Algebraic & Geometric Topology

Volume 23 (2023)

Issue 6 (pages 2415–2924)

msp

ALGEBRAIC & GEOMETRIC TOPOLOGY
msp.org/agt

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

The subscription price for 2023 is US $650/year for the electronic version, and $940/year (+$70, 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, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840. Periodical rate postage paid at Oakland, CA 94615-9651, and additional mailing offices. POSTMASTER: send address changes to Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840.

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

PUBLISHED BY
mathematical sciences publishers
nonprofit scientific publishing
http://msp.org/

Algebraic & Geometric Topology **23**:6 (2023) 2415–2448
DOI: 10.2140/agt.2023.23.2415
Published: 7 September 2023

An algorithmic definition of Gabai width

RICKY LEE

We define the Wirtinger width of a knot and prove that this equals its Gabai width. This leads to an efficient technique for establishing upper bounds on Gabai width. We demonstrate an application of this technique by calculating the Gabai width of 54 756 tabulated prime 4–bridge knots. This is done by writing code for a special category of prime 4–bridge tabulated knots to get upper bounds on Gabai width via the Wirtinger width, then comparing with the theoretical lower bound on Gabai width for prime 4–bridge knots. We also provide results showing the advantages our methods have over the obvious method of obtaining upper bounds on Gabai width via planar projections.

57M25, 57M27

1 Introduction

Gabai width is a geometric invariant of knots that was first used by Gabai in his proof of the property R conjecture [6]. Since then, the notion of Gabai width has played central roles in many important results in 3–manifold topology. Some examples are the resolution of the knot complement problem by Gordon and Luecke [8], the recognition problem for S^3 by Thompson [12], and the leveling of unknotting tunnels by Goda, Scharlemann and Thompson [7]. The importance of Gabai width is largely due to its deep connections with the topology of the knot exterior. For example, Gabai width can often be used to find incompressible surfaces; see Thompson [13] and Wu [15].

The bridge number of a knot is a closely related geometric invariant, defined as the minimal number of local maxima needed to construct an embedding of the knot. Roughly speaking, Gabai width depends on the number of critical points of a projection as well as their relative heights. Like most geometric invariants, both bridge number and Gabai width are notoriously difficult to calculate. However, there has been recent progress on finding algorithmically accessible definitions of bridge number. Blair, Kjuchukova, Velazquez and Villanueva [4] defined the Wirtinger number of a link and showed that it is equal to the bridge number. The Wirtinger number is calculated using a combinatorial coloring algorithm applied to a link diagram. Using ideas inspired by the Wirtinger number, we define the Wirtinger width of a knot and show it is equal to the Gabai width of a knot.

We now briefly summarize our procedure. The formal definition of Wirtinger width is given in Section 3. The Wirtinger width is also computed by coloring knot diagrams. Let D be a knot diagram. View D as the image of the knot $K \subset \mathbb{R}^3$ under the standard projection onto the xy–plane. Our goal, given the diagram D, is to obtain a knot K' in the same ambient isotopy class of K, but embedded so that K' realizes the Gabai width. Our coloring procedure allows us to obtain a knot $\widehat{K}$ from D such that $\widehat{K}$ is ambient isotopic to K, and the relative heights of the critical points of $\widehat{K}$ are controlled by combinatorial data attached to our coloring.

The coloring proceeds as follows. Suppose the knot diagram D has J strands. Then there are $J+1$ stages in the procedure. The knot diagram D begins uncolored at stage 0. To transition from one stage to the next, one can either add a new color to an uncolored strand, or extend an existing color to include another uncolored strand. The procedure terminates once all strands of D are colored.

In general, there are many different ways to color a knot diagram. Not all colorings will give data which corresponds to a thin position embedding of the knot. We assign a natural number to each coloring of a knot diagram, then let the Wirtinger width of the diagram D, denoted by $\mathbb{W}(D)$, be the minimum of these numbers over all colorings of D. Finally, for any ambient isotopy class of knots $\mathcal{K}$, we define the Wirtinger width of $\mathcal{K}$, denoted by $\mathbb{W}(\mathcal{K})$, to be the minimum of $\mathbb{W}(D)$ over all diagrams of knots in the ambient isotopy class $\mathcal{K}$. Letting $w(\mathcal{K})$ be the Gabai width of $\mathcal{K}$, we can state our main theorem as follows:

Theorem 1.1 *If $\mathcal{K}$ is an ambient isotopy class of knots, then $\mathbb{W}(\mathcal{K}) = w(\mathcal{K})$.*

The coloring can be viewed as an attempt to discretize the following process. Suppose now $K \subset \mathbb{R}^3$ is a knot in thin position with respect to the standard height function $h(x,y,z) := z$. Let $h^{-1}(r)$ be a level surface above K. The Gabai width of K is calculated by analyzing the intersection set $K \cap h^{-1}(r)$ as $r \to -\infty$ and $h^{-1}(r)$ sweeps across the maxima and minima of K. The addition of a new color to D represents $h^{-1}(r)$ sweeping across a maximum of K. The occurrence of a *multicolored crossing* (crossings where the over-strand is colored and both under-strands are assigned different colors) represents $h^{-1}(r)$ sweeping across a minimum of K. The order in which new colors and multicolored crossings appear in our coloring procedure dictates the ordering of the maxima and minima of $\widehat{K}$ by height.

There is an easy method of obtaining upper bounds on Gabai width. One can take a knot diagram, perform some planar isotopies if necessary, and use the original Gabai definition of width to obtain an upper bound in the obvious way. While our coloring procedure is less straightforward, it is more computationally accessible and enjoys the following advantage over any potential algorithm written to calculate upper bounds on Gabai width utilizing only planar isotopies on a knot diagram. Let $w_p(D)$ denote the planar width of a knot diagram D. A formal definition of planar width will be given in Section 2, but, roughly speaking, $w_p(D)$ is the upper bound on width one would get by applying the original Gabai definition to calculate width on D, after minimizing over all planar isotopies of D. We will prove:

Theorem 1.2 *For any ambient isotopy class $\mathcal{K}$ of knots and any positive integer n, there exist infinitely many diagrams D of knots in $\mathcal{K}$ such that $\mathbb{W}(D) = w(\mathcal{K})$ but $w_p(D) \geq \mathbb{W}(D) + n$.*

Colloquially, Theorem 1.2 states that, if a planar isotopy algorithm were to be implemented, there would still be an infinite number of cases where Wirtinger width performs better.

Since there are many different ways to completely color a knot diagram, the problem of finding a coloring which corresponds to a calculation of Gabai width is subtle. However, one can modify the Wirtinger number algorithm of Villanueva [14] to exhaust all possible colorings of a given diagram. This is possible because the rules for extending a coloring in the Wirtinger width procedure are the same as those for extending a coloring in the Wirtinger number procedure. We illustrate these ideas in Section 8, where we describe an algorithm that we implemented in Python [10] and used to calculate the Gabai width of 54 756 prime 4–bridge knots.

Our algorithm runs fast in practice, but depends on knowing beforehand that the inputted Gauss codes are of prime knots with bridge number 4 and such that the code from [14] can actually detect bridge number 4. The algorithm takes as input such a Gauss code, and outputs upper bounds on Wirtinger width. By Theorem 1.1, this gives upper bounds on Gabai width. It is known, and explained in Section 8, that the Gabai width of a prime 4–bridge knot must be 32 or 28. Of 86 981 knots tested, our code gave an upper bound of 28 on Wirtinger width for 54 756 knots. Since our upper bound equals the theoretical lower bound on Gabai width for such prime 4–bridge knots, this means we got the exact Gabai width in this case.

Structure of the paper In Section 2, we give preliminary definitions. In Section 3, we give the formal definition of Wirtinger width via a coloring procedure similar to the coloring algorithm of Wirtinger number in [4]. Section 4 contains results showing how Wirtinger number is related to Wirtinger width. In Section 5, we describe a specific coloring sequence, which, when performed on a projection of a knot in thin position, shows that $\mathbb{W}(\mathcal{K}) \leq w(\mathcal{K})$. In Section 6, we show how to use our coloring data to obtain Morse embeddings of knots from a colored knot diagram. This is used to show $\mathbb{W}(\mathcal{K}) \leq w(\mathcal{K})$. In Section 7, we use the results of the previous sections to prove Theorems 1.1 and 1.2. Many technical lemmas and results from Sections 4, 5 and 6 do not apply to diagrams of the unknot, so Section 7 handles this special case separately. In Section 8, we explain how we used Wirtinger width to write an algorithm in Python that obtained our numerical results, and present some open questions.

Acknowledgments The author would like to thank Ryan Blair for introducing this topic and for many helpful discussions, especially about Theorem 1.2. We also thank the referee for a close reading of our initial drafts, leading to many corrections and a great improvement to our exposition.

2 Preliminaries

Let $\mathcal{K}$ denote an ambient isotopy class of knots in $\mathbb{R}^3$. As stated in the introduction, let $h\colon \mathbb{R}^3 \to \mathbb{R}$ defined by $h(x,y,z) := z$ be the standard height function. Let $K \subset \mathbb{R}^3$ denote a knot in the ambient isotopy class $\mathcal{K}$. We will always assume that the embedding of K is such that $h|_K$ is a Morse function.

Let $p\colon \mathbb{R}^3 \to \mathbb{R}^2$ defined by $p(x,y,z) := (y,z)$ be the projection map onto the yz–plane. We will always assume K is embedded so that $p|_K$ is a regular projection. Then $p(K)$ is a finite four-valent graph in the yz–plane. We say that D is a *knot diagram* of K resulting from the projection p if D is the graph $p(K)$ together with labels at each vertex to indicate which edges are over and which are under. By convention, these labels take the form of deleting parts of the under-arc at every crossing. Thus, we can view D as a disjoint union of closed arcs in the plane. Let $\alpha_1, \dots, \alpha_J$ denote the connected components of D. For each α_i, we let s_i denote the union of all edges in

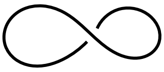


Figure 1: The unique knot diagram containing a strand adjacent to itself.

$p(K)$ whose interiors have nonempty intersection with α_i. We refer to each s_i as a *strand* and let $s(D)$ denote the set of strands of D. We refer to the vertices of $p(K)$ as *crossings* and denote the set of vertices by $v(D)$.

If $s \in s(D)$, then the two endpoints of s will be referred to as the crossings *incident* to s. If s_p and s_q are the under-strands of the same crossing $x \in v(D)$, then we say s_p and s_q are *adjacent at* x, or just *adjacent*. We say the subset $A \subseteq s(D)$ is *connected* if there exists a reordering of the strands $s_{i_1}, s_{i_2}, \dots, s_{i_{|A|}}$ in A such that s_{i_j} is adjacent to $s_{i_{j+1}}$ for all $1 \le j \le |A|$. Note that there is a unique knot diagram up to planar isotopy for which there exists a strand adjacent to itself (see Figure 1). In all cases considered, we assume that adjacent strands are distinct. We say a knot diagram is *trivial* if it is a diagram of the unknot.

For $s \in s(D)$, we define $h(s) := \max_{y \in s} h(y)$ and refer to $h(s)$ as the *height of the strand* s. For a crossing $x \in v(D)$, we refer to $h(x)$ as the *height of the crossing* x.

Note we do not consider the labels of the knot diagram when we calculate the height of a strand. It is therefore possible that a strand and a crossing have equal heights. In fact, if a strand is monotonic with respect to h, then it must have height equal to one of its incident crossings.

By *critical points* of D we will always be referring to images of the critical points of $h|_K$ under the projection p. We say that D is in *general position with respect to h* if all the critical points and crossings of D have distinct heights with respect to h, $h|_K$ is Morse, and $p(K)$ is a regular projection. Observe that, if the knot diagram D is in general position with respect to h, then all the strands must have different heights. See Figure 2.

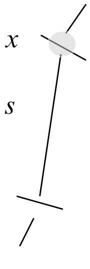


Figure 2: The strand s and the incident crossing x have equal heights ($h(s) = h(x)$).

Now we recall the definition of bridge number. We let $\beta(K)$ denote the number of maxima of $h|_K$. Then the bridge number $\beta(\mathcal{K})$ is defined as $\min_{K'\in\mathcal{K}} \beta(\mathcal{K})$, where the minimum is taken over all Morse embeddings of knots in the equivalence class $\mathcal{K}$.

We now recall the definition of Gabai width. Order the critical values of $h|_K$ by $c_1 > \cdots > c_N$. Let $r_i \in (c_{i+1}, c_i)$ denote arbitrarily chosen regular values of $h|_K$ for $1 \le i \le N-1$. For any $y \in \mathbb{R}$, define $w(y) := |K \cap h^{-1}(y)|$. Define $w(K) := \sum_{i=1}^{N-1} w(r_i)$. The Gabai width of $\mathcal{K}$ is defined as $\min_{K'\in\mathcal{K}} w(K')$, where the minimum is taken over all Morse embeddings of knots in the equivalence class $\mathcal{K}$. If K' is such that $w(K') = w(\mathcal{K})$, then we say K' is in *thin position.*

Finally, we give our formal definition of planar width. For any knot diagram D in the yz–plane that is in general position with respect to h, let $K_D \subset \mathbb{R}^3$ be any knot in the ambient isotopy class $\mathcal{K}$ such that $p(K_D) = D$. We define the *planar width* of D, denoted by $w_p(D)$, as

$$w_p(D) := \min w(K_D),$$

where the minimum is taken over all planar isotopies of D.

3 The coloring rules

In this section, we define Wirtinger width via a combinatorial method for coloring knot diagrams. Let D be a knot diagram. Let $s(D) = \{s_1, \dots, s_J\}$ denote the set of strands of D.

Definition 3.1 A *partial coloring* is a tuple (A, f), where A is a subset of $s(D)$ and $f\colon A \to Z$ is a function with $Z \subset \mathbb{Z}$.

Remark Set $A_0 := \varnothing$, $Z_0 := \varnothing$, and let f_0 be the empty function. Then (A_0, f_0) is a partial coloring. We fix (A_0, f_0) to denote this vacuous partial coloring.

We define two rules for extending partial colorings. Let (A_{t-1}, f_{t-1}) denote a partial coloring, where $t \in \mathbb{N}$ and $f\colon A_{t-1} \to Z_{t-1}$. See Figure 3 for examples of each rule.

Seed addition We say the partial coloring (A_t, f_t) is the result of a *seed addition* to (A_{t-1}, f_{t-1}), denoted by $(A_{t-1}, f_{t-1}) \to (A_t, f_t)$, if:

- $A_{t-1} \subset A_t$ and $A_t \setminus A_{t-1} = \{s_i\}$ for some strand $s_i \in s(D) \setminus A_{t-1}$.
- $Z_t := Z_{t-1} \cup \{t\}$.
- $f_t\colon A_t \to Z_t$ is defined by $f_t|_{A_{t-1}} = f_{t-1}$ and $f_t(s_i) := t$.

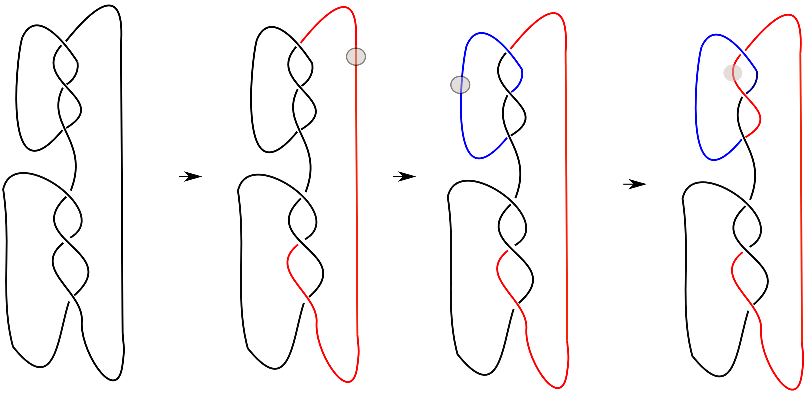


Figure 3: The first two transitions depict seed additions, the first adding the color red the second adding the color blue. The last transition depicts a coloring move extending the color red.

Coloring move We say (A_t, f_t) is the result of a *coloring move* on (A_{t-1}, f_{t-1}), denoted by $(A_{t-1}, f_{t-1}) \to (A_t, f_t)$, if:

- $A_{t-1} \subset A_t$ and $A_t \setminus A_{t-1} = \{s_q\}$ for some strand $s_q \in s(D) \setminus A_{t-1}$.
- s_q is adjacent to s_p at some crossing $x \in v(D)$ and $s_p \in A_{t-1}$.
- The over-strand s_v of x is an element of A_{t-1}.
- $Z_t := Z_{t-1}$.
- $f_t \colon A_t \to Z_t$ is defined by $f_t|_{A_{t-1}} := f_{t-1}$ and $f_t(s_q) := f_{t-1}(s_p)$.

There are two ways we refer to a coloring move. We say that s_q inherits its color from s_p, or that the coloring move was performed over the crossing x.

Remark We can always perform a seed addition to any uncolored strand. This allows us to use seed additions to extend the vacuous partial coloring (A_0, f_0).

Definition 3.2 If $(A_0, f_0) \to \cdots \to (A_t, f_t)$ is a sequence of coloring moves and seed additions on D, then we say the sequence is a *partial coloring sequence*. If we have a partial coloring sequence $(A_0, f_0) \to \cdots \to (A_J, f_J)$ such that $s(D) = A_J$, then we say the sequence is a *completed coloring sequence*. If t is an index of a partial coloring (A_t, f_t) in a specified coloring sequence, then we will refer to t as a *stage*.

Note that we can define a completed coloring sequence for any knot diagram since we can perform a seed addition to any strand.

Definition 3.3 If (A_t, f_t) is the result of a seed addition to (A_{t-1}, f_{t-1}) with $\{s_i\} = A_t \setminus A_{t-1}$, then we call s_i a *seed strand.*

Definition 3.4 Let $(A_0, f_0) \to \cdots \to (A_J, f_J)$ be a completed coloring sequence on the knot diagram D. Let $x \in v(D)$. Denote the over-strand of x by s_v and the under-strands of x by s_p and s_q. If there exists a stage t such that $s_p, s_q, s_v \in A_t$ and $f_t(s_p) \neq f_t(s_q)$, then we say x is a *multicolored crossing*. The smallest stage at which all previously stated conditions are satisfied will be referred to as the stage at which the crossing x becomes multicolored.

Completed coloring sequences allow us to extract geometric information from knot diagrams. To do this, we first record the order in which strands become colored, and crossings become multicolored.

Definition 3.5 Let $(A_0, f_0) \to \cdots \to (A_J, f_J)$ be a completed coloring sequence with multicolored crossing set $\mathcal{C}$. Let $\mathcal{C}_t$ denote the set of crossings that become multicolored at stage t. A Δ*–ordering* is an enumeration of the elements in $s(D) \cup \mathcal{C}$, $\Delta := (d_i)_{i=1}^{|s(D)|+|\mathcal{C}|}$, satisfying the following conditions:

(1) For all $0 \le t < u \le J$, all elements colored (or multicolored) at stage t are listed before any element colored (or multicolored) at stage u.

(2) For each stage $0 \le t \le J$, the element in $A_t \setminus A_{t-1}$ is listed, followed by all elements in $\mathcal{C}_t$ (if $\mathcal{C}_t \neq \varnothing$). That is, if at stage t a strand receives its color and a subset of crossings become multicolored, then we list the strand first, followed by all crossings that become multicolored at stage t.

Later, we use Δ–orderings to reconstruct an embedding of our knot in $\mathbb{R}^3$ from a colored knot diagram. Each seed strand will induce a single maximum and each multicolored crossing will induce a single minimum in our reconstructed embedding. The ordering of the critical points, by decreasing height with respect to h, is reflected in our Δ–ordering. We now show how to elevate this relationship into a calculation of Gabai width.

Definition 3.6 Let $(A_0, f_0) \to \cdots \to (A_J, f_J)$ be a completed coloring sequence. Let $\mathcal{S} \subseteq s(D)$, $\mathcal{C} \subseteq v(D)$ and Δ be the seed strands, multicolored crossings and Δ–ordering, respectively, of our completed coloring sequence. Let $\Delta' := (d_{i_j})_{j=1}^{|\mathcal{S}|+|\mathcal{C}|}$ denote the subsequence of Δ formed by restricting our Δ–ordering to the set $\mathcal{S} \cup \mathcal{C}$. We define the *attached sequence* $(a_i)_{i=0}^{|\Delta'|}$ to be the sequence created via the following rule:

- Set $a_0 := 0$.
- If $d_{i_j} \in \Delta'$ is a seed strand, then set $a_j := a_{j-1} + 2$.
- If $d_{i_j} \in \Delta'$ is a multicolored crossing, then set $a_j := a_{j-1} - 2$.

