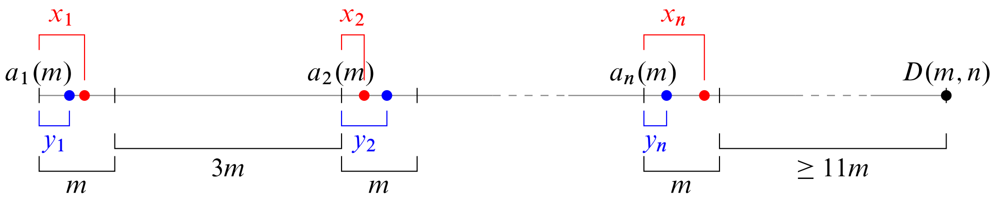


Figure 3: A representation of the images of two points $(x_i)_i, (y_i)_i \in [0,m]^n$ along φ_m^n. The subset $\varphi_m^n((x_i)_i)$ is given by the red dots and the black dot, while $\varphi_m^n((y_i)_i)$ consists of the blue dots and the black dot. In the picture, we can see that $d_{\mathrm{H}}\big(\varphi_m^n((x_i)_i), \varphi_m^n((y_i)_i)\big) = d_m^n((x_i)_i, (y_i)_i)$.

for every $(x_i)_i, (y_i)_i \in [0,m]^n$ (see also Figure 3). We want to prove that ρ_- is the other control function. Let $f \in \mathrm{Isom}(\mathbb{R}^d)$ be an isometry such that

$$d_{\mathrm{H}}\big(\varphi_m^n((x_i)_i), f\big(\varphi_m^n((y_i)_i)\big)\big) \leq d_m^n((x_i)_i, (y_i)_i) \leq m. \tag{3}$$

The isometry f is the composite of a translation g and a rotation h. If $n = 1$, then $|\varphi_m^n((x_i)_i)| = 2$, and so, without loss of generality, h can be taken as the identity.

Claim 4.5 *Assume that* $n \geq 2$. *Then*, $h = \mathrm{id}$.

Proof Suppose, by contradiction, that $h \neq \mathrm{id}$. According to (3), and because of the definition of φ_m^n,

$$|f(D(m,n)) - (a_1(m) + x_1)| \leq m. \tag{4}$$

Since $|(a_1(m) + x_1) - (a_2(m) + x_1)| \geq |a_1(m) - a_2(m)| - m > 2m$, (4) implies that

$$|f(D(m,n)) - (a_2(m) + x_2(m))| > m.$$

Therefore, because of (3),

$$|(a_2(m) + x_2) - f(a_n(m) + y_n)| \leq m. \tag{5}$$

Using the fact that f is an isometry, the triangular inequality, (4) and (5), the following chain of inequalities descends:

$$\begin{aligned} 11m &\leq |D(m,n) - a_n(m)| - m \leq |D(m,n) - (a_n(m) + y_n)| = |f(D(m,n)) - f(a_n(m) + y_n)| \\ &\leq |f(D(m,n)) - (a_1(m) + x_1)| + |(a_1(m) + x_1) - (a_2(m) + x_2)| + |(a_2(m) + x_2) - f(a_n(m) + y_n)| \\ &\leq m + 5m + m = 7m. \end{aligned}$$

Thus, we obtain a contradiction. □

We can now assume that the map $f = g$ is a translation. Using the triangular inequality, and since $a_{i+1}(m) - a_i(m) = 4m$ for every $i \in \{1, \dots, n\}$, it can be easily check that

$$|(a_i(m) + x_i) - f(a_i(m) + y_i)| \leq m \quad \text{for every } i \in \{1, \dots, n\} \qquad \text{and} \qquad |D(m,n) - f(D(m,n))| \leq m.$$

The point $D(m,n)$, common to both $\varphi_m^n((x_i)_i)$ and $\varphi_m^n((y_i)_i)$, misaligns as soon as f is nontrivial. Therefore, f creates a trade-off between the misalignment just described and a better alignment of the other n pairs of points ($a_i(m)+x_i$ and $a_i(m)+y_i$, for every $i\in\{1,\dots,n\}$). Thus, the best Hausdorff distance that can be achieved is at most $d_{\mathrm{H}}\bigl(\varphi_m^n((x_i)_i),\varphi_m^n((y_i)_i)\bigr)/2=d_m^n((x_i)_i,(y_i)_i)/2$, and so the claim.