If the first t stages of the completed coloring involve $|S|$ total seed additions, and $|C|$ total crossings become multicolored by stage t, then we say the partial coloring sequence $(A_0, f_0) \to \cdots \to (A_t, f_t)$ induces the first $|S| + |C|$ terms of the attached sequence $(a_i)_{i=0}^{|\Delta'|}$.

Definition 3.7 Define $\mathbb{W}(D) := \min \sum_{i=0}^{N} a_i$, where the minimum is taken over all possible completed coloring sequences defined for the diagram D. Let $\mathbb{W}(\mathcal{K}) := \min \mathbb{W}(D)$, where the minimum is taken over all possible knot diagrams of knots in the isotopy class $\mathcal{K}$. We define $\mathbb{W}(\mathcal{K})$ to be the *Wirtinger width* of $\mathcal{K}$.

Remark The Δ–ordering resulting from a completed coloring sequence need not be unique. For example, if at some stage in a coloring sequence the strand s becomes colored and the crossings x_i and x_j both become multicolored, then both

$$\Delta_1 := \{\dots, s, x_i, x_j, \dots\} \quad \text{and} \quad \Delta_2 := \{\dots, s, x_j, x_i, \dots\}$$

are Δ–orderings resulting from the same coloring. In the ultimate calculation of $\mathbb{W}(D)$, such nuances do not matter as both Δ_1 and Δ_2 would induce the same attached sequence $(a_i)_{i=0}^{\Delta'}$. This is because, in each possible Δ–ordering, the crossings that become multicolored at the same stage must always be listed consecutively by the second condition in Definition 3.5.

In order to prove statements about Wirtinger width, one often needs to specify a Δ–ordering to work with. The following definition allows us to do this:

Definition 3.8 Let $\Delta = \{d_i\}_{i=1}^{|s(D)|+|\mathcal{C}|}$ be a Δ–ordering resulting from a completed coloring sequence on the knot diagram D. We define the *height function* $h_o\colon \Delta \to \mathbb{Z}$ associated to Δ by $h_o(d_t) := -t$.

The function h_o retrieves the negative of the position of d_t in the Δ–ordering. We introduce a negative sign to allow us to focus on maxima instead of minima in later constructions. The main use of h_o in later proofs will be to compare the relative positions of strands and multicolored crossings in a Δ–ordering. If d_i and d_j represent strands of a knot diagram, then the inequality $h_o(d_i) > h_o(d_j)$ should be interpreted as "d_i is colored before d_j".

Remark The name Wirtinger width comes from the fact, proved in [4], that the minimum number of seed additions necessary to obtain a completed coloring sequence on the knot diagram D is equal to the minimum number of meridional generators needed in a Wirtinger presentation of the knot group from a diagram.

4 Connections to the Wirtinger number

In this section, we prove some preliminary results that will be needed for our proof of Theorem 1.1. These results are the Wirtinger width analogues of [4, Proposition 2.2]. Let $s(D) = \{s_1, \dots, s_J\}$ denote the strands of the knot diagram D.

Definition 4.1 Let $A := \{s_1, \dots, s_n\}$ be a connected subset of $s(D)$, ordered by adjacency. Let $g\colon A \to \mathbb{Z}$. We say g has a *local maximum* at s_j if $n > 1$ and

$$g(s_j) > \begin{cases} \max\{g(s_{j-1}), g(s_{j+1})\} & \text{if } 1 < j < n, \\ g(s_2) & \text{if } j = 1, \\ g(s_{n-1}) & \text{if } j = n. \end{cases}$$

If $n = 1$, then g has a maximum at s_1.

The following is an equivalent reformulation of being *k–meridionally colorable*, and the main theorem, from [4]:

Definition 4.2 D is *k–meridionally colorable* if there exists a completed coloring sequence $(A_0, f_0) \to \cdots \to (A_J, f_J)$ containing only k seed additions.

Theorem 4.3 *Let $\mu(\mathcal{K})$ denote the minimal k such that there exists a knot diagram D of a knot in the ambient isotopy class $\mathcal{K}$ which is k–meridionally colorable. Recall $\beta(\mathcal{K})$ denotes the bridge number of $\mathcal{K}$. Then $\mu(\mathcal{K}) = \beta(\mathcal{K})$.*

Proposition 4.4 *Let $(A_0, f_0) \to \cdots \to (A_J, f_J)$ be a completed coloring sequence on a knot diagram D. Let $\Delta := (d_i)_{i=1}^{M}$ be a Δ–ordering on $s(D) \cup \mathcal{C}$ induced by the completed coloring sequence on D. Let $h_o\colon \Delta \to \mathbb{Z}$ be the height function on Δ defined by $h_o(d_t) := -t$. Let $x \in v(D)$ be a crossing with under-strands s_p and s_q and over-strand s_v. Let s_p and s_r be the strands adjacent to s_q.*

1. *For all $u \in \{0, 1, \dots, J\}$ and $y \in f_u(A_u)$, $f_u^{-1}(y)$ is connected.*
2. *For all $y \in f_J(A_J)$, h_o has a unique local maximum on $f_J^{-1}(y)$ when the set $f_J^{-1}(y)$ is ordered sequentially by adjacency. The local maximum is the unique seed strand contained in $f_J^{-1}(y)$.*

(3) *Suppose now D is a nontrivial knot diagram and $f_J(s_p) = f_J(s_q) = f_J(s_r) = y$. If k is such that $\{s_q\} = A_k \setminus A_{k-1}$, then we cannot have $\{s_p, s_r\} \subset A_{k-1}$.*

(4) *If D is a nontrivial knot diagram and $x \notin \mathcal{C}$, then $h_o(s_v) > \min\{h_o(s_p), h_o(s_q)\}$.*

(5) *If D is any knot diagram and $x \in \mathcal{C}$, then $h_o(x) < \min\{h_o(s_p), h_o(s_q), h_o(s_v)\}$.*

Proof (1) This result is a reformulation of [4, Proposition 2.2(1)] in our notation. We induct on the stage u. Recall $A_0 = \varnothing$ and f_0 is the empty function, so the claim is vacuously true for f_0.

Suppose for induction that $f_u^{-1}(y)$ is connected for all $u < t$ and $y \in f_u(A_u)$. We will show that $f_t^{-1}(y)$ is connected for all $y \in f_t(A_t)$. Say $\{s_i\} = A_t \setminus A_{t-1}$ and $f_t(s_i) = r$. We consider two cases.

First suppose (A_t, f_t) is the result of a seed addition to (A_{t-1}, f_{t-1}). By our definition of seed addition, $f_t^{-1}(r) = \{s_i\}$ and $f_t^{-1}(y) = f_{t-1}^{-1}(y)$ for all $y \in f_t(A_t) \setminus \{r\}$. Since $f_t^{-1}(r)$ is a singleton, it is connected. By our induction hypothesis, $f_{t-1}^{-1}(y)$ is connected for all $y \neq r$.

Now suppose (A_t, f_t) is the result of a coloring move on (A_{t-1}, f_{t-1}). By our definition of coloring move, $f_t^{-1}(r) = f_{t-1}^{-1}(r) \cup \{s_i\}$ and s_i must be adjacent to a strand in $f_{t-1}^{-1}(r)$. Our induction hypothesis implies $f_{t-1}^{-1}(r)$ is connected. Therefore, $f_t^{-1}(r)$ must also be connected. For all $y \in f_t(A_t) \setminus \{r\}$, we have $f_t^{-1}(y) = f_{t-1}^{-1}(y)$. Therefore, our induction hypothesis also implies $f_t^{-1}(y)$ is connected for all $y \in f_t(A_t)$. This completes the induction.

(2) This result is a reformulation of [4, Proposition 2.2(2)] in our notation. The assertion comes from the following observation. For every color $y \in f_J(A_J)$ used in the coloring of D, the set $f_J^{-1}(y)$ contains a single seed strand s_e, which is the first strand assigned the color y. All other strands $s_j \in f_J^{-1}(y)$ assigned the color y occur after s_e in the sequence Δ.

We induct on the stage u. By definition, A_1 is a singleton and $f_1 \colon A_1 \to \{1\}$. Thus h_o trivially attains a unique local maximum on the set $A_1 = f^{-1}(1)$, which contains only a seed strand.

Suppose for induction that, for all $u < t$ and all $y \in f_u(A_u)$, the seed strand of $f_u^{-1}(y)$ is the unique local maximum of h_o on the set $f_u^{-1}(y)$ when ordered sequentially by adjacency. We claim the same holds for f_t. Say $\{s_i\} = A_t \setminus A_{t-1}$ and $f_t(s_i) = r$. We consider two cases.

First suppose (A_t, f_t) is the result of a seed addition to (A_{t-1}, f_{t-1}). By our definition of seed addition, $f_t^{-1}(r) = \{s_i\}$, so h_o trivially attains a unique local maximum on this set. For all $y \in f_t(A_t) \setminus \{r\}$, we have $f_t^{-1}(y) = f_{t-1}^{-1}(y)$, so our claim follows from the induction hypothesis.

Now suppose (A_t, f_t) is the result of a coloring move on (A_{t-1}, f_{t-1}). Then there exists a strand $s_l \in A_{t-1}$ such that $f_t(s_l) = r$ and s_l is adjacent to s_i. By our definition of coloring move and h_o, since s_l is adjacent to s_i but colored before s_i, $h_o(s_i) < h_o(s_l)$. Thus s_i is not a local maximum in $f_t^{-1}(r)$. Since $f_t^{-1}(r) = f_{t-1}^{-1}(r) \cup \{s_i\}$ and $f_t^{-1}(y) = f_{t-1}^{-1}(y)$ for all $y \in f_t(A_t) \setminus \{r\}$, our claim follows from the induction hypothesis. This completes the induction.

(3) Colloquially, our assertion is that, if D is not a diagram of the unknot, then at no stage in the coloring process can we have an uncolored strand s_q adjacent to two strands s_p and s_r that were assigned the same color. Suppose for contradiction that $s_p, s_r \in A_{k-1}$. By assumption, $s_q \notin A_{k-1}$. By part (1) of this proposition, $f_{k-1}^{-1}(y)$ is connected. Since $f_J(s_p) = f_J(s_r)$, we have $\{s_p, s_r\} \subset f_{k-1}^{-1}(y)$. Since D is a knot diagram, the connectivity of $f_{k-1}^{-1}(y)$ and the inclusion $\{s_p, s_r\} \subset f_{k-1}^{-1}(y)$ implies $s(D) \setminus \{s_q\} = f_{k-1}^{-1}(y)$. Thus $s(D) = f_J^{-1}(y)$ and so our completed coloring sequence has a single seed strand. By Theorem 4.3, this implies D is a diagram of a knot with bridge number 1. But the unknot is the only knot with bridge number 1. This contradicts the nontriviality of D.

(4) Colloquially, the claim states that, if D is nontrivial and x is not multicolored, then the over-strand of x is colored before one of its under-strands. Hence, the x comes earlier in the sequence Δ than at least one of s_p or s_q.

Assume for contradiction that $h_o(s_v) < \min\{h_o(s_p), h_o(s_q)\}$. That is, the over-strand of x is colored after both under-strands s_p and s_q have been colored. Since D is a nontrivial knot diagram, the adjacent strands s_p and s_q are distinct. Without loss of generality, say s_p is colored before s_q. Let k be the stage that s_q receives its color, so $\{s_q\} = A_k \setminus A_{k-1}$.

Since $h_o(s_v) < \min\{h_o(s_p), h_o(s_q)\}$ and k is the stage at which s_q receives its color, s_v has not been colored by stage k. Therefore, no coloring move was performed over x in the completed coloring sequence.

Let s_p and s_r be the strands adjacent to s_q. By assumption, $x \notin \mathcal{C}$. That is, x is not multicolored, so s_p and s_q have been assigned the same color. Since s_p and s_q have been assigned the same color and are adjacent at x, but no coloring move was performed at x, s_q must have inherited its color from s_r via a coloring move. But s_p was colored before s_q. Therefore, $\{s_p, s_r\} \subset A_{k-1}$.

Since s_p and s_q were assigned the same color and s_q inherited its color from s_r, we have $f_J(s_p) = f_J(s_q) = f_J(s_r)$. But we have also showed $\{s_q\} = A_k \setminus A_{k-1}$ and $\{s_p, s_r\} \subset A_{k-1}$. Since D is a nontrivial knot diagram, we get the desired contradiction by part (3) of this proposition.

(5) The inequality is a reformulation of condition (2) in Definition 3.5 in terms of the height function h_o. In words, it states that, in the definition of sequence Δ, at each stage, the strand is listed before any crossings that become multicolored, as such a crossing does not become multicolored at stage t unless all of s_p, s_q, and s_v are in A_t. □

5 Coloring by height

In this section we describe a specific procedure for coloring diagrams of knots in thin position. It will be used to establish the inequality $\mathbb{W}(\mathcal{K}) \leq w(\mathcal{K})$. Our goal is to obtain a coloring sequence that induces a Δ–ordering which respects the ordering of the critical points of $h|_D$ by height.

For the rest of this section, let K be an embedding of the knot $\mathcal{K}$ in $\mathbb{R}^3$ that is in thin position with respect to h. Furthermore, let K be such that the knot diagram $D \subset \{yz\text{–plane}\}$, resulting from the projection p into the yz–plane is in general position with respect to h. Let $c_1 > c_2 > \cdots > c_N$ be the critical values of $h|_K$ ordered by decreasing height with respect to h. We also assume that $\mathcal{K}$ is not the ambient isotopy class of the unknot, so that D is a nontrivial diagram.

Definition 5.1 Let L be any knot diagram embedded in the yz–plane that is in general position with respect to h. Let $x \in v(L)$. Denote the under-strands of x by s_f and s_r. If $h|_{s_f}$ has a local maximum at x, then we say s_f is *the falling strand of* x. If $h|_{s_r}$ has a local minimum at x, then we say s_r is *the rising strand of* x.

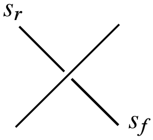


Figure 4: The rising strand and falling strand of the pictured crossing are denoted by s_f and s_r.

Recall that, for a strand s, we have defined the height of the strand to be $h(s) = \max_{y \in s} h(y)$. The assumption that D is in general position with respect to h means that all strands have distinct heights. This enables the following definition:

Definition 5.2 We say that we *color D by height* if we obtain a completed coloring sequence $(A_0, f_0) \to \cdots \to (A_J, f_J)$ by the following procedure:

Step 1 Write $s(D) = \{s_1, \dots, s_{|s(D)|}\}$, where $h(s_1) > \cdots > h(s_{|s(D)|})$.

Step 2 Let (A_1, f_1) be the result of a seed addition to (A_0, f_0) such that $\{s_1\} = A_1 \backslash A_0$.

Step 3 Suppose we have a partial coloring sequence $(A_0, f_0) \to \cdots \to (A_{n-1}, f_{n-1})$ defined, where $A_{n-1} = \{s_1, \dots, s_{n-1}\}$. Let x_i and x_j be the crossings incident to s_n. Say $h(x_i) < h(x_j)$. We consider two cases:

Case 1 Suppose $h|_{s_n}$ is maximized in $\operatorname{int}(s_n)$. Then we let (A_n, f_n) be the result of a seed addition to (A_{n-1}, f_{n-1}) such that $\{s_n\} = A_n \setminus A_{n-1}$.

Case 2 Suppose $h|_{s_n}$ is maximized in ∂s_n (so s_n is the falling strand of x_j). Then we let (A_n, f_n) be the result of a coloring move over x_j.

Remark When a coloring move is performed over a crossing x during the color by height process, colors must extend from the rising strand of x to the falling strand of x. Recall that, since D is assumed to be a nontrivial knot diagram, adjacent strands are distinct, so the rising and falling strands of x will always be distinct.

We first verify that knot diagrams in general position can always be colored by height.

Proposition 5.3 *If D is a knot diagram in general position with respect to h, then D can be colored by height.*

Proof We verify that each step of the color by height procedure can always be performed on D. Since D is in general position with respect to h, all strands have distinct heights. Thus, they can be ordered by decreasing height. By definition, we can always perform seed addition moves at any stage. What remains to be verified is that we can perform the coloring move stated in Step 3, Case 2 of Definition 5.2.

Let (A_n, f_n), s_n, x_i and x_j be as stated in Step 3, Case 2 of Definition 5.2. Let s_v and s_r denote the over-strand and rising strand of the crossing x_j, respectively. Since $h|_{s_n}$ is maximized in ∂s_n, we have $h(s_n) = h(x_j)$. By assumption, D is in general position with respect to h. Therefore, $h(s_n) = h(x_j) < \min\{h(s_v), h(s_r)\}$. Since the strands

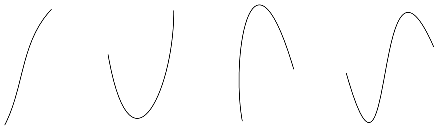


Figure 5: It will be shown that, since K is in thin position and D is in general position with respect to h, the strands of D can have at most two critical points. Moreover, if a strand has two critical points, then one must be a maximum and the other must be a minimum. This figure illustrates the stated possibilities.

were ordered by decreasing height, this implies $\{s_v, s_r\} \subset A_{n-1}$, so we can perform the desired coloring move. □

Our goal now is to show that, when we color D by height, we will get $\mathbb{W}(D) \leq w(K)$. The idea behind the upcoming technical results is that, since K is in thin position and the resulting diagram D is in general position with respect to h, the strands of D can be classified by how many critical points they contain. Figure 5 illustrates the classification, which will be used to show that the number of seed additions that occur when we color by height is equal to the number of maxima in K. Moreover, the number of multicolored crossings that occur is equal to the number of minima in K.

Lemma 5.4 *If $s \in s(D)$ and $r \in \mathbb{R}$ is a regular value of $h|_D$, then $|s \cap h^{-1}(r)| \leq 2$.*

Proof Suppose for contradiction we have a strand $s \in s(D)$ and a regular value $r \in \mathbb{R}$ of $h|_D$ such that $|s \cap h^{-1}(r)| \geq 3$. (See eg Figure 6.)

Recall that $c_1 > c_2 > \cdots > c_N$ are the critical values of $h|_K$, and say $r \in (c_{j+1}, c_j)$. Choose regular values $r_i \in (c_{i+1}, c_i)$ for $1 \leq i \leq N-1$ with $r_j = r$. Recall K is in thin position, so $w(K) = w(\mathcal{K})$. To obtain our desired contradiction, we will exhibit an isotopy on K to produce another embedding of $\mathcal{K}$ with strictly lower width.

Take three points a, b and c in $s \cap h^{-1}(r)$ that are consecutive in the strand s with respect to some orientation on s. Let $s_{a,b}$ denote the subarc of s in the yz–plane with boundary

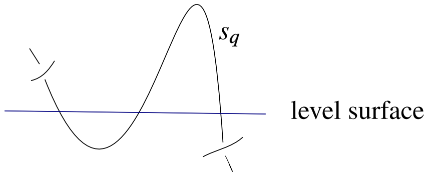


Figure 6: An example of a violation of Lemma 5.4.

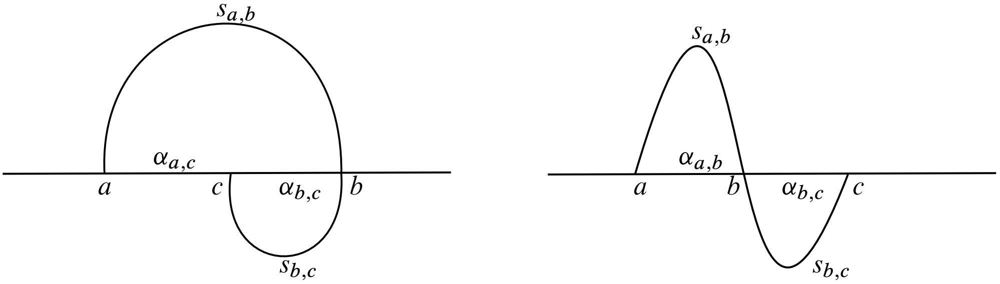


Figure 7: The setup for Cases 1 and 2 in the proof of Lemma 5.4 are on the left and right, respectively.

set $\{a, b\}$. Define $s_{a,c}$ and $s_{b,c}$ similarly. Let $\alpha_{a,b}$ be the arc in yz–plane $\cap\, h^{-1}(r)$ with boundary set $\{a, b\}$. Define $\alpha_{a,c}$ and $\alpha_{b,c}$ similarly.

Before describing the isotopy, we must consider cases based on the order of the points $\{a, b, c\}$ in yz–plane $\cap\, h^{-1}(r)$. The ordering is by the y–coordinates of the points. Up to symmetry, there are two cases to consider, as depicted in Figure 7.

Case 1 Suppose $a < c < b$. Let $D_{a,c}$ be the disk cobounded by $s_{a,c}$ and $\alpha_{a,c}$ in the yz–plane. We now define the steps of the isotopy. Let $\hat{s}_{a,c}$ be the arc component of $K \cap p^{-1}(s_{a,c})$.

- **Step 1** Perform an isotopy on K that fixes the y– and z–coordinates of all points on K, and arranges that $\hat{s}_{a,c} = p(\hat{s}_{a,c}) = s_{a,c}$ and all points in $K \setminus \hat{s}_{a,c}$ have negative x–coordinate. Note now $\hat{s}_{a,c}$ cobounds the disk $D_{a,c}$ with $\alpha_{a,c}$ in the yz–plane.
- **Step 2** Perform an isotopy on $\hat{s}_{a,c}$ that fixes a and c and pushes $\hat{s}_{a,c}$ across $D_{a,c}$ onto $\alpha_{a,c}$.
- **Step 3** After performing the isotopy, perturb the portion of K in a neighborhood of $\alpha_{a,c}$ so that $h|_K$ is Morse and has two fewer critical points than it had originally.

Let $s'_{a,c}$ and K' denote the image of $s_{a,c}$ and K, respectively, after the isotopy and perturbation. Let D' denote the diagram of K' given by projection into the yz–plane. Let $s'_{a,c}$ denote the image of $\hat{s}_{a,c}$ in D'.

Case 2 Suppose $a < b < c$. Then $s_{a,b}$ and $s_{b,c}$ cobound disks with $\alpha_{a,b}$ and $\alpha_{b,c}$, respectively, in the yz–plane. We obtain $s'_{a,c}$, K' and D' from a procedure analogous to that in Case 1. The only modification is that, in Step 2, we push across two disks instead of one.

We now claim $w(K') < w(K)$. By construction,

$$|s'_{a,c} \cap h^{-1}(r_j)| < |\hat{s}_{a,c} \cap h^{-1}(r_j)|.$$

Our procedure fixed the height of all points in K outside of a small neighborhood of $\hat{s}_{a,c}$ and did not introduce any new critical points. Therefore,

$$\sum_{i=1}^{N-1} |K' \cap h^{-1}(r_i)| < \sum_{i=1}^{N-1} |K \cap h^{-1}(r_i)| = w(\mathcal{K}).$$

The above inequality shows $w(K') < w(\mathcal{K})$. Since K was assumed to be in thin position, we get our desired contradiction. □

Proposition 5.5 *Let $(A_0, f_0) \to \cdots \to (A_J, f_J)$ be a completed coloring sequence obtained from coloring D by height.*

(1) *A seed addition is performed on the strand s if and only if $h|_s$ is maximized in the interior of s.*

(2) *Let x_i be a crossing with falling strand s_q, where x_i and x_j are the crossings incident to s_q. Then x_i is multicolored if and only if $h|_{s_q}$ is minimized in the interior of s_q and $h(x_i) < h(x_j)$.*

Proof (1) By Definition 5.2, a seed addition is performed on a strand if and only if that strand has a maximum in its interior.

(2) Let t be the stage at which s_q receives its color, so $\{s_q\} = A_t \setminus A_{t-1}$.

Suppose x_i is a multicolored crossing. Then $h|_{s_q}$ must be minimized in the interior of s_q, for otherwise, as s_q is the falling strand of x_i, it would be minimized at x_j. But, if s_q is the falling strand of x_i and $h|_{s_q}$ is minimized at x_j, then $h|_{s_q}$ would also have to be maximized at x_i, for otherwise we could find a regular value r such that $|s_q \cap h^{-1}(r)| \geq 3$, which would violate Lemma 5.4. In other words, s_q would be monotonic with respect to h. But this would mean (A_t, f_t) was the result of a coloring move on (A_{t-1}, f_{t-1}) over x_i, which is impossible because x_i is assumed to be multicolored.

In addition, if $h(x_i) > h(x_j)$, then s_q would have been colored via a seed addition, because the assumption that x_i is multicolored forbids any coloring move from being performed over x_i. The inequality $h(x_i) > h(x_j)$ would mean no coloring move was performed over x_j because we are coloring by height. By part (1) of this proposition,

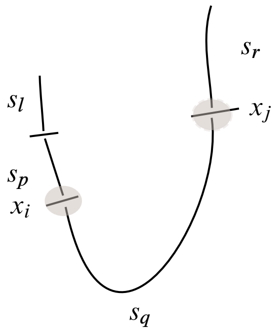


Figure 8: The setup for the proof of Proposition 5.5(2), where we want to show x_i is multicolored. It is assumed that s_q has a minimum in its interior and x_i is the lower incident crossing of s_q. The strands adjacent to s_q are s_p and s_r. The strands adjacent to s_p are s_l and s_q.

$h|_{s_q}$ would be maximized in the interior of s_q. But it was shown that $h|_{s_q}$ is also minimized in the interior of s_q. Since s_q is the falling strand of x_i and contains both a maximum and a minimum of $h|_{s_q}$ in its interior, the inequality $h(x_i) > h(x_j)$ would imply the existence of a regular value r such that $|s_q \cap h^{-1}(r)| \geq 3$, which would violate Lemma 5.4. We conclude $h(x_i) < h(x_j)$.

Conversely, suppose that $h|_{s_q}$ is minimized in the interior of s_q and $h(x_i) < h(x_j)$. We will show x_i is a multicolored crossing. Let s_p and s_r be the strands adjacent to s_q at the crossings x_i and x_j, respectively. Let s_l be the other strand adjacent to s_p. See Figure 8 for a diagram of this setup. Let u be the stage at which s_p is colored, so that $\{s_p\} = A_u \setminus A_{u-1}$.

Suppose for contradiction that x_i is not multicolored. Observe that, since $h(x_i) < h(x_j)$ and s_q is the falling strand of x_i, no coloring move could have been performed at x_i when we color D by height. We consider two cases.

Recall $\{s_q\} = A_t \setminus A_{t-1}$. First suppose (A_t, f_t) was the result of a seed addition to (A_{t-1}, f_{t-1}). By assumption, $x_i \notin \mathcal{C}$, so $f_J(s_p) = f_J(s_q)$. Thus s_p cannot also be a seed strand. Hence, s_p must have inherited its color from s_l because no coloring move could have been performed over x_i when we colored D by height. But this means $f_J(s_l) = f_J(s_p) = f_J(s_q)$ and $\{s_l, s_q\} \subseteq A_{u-1}$ must hold. This contradicts Proposition 4.4(3).

Now suppose (A_t, f_t) was the result of a coloring move on (A_{t-1}, f_{t-1}). No coloring move could have been performed over x_i when we colored D by height, so s_q must have inherited its color from s_r. But $x_i \notin \mathcal{C}$. Therefore, $f_J(s_p) = f_J(s_q) = f_J(s_r)$. If $u < t$ (that is, if s_p was colored before s_q), then $\{s_p, s_r\} \subset A_{t-1}$ and we have a contradiction to Proposition 4.4(3).

Now say $t < u$ (that is, s_q was colored before s_p). We still have $f_J(s_p) = f_J(s_q)$, so s_p cannot be a seed strand under the current assumptions. Thus s_p must have inherited its color from s_l since no coloring move could have been performed over x_i when we colored D by height. This forces $f_J(s_l) = f_J(s_p) = f_J(s_q)$ and $\{s_l, s_q\} \subset A_{u-1}$, contradicting Proposition 4.4(3).

We conclude x_i is multicolored. □

Recall that K is in thin position and D, which is the diagram of K obtained by projection into the yz–plane, has N critical points.

Corollary 5.6 *If $\mathcal{S}$ and $\mathcal{C}$ are the sets of seed strands and multicolored crossings resulting from a coloring of D by height, then $|\mathcal{S}| + |\mathcal{C}| = N$.*

Proof Proposition 5.5 implies that $\mathcal{S}$ and $\mathcal{C}$ are in bijective correspondence with the set of local maxima and the set of local minima of $h|_K$, respectively. This follows because K is assumed to be such that D is in general position with respect to h. □

Theorem 5.7 *If $\mathcal{K}$ is an ambient isotopy class of knots that does not contain the unknot, then $\mathbb{W}(\mathcal{K}) \le w(\mathcal{K})$.*

Proof Since D is a diagram of the knot K in $\mathcal{K}$, it suffices to show $\mathbb{W}(D) \le w(K)$. Let $(A_0, f_0) \to \cdots \to (A_J, f_J)$ be a completed coloring sequence on D obtained from coloring D by height. Let $(a_i)_{i=0}^N$ be the attached sequence of the coloring. We claim $\sum_{i=0}^N a_i \le w(K)$.

Note that Corollary 5.6 verifies that the number of critical points of K is equal to N, where the attached sequence $(a_i)_{i=0}^N$ resulting from coloring D by height contains $N+1$ terms. Let $r_n \in (c_{n+1}, c_n)$ denote a regular value of $h|_D$. It suffices to show $a_n \le w(r_n)$ for $1 \le n \le N$. Recall that we always have $a_0 = 0$ by definition. Fix one such n.

First we fix some notation. For all critical values c_i, let γ_i be the unique strand at which $h^{-1}(c_i)$ fails to intersect D transversely. Set $w(r_0) := 0$ for notational convenience. Write

$$a_n = \sum_{i=1}^{n} a_i - a_{i-1}, \quad w(r_n) = \sum_{i=1}^{n} w(r_i) - w(r_{i-1}),$$

so that our goal is to show

$$\sum_{i=1}^{n} a_i - a_{i-1} \le \sum_{i=1}^{n} w(r_i) - w(r_{i-1}). \tag{1}$$

Observe that $a_i - a_{i-1} \in \{-2, 2\}$ and $w(r_i) - w(r_{i-1}) \in \{-2, 2\}$ for each i. Thus, it suffices to show that the number of positive terms in the left sum is bounded above by the number of positive terms in the right sum in equation (1).

Let t be the stage such that $s \in A_t$ if and only if $r_n < h(s)$. That is, a strand is colored by stage t if and only if its height is greater than r_n. We can acquire such a t because our completed coloring sequence was obtained from coloring D by height. To count the number of positive terms in the sums for equation (1), define

$$S_n := \{i \mid a_i - a_{i-1} = 2,\ 1 \le i \le n\}, \quad M_n := \{i \mid w(r_i) - w(r_{i-1}) = 2,\ 1 \le i \le n\}.$$

The value $|M_n|$ is the number of maxima above r_n. The value $|S_n|$ is related to the number of seed additions that have been performed by stage t. When coloring by height, it is possible that the lower incident crossing corresponding to a minimum below r_n becomes multicolored by stage t. Therefore, we cannot guarantee the equality of $|S_n|$ and $|M_n|$. However, we have the following claim, which suffices for our desired result:

Claim $$|S_n| \le |M_n|.$$

Proof By Proposition 5.5(1), each strand containing a maximum with height above r_n must have been colored via a seed addition by stage t. Since D is in general position with respect to h, for all c_j above r_n corresponding to a minimum of a strand γ_j, the over- and under-strands of the lower incident crossing of γ_j have height greater then c_j, and hence r_n. Therefore, by Proposition 5.5(2), each minimum above r_n corresponds to a crossing that becomes multicolored by stage t. Since there are n critical points above r_n, we conclude that $(A_0, f_0) \to \cdots \to (A_t, f_t)$ induces at least the first $n+1$ terms $(a_i)_{i=0}^{n}$ in the attached sequence $(a_i)_{i=0}^{N}$.

By Definition 5.2, of coloring by height, $|M_n|$ is the number of seed additions in the partial coloring sequence $(A_0, f_0) \to \cdots \to (A_t, f_t)$. Since $(A_0, f_0) \to \cdots \to (A_t, f_t)$ induces at least the first $n+1$ terms $(a_i)_{i=0}^{n}$ in the attached sequence $(a_i)_{i=0}^{N}$, $|S_n|$ is bounded above by the number of seed additions in $(A_0, f_0) \to \cdots \to (A_t, f_t)$. Therefore, $|S_n| \le |M_n|$, as desired. □

This claim shows that the number of positive terms in $\sum_{i=1}^{n} a_i - a_{i-1}$ is bounded above by the number of positive terms in $\sum_{i=1}^{n} w(r_i) - w(r_{i-1})$, which verifies the inequality in equation (1). □

6 Lifting a colored diagram

In this section we give a method for obtaining a Morse embedding of a knot from a colored knot diagram such that the ordering of the maxima and minima by height matches the Δ–ordering of seed strands and multicolored crossings. Then we use this method to show $\mathbb{W}(\mathcal{K}) \geq w(\mathcal{K})$.

For the rest of this section, let D be a diagram of a knot in the ambient isotopy class $\mathcal{K}$ such that $\mathbb{W}(D) = \mathbb{W}(\mathcal{K})$. Assume $\mathcal{K}$ is not the ambient isotopy class of the unknot, so that D is a nontrivial diagram. Let $(A_0, f_0) \to \cdots \to (A_J, f_J)$ be a completed coloring sequence on D with attached sequence $(a_i)_{i=0}^N$. Let $\mathcal{S}$, $\mathcal{C}$ and $\Delta = \{d_i\}_{i=1}^M$ denote the set of seed strands, multicolored crossings and the Δ–ordering on $s(D) \cup \mathcal{C}$ induced by our completed coloring sequence, respectively. Let $\Delta' := \{d_{i_j}\}_{j=1}^N$ be the subsequence of Δ formed by restricting our Δ–ordering to $\mathcal{S} \cup \mathcal{C}$. Let $h_o \colon \Delta \to \mathbb{Z}$ be the height function associated to Δ, defined by $h_o(d_t) := -t$.

In this section, we embed our diagram into the plane $z = -M - 1$. Recall that D is defined as a four-valent graph with labels at each vertex containing over/under information. The labels take the form of deleting parts of the edges in the graph corresponding to under-strands. We now want to view D as a disjoint union of arcs in the plane. To this end, for all $d_i \in \Delta$ representing a strand, let d_i^* be the strand d_i with neighborhoods of the boundary of d_i removed, as dictated by the labels on the vertices of D. For each $d_i \in \Delta$ representing a multicolored crossing, let $d_i^* := d_i$. This switch in perspective on knot diagrams, from a four-valent graph to a disjoint union of arcs in the plane, is necessary to adapt the proof of the main theorem in [4] to our situation.

Theorem 6.1 *There exists a knot $\widehat{K}$ in the ambient isotopy class $\mathcal{K}$ embedded so that $h|_{\widehat{K}}$ has N critical values $c_1 > c_2 > \cdots > c_N$. For all critical values, c_j is a maximum if and only if d_{i_j} is a seed strand. In addition, c_j is a minimum if and only if d_{i_j} is a multicolored crossing.*

Proof For all $d_t \in \Delta$, let $\hat{d}_t$ denote the copy of d_t^* embedded in the plane $z = h_o(d_t)$ so that the orthogonal projection of $\hat{d}_t$ onto the plane $z = -M - 1$ is d_t^*. Recall that the crossings of a knot diagram are by definition just points on the plane, so, if d_t is a crossing, then d_t^* is the point in the plane $z = h_0(d_t)$ projecting orthogonally onto d_t. We call $\hat{d}_t$ the *lift* of d_t.

In what follows, we show that the lifts $\hat{d}_t$ can be connected in such a way that the resulting knot has D as the diagram of its projection onto the plane $z = -M - 1$. Let

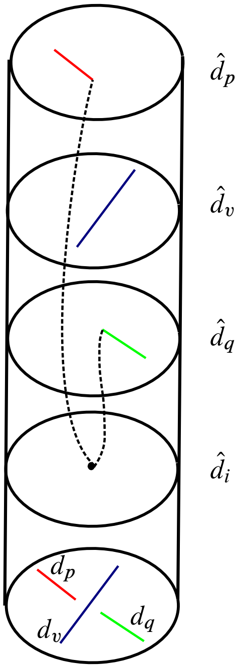


Figure 9: The construction of s_{pq} (the black dashed line) at the multicolored crossing d_i.

d_p and d_q be strands adjacent at the crossing x. Let d_v be the over-strand of x. Let $\epsilon > 0$ be such that the ball, denoted by $B(x,\epsilon)$, in the plane $z = -M-1$ has nonempty connected intersection with the strands d_p, d_q and d_v and empty intersection with all other strands. Then the cylinder $B(x,\epsilon)\times\mathbb{R}$ (where $\mathbb{R}$ denotes the z–direction) has nonempty connected intersection with $\hat{d}_p$, $\hat{d}_q$, and $\hat{d}_v$. The cylinder $B(x,\epsilon)\times\mathbb{R}$ is disjoint from all other lifts. At the crossing x, we embed an arc connecting the lifts $\hat{d}_p$ and $\hat{d}_q$, denoted by s_{pq}, via the following rule based on whether or not x is multicolored:

Connection case 1 Suppose x is a multicolored crossing. Say $x = d_i$. By Proposition 4.4(5), $h_o(d_i) < \min\{h_o(d_p), h_o(d_q), h_o(d_v)\}$. This means the plane $z = h_o(d_i)$ is below the planes containing the lifted under- and over-strands of x. Therefore, we can let s_{pq} be the union of two smooth monotone arcs connecting the endpoints of $\hat{d}_p$ and $\hat{d}_q$ in $B(x,\epsilon)\times\mathbb{R}$ to the point $\hat{d}_i$. This means $\hat{d}_i$ is the unique minimum of $h|_{s_{pq}}$. Moreover, we can choose s_{pq} such that it is contained in $B(x,\epsilon)\times\mathbb{R}$, disjoint from $\operatorname{int}(\hat{d}_v)$, and such that the orthogonal projection of

$$(\hat{d}_p \cup s_{pq} \cup \hat{d}_q \cup \hat{d}_v \cup \hat{d}_i) \cap (B(x,\epsilon)\times\mathbb{R})$$

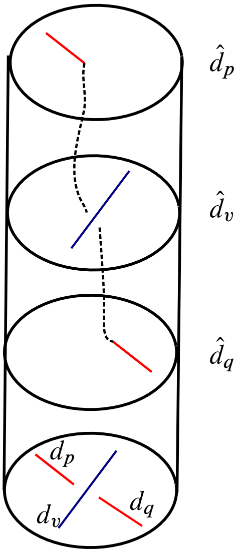


Figure 10: The construction of s_{pq} (the black dashed line) at crossings that are not multicolored.

onto the plane $z=-M-1$ is $B(x,\epsilon)\cap D$, where s_{pq} projects to the deleted portions of the under-strands of x in D. See Figure 9 for a diagram of this construction.

Connection case 2 Suppose x is not a multicolored crossing. By Proposition 4.4(4), $h_o(s_v)>\min\{h_o(s_p),h_o(s_q)\}$. This means the plane $z=h_o(d_v)$ containing the lifted over-strand of x is above at least one of the planes containing the lifted under-strands of x. Therefore, we can let s_{pq} be a smooth monotone arc that connects the endpoints of $\hat{d}_p$ and $\hat{d}_q$ that intersect $B(x,\epsilon)\times\mathbb{R}$. Moreover, we can choose s_{pq} such that it is contained in $B(x,\epsilon)\times\mathbb{R}$, disjoint from $\operatorname{int}(\hat{d}_v)$, and such that the orthogonal projection of

$$(\hat{d}_p\cup s_{pq}\cup\hat{d}_q\cup\hat{d}_v)\cap(B(x,\epsilon)\times\mathbb{R})$$

onto the plane $z=-M-1$ is $B(x,\epsilon)\cap D$, where s_{pq} projects to the deleted portions of the under-strand of x in D. See Figure 10 for a diagram of this construction.

Performing the above procedure at each crossing of D to connect all the lifts gives us a knot. Let $\widetilde{K}:=\{\bigcup_t \hat{d}_t\}\cup\{\bigcup_{p,q} s_{pq}\}$. Since we respected the crossings under projection when defining each s_{pq}, D is a diagram of $\widetilde{K}$ under orthogonal projection onto the plane $z=-M-1$. Hence, $\widetilde{K}$ is in the ambient isotopy class $\mathcal{K}$. However, $\widetilde{K}$ does not have the desired local extrema because the lifted strands are parallel to the xy–plane.