The final assertion trivially follows from the definition of φ_m^n. □

The following result is immediate, but it is an important step in the proof of Theorem 4.2.

Lemma 4.6 *If $\mathcal{Y}$ and $\mathcal{Z}$ are two families of finite subsets of a metric space X,*

$$\operatorname{dist}_{d_{\mathrm{H}}}(\mathcal{Y},\mathcal{Z})\geq\tfrac{1}{2}\inf_{\substack{Y\in\mathcal{Y}\\ Z\in\mathcal{Z}}}|\operatorname{diam}Y-\operatorname{diam}Z|.$$

Furthermore, if $X=\mathbb{R}^d$, then

$$\operatorname{dist}_{d_{\mathrm{EH}}}(\mathcal{Y},\mathcal{Z})\geq\tfrac{1}{2}\inf_{\substack{Y\in\mathcal{Y}\\ Z\in\mathcal{Z}}}|\operatorname{diam}Y-\operatorname{diam}Z|.$$

Proof Since, for every pair of subsets Y and Z of X,

$$d_{\mathrm{H}}(Y,Z)\geq\tfrac{1}{2}|\operatorname{diam}Y-\operatorname{diam}Z|,$$

the first inequality is immediate. The second one follows from the fact that an isometry's action does not change the diameter of a subset. □

As a consequence of Lemma 4.6, if $\mathcal{Y}$ and $\mathcal{Z}$ are two families of finite subsets of $\mathbb{R}^d$ with the property that, for every $Y\in\mathcal{Y}$ and $Z\in\mathcal{Z}$, $\operatorname{diam}Y\geq\operatorname{diam}Z$,

$$\operatorname{dist}_{d_{\mathrm{EH}}}(\mathcal{Y},\mathcal{Z})\geq\tfrac{1}{2}\bigl(\inf_{Y\in\mathcal{Y}}\operatorname{diam}Y-\sup_{Z\in\mathcal{Z}}\operatorname{diam}Z\bigr).\tag{6}$$

We now have the tools to show the main result of the section, Theorem 4.2, and its consequence, Theorem B.

Proof of Theorem 4.2 We intend to use Lafforgue's result (Theorem 4.1), and, thanks to Lemma 4.3, we need to show that $\mathcal{EH}_1^{<\omega}$ contains a coarse disjoint union of $\bigl\{([0,m]^n,d_m^n)\bigr\}_{m,n\in\mathbb{N}\setminus\{0\}}$. According to Lemma 4.4, $\{\varphi_m^n\colon[0,m]^n\to\mathcal{EH}_1^{<\omega}\}$ is a coarse embedding. It remains to show that

$$\operatorname{dist}_{d_{\mathrm{EH}}}\bigl(\varphi_m^n([0,m]^n),\varphi_{m'}^{n'}([0,m']^{n'})\bigr)\xrightarrow{T(m,n)+T(m',n')\to\infty}\infty.$$

Without loss of generality, we assume that $(m',n')\prec(m,n)$. Then, according to Lemma 4.4, (6), and the definition of $D(m,n)$,

$$\operatorname{dist}_{d_{\mathrm{EH}}}\bigl(\varphi_m^n([0,m]^n),\varphi_{m'}^{n'}([0,m']^{n'})\bigr)\geq\tfrac{1}{2}(D(m,n)-D(m',n')-m)\geq2^{T(m,n)-1}.\tag{7}$$

From (7), the desired result descends. □

5 Bi-Lipschitz nonembeddability into finite-dimensional spaces

Let us consider the isometry classes of all finite subsets of the real line with diameter at most R. This set can be identified with the isometry classes of elements in $[[0,R]]^{<\omega}$. Let us denote by $\mathcal{EH}^{<\omega}_{[0,R]}$ this space equipped with the Euclidean–Hausdorff distance.

Since, for every pair of metric spaces $X, Y \in \mathcal{GH}^{<\omega}_{\leq R}$,

$$d_{\mathrm{GH}}(X,Y) \leq \tfrac{1}{2}\max\{\operatorname{diam} X, \operatorname{diam} Y\} \leq \tfrac{1}{2}R,$$

the diameter of $\mathcal{GH}^{<\omega}_{\leq R}$ is finite. Therefore, it can be trivially coarsely embedded into a one-point metric space. However, we can still provide nonembeddability results if we restrict the class of embeddings. More precisely, we prove the following.