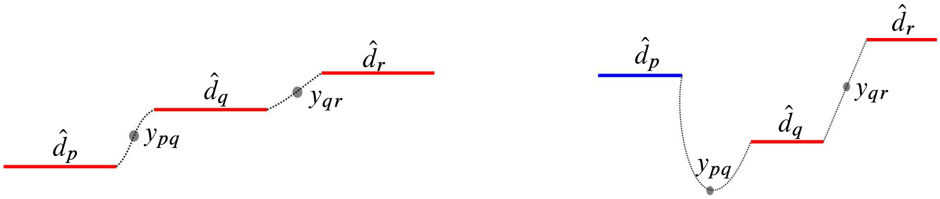


Figure 11: The setup of perturbation case 1, divided into subcases based on whether y_{pq} does (right) or does not (left) orthogonally project onto a multicolored crossing. Here d_q is not a seed strand. The idea is to perturb $[y_{pq}, y_{qr}]$, the subarc from y_{pq} to y_{qr} containing $\hat{d}_q$, into a monotonic arc with endpoints y_{pq} and y_{qr}.

We now show how to perturb the lifted strands contained in $\widetilde{K}$ so that we have the desired local extrema. For all s_{pq}, let y_{pq} denote the point in ∂s_{pq} that orthogonally projects to the corresponding crossing. Let d_p and d_r be the strands adjacent to d_q. Let $[y_{pq}, y_{qr}]$ denote the subarc of $s_{pq} \cup \hat{d}_q \cup s_{qr}$ from y_{pq} to y_{qr}. We consider cases based on whether d_q is a seed strand.

Perturbation case 1 Suppose d_q is not a seed strand. See Figure 11 for diagrams of what the lifts and $[y_{pq}, y_{qr}]$ could look like in this case. By Proposition 4.4(2), d_q is not the local maximum of h_o on $f_J^{-1}(f_J(d_q))$.

Claim $$\min\{y_{pq}, y_{qr}\} < h_o(d_q) < \max\{y_{pq}, y_{qr}\}.$$

Proof We consider cases based on whether the points y_{pq} and y_{qr} orthogonally project onto multicolored crossings. First suppose neither y_{pq} nor y_{qr} orthogonally projects onto multicolored crossings. Then d_p, d_q and d_r have all been assigned the same color. That is, $d_p, d_q, d_r \in f_J^{-1}(f_J(d_q))$. Since D is assumed to be nontrivial, if k denotes the stage at which d_q receives its color, then Proposition 4.4(3) asserts that $\{d_p, d_r\} \not\subset A_{k-1}$. That is, either d_p or d_r is uncolored at stage k. This implies $\min\{h_o(d_p), h_o(d_r)\} < h_o(d_q)$. But d_q is not the local maximum of h_o. Therefore, $h_o(d_q) < \max\{h_o(d_p), h_o(d_r)\}$. By the proof of connection case 2 of this theorem, the strands s_{pq} and s_{qr} are monotonic, so

$$\begin{aligned}\min\{h_o(d_p), h_o(d_r)\} < \min\{y_{pq}, y_{qr}\} < h_o(d_q) < \max\{y_{pq}, y_{qr}\} \\ < \max\{h_o(d_p), h_o(d_r)\},\end{aligned}$$

which gives the claim in this case.

Now say y_{pq} orthogonally projects onto a multicolored crossing. Then there exists some d_i such that $\hat{d}_i = y_{pq}$ and $h_o(d_i) = y_{pq}$. Proposition 4.4(5) implies

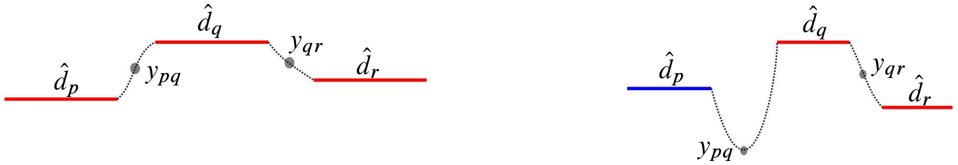


Figure 12: The setup of perturbation case 2, divided into subcases based on whether y_{pq} does (right) or does not (left) orthogonally project onto a multicolored crossing. Here d_q is a seed strand. The idea is to perturb $[y_{pq}, y_{qr}]$, the subarc from y_{pq} to y_{qr} containing $\hat{d}_q$, into an arc with a single maximum at the midpoint of $[y_{pq}, y_{qr}]$.

$y_{pq} = h_o(d_i) < h_o(d_q)$. Since $f_J^{-1}(f_J(d_q))$ is connected by Proposition 4.4(1) and d_q is not a seed strand, y_{qr} does not orthogonally project onto a multicolored crossing. Therefore, d_p must have inherited its color from d_r via a coloring move, so $h_o(d_q) < h_o(d_r)$. Since $h_o(d_q) < y_{qr} < h_o(d_r)$, we get the claim in this case. The argument for if y_{qr} orthogonally projects onto a multicolored crossing is similar. □

By the above claim, we can let the subarc $[y_{pq}, y_{qr}]'$ be an arbitrarily small perturbation of $[y_{pq}, y_{qr}]$ into a smooth monotonic arc, strictly increasing or decreasing as dictated by the values of $h_o(d_p)$ and $h_o(d_r)$. The perturbation is assumed to fix y_{pq}, y_{qr} and the projection to the plane $z = -M - 1$.

Perturbation case 2 Suppose d_q is a seed strand. See Figure 12 for diagrams of what the lifts and $[y_{pq}, y_{qr}]$ could look like in this case. By Proposition 4.4(2), d_q is the unique local maximum of h_o on $f_J^{-1}(f_J(d_q))$.

Claim $$\max\{y_{pq}, y_{qr}\} < h_o(d_q).$$

Proof If y_{pq} orthogonally projects onto a multicolored crossing, then $y_{pq} < h_o(d_q)$ by the same reasoning as in the proof of the claim for perturbation case 1. So suppose y_{pq} does not orthogonally project onto a multicolored crossing. Then d_p and d_q received the same color. That is, $d_p \in f_J^{-1}(f_J(d_q))$. Since d_q is the unique local maximum of h_o on $f_J^{-1}(f_J(d_q))$, the plane $z = h_o(d_p)$ containing $\hat{d}_p$ lies below the plane $z = h_o(d_q)$ containing $\hat{d}_q$. Hence, $y_{pq} < h_o(d_q)$. We have $y_{qr} < h_o(d_q)$ by similar reasoning. □

Let m_q be the midpoint of $\hat{d}_q$. By the previous claim, we can let $[y_{pq}, y_{qr}]'$ be an arbitrarily small perturbation of $[y_{pq}, y_{qr}]$ that fixes y_{pq}, m_q and y_{qr}. In addition, we arrange $[y_{pq}, y_{qr}]'$ so that $h|_{[y_{pq}, y_{qr}]'}$ strictly increases from y_{pq} to m_q and strictly decreases from m_q to y_{qr} while fixing the projection to the plane $z = -M - 1$.

Perform a perturbation on the set of subarcs $\{[y_{pq}, y_{qr}]\}$ of $\widetilde{K}$ as dictated above. Let $\widehat{K}$ denote the resulting knot. Note $\widehat{K}$ is ambient isotopic to $\widetilde{K}$. Recall $\Delta' := \{d_{i_j}\}$ is the restriction of our Δ–ordering to $\mathcal{S} \cup \mathcal{C}$.

By perturbation case 2, each lifted seed strand $\hat{d}_{i_j}$ results in a maximum of $\widehat{K}$. The critical point corresponding to this maximum is the midpoint m_{i_j} of $\hat{d}_{i_j}$. Therefore, $\widehat{K}$ has a single maximum for every seed strand d_{i_j} with height $h_o(d_{i_j})$. By perturbation case 1, all other lifted strands become monotonic after perturbation.

By connection case 1, each multicolored crossing results in a minimum of $\widehat{K}$. The critical point corresponding to this minimum is the lifted multicolored crossing. Therefore, $\widehat{K}$ has a single minimum for every multicolored crossing d_{i_j} with height $h_o(d_{i_j})$.

Since the monotonicity of the subarcs of s_{pq} from y_{pq} to $\partial\hat{d}_q$ is preserved by our perturbation, $\widehat{K}$ has only $|\mathcal{S} \cup \mathcal{C}| = N$ local extrema. Ordering the critical values $c_1 > c_2 > \cdots > c_N$ of $h|_{\widehat{K}}$ by decreasing height for each j between 1 and N, c_j is a maximum if and only if d_{i_j} is a seed strand and c_j is a minimum if and only if d_{i_j} is a multicolored crossing, as desired. □

Corollary 6.2 *If $\mathcal{K}$ is an ambient isotopy class of knots that does not contain the unknot, then $\mathbb{W}(\mathcal{K}) \geq w(\mathcal{K})$*

Proof Let D be a diagram of a knot in the ambient isotopy class $\mathcal{K}$ such that $\mathbb{W}(D) = \mathbb{W}(\mathcal{K})$. Then there exists a completed coloring sequence on D with attached sequence $(a_i)_{i=0}^N$ such that $\sum_{i=0}^N a_i = \mathbb{W}(\mathcal{K})$. Let $\Delta' = \{d_{i_j}\}_{j=1}^N$ denote the Δ–ordering resulting from this coloring, restricted to the resulting seed strands and multicolored crossings. By Theorem 6.1, there exists a knot $\widehat{K}$ in the ambient isotopy class $\mathcal{K}$ with N local extrema that satisfy the following property: if $c_1 > c_2 > \cdots > c_N$ are the critical values of $h|_{\widehat{K}}$ ordered by decreasing height, then c_j is a maximum if and only if d_{i_j} is a seed strand and c_j is a minimum if an only if d_{i_j} is a multicolored crossing. This property ensures that, if $r_i \in (c_{i+1}, c_i)$ is a regular value of $h|_{\widehat{K}}$, then $a_i = |\widehat{K} \cap h^{-1}(r_i)|$. Therefore,

$$\mathbb{W}(\mathcal{K}) = \mathbb{W}(D) = w(\widehat{K}) \geq w(\mathcal{K}). \qquad \square$$

7 Proof of the main theorems

In this section we summarize previous results to prove our main theorems. Note that most results of Sections 5 and 6 do not apply to the unknot, so we must handle that case separately.

Before proving Theorem 1.1, we need one more technical lemma. Colloquially, it states that, at any stage of a coloring sequence, the number of multicolored crossings that have occurred is bounded above by the number of colors (seed strands) that have appeared.

Lemma 7.1 *Let $(A_0, f_0) \to \cdots \to (A_t, f_t)$ be a partial coloring sequence on the knot diagram D. Let $C := \{x_1, \dots, x_m\} \subset v(D)$ be the set of crossings of D that have become multicolored by stage t. Then $|C| \le |f_t(A_t)|$.*

Proof We define a graph associated to the partial coloring sequence. Let $V := \{v_1, \dots, v_m\}$ be the vertex set, where we have one vertex for every multicolored crossing. Recall from Proposition 4.4(1) that, for all $y \in f_t(A_t)$, the set $f_t^{-1}(y)$ is connected. This means that, for all $y \in f_t(A_t)$, there are at most two multicolored crossings with under-strands assigned the color y. That is, the set $f_t^{-1}(y)$ contains the under-strands of at most two multicolored crossings. For each $y \in f_t(A_t)$ where $f_t^{-1}(y)$ contains the under-strands of two distinct multicolored crossings $x_i, x_j \in C$ (so $i \neq j$), let e_{ij} be an edge that joins the vertices v_i and v_j. For each $y \in f_t(A_t)$ where $f_t^{-1}(y)$ contains the under-strand of a single multicolored crossing $x_i \in C$, let e_{ii} be a loop based at the vertex v_i. That is, e_{ii} is an edge with both endpoints at v_i. Let E be the set of all edges obtained by this procedure.

Let $G := (V, E)$ denote the resulting graph. From the definition of G, it is clear that $|E| \le |f_t(A_t)|$ and $|C| = |V|$. Let $\deg(v)$ denote the number of edges incident to v, where any loop based at v is counted twice. The *handshaking lemma*, which is a standard result in graph theory, states that

$$\sum_{v \in V} \deg(v) = 2|E|.$$

The under-strands of each multicolored crossing must be assigned different colors, and loops based at v are counted twice in the definition of $\deg(v)$, so $2 \le \deg(v) \le 4$ for all $v \in V$. Therefore,

$$2|V| \le \sum_{v \in V} \deg(v).$$

But $|C| = |V|$ and $|E| \le |f_t(A_t)|$. Therefore,

$$2|C| = 2|V| \le \sum_{v \in V} \deg(v) = 2|E| \le 2|f_t(A_t)|,$$

which gives the desired inequality. □

We now restate and prove our main theorems.

Theorem 1.1 *If $\mathcal{K}$ is an ambient isotopy class of knots, then $\mathbb{W}(\mathcal{K}) = w(\mathcal{K})$.*

Proof We begin with the case where $\mathcal{K}$ is not the ambient isotopy class of the unknot. Theorem 5.7 gives $\mathbb{W}(\mathcal{K}) \le w(\mathcal{K})$. Corollary 6.2 gives $w(\mathcal{K}) \le \mathbb{W}(\mathcal{K})$, so we get the desired equality.

Now suppose that $\mathcal{K}$ is the ambient isotopy class of the unknot. Then $w(\mathcal{K}) = 2$. We can obtain a completed coloring sequence on the standard diagram of the unknot, with no crossings, by performing a single seed addition. This shows $\mathbb{W}(\mathcal{K}) \le 2$. We now verify that $\mathbb{W}(\mathcal{K}) \ge 2$. Let U be a diagram of the unknot. Let $(A_0, f_0) \to \cdots \to (A_J, f_J)$ be a completed coloring sequence on U with attached sequence $(a_i)_{i=0}^N$.

Let $a_n := \min\{a_i\}_{i=0}^N$. Then there exists a stage t such that the partial coloring sequence $(A_0, f_0) \to \cdots \to (A_t, f_t)$ induces the first n terms, $(a_i)_{i=0}^n$, in our attached sequence. Write

$$a_n = \sum_{i=1}^{n} a_i - a_{i-1}. \tag{2}$$

Define

$$S := \{i \mid a_i - a_{i-1} = 2,\ 1 \le i \le n\}, \quad C := \{i \mid a_i - a_{i-1} = -2,\ 1 \le i \le n\}.$$

The quantity $|S|$ is equal to the number of seed additions that have been performed by stage t. Thus, $|S| = |f_t(A_t)|$. The quantity $|C|$ is the number of crossings that have become multicolored by stage t, because $a_n = \min\{a_i\}_{i=0}^N$. By Lemma 7.1, $|C| \le |f_t(A_t)| = |S|$. We have $a_i - a_{i-1} \in \{-2, 2\}$ for all i between 1 and n, so $|S|$ is also the number of positive terms in equation (2), and $|C|$ is also the number of negative terms in equation (2). Therefore, Lemma 7.1 implies that the number of negative terms is bounded above by the number of positive terms in equation (2). We conclude $a_n \ge 0$.

Since $a_n = \min\{a_i\}_{i=0}^N$, all terms in the attached sequence are nonnegative. Any completed coloring sequence on a knot diagram must start with a seed addition. Therefore, $a_0 = 0$ and $a_1 = 2$. Hence, our conclusion verifies that $\mathbb{W}(U) \ge 2$. But U was arbitrary, so $\mathbb{W}(\mathcal{K}) \ge 2$. Therefore, $\mathbb{W}(\mathcal{K}) = 2 = w(\mathcal{K})$. □

Theorem 1.2 *For any ambient isotopy class $\mathcal{K}$ of knots and any positive integer n, there exist infinitely many diagrams D of knots in $\mathcal{K}$ such that $\mathbb{W}(D) = w(\mathcal{K})$ but $w_p(D) \ge \mathbb{W}(D) + n$.*

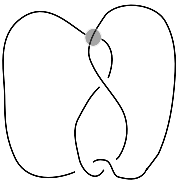


Figure 13: The diagram of the unknot U with a highlighted crossing.

Proof Let U be the diagram of the unknot depicted in Figure 13, contained in the yz–plane. Let E be the diagram obtained by performing a crossing change to the highlighted crossing in Figure 13. See Figure 14. Let $\mathcal{E}$ denote the ambient isotopy class of the figure 8 knot and K_E denote a knot in $\mathcal{E}$ such that $p(K_E) = E$. (Recall $p\colon \mathbb{R}^3 \to \mathbb{R}^2$ is the standard projection into the yz–plane.)

By Theorem 1.1, there exists a diagram D' of a knot $K_{D'}$ in $\mathcal{K}$ such that $\mathbb{W}(D') = w(\mathcal{K})$. Let

$$D = D' \# U \# \cdots \# U,$$

where there are m terms in the connected sum, and the connected sum is performed as shown in Figure 15.

We take the strand of D' on which we surger to form D to be a seed strand of a completed coloring sequence on D' which realizes the equality $\mathbb{W}(D') = w(\mathcal{K})$. After performing a seed addition to the strand of D labeled s in Figure 15, we can use coloring moves to extend the color to all other strands of D which correspond to components of U. Since D was formed by surgering the aforementioned seed strand of D', it is easy to see $\mathbb{W}(D) = \mathbb{W}(D') = w(\mathcal{K})$. These equalities are independent of m.

By performing a crossing change at each crossing of D highlighted in Figure 15, we get a diagram of the knot $K_{D'} \# K_E \# \cdots \# K_E$. See Figure 16.

Without loss of generality, we can perform an arbitrarily small perturbation on the knot $K_{D'} \# K_E \# \cdots \# K_E$, which descends to a planar isotopy on $D' \# E \# \cdots \# E$, such that

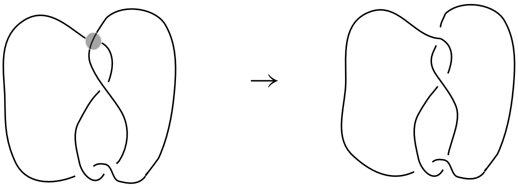


Figure 14: The crossing change performed on U (left) at the highlighted crossing to obtain E (right).

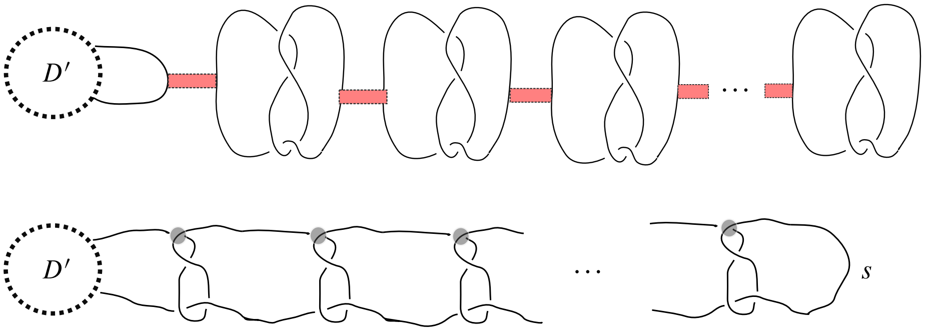


Figure 15: Top: D' with m copies of U. The rectangles along which we surger to form D are in red. Our calculations of width are independent of the orientations of the diagrams, so we assume each diagram is oriented to make the depicted connected sum well defined. Bottom: D with some crossings highlighted and a strand labeled s.

$h|_{K_{D'}\#K_E\#\cdots\#K_E}$ is Morse. Since planar width is unaffected by crossing changes, we get

$$w_p(D) = w_p(D' \# E \# \cdots \# E) \geq w(K_{D'} \# K_E \# \cdots \# K_E).$$

Recall Schubert's theorem on the additivity of bridge number (see [11, Theorem 1]), which states that, for any two ambient isotopy classes of knots $\mathcal{K}_1$ and $\mathcal{K}_2$,

$$\beta(\mathcal{K}_1 \# \mathcal{K}_2) = \beta(\mathcal{K}_1) + \beta(\mathcal{K}_2) - 1.$$

For any ambient isotopy class of knots, bridge number is a lower bound on Gabai width. By inductively applying Schubert's theorem with this observation, and the fact that $\beta(\mathcal{E}) = 2$ (recall $\mathcal{E}$ is the ambient isotopy class of the figure 8 knot), we get

$$w(K_{D'} \# K_E \# \cdots \# K_E) \geq \beta(K_{D'} \# K_E \# \cdots \# K_E) \geq \beta(\mathcal{K}) + m\beta(\mathcal{E}) - m = \beta(\mathcal{K}) + m,$$

where we got the second inequality because m is just the number of copies of E that we used in the connected sum to form D. Since the equalities $\mathbb{W}(D) = \mathbb{W}(D') = w(\mathcal{K})$ are independent of m, we can take m arbitrary large. Taking $m = \mathbb{W}(D) + n - \beta(\mathcal{K})$ in particular gives $w_p(D) \geq \mathbb{W}(D) + n$. □

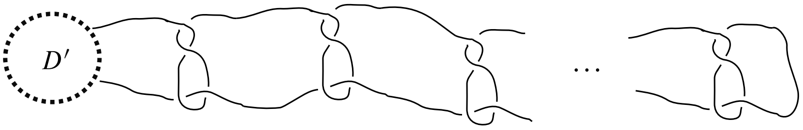


Figure 16: The resulting diagram of the knot $K_{D'} \# K_E \# \cdots \# K_E$ after performing a crossing change at each highlighted crossing in Figure 15.

8 Applications and further questions

In this section, we demonstrate how Theorem 1.1 can be used to write algorithms for calculating Gabai width. We will describe an algorithm we wrote that calculated the Gabai width of a large subset of tabulated knots from [9]. The data and code for our calculation are available at [10].

Our strategy was to modify the code in [14], which is the original algorithm for calculating Wirtinger number developed by the authors in [4], so that, given a Gauss code, it will output a completed coloring sequence for Wirtinger width. The modification is easy because the coloring moves for Wirtinger number and Wirtinger width are the same. Our modifications were motivated by the following lemma:

Lemma 8.1 *If $K \subset S^3$ is a 4–bridge prime knot in thin position, and thin position for K is not bridge position, then K has Gabai width 28.*

Proof Consider $\mathbb{R}^3$ now as in $S^3 = \mathbb{R}^3 \cup \{\infty\}$, with h the same height function as before. A thin position embedding of a 4–bridge knot must have four maxima and four minima. Since K is prime, $S^3 \setminus \eta(K)$ does not contain any essential 2–punctured spheres, where $\eta(K)$ is a tubular neighborhood of K. Wu [15] showed that the thinnest thin level of a knot that is in thin position but not bridge position is an essential surface in $S^3 \setminus \eta(K)$. Therefore, $|K \cap h^{-1}(r)| \neq 2$ for any regular value r of $h|_K$. For any regular value r of $h|_K$ at the thinnest level, the number of maxima above $h^{-1}(r)$ must be greater than or equal to the number of minima above $h^{-1}(r)$. These facts mean that the only possible orderings of the critical points of a prime 4–bridge knot are

$$M > M > M > M > m > m > m > m \quad \text{and} \quad M > M > M > m > M > m > m > m,$$

where the M's represent maxima and m's represent minima. The first ordering corresponds to a Gabai width of 32 while the second corresponds to a Gabai width of 28. However, the first ordering also corresponds to a bridge position embedding of a 4–bridge knot. Since bridge position of K is not thin position, the ordering of the critical points of K must be as in the second ordering above, so K has Gabai width 28. □

We focused on a subset of tabulated knots from [9] that are known to be prime with bridge number 4, with Gauss codes such that the code in [14] can actually detect bridge number 4. A prime knot with bridge number 4 such that thin position is bridge position must have Gabai width 32. Therefore, given a Gauss code representing a prime knot with bridge number 4, Lemma 8.1 implies that such a knot must have Gabai width 32 or 28. By Theorem 1.1, such a knot must have Wirtinger width 32 or 28. So every time we can find a completed coloring sequence on such a Gauss code giving Wirtinger width 28, we know the Gauss code represents a knot with Gabai width 28. Whenever our algorithm outputs an upper bound of 32 on the Wirtinger width for a given Gauss code, we unfortunately do not get any new information about Gabai width for the corresponding knot.

In light of these observations, we modified the code in [14] to search for a completed coloring sequence that starts with three seed additions, followed by coloring moves until we get a multicolored crossing, then finishes coloring the diagram with a seed addition that comes before three more multicolored crossings appear. Recall that seed strands correspond to maxima and the multicolored crossings correspond to minima, so such a coloring sequence corresponds to an embedding of the knot with Gabai width 28.

Our code implemented the above strategy and was able to verify that 54 756 tabulated knots have Gabai width 28, out of 86 981 knots that were tested. This is the first time a systematic calculation of Gabai width has been performed on this collection of Gauss codes. The appendix of [4] states that the code we modified in [14] for our algorithm runs in factorial time. Our modifications are such that our algorithm also runs in factorial time. However, our algorithm ran fast in practice since we had such specific information about the ordering of the seed strands and multicolored crossings in the completed coloring sequence we desired. In general, whenever bridge number is much less than the crossing number, the code in [14] runs fast in practice.

We remark that it was important to know the Gauss codes we were working on had diagrams such that the code in [14] can actually detect Wirtinger number 4 (and hence bridge number 4). In general, this does not always happen. In [3], the authors give examples of prime, reduced, alternating diagrams of a knot such that the Wirtinger number is strictly greater then the bridge number.

We briefly describe how we knew the bridge number. In [1], the authors give a method of establishing bridge number based on homomorphisms from the knot group to Coxeter groups. In ongoing work [2], the authors use computational methods to find homomorphisms as described in [1] to verify that each of the knots tested in our code [10] have bridge number 4.

Our implementation depended heavily on the Wirtinger number of a knot diagram. In general, the search for the minimum $\mathbb{W}(D)$ over all possible diagrams D is subtle. We took great advantage of the fact that the diagrams we worked on actually realized the Wirtinger number $\mu(D)$. In order to find a more robust implementation of our notions, it is important to understand how Wirtinger number and Wirtinger width interact. This leads to the following natural questions:

Question How can we determine whether or not a diagram D realizes the minimal $\mathbb{W}(D)$ without knowing beforehand that it realizes the minimal $\mu(D)$, the Wirtinger number?

Question If the knot diagram D realizes the Wirtinger number, then does D also realize the Wirtinger width?

One expects the answer to the second question to be no, since in [5] the authors exhibit a knot $\mathcal{K}$ such that the thin position embedding has more that $\beta(\mathcal{K})$ many maxima. However, finding a knot diagram which disproves our question seems difficult. An obvious first step is to check our knot data for a knot such that our algorithm outputs an upper bound of 32 for Gabai width, and try to show that the Gabai width of such a knot is actually 28.

References

[1] **S Baader**, **R Blair**, **A Kjuchukova**, *Coxeter groups and meridional rank of links*, Math. Ann. 379 (2021) 1533–1551 MR Zbl

[2] **R Blair**, **A Kjuchukova**, **N Morrison**, *Coxeter quotients of knot groups through* 16 *crossings*, preprint (2022) arXiv 2208.09032

[3] **R Blair**, **A Kjuchukova**, **M Ozawa**, *The incompatibility of crossing number and bridge number for knot diagrams*, Discrete Math. 342 (2019) 1966–1978 MR Zbl

[4] **R Blair**, **A Kjuchukova**, **R Velazquez**, **P Villanueva**, *Wirtinger systems of generators of knot groups*, Comm. Anal. Geom. 28 (2020) 243–262 MR Zbl

[5] **R Blair**, **M Tomova**, *Width is not additive*, Geom. Topol. 17 (2013) 93–156 MR Zbl

[6] **D Gabai**, *Foliations and the topology of* 3*–manifolds, III*, J. Differential Geom. 26 (1987) 479–536 MR Zbl

[7] **H Goda**, **M Scharlemann**, **A Thompson**, *Levelling an unknotting tunnel*, Geom. Topol. 4 (2000) 243–275 MR Zbl

[8] **C M Gordon**, **J Luecke**, *Knots are determined by their complements*, J. Amer. Math. Soc. 2 (1989) 371–415 MR Zbl

[9] **J Hoste**, **M Thistlethwaite**, **J Weeks**, *The first* 1 701 936 *knots*, Math. Intelligencer 20 (1998) 33–48 MR Zbl

[10] **R Lee**, *Wirtinger width*, Python code (2019) Available at `https://github.com/LeeRicky/Wirtinger-Width`

[11] **J Schultens**, *Additivity of bridge numbers of knots*, Math. Proc. Cambridge Philos. Soc. 135 (2003) 539–544 MR Zbl

[12] **A Thompson**, *Thin position and the recognition problem for S^3*, Math. Res. Lett. 1 (1994) 613–630 MR Zbl

[13] **A Thompson**, *Thin position and bridge number for knots in the* 3*–sphere*, Topology 36 (1997) 505–507 MR Zbl

[14] **P Villanueva**, *Wirtinger number*, Python code (2018) Available at `https://github.com/pommevilla/calc_wirt`

[15] **Y-Q Wu**, *Thin position and essential planar surfaces*, Proc. Amer. Math. Soc. 132 (2004) 3417–3421 MR Zbl

Department of Mathematics, UC Santa Barbara
Goleta, CA, United States

rickylee@ucsb.edu

Received: 10 December 2019 Revised: 10 April 2022

Algebraic & Geometric Topology **23**:6 (2023) 2449–2518
DOI: 10.2140/agt.2023.23.2449
Published: 7 September 2023

Classification of torus bundles that bound rational homology circles

JONATHAN SIMONE

We completely classify orientable torus bundles over the circle that bound smooth 4–manifolds with the rational homology of the circle. Along the way, we classify certain integral surgeries along chain links that bound rational homology 4–balls and explore a connection to 3–braid closures whose double branched covers bound rational homology 4–balls.

57K30, 57K40, 57K41

1 Introduction

In [13], we showed that two infinite families of T^2–bundles over S^1 bound (smooth) rational homology circles ($\mathbb{Q}S^1 \times B^3$'s). As an application, the $\mathbb{Q}S^1 \times B^3$'s were used to construct infinite families of rational homology 3–spheres ($\mathbb{Q}S^3$'s) that bound rational homology 4–balls ($\mathbb{Q}B^4$'s). The main purpose of this article is to show that the two families of torus bundles used in [13] are the only torus bundles that bound smooth $\mathbb{Q}S^1 \times B^3$'s.

After endowing $T^2 \times [0,1] = \mathbb{R}^2/\mathbb{Z}^2 \times [0,1]$ with the coordinates $(\boldsymbol{x}, t) = (x, y, t)$, any orientable torus bundle over S^1 is of the form $T^2 \times [0,1]/(\boldsymbol{x}, 1) \sim (\pm A\boldsymbol{x}, 0)$, where $A \in \mathrm{SL}(2,\mathbb{Z})$. The matrix A is called the *monodromy* of the torus bundle and is defined up to conjugation. Throughout, we will express the monodromy in terms of the generators $T = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ and $S = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$. A torus bundle is called *elliptic* if $|\mathrm{tr}\, A| < 2$, *parabolic* if $|\mathrm{tr}\, A| = 2$, and *hyperbolic* if $|\mathrm{tr}\, A| > 2$. Moreover, a torus bundle is called *positive* if $\mathrm{tr} A > 0$ and *negative* if $\mathrm{tr} A < 0$. Torus bundles naturally arise as the boundaries of plumbings of D^2–bundles over S^2 (see Neumann [11, Section 6] for details). Using these plumbing descriptions, it is easy to draw surgery diagrams for torus bundles. Table 1 gives a complete list of torus bundles over S^1, along with their monodromies (up to conjugation) and surgery diagrams. To simplify notation, $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$ will always denote the hyperbolic torus bundle with monodromy $\pm A(\boldsymbol{a}) = \pm T^{-a_1} S \cdots T^{-a_n} S$, where $\boldsymbol{a} = (a_1, \dots, a_n)$, $a_1 \geq 3$, and $a_i \geq 2$ for all i.

Theorem 1.1 *A torus bundle over S^1 bounds a $\mathbb{Q}S^1 \times B^3$ if and only if*

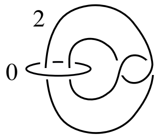
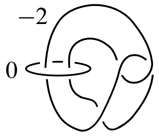
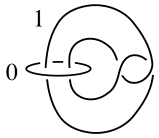
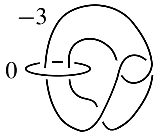
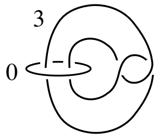
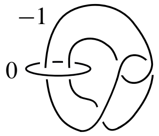
- *it is negative parabolic, or*
- *it is positive hyperbolic of the form $\boldsymbol{T}_{A(\boldsymbol{a})}$, where*

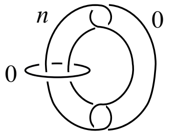
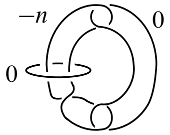
  $$\boldsymbol{a} = (3 + x_1, 2^{[x_2]}, \dots, 3 + x_{2m+1}, 2^{[x_1]}, 3 + x_2, 2^{[x_3]}, \dots, 3 + x_{2m}, 2^{[x_{2m+1}]}),$$

  $m \geq 0$, and $x_i \geq 0$ for all i.

Elliptic torus bundles and parabolic torus bundles that bound $\mathbb{Q}S^1 \times B^3$'s are rather simple to classify. Classifying hyperbolic torus bundles, which make up the "generic" class of torus bundles, is much more involved and includes the bulk of the technical work. In [13], it is shown that $\boldsymbol{T}_{A(\boldsymbol{a})}$ indeed bounds a $\mathbb{Q}S^1 \times B^3$ when $\boldsymbol{a} = (3 + x_1, 2^{[x_2]}, \dots, 3 + x_{2m+1}, 2^{[x_1]}, 3 + x_2, 2^{[x_3]}, \dots, 3 + x_{2m}, 2^{[x_{2m+1}]})$. To obstruct all other hyperbolic torus bundles from bounding $\mathbb{Q}S^1 \times B^3$'s, we first consider a related class of $\mathbb{Q}S^3$'s.

Let L_n^t denote the n–component link shown in Figure 1, where t denotes the number of half-twists. We call L_n^t the *n–component, t–half-twisted chain link.* If $t = 0$, we call the chain link *untwisted.* Consider the surgery diagram for the hyperbolic torus bundle $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$ given in Table 1. Now perform m–surgery along a meridian of the 0–framed unknot as in the left side of each of the four diagrams in Figure 2. Next, slide the unknot with framing $-a_1$ (or $-a_1 \pm 2$) twice over the blue m–framed unknot so that it no longer passes through the 0–framed unknot. Then cancel the 0–framed and m–framed unknots. When $n \geq 2$, the resulting 3–manifolds are obtained by

elliptic torus bundles			
monodromy	surgery diagram	monodromy	surgery diagram
S		$-S$	
$T^{-1}S$		$-T^{-1}S$	
$(T^{-1}S)^2$		$-(T^{-1}S)^2$	

parabolic torus bundles			
monodromy	surgery diagram	monodromy	surgery diagram
T^n $(n \in \mathbb{Z})$		$-T^n$ $(n \in \mathbb{Z})$	

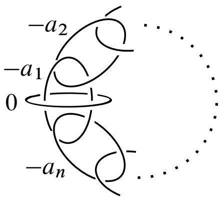
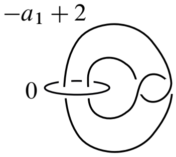
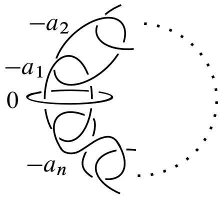
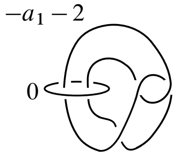
hyperbolic torus bundles $T_{\pm A(a_1,\dots,a_n)}$		
monodromy	surgery diagram	
	$n > 1$	$n = 1$
$T^{-a_1}S\cdots T^{-a_n}S$ $(a_1 \geq 3, a_i \geq 2$ for all $i)$		
$-T^{-a_1}S\cdots T^{-a_n}S$ $(a_1 \geq 3, a_i \geq 2$ for all $i)$		

Table 1: Monodromy and surgery diagrams of parabolic, elliptic and hyperbolic T^2–bundles over S^1.

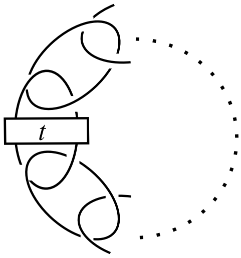


Figure 1: The n–component, t–half-twisted chain link, L_n^t. The box labeled t denotes t half-twists.

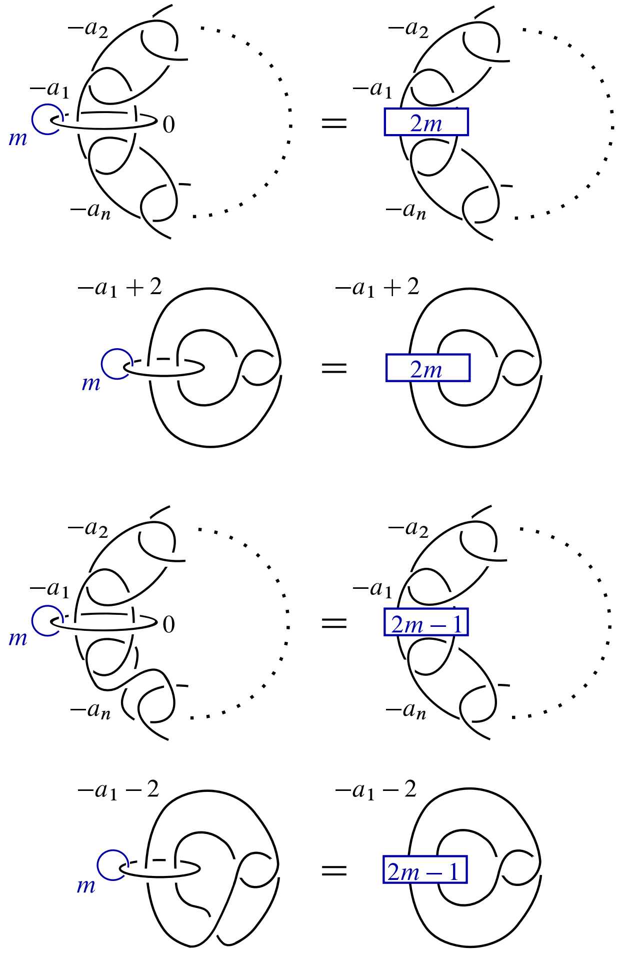


Figure 2: Surgering the hyperbolic torus bundle $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$, where $\boldsymbol{a} = (a_1, \dots, a_n)$, to obtain the rational homology sphere $Y_{\boldsymbol{a}}^t$. The blue boxes labeled $2m$ and $2m-1$ indicate the number of half-twists.

$(-a_1,\dots,-a_n)$–surgery along the chain link L^t_n, where $t=2m$ or $2m-1$. We denote these 3–manifolds by $Y^t_{\boldsymbol{a}}=S^3_{(-a_1,\dots,-a_n)}(L^t_n)$, where $\boldsymbol{a}=(a_1,\dots,a_n)$, $a_1\geq 3$, and $a_i\geq 2$ for all i. Note that, by cyclically reordering or reversing the surgery coefficients, we obtain the same 3–manifold. When $n=1$, the resulting 3–manifolds are obtained by $-(a_1\pm 2)$–surgery along L^t_1, where $t=2m+(1\pm 1)$; we denote them by $Y^t_{\boldsymbol{a}}=Y^t_{(a_1)}$. Note that $Y^t_{(a_1)}=S^3_{-a_1+2}(L^t_1)$ when t is even, and $Y^t_{(a_1)}=S^3_{-a_1-2}(L^t_1)$ when t is odd. Finally note that $Y^t_{\boldsymbol{a}}$ is a $\mathbb{Q}S^3$ for all $\boldsymbol{a}$ and t; this follows from the fact that $|H_1(Y^t_{\boldsymbol{a}})|=|\mathrm{Tor}(H_1(\boldsymbol{T}_{\pm A(\boldsymbol{a})}))|$ is finite (see Lemma A.1).

Lemma 1.2 [13] *Let Y be a $\mathbb{Q}S^1\times S^2$ that bounds a $\mathbb{Q}S^1\times B^3$ and let K be a knot in Y such that $[K]$ has infinite order in $H_1(Y;\mathbb{Z})$. Then any integer surgery on Y along K yields a $\mathbb{Q}S^3$ that bounds a $\mathbb{Q}B^4$.*

By Lemma 1.2, if $\boldsymbol{T}_{A(\boldsymbol{a})}$ bounds a $\mathbb{Q}S^1\times B^3$, then $Y^t_{\boldsymbol{a}}$ bounds a $\mathbb{Q}B^4$ for all even t, and if $\boldsymbol{T}_{-A(\boldsymbol{a})}$ bounds a $\mathbb{Q}S^1\times B^3$, then $Y^t_{\boldsymbol{a}}$ bounds a $\mathbb{Q}B^4$ for all odd t. Thus, if $Y^t_{\boldsymbol{a}}$ does not bound a $\mathbb{Q}B^4$ for some even (or odd) t, then $\boldsymbol{T}_{A(\boldsymbol{a})}$ (or $\boldsymbol{T}_{-A(\boldsymbol{a})}$) does not bound a $\mathbb{Q}S^1\times B^3$. Using this fact, we will obstruct most hyperbolic torus bundles from bounding $\mathbb{Q}S^1\times B^3$'s by identifying the strings $\boldsymbol{a}$ for which $Y^0_{\boldsymbol{a}}$ and $Y^{-1}_{\boldsymbol{a}}$ do not bound $\mathbb{Q}B^4$'s. Before writing down the result, we first recall and introduce some useful terminology.