Theorem 5.1 *$\mathcal{EH}^{<\omega}_{[0,R]}$ cannot be bi-Lipschitz embedded into a finite-dimensional Hilbert space.*

Immediately, we can deduce Theorem C since Theorem 2.8 provides a bi-Lipschitz embedding.

To prove Theorem 5.1, we use the Assouad dimension.

Definition 5.2 Given a metric space (X,d), a subset $E \subseteq X$ and $r > 0$, we denote by $N_r(E)$ the least number of open balls of radius less or equal to r that cover E. Then, the *Assouad dimension of X* [5; 10] is

$$\dim_{\mathrm{A}} X = \inf\bigl\{\alpha > 0 \mid \exists C > 0 : \forall r > 0, \forall \beta \in (0,1], \sup_{x\in X} N_{\beta r}(B^o_d(x,r)) < C\beta^{-\alpha}\bigr\},$$

where $B^o_d(x,r)$ denotes the open ball centred in x with radius r.

This dimension notion was introduced precisely to prove obstructions to bi-Lipschitz embed metric spaces into an Euclidean space. In particular, the following properties lead to the desired conclusion (see, for example, [57]):

- If $\varphi\colon X \to Y$ is a bi-Lipschitz embedding between metric spaces, $\dim_{\mathrm{A}} X = \dim_{\mathrm{A}} \operatorname{im}(\varphi) \leq \dim_{\mathrm{A}} Y$.
- For every $n \in \mathbb{N}$, $\dim_{\mathrm{A}} \mathbb{R}^n = n$.

Therefore, once we prove that $\dim_{\mathrm{A}} \mathcal{EH}^{<\omega}_{[0,R]} = \infty$ (Proposition 5.3), Theorem 5.1 immediately follows.

Let us mention that the same strategy was used in [16] to prove that spaces of persistence diagrams cannot be bi-Lipschitz embedded into a finite-dimensional Hilbert space.

Proposition 5.3 *For every $R > 0$, $\dim_{\mathrm{A}} \mathcal{EH}^{<\omega}_{[0,R]} = \infty$.*

Proof We want to provide, for every $\alpha > 0$ and every $C > 0$, a radius $r > 0$, a constant $\beta \in (0,1]$, a finite subset A of $[0,R]$ and $M = \lceil C\beta^{-\alpha} + 1\rceil$-many finite subsets $A_1, \dots, A_M$ of $[0,R]$ with the following properties: $d_{\mathrm{EH}}(A, A_i) < r$ and $d_{\mathrm{EH}}(A_i, A_j) > 2\beta r$ for every $i, j \in \{1, \dots, M\}$. Therefore, the open ball centred in A with radius r cannot be covered by fewer than M-many open balls with radius βr (A_i and A_j are contained in the same ball only if $i = j$), which will conclude the proof since $M > C\beta^{-\alpha}$.

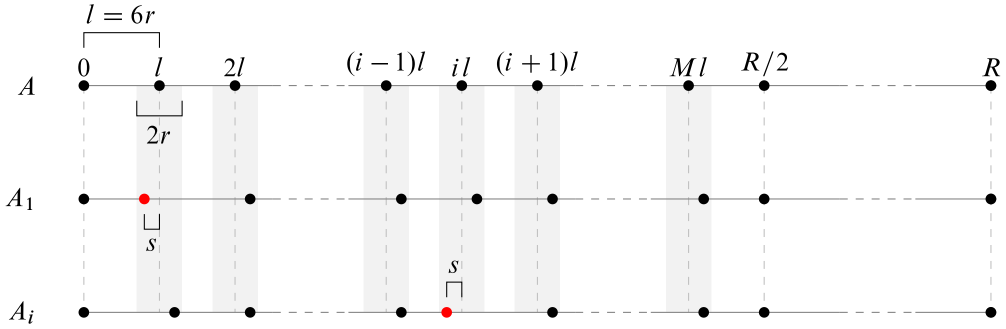


Figure 4: A representation of the subsets A, A_1 and A_i defined in the proof of Proposition 5.3. The distinctive points $l-s \in A_1$ and $il-s \in A_i$ are emphasised in red. The light grey strips are meant to visualise the fact that $d_{\mathrm{H}}(A, A_1) < r$ and $d_{\mathrm{H}}(A, A_i) < r$.

Define $l = \frac{R}{2(M+1)}$, $r = \frac{l}{6}$, $s = \frac{2r}{3}$ and $\beta = \frac{1}{2}$. We construct the subsets A and A_i as

$$\begin{aligned} A &= \{j \cdot l \mid j \in \{0, \dots, M+1\}\} \cup \{R\}, \\ A_i &= \{j \cdot l + s \mid j \in \{1, \dots, M\} \setminus \{i\}\} \cup \{i \cdot l - s\} \cup \{0, \tfrac{1}{2}R, R\}. \end{aligned}$$

The subset A and A_i are represented in Figure 4. It is clear that $d_{\mathrm{EH}}(A, A_i) < r$ since $s < r$ (actually, $d_{\mathrm{H}}(A, A_i) = s < r$). It remains to show that, for every pair of distinct indices $i, j \in \{1, \dots, M\}$, $d_{\mathrm{EH}}(A_i, A_j) \geq 2s > r$.

Let $i, j \in \{1, \dots, M\}$ be two distinct indices. Assume, by contradiction, that $d_{\mathrm{EH}}(A_i, A_j) < 2s$, and let f be an isometry of $\mathbb{R}$ such that $d_{\mathrm{H}}(A_i, f(A_j)) < 2s$. Adapting the argument of Claim 4.5, we can assume that f is a translation.

For the sake of simplicity, for every $k \in \{1, \dots, n\}$, we name the points of A_k as

$$\begin{gathered} a_0^k = 0, \quad a_1^k = l + s, \quad \dots, \quad a_{k-1}^k = (k-1)l + s, \quad a_k^k = kl - s, \\ a_{k+1}^k = (k+1)l + s, \quad \dots, \quad a_M^k = Ml + s, \quad a_{M+1}^k = \tfrac{1}{2}R, \quad a_{M+2}^k = R. \end{gathered}$$

Following the strategy used in the proof of Lemma 4.4, we can show that, for every $k \in \{0, \dots, M+2\}$,

$$|a_k^i - f(a_k^j)| < 2s. \tag{8}$$

Assume, without loss of generality, that $i < j$. Then, in particular,

$$|a_i^j - a_j^j| = (j-i)l - 2s \quad \text{and} \quad |a_i^i - a_j^i| = (j-i)l + 2s. \tag{9}$$

Using the triangular inequality, (8), (9) and the fact that f is an isometry we obtain

$$\begin{aligned} (j-i)l + 2s = |a_i^i - a_j^i| &\leq |a_i^i - f(a_i^j)| + |f(a_i^j) - f(a_j^j)| + |f(a_j^j) - a_j^i| \\ &< |a_i^j - a_j^j| + 4s = (j-i)l + 2s. \end{aligned}$$

Hence, we have found a contradiction. □

Remark 5.4 As a byproduct of the proof of Proposition 5.3, we obtain that the Assouad dimension of the family of all finite subsets of an interval equipped with the Hausdorff distance is infinite. Indeed, in the notation of the mentioned proof, $d_{\mathrm{H}}(A, A_i) = s < r$, but $d_{\mathrm{H}}(A_i, A_j) = 2s > r$ for every pair of distinct indices $i, j \in \{1, \dots, M\}$.

5.1 Questions about bi-Lipschitz embeddability and Assouad dimension

Let us conclude the presentation with a discussion about potential future research directions concerning the bi-Lipschitz embeddability of the Gromov–Hausdorff space. First of all, Theorem C leaves the following question open.

Question 5.5 *Can $\mathcal{GH}^{<\omega}_{\leq R}$ be bi-Lipschitz embedded into an infinite-dimensional Hilbert space?*

Furthermore, it is natural to ask what the embeddability properties are if we bound the cardinality of the metric spaces as in Section 3. Let us define $\mathcal{GH}^{\leq n}_{\leq R} = \mathcal{GH}^{\leq n} \cap \mathcal{GH}^{<\omega}_{\leq R}$ for $n \in \mathbb{N}$ and $R > 0$.