Let $(b_1,\dots,b_k)$ be a string of integers such that $b_i\geq 2$ for all i. If $b_j\geq 3$ for some j, then we can write this string in the form $(2^{[m_1]},3+n_1,\dots,2^{[m_j]},2+n_j)$, where $m_i,n_i\geq 0$ for all i and $2^{[t]}$ denotes a string $2,\dots,2$ of t 2's. The string $(c_1,\dots,c_l)=(2+m_1,2^{[n_1]},3+m_2,\dots,3+m_j,2^{[n_j]})$ is called the *linear-dual string* of $(b_1,\dots,b_k)$. If $b_i=2$ for all $1\leq i\leq k$, then we define its linear-dual string to be $(k+1)$. Linear-dual strings have a topological interpretation. If Y is obtained by $(-b_1,\dots,-b_k)$–surgery along a linear chain of unknots, then the reversed-orientation manifold $\overline{Y}$ can be obtained by $(-c_1,\dots,-c_l)$–surgery along a linear chain of unknots (see Neumann [11, Theorem 7.3]). Finally, we define the linear-dual string of (1) to be the empty string.

Suppose $\boldsymbol{a}=(a_1,\dots,a_n)$ is of the form $(2^{[m_1]},3+n_1,\dots,2^{[m_j]},3+n_j)$, where $m_i,n_i\geq 0$ for all i; we define its *cyclic-dual* to be the string $\boldsymbol{d}=(d_1,\dots,d_m)=(3+m_1,2^{[n_1]},\dots,3+m_j,2^{[n_j]})$. In particular, a string of the form (x) with $x\geq 3$ has cyclic-dual $(2^{[x-3]},3)$. Notice that this definition only slightly differs from the definition of the linear-dual string. As a topological interpretation of cyclic-dual strings, the reversed-orientation of $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$ is given by $\overline{\boldsymbol{T}}_{\pm A(\boldsymbol{a})}=\boldsymbol{T}_{\pm A(\boldsymbol{d})}$ (see Neumann [11, Theorem 7.3]). Finally, $(a_n,\dots,a_1)$ is called the *reverse* of $(a_1,\dots,a_n)$.

Example 1.3 Consider the strings in Theorem 1.7,

$$\boldsymbol{a}=(3+x_1,2^{[x_2]},\dots,3+x_{2m+1},2^{[x_1]},3+x_2,2^{[x_3]},\dots,3+x_{2m},2^{[x_{2m+1}]}).$$

It is easy to see that the cyclic-dual of $\boldsymbol{a}$ is simply $\boldsymbol{a}$. Moreover, $\boldsymbol{a}$ is of the above form if and only if it can be expressed in the form $\boldsymbol{a}=(b_1+1,b_2,\dots,b_{k-1},b_k+1,c_1,\dots,c_l)$ if $k\geq 2$, where $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$ are linear-dual strings, or $\boldsymbol{a}=(b_1+2,2^{[b_1-1]})$ if $k=1$.

To remove the necessity of multiple cases, from now on, if $\boldsymbol{a}$ contains a substring of the form $(b_1+1,b_2,\dots,b_{k-1},b_k+1)$ and $k=1$, then we will understand this substring to simply be (b_1+2), as in Example 1.3.

Definition 1.4 Two strings are considered to be equivalent if one is a cyclic reordering and/or reverse of the other. Each string in the following sets is defined up to this equivalence. Moreover, strings of the form $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$ are assumed to be linear-dual. We define

$$\begin{aligned}
\mathcal{S}_{1a}&=\{(b_1,\dots,b_k,2,c_l,\dots,c_1,2)\mid k+l\geq 3\},\\
\mathcal{S}_{1b}&=\{(b_1,\dots,b_k,2,c_l,\dots,c_1,5)\mid k+l\geq 2\},\\
\mathcal{S}_{1c}&=\{(b_1,\dots,b_k,3,c_l,\dots,c_1,3)\mid k+l\geq 2\},\\
\mathcal{S}_{1d}&=\{(2,b_1+1,b_2,\dots,b_{k-1},b_k+1,2,2,c_l+1,c_{l-1},\dots,c_2,c_1+1,2)\mid k+l\geq 2\},\\
\mathcal{S}_{1e}&=\{(2,3+x,2,3,3,2^{[x-1]},3,3)\mid x\geq 0\text{ and }(3,2^{[-1]},3):=(4)\},\\
\mathcal{S}_{2a}&=\{(b_1+3,b_2,\dots,b_k,2,c_l,\dots,c_1)\},\\
\mathcal{S}_{2b}&=\{(3+x,b_1,\dots,b_{k-1},b_k+1,2^{[x]},c_l+1,c_{l-1},\dots,c_1)\mid x\geq 0\text{ and }k+l\geq 2\},\\
\mathcal{S}_{2c}&=\{(b_1+1,b_2,\dots,b_{k-1},b_k+1,c_1,\dots,c_l)\mid k+l\geq 2\},\\
\mathcal{S}_{2d}&=\{(2,2+x,2,3,2^{[x-1]},3,4)\mid x\geq 0\text{ and }(3,2^{[-1]},3):=(4)\},\\
\mathcal{S}_{2e}&=\{(2,b_1+1,b_2,\dots,b_k,2,c_l,\dots,c_2,c_1+1,2),(2,2,2,3)\mid k+l\geq 2\},\\
\mathcal{O}&=\{(6,2,2,2,6,2,2,2),(4,2,4,2,4,2,4,2),(3,3,3,3,3,3)\},\\
\mathcal{S}_1&=\mathcal{S}_{1a}\cup\mathcal{S}_{1b}\cup\mathcal{S}_{1c}\cup\mathcal{S}_{1d}\cup\mathcal{S}_{1e},\\
\mathcal{S}_2&=\mathcal{S}_{2a}\cup\mathcal{S}_{2b}\cup\mathcal{S}_{2c}\cup\mathcal{S}_{2d}\cup\mathcal{S}_{2e},\\
\mathcal{S}&=\mathcal{S}_1\cup\mathcal{S}_2.
\end{aligned}$$

Definition 1.5 Let $\boldsymbol{a}=(a_1,\dots,a_n)$, where $a_i\geq 2$ for all i. Define $I(\boldsymbol{a})$ to be the integer $I(\boldsymbol{a})=\sum_{i=1}^{n}(a_i-3)$.

Remark 1.6 If $\boldsymbol{b}$ and $\boldsymbol{c}$ are linear-dual strings, it is easy to see that $I(\boldsymbol{b})+I(\boldsymbol{c})=-2$. Using this observation, it easy to check that, if $\boldsymbol{a}\in\mathcal{S}_1$, then $-4\le I(\boldsymbol{a})\le -1$, and if $\boldsymbol{a}\in\mathcal{S}_2$, then $-3\le I(\boldsymbol{a})\le 0$. In the same vein, if $\boldsymbol{a}$ and $\boldsymbol{d}$ are cyclic-dual strings, then $I(\boldsymbol{a})+I(\boldsymbol{d})=0$. Consequently, if $\boldsymbol{a},\boldsymbol{d}\in\mathcal{S}$, then $I(\boldsymbol{a})=I(\boldsymbol{d})=0$. Moreover, $\boldsymbol{a}\in\mathcal{S}$ and $I(\boldsymbol{a})=0$ if and only if $\boldsymbol{a}\in\mathcal{S}_{2a}\cup\mathcal{S}_{2b}\cup\mathcal{S}_{2c}$.

Theorem 1.7 *Let $\boldsymbol{a}=(a_1,\dots,a_n)$, where $n\ge 1$, $a_i\ge 2$ for all i, and $a_j\ge 3$ for some j, and let $\boldsymbol{d}$ be the cyclic-dual of $\boldsymbol{a}$.*

(1) *Suppose $\boldsymbol{d}\notin\mathcal{S}_{1a}\cup\mathcal{O}$. Then $Y_{\boldsymbol{a}}^{-1}$ bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{a}\in\mathcal{S}_1$ or $\boldsymbol{d}\in\mathcal{S}_{1b}\cup\mathcal{S}_{1c}\cup\mathcal{S}_{1d}\cup\mathcal{S}_{1e}$.*

(2) *Suppose $\boldsymbol{a}\notin\mathcal{S}_{1a}\cup\mathcal{O}$. Then $Y_{\boldsymbol{a}}^{1}$ bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{d}\in\mathcal{S}_1$ or $\boldsymbol{a}\in\mathcal{S}_{1b}\cup\mathcal{S}_{1c}\cup\mathcal{S}_{1d}\cup\mathcal{S}_{1e}$.*

(3) *$Y_{\boldsymbol{a}}^{0}$ bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{a}\in\mathcal{S}_2$ or $\boldsymbol{d}\in\mathcal{S}_2$.*

Remark 1.8 The hypothesis "$a_j\ge 3$ for some j" in Theorem 1.7 ensures that $T_{\pm A(\boldsymbol{a})}$ is a hyperbolic torus bundle. If we remove this condition from the theorem, then we would have an additional case: $a_i=2$ for all i. In this case, $Y_{\boldsymbol{a}}^{-1}$ bounds a $\mathbb{Q}B^4$ and $Y_{\boldsymbol{a}}^{0}$ does not bound a $\mathbb{Q}B^4$. This follows from Lemma 1.2 and Theorem 1.1 and the fact that the corresponding torus bundles are the parabolic torus bundles with respective monodromies $-T^n$ and T^n (see [13]).

Remark 1.9 We will see in Lemma 4.2 that, for certain strings $\boldsymbol{d}$ that are the cyclic-duals of $(b_1,\dots,b_k,2,c_l,\dots,c_1,2)$, $Y_{\boldsymbol{d}}^{-1}$ does not bound a $\mathbb{Q}B^4$ (see Theorem 1.7(1)). However, we are unable to prove this fact for all such strings. Moreover, for each $\boldsymbol{a}\in\mathcal{O}$, we are unable to obstruct $Y_{\boldsymbol{a}}^{\pm 1}$ from bounding a $\mathbb{Q}B^4$ or show that it indeed bounds a $\mathbb{Q}B^4$. These strings are outliers that are unobstructed by the analysis we present here.

Combined with Lemma 1.2, Theorem 1.7 obstructs most hyperbolic torus bundles from bounding $\mathbb{Q}S^1\times B^3$'s. In Section 3, we will obstruct the rest by considering certain cyclic covers of $\mathbb{Q}S^1\times B^3$'s. The proof of Theorem 1.7 relies on Donaldson's diagonalization theorem [6] and lattice analysis. From this analysis, it follows that, if $\boldsymbol{a}\notin\mathcal{S}_1\cup\mathcal{O}$, then $Y_{\boldsymbol{a}}^{t}$ does not bound a $\mathbb{Q}B^4$ for all odd t, and if $\boldsymbol{a}\notin\mathcal{S}_2$, then $Y_{\boldsymbol{a}}^{t}$ does not bound a $\mathbb{Q}B^4$ for all even t. Moreover, by Lemma 1.2 and Theorem 1.1, if $\boldsymbol{a}\in\mathcal{S}_{2c}$, then $Y_{\boldsymbol{a}}^{t}$ bounds a $\mathbb{Q}B^4$ for all even t. This leads to the following question:

Question 1.10 For what values of t and for which strings $\boldsymbol{a}\in\mathcal{S}\setminus\mathcal{S}_{2c}$ does $Y_{\boldsymbol{a}}^{t}$ bound a $\mathbb{Q}B^4$?

1.1 Connection to 3–braids

There is an intimate connection between the rational homology 3–spheres $Y_{\boldsymbol{a}}^t$ and 3–braid closures; we will show that $Y_{\boldsymbol{a}}^t$ is the double cover of S^3 branched over the link given by the closure of the 3–braid word $(\sigma_1\sigma_2)^{3t}\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$, where σ_1 and σ_2 are the standard generators of the braid group on three strands.

Let $\boldsymbol{a} = (a_1, \dots, a_n)$ and consider $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^0$, as shown in the top of Figure 3. Using the techniques of Akbulut and Kirby [2], it is clear that $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^0$ are the double covers of S^3 branched over the links shown in the middle of Figure 3. The $\mathbb{Z}_2$–action inducing these covers are the 180° rotations shown in Figure 3. By isotoping these

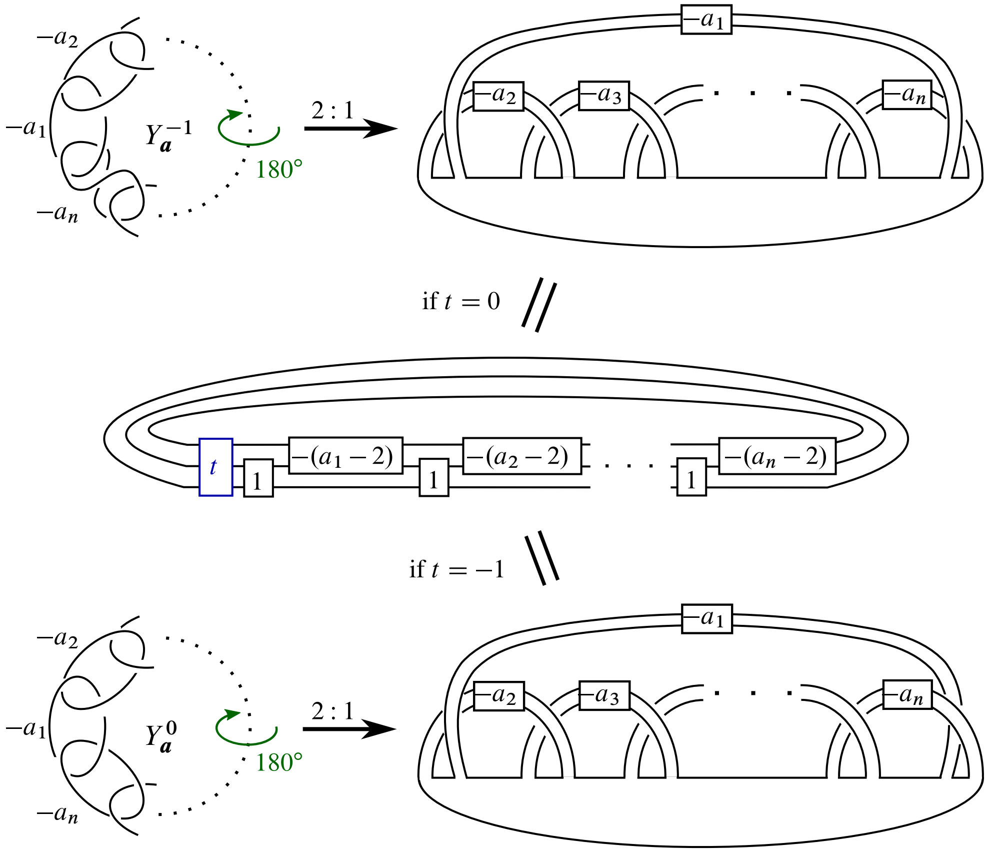


Figure 3: $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^0$ are the double covers of S^3 branched over the closure of the 3–braid word $(\sigma_1\sigma_2)^{3t}\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$, where $t = -1$ and $t = 0$, respectively. The blue box labeled t indicates the number of full-twists, while all other boxes in all other diagrams indicated the number of half-twists.

links, we obtain the closures of the 3–braid words $(\sigma_1\sigma_2)^{-3}\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$ and $\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$, respectively, as shown in Figure 3. Note that, in the figure, the blue box labeled t indicates the number of full-twists, while all other boxes indicate the number of half-twists.

Using Kirby calculus, we can argue that, for any t, $Y_{\boldsymbol{a}}^t$ is the double cover of S^3 branched over the closure of the 3–braid word $(\sigma_1\sigma_2)^{3t}\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$. Notice that, if $t=2m-1\geq -1$ is odd, then $Y_{\boldsymbol{a}}^t$ can be realized as $(-1^{[m]})$–surgery along a link in $Y_{\boldsymbol{a}}^{-1}$, as shown in the top left of Figure 4, top, and if $t=2m\geq 0$ is even, then $Y_{\boldsymbol{a}}^t$ can be realized as $(-1^{[m]})$–surgery along a link in $Y_{\boldsymbol{a}}^0$, as shown in the top left of Figure 4, bottom. Under the $\mathbb{Z}_2$–action, each of these surgery curves double covers a curve isotopic to the braid axis of the 3–braid. Thus each -1–surgery curve maps to a $-\frac{1}{2}$–surgery curve isotopic to the braid axis, as shown in the intermediate stages in Figure 4. By blowing down these curves, we obtain the desired 3–braid closures at the bottom of the figures. Note that the same argument can be used when $t<-1$; the only difference is that the surgery curves would all have positive coefficients.

Coupling this characterization with Theorems 1.7 and 1.1 and Lemma 1.2, we can classify certain families of 3–braid closures admitting double branched covers bounding $\mathbb{Q}B^4$'s.

Corollary 1.11 *Let $\boldsymbol{a}=(a_1,\dots,a_n)$, where $n\geq 1$, $a_i\geq 2$ for all i, and $a_j\geq 3$ for some j, and let $\boldsymbol{d}$ be the cyclic-dual of $\boldsymbol{a}$.*

- *Suppose $\boldsymbol{d}\notin\mathcal{S}_{1a}\cup\mathcal{O}$. Then the double cover of S^3 branched over the closure of the 3–braid word $(\sigma_1\sigma_2)^{-3}\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$ bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{a}\in\mathcal{S}_1$ or $\boldsymbol{d}\in\mathcal{S}_{1b}\cup\mathcal{S}_{1c}\cup\mathcal{S}_{1d}\cup\mathcal{S}_{1e}$.*
- *Suppose $\boldsymbol{a}\notin\mathcal{S}_{1a}\cup\mathcal{O}$. Then the double cover of S^3 branched over the closure of the 3–braid word $(\sigma_1\sigma_2)^{3}\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$ bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{d}\in\mathcal{S}_1$ or $\boldsymbol{a}\in\mathcal{S}_{1b}\cup\mathcal{S}_{1c}\cup\mathcal{S}_{1d}\cup\mathcal{S}_{1e}$.*
- *The double cover of S^3 branched over the closure of the 3–braid word*
$$\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$$
bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{a}\in\mathcal{S}_2$.
- *If $\boldsymbol{a}\in\mathcal{S}_{2c}$, then the double cover of S^3 branched over the closure of the 3–braid word $(\sigma_1\sigma_2)^{3t}\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$ bounds a $\mathbb{Q}B^4$ for all even t.*

The 3–braid knots corresponding to strings in $\mathcal{S}_{1a}\cup\mathcal{S}_{2a}\cup\mathcal{S}_{2b}\cup\mathcal{S}_{2c}$ (and their mirrors) were shown by Lisca [10] to be 3–braid knots of finite concordance order. Moreover,

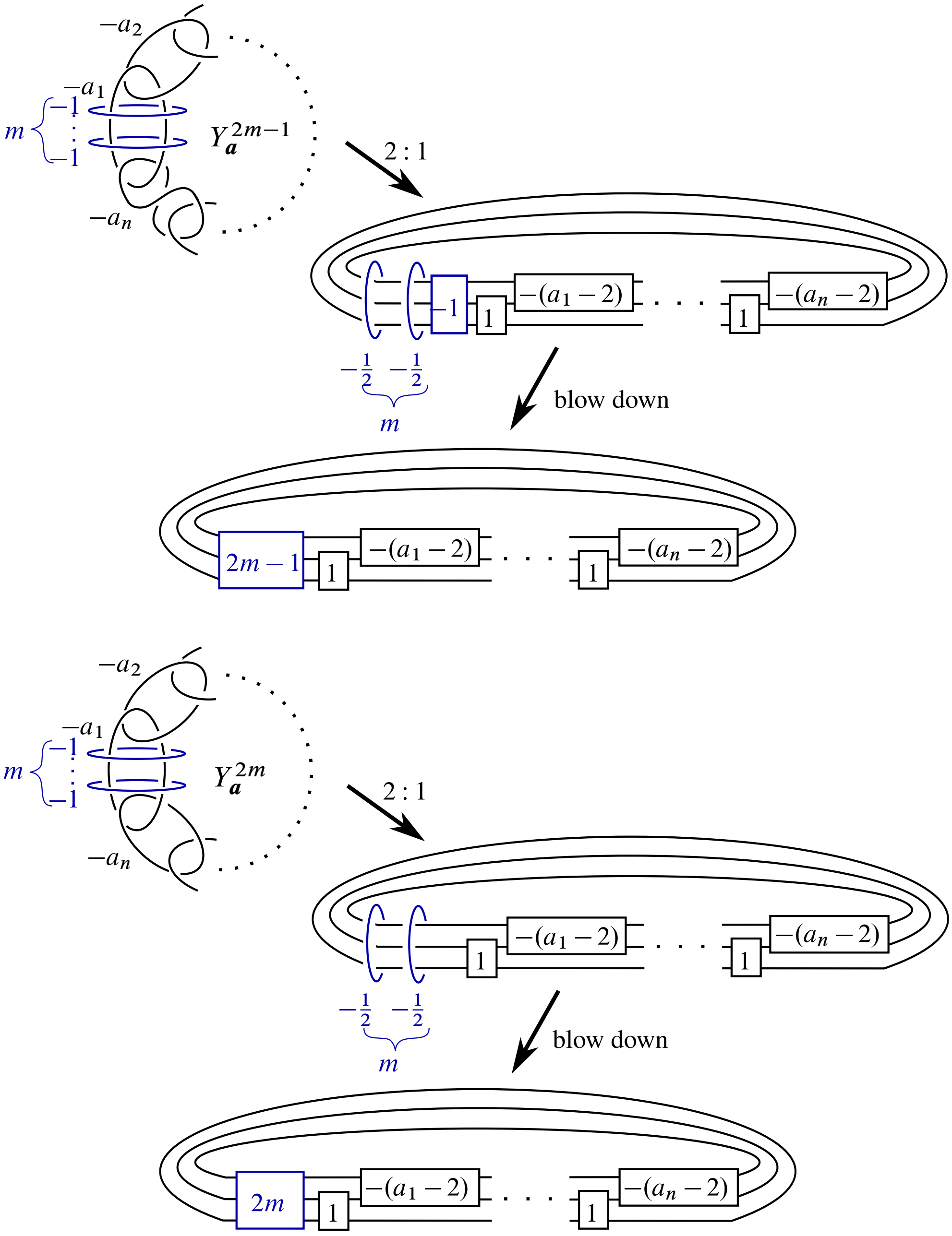


Figure 4: When $t \geq -1$, $Y_{\boldsymbol{a}}^{t}$ is the double cover of S^3 branched over the closure of the 3–braid word $(\sigma_1\sigma_2)^{3t}\sigma_1\sigma_2^{-(a_1-2)}\cdots\sigma_1\sigma_2^{-(a_n-2)}$. The same is true when $t < -1$.

some of them were shown be slice knots and so for these the corresponding double branched covers are already known to bound $\mathbb{Q}B^4$'s. Furthermore, by the classification in [10], many of the remaining strings in $\mathcal{S}$ correspond to infinite concordance order 3–braid knots. Thus, these give examples of infinite concordance order knots whose double branched covers bound $\mathbb{Q}B^4$'s. Rewording Question 1.10 in terms of 3–braids, a natural question is the following:

Question 1.12 Which other 3–braid closures admit double branched covers bounding $\mathbb{Q}B^4$'s?