Question 5.6 *Can $\mathcal{GH}^{\leq n}$ and $\mathcal{GH}^{\leq n}_{\leq R}$ be bi-Lipschitz embedded into a (finite-dimensional) Hilbert space?*

To approach Question 5.6, we may investigate the Assouad dimension of those spaces. The inequalities

$$\dim_{\mathrm{A}} \mathcal{GH}^{\leq n} \geq \dim_{\mathrm{A}} \mathcal{GH}^{\leq n}_{\leq R} \geq \tfrac{1}{2}n(n-1)$$

can be derived similarly to the proof of Lemma 3.4. Indeed, according to [35, Theorem 4.1], $\mathcal{GH}^{\leq n}_{\leq R}$ contains a subspace isometric to an open ball in $\mathbb{R}^{n(n-1)/2}$ of sufficiently small radius, which has Assouad dimension $n(n-1)/2$ [57, Lemma 9.6(iii)]. Hence, monotonicity and bi-Lipschitz invariance of the dimension imply the desired estimate. However, to the best of the author's knowledge, no upper bounds are known for the dimension of those spaces, and the following questions remain open.

Question 5.7 *What are $\dim_{\mathrm{A}} \mathcal{GH}^{\leq n}$ and $\dim_{\mathrm{A}} \mathcal{GH}^{\leq n}_{\leq R}$? Are they finite?*

A connection between Questions 5.7 and 5.6 has already been exploited to deduce Theorem C, namely, infinite Assouad dimension prevents the existence of bi-Lipschitz embeddings into finite-dimensional Hilbert spaces. However, unlike the situation described for asymptotic dimension and coarse embeddings, a positive answer to the second part of Question 5.7 does not imply the existence of a bi-Lipschitz embedding into some $\mathbb{R}^n$. Indeed, having finite Assouad dimension is not a sufficient condition for the existence of a bi-Lipschitz embedding even into some infinite-dimensional Hilbert space [38; 40; 53; 60]. However, some positive results can be proved at the cost of modifying the original metric space. For a metric space (X, d) and $0 < \varepsilon < 1$, the *ε-snowflaking of X* is the metric space (X, d^{ε}), where $d^{\varepsilon}(x, y) = (d(x, y))^{\varepsilon}$.

Theorem 5.8 (Assouad embedding theorem [5]) *If (X, d) is a metric space with finite Assouad dimension, then, for every $0 < \varepsilon < 1$, there exists a bi-Lipschitz embedding of (X, d^{ε}) into $\mathbb{R}^n$ for some n depending only on $\dim_{\mathrm{A}} X$ and ε.*

Let us also mention that, if $\frac{1}{2} < \varepsilon < 1$, the parameter n in the Assouad embedding theorem can be chosen independently from ε ([51]; see also [21] for an explicit map construction). We address the interested reader to the monographs [28; 57] for more details. A positive answer to the second part of Question 5.7 could then motivate the search for computable bi-Lipschitz embeddings of the ε-snowflaking of $\mathcal{GH}^{\leq n}$ or $\mathcal{GH}^{\leq n}_{\leq R}$ into some finite-dimensional Hilbert space, further tightening the connection between computational topology and dimension theory.

References

[1] **H Adams**, **J Bush**, **N Clause**, **F Frick**, **M Gómez**, **M Harrison**, **R A Jeffs**, **E Lagoda**, **S Lim**, **F Mémoli**, **M Moy**, **N Sadovek**, **M Superdock**, **D Vargas**, **Q Wang**, **L Zhou**, *Gromov–Hausdorff distances*, *Borsuk–Ulam theorems*, *and Vietoris–Rips complexes*, preprint (2023) arXiv 2301.00246

[2] **P K Agarwal**, **K Fox**, **A Nath**, **A Sidiropoulos**, **Y Wang**, *Computing the Gromov–Hausdorff distance for metric trees*, ACM Trans. Algorithms 14 (2018) art. id. 24 MR

[3] **S A Antonyan**, *The Gromov–Hausdorff hyperspace of a Euclidean space*, Adv. Math. 363 (2020) art. id. 106977 MR

[4] **S A Antonyan**, *The Gromov–Hausdorff hyperspace of a Euclidean space, II*, Adv. Math. 393 (2021) art. id. 108055 MR

[5] **P Assouad**, *Plongements lipschitziens dans $\mathbb{R}^n$*, Bull. Soc. Math. France 111 (1983) 429–448 MR

[6] **D Bate**, **A L Garcia Pulido**, *Bi-Lipschitz embeddings of the space of unordered m-tuples with a partial transportation metric*, Math. Ann. 390 (2024) 3109–3131 MR

[7] **G Bell**, **A Dranishnikov**, *Asymptotic dimension*, Topology Appl. 155 (2008) 1265–1296 MR

[8] **G S Bloom**, *A counterexample to a theorem of S Piccard*, J. Combinatorial Theory Ser. A 22 (1977) 378–379 MR

[9] **L M Blumenthal**, *Theory and applications of distance geometry*, Clarendon, Oxford (1953) MR

[10] **M G Bouligand**, *Ensembles impropres et nombre dimensionnel*, Bull. Sci. Math. 52 (1928) 320–344, 361–376

[11] **P Bubenik**, **A Wagner**, *Embeddings of persistence diagrams into Hilbert spaces*, J. Appl. Comput. Topol. 4 (2020) 339–351 MR

[12] **D Burago**, **Y Burago**, **S Ivanov**, *A course in metric geometry*, Graduate Studies in Math. 33, Amer. Math. Soc., Providence, RI (2001) MR

[13] **G Carlsson**, *Topology and data*, Bull. Amer. Math. Soc. 46 (2009) 255–308 MR

[14] **G Carlsson**, **F Mémoli**, *Persistent clustering and a theorem of J Kleinberg*, preprint (2008) arXiv 0808.2241

[15] **G Carlsson**, **F Mémoli**, *Characterization, stability and convergence of hierarchical clustering methods*, J. Mach. Learn. Res. 11 (2010) 1425–1470 MR

[16] **M Carrière**, **U Bauer**, *On the metric distortion of embedding persistence diagrams into separable Hilbert spaces*, from "35th International Symposium on Computational Geometry" (G Barequet, Y Wang, editors), Leibniz Int. Proc. Inform. 129, Schloss Dagstuhl, Wadern (2019) art. id. 21 MR

[17] **F Chazal**, **D Cohen-Steiner**, **L J Guibas**, **F Mémoli**, **S Y Oudot**, *Gromov–Hausdorff stable signatures for shapes using persistence*, Computer Graphics Forum 28 (2009) 1393–1403

[18] **F Chazal**, **V de Silva**, **S Oudot**, *Persistence stability for geometric complexes*, Geom. Dedicata 173 (2014) 193–214 MR

[19] **S Chowdhury**, **F Mémoli**, *A functorial Dowker theorem and persistent homology of asymmetric networks*, J. Appl. Comput. Topol. 2 (2018) 115–175 MR

[20] **S Chowdhury**, **F Mémoli**, *Distances and isomorphism between networks: stability and convergence of network invariants*, J. Appl. Comput. Topol. 7 (2023) 243–361 MR

[21] **G David**, **M Snipes**, *A constructive proof of the Assouad embedding theorem with bounds on the dimension*, preprint (2012) arXiv 1211.3223

[22] **A N Dranishnikov**, **G Gong**, **V Lafforgue**, **G Yu**, *Uniform embeddings into Hilbert space and a question of Gromov*, Canad. Math. Bull. 45 (2002) 60–70 MR

[23] **J Dydak**, **Ž Virk**, *Preserving coarse properties*, Rev. Mat. Complut. 29 (2016) 191–206 MR

[24] **J Dydak**, **T Weighill**, *Monotone-light factorizations in coarse geometry*, Topology Appl. 239 (2018) 160–180 MR

[25] **H Edelsbrunner**, **J L Harer**, *Computational topology: an introduction*, Amer. Math. Soc., Providence, RI (2010) MR

[26] **H Edelsbrunner**, **T Heiss**, **V Kurlin**, **P Smith**, **M Wintraecken**, *The density fingerprint of a periodic point set*, from "37th International Symposium on Computational Geometry" (K Buchin, E Colin de Verdière, editors), Leibniz Int. Proc. Inform. 189, Schloss Dagstuhl, Wadern (2021) art. id. 32 MR