Organization

In Section 2, we will highlight some simple obstructions to $\mathbb{Q}S^1 \times S^2$'s bounding $\mathbb{Q}S^1 \times B^3$'s, recall Heegaard Floer homology calculations of 3–braid closures due to Baldwin, and use these calculations to explore the orientation reversal of the 3–manifold $Y^t_{\boldsymbol{a}}$. These obstructions and calculations will be used in Sections 3 and 4. In particular, in Section 3, we will use the obstructions and other techniques to prove Theorem 1.1, and in Section 4, we will show that the $\mathbb{Q}S^3$'s of Theorem 1.7 do indeed bound $\mathbb{Q}B^4$'s by explicitly constructing them. In Sections 5–7, we will use lattice analysis to prove that the $\mathbb{Q}S^3$'s of Theorem 1.7 are the only such $\mathbb{Q}S^3$'s that bound $\mathbb{Q}B^4$'s. Finally, the appendix provides some continued fraction calculations that are used in Sections 2 and 4.

Acknowledgements

Thanks to Vitalijs Brejevs for pointing out a missing case in Lemma 6.1 and thanks to the referee for carefully reading through the technical aspects of the paper and suggesting ways to greatly improve the flow of the paper.

2 Obstructions

In this section, we highlight some simple ways to obstruct a $\mathbb{Q}S^1 \times S^2$ from bounding a $\mathbb{Q}S^1 \times B^3$, recall Baldwin's calculations of the Heegaard Floer homology of double covers of S^3 branched over certain 3–braid closures [3] (ie the rational homology 3–spheres $Y^t_{\boldsymbol{a}}$), and show that reversing the orientation of the rational homology sphere $Y^t_{\boldsymbol{a}}$ yields $Y^{-t}_{\boldsymbol{d}}$, where $\boldsymbol{d}$ is the cyclic-dual of $\boldsymbol{a}$. The first obstruction is a consequence of [5, Proposition 1.5 and Corollary 1.6].

Lemma 2.1 [5] *If $K \subset S^3$ is an alternating knot and $S^3_0(K)$ bounds a $\mathbb{Q}S^1 \times B^3$, then $\sigma(K) = 0$.*

The next obstruction is akin to a well-known homology obstruction of $\mathbb{Q}S^3$'s bounding $\mathbb{Q}B^4$'s [4, Lemma 3].

Lemma 2.2 *If Y bounds a $\mathbb{Q}S^1 \times B^3$, then the torsion part of $H_1(Y)$ has square order.*

Proof It is well known that, if a $\mathbb{Q}S^3$ bounds a $\mathbb{Q}B^4$, then its first homology group has square order [4, Lemma 3]. A similar but more complicated argument will prove the lemma.

Let $A = \mathrm{Tor}(H_1(Y))$. We aim to show that $|A|$ is a perfect square. Let W be a $\mathbb{Q}S^1 \times B^3$ bounded by Y. Then

$$H_i(W) \cong \begin{cases} T_2 & \text{if } i = 2, \\ \mathbb{Z} \oplus T_1 & \text{if } i = 1, \\ \mathbb{Z} & \text{if } i = 0, \end{cases}$$

where T_1 and T_2 are torsion groups. By duality and the universal coefficient theorem,

$$H_i(W, Y) \cong \begin{cases} \mathbb{Z} & \text{if } i = 3, \\ T_1 & \text{if } i = 2, \\ T_2 & \text{if } i = 1. \end{cases}$$

Consider the long exact sequence

$$\begin{tikzcd}
H_3(W,Y) \arrow[r, "f"] \arrow[d, phantom, "\wr\parallel"] & H_2(Y) \arrow[r] \arrow[d, phantom, "\wr\parallel"] & H_2(W) \arrow[r] \arrow[d, phantom, "\wr\parallel"] & H_2(W,Y) \arrow[r] \arrow[d, phantom, "\wr\parallel"] & H_1(Y) \arrow[r, "g"] \arrow[d, phantom, "\wr\parallel"] & H_1(W) \arrow[r, "h"] \arrow[d, phantom, "\wr\parallel"] & H_1(W,Y) \arrow[d, phantom, "\wr\parallel"] \\
\mathbb{Z} & \mathbb{Z} & T_2 & T_1 & \mathbb{Z} \oplus A & \mathbb{Z} \oplus T_1 & T_2
\end{tikzcd}$$

Since $H_3(W)$ and $H_1(W, Y)$ are torsion groups, and $H_3(W, Y) \cong H_0(Y) \cong \mathbb{Z}$, the maps $H_3(W) \to H_3(W, Y)$ and $H_1(W, Y) \to H_0(Y)$ in the long exact sequence of the pair (W, Y) are trivial; hence, f is injective and g is surjective. Express the map g as $g = g_1 + g_2$, where $g_1\colon \mathbb{Z} \to \mathbb{Z} \oplus T_1$ and $g_2\colon A \to \{0\} \oplus T_1$. Notice that $\mathrm{Im}\, g \cong \mathrm{Im}\, g_1 \oplus \mathrm{Im}\, g_2$ and g_1 is injective. Thus $\mathrm{Im}\, g_2$ can be identified with a subgroup of $\mathrm{coker}\, g_1$ and $T_2 \cong \mathrm{coker}\, g \cong \mathrm{coker}\, g_1 / \mathrm{Im}\, g_2$. Moreover, it follows from duality that, if f is given by multiplication by n, then g_1 is of the form $g_1(x) = \pm nz + \sum \lambda_i b_i$, where x is a generator of the domain of g_1 and $\{z, b_i\}$ is a basis for $\mathbb{Z} \oplus T_1$ such that z is an infinite order element and the b_i are torsion elements. Thus $|\mathrm{coker}\, g_1| = n|T_1| = |\mathrm{coker}\, f||T_1|$.

By exactness, we can reduce the above sequence to the short exact sequence

$$0 \to T_1/(T_2/\operatorname{coker} f) \xrightarrow{i} \mathbb{Z} \oplus A \xrightarrow{g} \operatorname{Im} g \to 0,$$

where we identify $\operatorname{coker} f$ with its image in T_2 and $T_2/\operatorname{coker} f$ with its image in T_1. Since $g_1\colon \mathbb{Z} \to \operatorname{Im} g_1$ is an isomorphism, we have the short exact sequence of finite groups

$$0 \to T_1/(T_2/\operatorname{coker} f) \xrightarrow{i} A \xrightarrow{g_2} \operatorname{Im} g_2 \to 0.$$

Consequently, $|A| = |T_1/(T_2/\operatorname{coker} f)| \cdot |\operatorname{Im} g_2|$.

Moreover,

$$\left|\frac{T_1}{T_2/\operatorname{coker} f}\right| = \frac{|T_1|\,|\operatorname{coker} f|}{|T_2|} = \frac{|\operatorname{coker} g_1|}{|\operatorname{coker} g_1|/|\operatorname{Im} g_2|} = |\operatorname{Im} g_2|.$$

Thus, $|A| = |\operatorname{Im} g_2|^2$ is a square. □

2.1 Heegaard Floer homology calculations

Let $\boldsymbol{a} = (a_1, \dots, a_n)$, where $a_i \geq 2$ for all $1 \leq i \leq n$ and $a_j \geq 3$ for some j. As mentioned in Section 1.1, the rational sphere $Y_{\boldsymbol{a}}^t$ is the double cover of S^3 branched over the closure of the 3–braid represented by the word $(\sigma_1\sigma_2)^{3t}\sigma_1\sigma_2^{-(a_1-2)} \cdots \sigma_1\sigma_2^{-(a_n-2)}$. In [3], Baldwin calculated the Heegaard Floer homology of these 3–manifolds equipped with a canonical spin^c structure $\mathfrak{s}_0$. In particular, he showed that

$$HF^+(Y_{\boldsymbol{a}}^{2m}, \mathfrak{s}_0) = \begin{cases} (\mathcal{T}_0^+ \oplus \mathbb{Z}_0^m)\{\frac{1}{4}(3n - \sum a_i)\} & \text{if } m \geq 0, \\ (\mathcal{T}_0^+ \oplus \mathbb{Z}_{-1}^{-m})\{\frac{1}{4}(3n - \sum a_i)\} & \text{if } m < 0, \end{cases}$$

$$HF^+(Y_{\boldsymbol{a}}^{2m+1}, \mathfrak{s}_0) = \begin{cases} (\mathcal{T}_0^+ \oplus \mathbb{Z}_{-1}^m)\{\frac{1}{4}(3n + 4 - \sum a_i)\} & \text{if } m \geq 0, \\ (\mathcal{T}_{-2}^+ \oplus \mathbb{Z}_{-2}^{-(m+1)})\{\frac{1}{4}(3n + 4 - \sum a_i)\} & \text{if } m < 0, \end{cases}$$

and

$$\{d(Y_{\boldsymbol{a}}^t, \mathfrak{s}) \mid \mathfrak{s} \neq \mathfrak{s}_0\} = \{d(Y_{\boldsymbol{a}}^s, \mathfrak{s}) \mid \mathfrak{s} \neq \mathfrak{s}_0\} \quad \text{for all } s, t \in \mathbb{Z}.$$

2.2 Reversing orientation

Let $\boldsymbol{a} = (a_1, \dots, a_n)$, where $a_i \geq 2$ for all $1 \leq i \leq n$ and $a_j \geq 3$ for some j. As discussed in the introduction, reversing the orientation of the hyperbolic torus bundle $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$ yields the hyperbolic torus bundle $\overline{\boldsymbol{T}}_{\pm A(\boldsymbol{a})} = \boldsymbol{T}_{\pm A(\boldsymbol{d})}$, where $\boldsymbol{d} = (d_1, \dots, d_m)$ is the cyclic-dual of $\boldsymbol{a}$ [11]. Therefore, by construction, reversing the orientation on $Y_{\boldsymbol{a}}^t$ yields $\overline{Y}_{\boldsymbol{a}}^t = Y_{\boldsymbol{d}}^s$ for some integer s. The following lemma shows that $s = -t$:

Lemma 2.3 *Let $\boldsymbol{a} = (a_1, \dots, a_n)$ and $\boldsymbol{d} = (d_1, \dots, d_m)$ be cyclic-dual. Then $\overline{Y}_{\boldsymbol{a}}^t = Y_{\boldsymbol{d}}^{-t}$.*

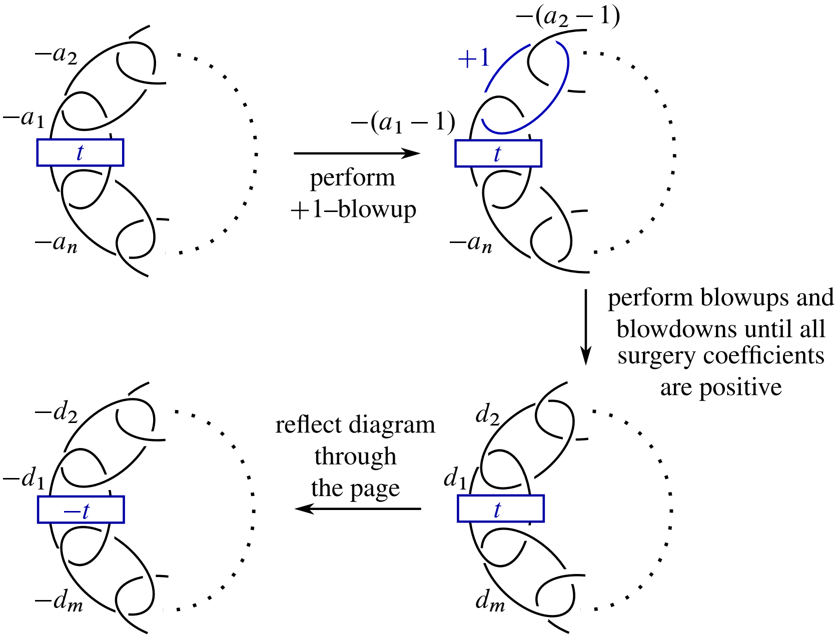


Figure 5: Proving that $\overline{Y}^t_{\boldsymbol{a}} = Y^{-t}_{\boldsymbol{d}}$, where $(d_1, \dots, d_m)$ is the cyclic-dual of $\boldsymbol{a} = (a_1, \dots, a_n)$ and $n > 1$.

Proof This is an exercise in Kirby calculus. We will focus on the case $n > 1$. The case $n = 1$ is similar, but much simpler. Start with the surgery diagram of $Y^t_{\boldsymbol{a}}$ that is made up of a t–half-twisted chain link with surgery coefficients $(-a_1, \dots, -a_n)$, as in the top left of Figure 5. We will produce a different surgery diagram for $Y^t_{\boldsymbol{a}}$ using blowups and blowdowns. Without loss of generality, assume that $a_1 \geq 3$. Let $i > 1$ be the smallest integer such that $a_i \geq 3$ and let K_i denote the unknot with surgery coefficient $-a_i$. If $a_i = 2$ for all $2 \leq i \leq n$, then set $i = n + 1$, with the understanding that $a_{n+1} = a_1$ and $K_{n+1} = K_1$. We will prove the lemma in the case $i \leq n$. The case of $i = n + 1$ is similar and requires fewer steps. Blow up the linking of the $-a_1$– and $-a_2$–framed unknots with a $+1$–framed unknot to obtain the second diagram in Figure 5. We can now perform $i - 2$ successive blowdowns of -1–framed unknots (with $i - 2 = 0$ a possibility). Next, perform $a_i - 2$ successive $+1$–blowups of the linking between K_i and the adjacent positively framed unknot; the resulting framing on K_i is -1. Continue to perform blowdowns and blowups in this way until every surgery coefficient is a positive number; we obtain the surgery diagram for $Y^t_{\boldsymbol{a}}$ made up of a chain link with positive surgery coefficients $(d_1, \dots, d_m)$, as in the third diagram of Figure 5, where $\boldsymbol{d} = (d_1, \dots, d_m)$ is the cyclic-dual of $\boldsymbol{a}$. Now we can change the orientation of $Y_{\boldsymbol{a}}^t$ by reflecting this new surgery diagram through the page. This yields a surgery diagram of $\overline{Y}_{\boldsymbol{a}}^t$ that is made up of a $-t$–half-twisted chain link with surgery coefficients $(-d_1, \dots, -d_n)$, as shown in the final diagram of Figure 5. Thus $\overline{Y}_{\boldsymbol{a}}^t = Y_{\boldsymbol{d}}^{-t}$. □

3 Torus bundles over S^1 that bound rational homology circles

In this section, we will prove Theorem 1.1. By considering the obvious handlebody diagrams of the plumbings shown in Table 1, it is rather straightforward to classify elliptic and parabolic torus bundles over S^1 that bound $\mathbb{Q}S^1 \times B^3$'s. In fact, through Kirby calculus, we will explicitly construct $\mathbb{Q}S^1 \times B^3$'s bounded by negative parabolic torus bundles and use the obstructions in Section 2 to obstruct positive parabolic torus bundles and elliptic torus bundles from bounding $\mathbb{Q}S^1 \times B^3$'s.

Proposition 3.1 *No elliptic torus bundle bounds a $\mathbb{Q}S^1 \times B^3$.*

Proof According to Table 1, there are only six elliptic torus bundles; they have monodromies $\pm S$, $\pm T^{-1}S$, and $\pm(T^{-1}S)^2$. By Lemma 2.2, if one of these torus bundles bounds a $\mathbb{Q}B^4$, then the torsion part of its first homology group must be a square. By considering the surgery diagrams in Table 1, it is easy to see that the only elliptic torus bundles that have the correct first homology are those with monodromy $T^{-1}S$ or $-(T^{-1}S)^2$. Moreover, note that, by reversing the orientation on the torus bundle with monodromy $T^{-1}S$, we obtain the torus bundle with monodromy $-(T^{-1}S)^2$. Thus we need only show that one of these torus bundles does not bound a $\mathbb{Q}S^1 \times B^3$. Consider the leftmost surgery diagram of the elliptic torus bundle with monodromy $T^{-1}S$ in Figure 6. By blowing down the 1–framed unknot, we obtain 0–surgery on the right-handed trefoil. Since the signature of the right-handed trefoil is 2, by Lemma 2.1, the elliptic torus bundle does not bound a $\mathbb{Q}S^1 \times B^3$. □

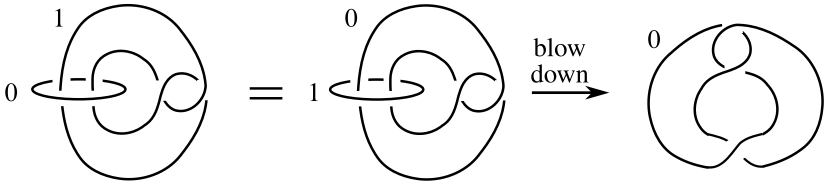


Figure 6: The elliptic torus bundle with monodromy $T^{-1}S$ does not bound a rational homology circle.

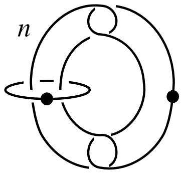


Figure 7: A $\mathbb{Q}S^1 \times B^3$ bounded by the negative parabolic torus bundle with monodromy $-T^n$.

Proposition 3.2 *Every negative parabolic torus bundle bounds a $\mathbb{Q}S^1 \times B^3$. No positive parabolic torus bundle bounds a $\mathbb{Q}S^1 \times B^3$.*

Proof By considering the surgery diagrams of the parabolic torus bundles in Table 1, it is easy to see that positive parabolic torus bundles, which have monodromy T^n, satisfy $b_1 = 2$. Thus, by the homology long exact sequence of the pair, it is easy to see that no such torus bundle can bound a $\mathbb{Q}S^1 \times B^3$. On the other hand, the negative parabolic torus bundles with monodromy $-T^n$ bound obvious $\mathbb{Q}S^1 \times B^3$'s, as shown in Figure 7. □

Classifying hyperbolic torus bundles that bound $\mathbb{Q}S^1 \times B^3$'s is not as simple as the elliptic and parabolic cases. The hyperbolic torus bundles listed in Theorem 1.1 were shown to bound $\mathbb{Q}S^1 \times B^3$'s in [13].