[27] **D A Edwards**, *The structure of superspace*, from "Studies in topology" (N M Stavrakas, K R Allen, editors), Academic, New York (1975) 121–133 MR

[28] **J M Fraser**, *Assouad dimension and fractal geometry*, Cambridge Tracts in Mathematics 222, Cambridge Univ. Press (2021) MR

[29] **A Garber**, **Ž Virk**, **N Zava**, *On the metric spaces of lattices and periodic point sets*, preprint (2023) arXiv 2310.07594

[30] **B Giunti**, **J Lazovskis**, **B Rieck**, *DONUT: database of original and non-theoretical uses of topology* (2022) Available at `https://donut.topology.rocks`

[31] **M Gromov**, *Structures métriques pour les variétés riemanniennes*, Textes Mathématiques 1, CEDIC, Paris (1981) MR

[32] **M Gromov**, *Asymptotic invariants of infinite groups*, from "Geometric group theory, II" (G A Niblo, M A Roller, editors), London Math. Soc. Lecture Note Ser. 182, Cambridge Univ. Press (1993) 1–295 MR

[33] **K Hess**, *Topological adventures in neuroscience*, from "Topological data analysis — the Abel Symposium 2018" (N A Baas, G E Carlsson, G Quick, M Szymik, M Thaule, editors), Abel Symp. 15, Springer (2020) 277–305 MR

[34] **N Higson**, **J Roe**, *Amenable group actions and the Novikov conjecture*, J. Reine Angew. Math. 519 (2000) 143–153 MR

[35] **S Iliadis**, **A O Ivanov**, **A A Tuzhilin**, *Local structure of Gromov–Hausdorff space, and isometric embeddings of finite metric spaces into this space*, Topology Appl. 221 (2017) 393–398 MR

[36] **N J Kalton**, **M I Ostrovskii**, *Distances between Banach spaces*, Forum Math. 11 (1999) 17–48 MR

[37] **J Kucab**, **M Zarichnyi**, *On asymptotic power dimension*, Topology Appl. 201 (2016) 124–130 MR

[38] **T J Laakso**, *Plane with A_∞-weighted metric not bi-Lipschitz embeddable to $\mathbb{R}^N$*, Bull. London Math. Soc. 34 (2002) 667–676 MR

[39] **V Lafforgue**, *Un renforcement de la propriété (T)*, Duke Math. J. 143 (2008) 559–602 MR

[40] **U Lang**, **C Plaut**, *Bilipschitz embeddings of metric spaces into space forms*, Geom. Dedicata 87 (2001) 285–307 MR

[41] **S Lim**, **F Mémoli**, **Z Smith**, *The Gromov–Hausdorff distance between spheres*, Geom. Topol. 27 (2023) 3733–3800 MR

[42] **S Majhi**, **J Vitter**, **C Wenk**, *Approximating Gromov–Hausdorff distance in Euclidean space*, Comput. Geom. 116 (2024) art. id. 102034 MR

[43] **F Mémoli**, *On the use of Gromov–Hausdorff distances for shape comparison*, from "Eurographics Symposium on Point-Based Graphics" (M Botsch, R Pajarola, B Chen, M Zwicker, editors), The Eurographics Association (2007) 81–90

[44] **F Mémoli**, *Gromov–Hausdorff distances in Euclidean spaces*, from "2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops" (2008) 1–8

[45] **F Mémoli**, *Some properties of Gromov–Hausdorff distances*, Discrete Comput. Geom. 48 (2012) 416–440 MR

[46] **F Mémoli**, **G Sapiro**, *Comparing point clouds*, from "SGP '04: Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing", ACM, New York (2004) 32–40

[47] **F Mémoli**, **G Sapiro**, *A theoretical and computational framework for isometry invariant recognition of point cloud data*, Found. Comput. Math. 5 (2005) 313–347 MR

[48] **A Mitra**, **Ž Virk**, *The space of persistence diagrams on n points coarsely embeds into Hilbert space*, Proc. Amer. Math. Soc. 149 (2021) 2693–2703 MR

[49] **A Mitra**, **Ž Virk**, *Geometric embeddings of spaces of persistence diagrams with explicit distortions*, preprint (2024) arXiv 2401.05298

[50] **T Miyata**, **Ž Virk**, *Dimension-raising maps in a large scale*, Fund. Math. 223 (2013) 83–97 MR

[51] **A Naor**, **O Neiman**, *Assouad's theorem with dimension independent of the snowflaking*, Rev. Mat. Iberoam. 28 (2012) 1123–1142 MR

[52] **P W Nowak**, **G Yu**, *Large scale geometry*, European Math. Society, Zürich (2012) MR

[53] **P Pansu**, *Métriques de Carnot–Carathéodory et quasiisométries des espaces symétriques de rang un*, Ann. of Math. 129 (1989) 1–60 MR

[54] **P Petersen**, *Riemannian geometry*, Graduate Texts in Mathematics 171, Springer (1998) MR

[55] **N Pritchard**, **T Weighill**, *Coarse embeddability of Wasserstein space and the space of persistence diagrams*, Discrete Comput. Geom. 74 (2025) 358–373 MR

[56] **T M Radul**, **O Shukel**, *Functors of finite degree and asymptotic dimension*, Mat. Stud. 31 (2009) 204–206 MR

[57] **J C Robinson**, *Dimensions, embeddings, and attractors*, Cambridge Tracts in Mathematics 186, Cambridge Univ. Press (2011) MR

[58] **J Roe**, *Lectures on coarse geometry*, University Lecture Series 31, Amer. Math. Soc., Providence, RI (2003) MR

[59] **F Schmiedl**, *Computational aspects of the Gromov–Hausdorff distance and its application in non-rigid shape matching*, Discrete Comput. Geom. 57 (2017) 854–880 MR

[60] **S Semmes**, *On the nonexistence of bi-Lipschitz parameterizations and geometric problems about A_∞-weights*, Rev. Mat. Iberoamericana 12 (1996) 337–410 MR

[61] **A A Tuzhilin**, *Lectures on Hausdorff and Gromov–Hausdorff distance geometry*, preprint (2020) arXiv 2012.00756v1

[62] **A Wagner**, *Nonembeddability of persistence diagrams with $p > 2$ Wasserstein metric*, Proc. Amer. Math. Soc. 149 (2021) 2673–2677 MR

[63] **T Weighill**, **T Yamauchi**, **N Zava**, *Coarse infinite-dimensionality of hyperspaces of finite subsets*, Eur. J. Math. 8 (2022) 335–355 MR

[64] **G Yu**, *The Novikov conjecture for groups with finite asymptotic dimension*, Ann. of Math. 147 (1998) 325–355 MR

[65] **G Yu**, *The coarse Baum–Connes conjecture for spaces which admit a uniform embedding into Hilbert space*, Invent. Math. 139 (2000) 201–240 MR

[66] **N Zava**, *Stability of the q-hyperconvex hull of a quasi-metric space*, preprint (2022) arXiv 2208.10619

Institute of Science and Technology Austria
Klosterneuburg, Austria

nicolo.zava@gmail.com

Received: 30 June 2024 Revised: 4 October 2024

Guidelines for Authors

Submitting a paper to Algebraic & Geometric Topology

Papers must be submitted using the upload page at the AGT website. You will need to choose a suitable editor from the list of editors' interests and to supply MSC codes.

The normal language used by the journal is English. Articles written in other languages are acceptable, provided your chosen editor is comfortable with the language and you supply an additional English version of the abstract.

Preparing your article for Algebraic & Geometric Topology

At the time of submission you need only supply a PDF file. Once accepted for publication, the paper must be supplied in LaTeX, preferably using the journal's class file. More information on preparing articles in LaTeX for publication in AGT is available on the AGT website.

`arXiv` papers

If your paper has previously been deposited on the `arXiv`, we will need its `arXiv` number at acceptance time. This allows us to deposit the DOI of the published version on the paper's `arXiv` page.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Fuzzy or sloppily drawn figures will not be accepted. For labeling figure elements consider the pinlabel LaTeX package, but other methods are fine if the result is editable. If you're not sure whether your figures are acceptable, check with production by sending an email to graphics@msp.org.

Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

ALGEBRAIC & GEOMETRIC TOPOLOGY

Volume 25 Issue 8 (pages 4437–5174) 2025