Proposition 3.3 [13] *Let*

$$\boldsymbol{a} = (3 + x_1, 2^{[x_2]}, \dots, 3 + x_{2m+1}, 2^{[x_1]}, 3 + x_2, 2^{[x_3]}, \dots, 3 + x_{2m}, 2^{[x_{2m+1}]}) \in \mathcal{S}_{2c},$$

where $m \geq 0$ and $x_i \geq 0$ for all i. Then $T_{A(\boldsymbol{a})}$ bounds a $\mathbb{Q}S^1 \times B^3$.

It remains to obstruct all other hyperbolic torus bundles from bounding $\mathbb{Q}S^1 \times B^3$'s. A major ingredient towards proving this fact is Theorem 1.7, which we assume to be true throughout the remainder of this section. The proof of Theorem 1.7 will be covered in Sections 4–7. Note that "most" hyperbolic torus bundles are obstructed by Theorem 1.7. In particular, by Theorem 1.7, if $\boldsymbol{a}, \boldsymbol{d} \notin \mathcal{S}_1 \cup \mathcal{O}$, then $T_{-A(\boldsymbol{a})}$ does not bound a $\mathbb{Q}S^1 \times B^3$, and if $\boldsymbol{a}, \boldsymbol{d} \notin \mathcal{S}_2$, then $T_{A(\boldsymbol{a})}$ does not bound a $\mathbb{Q}S^1 \times B^3$ (where $\boldsymbol{d}$ is the cyclic-dual of $\boldsymbol{a}$). Thus, it remains to prove that, if $\boldsymbol{a}$ or $\boldsymbol{d} \in \mathcal{S}_1 \cup \mathcal{O}$, then $T_{-A(\boldsymbol{a})}$ does not bound a $\mathbb{Q}S^1 \times B^3$, and if $\boldsymbol{a}$ or $\boldsymbol{d} \in \mathcal{S}_2 \setminus \mathcal{S}_{2c}$, then $T_{A(\boldsymbol{a})}$ does not bound a $\mathbb{Q}S^1 \times B^3$ (recall that $\boldsymbol{a} \in \mathcal{S}_{2c}$ if and only if $\boldsymbol{d} \in \mathcal{S}_{2c}$ by Example 1.3). We will prove this by considering cyclic covers of these torus bundles. But first we need to better understand the set $\mathcal{S}$. In the upcoming subsection, we will round up some necessary technical results regarding $\mathcal{S}$, and in the subsequent subsection, we will explore cyclic covers and finish the proof of Theorem 1.1.

3.1 Analyzing $\mathcal{S}$

The first technical lemma shows that the sets $\mathcal{S}_1$ and $\mathcal{S}_2$ are disjoint.

Lemma 3.4 *For a fixed string $\boldsymbol{a}$, $Y_{\boldsymbol{a}}^0$ and $Y_{\boldsymbol{a}}^{-1}$ do not both bound $\mathbb{Q}B^4$'s (and consequently $\boldsymbol{T}_{A(\boldsymbol{a})}$ and $\boldsymbol{T}_{-A(\boldsymbol{a})}$ do not both bound $\mathbb{Q}S^1 \times B^3$'s). It follows that $\mathcal{S}_1 \cap \mathcal{S}_2 = \varnothing$.*

Proof By construction,

$$|H_1(Y_{\boldsymbol{a}}^0)| = |\mathrm{Tor}(H_1(\boldsymbol{T}_{A(\boldsymbol{a})}))| \quad \text{and} \quad |H_1(Y_{\boldsymbol{a}}^{-1})| = |\mathrm{Tor}(H_1(\boldsymbol{T}_{-A(\boldsymbol{a})}))|.$$

By Lemma A.1, $|\mathrm{Tor}(H_1(\boldsymbol{T}_{A(\boldsymbol{a})}))| = |\mathrm{Tor}(H_1(\boldsymbol{T}_{-A(\boldsymbol{a})}))| - 4$. Thus $|H_1(Y_{\boldsymbol{a}}^0)|$ and $|H_1(Y_{\boldsymbol{a}}^{-1})|$ cannot simultaneously be squares and so, by [4, Lemma 3], $Y_{\boldsymbol{a}}^0$ and $Y_{\boldsymbol{a}}^{-1}$ do not both bound $\mathbb{Q}B^4$'s. Now suppose $\boldsymbol{a} \in \mathcal{S}_1 \cap \mathcal{S}_2$. Then, by Theorem 1.7, $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^0$ both bound $\mathbb{Q}B^4$'s, which is not possible. Therefore, $\mathcal{S}_1 \cap \mathcal{S}_2 = \varnothing$. □

Recall from Example 1.3 that a string $\boldsymbol{a} \in \mathcal{S}_{2c}$ can be expressed in two different, but equivalent, ways, namely

(1) $\boldsymbol{a} = (3 + x_1, 2^{[x_2]}, \dots, 3 + x_{2m+1}, 2^{[x_1]}, 3 + x_2, 2^{[x_3]}, \dots, 3 + x_{2m}, 2^{[x_{2m+1}]})$,

(2) $\boldsymbol{a} = (b_1 + 1, b_2, \dots, b_{k-1}, b_k + 1, c_1, \dots, c_l)$,

where $m \geq 0$, $x_i \geq 0$ for all i, and $(b_1, \dots, b_k)$ and $(c_1, \dots, c_l)$ are linear-dual strings with $k + l \geq 2$. This relationship is easy to see:

$$\begin{aligned}(b_1 + 1, b_2, \dots, b_{k-1}, b_k + 1) &= (3 + x_1, 2^{[x_2]}, \dots, 3 + x_{2m+1}),\\ (c_1, \dots, c_l) &= (2^{[x_1]}, 3 + x_2, 2^{[x_3]}, \dots, 3 + x_{2m}, 2^{[x_{2m+1}]}).\end{aligned}$$

Also recall that $\mathcal{S}$ is defined up to cyclic reordering and reversing strings. Thus a string $\boldsymbol{a} = (a_1, \dots, a_n) \in \mathcal{S}_{2c}$ may not be of the form (1) written above. However, by a cyclic reordering of $\boldsymbol{a}$, we can put $\boldsymbol{a}$ in the form (1), which is equivalent to (2). Moreover, it is clear that, if $a_1 \geq 3$, then $\boldsymbol{a}$ is already in the form (1) and thus already in the form (2). This simple observation will be used throughout the rest of this subsection.

Definition 3.5 Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be strings. Then $\boldsymbol{ab}$ denotes the string obtained by concatenating $\boldsymbol{a}$ and $\boldsymbol{b}$, and $\boldsymbol{a}^p$ denotes the string obtained by concatenating $\boldsymbol{a}$ with itself p times.

The next lemma follows directly from the definitions of linear-dual and cyclic-dual strings. We leave the proof to the reader.

Lemma 3.6 (a) *Suppose* $\boldsymbol{a}$ *has linear-dual* $\boldsymbol{x} = (x_1, \dots, x_p)$ *and* $\boldsymbol{b}$ *has linear-dual* $\boldsymbol{y} = (y_1, \dots, y_q)$*. Then*

- (i) $\boldsymbol{ab}$ *has linear-dual* $(x_1, \dots, x_{p-1}, x_p - 1 + y_1, y_2, \dots, y_q)$*, and*
- (ii) $\boldsymbol{ab}$ *has cyclic-dual* $(x_2, \dots, x_{p-1}, x_p - 1 + y_1, y_2, \dots, y_{q-1}, y_q - 1 + x_1)$ *(up to cyclic reordering).*

(b) *If* $\boldsymbol{a}$ *has cyclic-dual* $\boldsymbol{d}$*, then* $\boldsymbol{a}^p$ *has cyclic-dual* $\boldsymbol{d}^p$*.*

Definition 3.7 We call a string $(a_1, \dots, a_n)$ a *palindrome* if $a_i = a_{n-(i-1)}$ for all $1 \le i \le n$.

Lemma 3.8 *Consider the strings* $\boldsymbol{a} = (b_1 + 3, b_2, \dots, b_k, 2, c_l, \dots, c_1) \in \mathcal{S}_{2a}$ *and* $\boldsymbol{b} = (3 + x, b_1, \dots, b_{k-1}, b_k + 1, 2^{[x]}, c_l + 1, c_{l-1}, \dots, c_1) \in \mathcal{S}_{2b}$*.*

(a) $\boldsymbol{a} \in \mathcal{S}_{2c}$ *if and only if* $(b_1 + 1, b_2, \dots, b_k)$ *is a palindrome.*

(b) $\boldsymbol{b} \in \mathcal{S}_{2c}$ *if and only if* $(b_1, \dots, b_k)$ *is a palindrome.*

Proof (a) Since $(c_1, \dots, c_l)$ is the linear-dual of $(b_1, \dots, b_k)$, $(2, c_1, \dots, c_l)$ is the linear-dual of $(b_1 + 1, b_2, \dots, b_k)$. Consequently, $(b_1 + 1, b_2, \dots, b_k)$ is a palindrome if and only if $(2, c_1, \dots, c_l)$ is a palindrome if and only if $c_l = 2$ and $c_i = c_{l-i}$ for all $1 \le i \le l - 1$.

Assume that $(b_1 + 1, b_2, \dots, b_k)$ is a palindrome. Then $b_k = b_1 + 1 \ge 3$ and, consequently, $c_l = 2$. Let $d_1 = b_1 + 2$, $d_k = b_k - 1$, and $d_i = b_i$ for all $2 \le i \le k-1$, so that $\boldsymbol{a} = (d_1 + 1, d_2, \dots, d_{k-1}, d_k + 1, 2, c_l, \dots, c_1)$. By Lemma 3.6, $(2, 2, c_1, c_2, \dots, c_{l-1})$ has linear-dual $(b_1 + 2, b_2, \dots, b_{k-1}, b_k - 1) = (d_1, \dots, d_k)$. On the other hand, since $(2, c_1, \dots, c_l)$ is a palindrome, $(2, 2, c_1, c_2, \dots, c_{l-1}) = (2, c_l, c_{l-1}, c_{l-2}, \dots, c_1)$. Set $e_1 = e_2 = 2$ and $e_i = c_{i-2}$ for all $3 \le i \le l + 1$. Then $(d_1, \dots, d_k)$ has linear-dual $(e_1, \dots, e_{l+1})$ and thus

$$(b_1 + 3, b_2, \dots, b_k, 2, c_l, \dots, c_1) = (d_1 + 1, d_2, \dots, d_{k-1}, d_k + 1, e_1, \dots, e_{l+1}) \in \mathcal{S}_{2c}.$$

Now assume $\boldsymbol{a} \in \mathcal{S}_{2c}$. Since $b_1 + 3 > 3$, $\boldsymbol{a}$ is of the form

$$\boldsymbol{a} = (d_1 + 1, d_2, \dots, d_{p-1}, d_p + 1, e_1, \dots, e_q),$$

where $(d_1, \dots, d_p)$ and $(e_1, \dots, e_q)$ are linear-dual. Thus $d_1 = b_1 + 2$ and $e_q = c_1$. Note that the length of $\boldsymbol{a}$ is $k + l + 1 = p + q$. We claim that $p = k$. Indeed, if $p > k$, then $(d_1,\dots,d_k)=(b_1+2,b_2,\dots,b_k)$ has linear-dual $(2,2,c_1,\dots,c_l)$, implying that the length of $\boldsymbol{a}$ is greater than $k+l+1$, a contradiction; if $p<k$, we arrive at a similar contradiction. Therefore $p=k$ and $q=l+1$; consequently, $e_1=2$ and $e_i=c_{l-i+2}$ for all $2\le i\le l+1$. On the other hand, by Lemma 3.6, the linear-dual of $(d_1,\dots,d_p)=(b_1+2,b_2,\dots,b_k-1)$ is $(e_1,\dots,e_q)=(2,2,c_1,\dots,c_{l-1})$. Thus $c_l=e_2=2$ and $c_i=c_{l-i}$ for all $1\le i\le l-1$. As mentioned above, this implies that $(b_1+1,b_2,\dots,b_k)$ is a palindrome.

(b) Note that $(b_1,\dots,b_k)$ is a palindrome if and only if $(c_1,\dots,c_l)$ is a palindrome.

Assume $(b_1,\dots,b_k)$ is a palindrome. Let $d_1=2+x$ and $d_i=b_{i-1}$ for all $2\le i\le k+1$. By Lemma 3.6, the linear-dual of $(d_1,\dots,d_{k+1})=(2+x,b_1,\dots,b_{k-1},b_k)$ is $(2^{[x]},c_1+1,c_2,\dots,c_l)=(2^{[x]},c_l+1,c_{l-1},\dots,c_1)$ since $(c_1,\dots,c_l)$ is a palindrome. Relabel this string as $(e_1,\dots,e_q)$. Then

$$\boldsymbol{b}=(d_1+1,d_2,\dots,d_k,d_{k+1}+1,e_1,\dots,e_q)\in\mathcal{S}_{2c}.$$

Now assume $\boldsymbol{b}\in\mathcal{S}_{2c}$. Since $3+x\ge 3$, $\boldsymbol{b}$ is of the form

$$\boldsymbol{b}=(d_1+1,d_2,\dots,d_{p-1},d_p+1,e_1,\dots,e_q),$$

where $(d_1,\dots,d_p)$ and $(e_1,\dots,e_q)$ are linear-dual. Thus $d_1+1=3+x$ and $e_q=c_1$. Following as in the proof of the first part, $p=k+1$ and $q=l+x$. Consequently, $e_{x+1}=c_l+1$ and $e_{x+j}=c_{l-j+1}$ for all $l\le j\le l$. On the other hand, the linear-dual of $(d_1,\dots,d_p)=(2+x,b_1,\dots,b_k)$ is $(e_1,\dots,e_q)=(2^{[x]},c_1+1,c_2,\dots,c_l)$. Thus $c_1=e_{x+1}-1=c_l$ and $c_j=e_{x+j}=c_{l-j+1}$ for all $2\le j\le l$. That is, $(c_1,\dots,c_l)$ is a palindrome and thus so is $(b_1,\dots,b_k)$. □

Lemma 3.9 *Let $\boldsymbol{b}\in\mathcal{S}_{2a}\cup\mathcal{S}_{2b}$ and $p\ge 4$. Then there does not exist some proper substring $\boldsymbol{a}$ of $\boldsymbol{b}$ such that $\boldsymbol{a}^p=\boldsymbol{b}$.*

Proof Let $\boldsymbol{b}=(3+x,b_1,\dots,b_{k-1},b_k+1,2^{[x]},c_l+1,c_{l-1},\dots,c_1)\in\mathcal{S}_{2b}$. Suppose that $\boldsymbol{a}$ is a proper substring of $\boldsymbol{b}$ satisfying $\boldsymbol{a}^p=\boldsymbol{b}$ for some $p\ge 4$. Then $\boldsymbol{a}=(3+x,b_1,\dots,b_m)$ for some m. If $m=0$, then $\boldsymbol{a}=(3+x)$ and every entry of $\boldsymbol{b}$ equals $3+x$. The only such string satisfies $x=0$ and $(b_1,\dots,b_k)=(2)=(c_1,\dots,c_l)$; that is, $\boldsymbol{b}=(3,3,3)$. But then $p=3$, a contradiction.

Assume $m\ge 1$. Since $\boldsymbol{a}^p=\boldsymbol{b}$, we have that $b_{m+1}=3+x\ge 3$; consequently, either $m\le k$ or $m\ge k+x$. If $m\ge k+x$, then $m\le l$. Thus, up to switching the roles of $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$, we may assume without loss of generality that $m\le k$.

By Lemma 3.6, the linear-dual of $(b_1,\dots,b_m)$ is of the form $(c_1,\dots,c_{n-1},c'_n)$, where $n\le l$ and $c'_n\le c_n$. We claim that $m=n$. First suppose $m<n$. Then, since $\boldsymbol{a}^p=\boldsymbol{b}$, we have $b_m=c_1,b_{m-1}=c_2,\dots,b_2=c_{m-1},b_1=c_m$; that is, $(b_1,\dots,b_m)$ is a proper substring of $(c_1,\dots,c_{n-1},c'_n)$. But then the linear-dual of $(b_1,\dots,b_m)$ (ie $(c_1,\dots,c_{n-1},c'_n)$) is a proper substring of the linear-dual of $(c_1,\dots,c_{n-1},c'_n)$ (ie $(b_1,\dots,b_m)$), which is a contradiction. A similar argument shows that $n<m$ is also not possible. Thus $m=n$.

Since $m=n$ and $\boldsymbol{a}^p=\boldsymbol{b}$, we have that $b_m=c_1,b_{m-1}=c_2,\dots,b_2=c_{m-1},b_1=c_m$, and $c_{m+1}=3+x\ge 3$. If $m=k$, then, since $c_{m+1}\ge 3$, we necessarily have that $x=0$ and $p=2$, a contradiction. If $m=k-1$, then $b_k+1=b_{m+1}=3+x$ and, by Lemma 3.6, $(c_1,\dots,c_l)=(c_1,\dots,c'_m+1,2^{[x]})$; since $c_{m+1}\ge 3$, we once again have $x=0$ and $p=2$, a contradiction. Thus either $x=0$ or $m\le k-2$. In the latter case, since $(b_1,\dots,b_k)$ has linear-dual $(c_1,\dots,c_{m-1},c'_m)$, by Lemma 3.6, $(b_1,\dots,b_m,3+x,b_1)$ has linear-dual $(c_1,\dots,c_{m-1},c'_m+1,2^{[x]},3,2^{[b_1-2]})$; since $c_{m+1}=3+x\ge 3$, we necessarily have that $x=0$. Thus $c_{m+1}=b_{m+1}=3$. Moreover, since $(b_1,\dots,b_m)$ has linear-dual $(c_1,\dots,c_{m-1},c'_m)$, by Lemma 3.6, $(b_1,\dots,b_m,3)$ has linear-dual $(c_1,\dots,c_{m-1},c'_m+1,2)$. Therefore, $c_m=c'_m+1$.

Since $p\ge 4$, it follows that either $2m+2\le k$ or $2m+2\le l$. Without loss of generality, assume $2m+2\le k$. Then $(b_1,\dots,b_m,3,b_1,\dots,b_m,3)$ is a substring of $(b_1,\dots,b_k)$ and its linear-dual is a substring of $(c_1,\dots,c_l)$. By Lemma 3.6, $(b_1,\dots,b_m,3)$ has linear-dual $(c_1,\dots,c_m,2)$ and consequently $(b_1,\dots,b_m,3,b_1,\dots,b_m,3)$ has linear-dual $(c_1,\dots,c_m,c_1+1,c_2,\dots,c_m,2)$. But, since $\boldsymbol{a}^p=\boldsymbol{b}$, the latter string is also of the form $(b_m,\dots,b_1,3,b_m,\dots,b_2,b_1)$. Thus $c_1=2$ and $b_1=2$. But, since $(b_1,\dots,b_m)$ and $(c_1,\dots,c'_m)$ are linear-dual and $c_1=b_1=2$, we necessarily have $(b_1,\dots,b_k)=(2)=(c_1,\dots,c_l)$; therefore, $\boldsymbol{b}=(3,3,3)$ and $p=3$, a contradiction. We have thus shown that there does not exist a proper substring $\boldsymbol{a}$ of $\boldsymbol{b}$ such that $\boldsymbol{b}=\boldsymbol{a}^p$ for some $p\ge 4$.

Next suppose $\boldsymbol{b}=(b_1+3,b_2,\dots,b_k,2,c_l,\dots,c_1)\in\mathcal{S}_{2a}$. Let $\boldsymbol{a}=(b_1+3,b_2,\dots,b_m)$ be a substring of $\boldsymbol{b}$ such that $\boldsymbol{a}^p=\boldsymbol{b}$, where $p\ge 4$. We first claim that $m<k$. Assume otherwise. Then $m\le l$ and since $\boldsymbol{a}^p=\boldsymbol{b}$, $(b_1+3,b_2,\dots,b_k)$ is a substring of $(c_1,\dots,c_l)$. Consequently, the linear-dual of $(b_1+3,b_2,\dots,b_k)$ (ie $(2,2,2,c_1,\dots,c_l)$) is a substring of the linear-dual of $(c_1,\dots,c_l)$ (ie $(b_1,\dots,b_k)$), implying that $l<k<m$, a contradiction. Thus $m\le k$. If $m=k$, then $b_{m+1}=b_1+3\ge 3$; on the other hand, $b_{m+1}=b_{k+1}=2$, a contradiction. Thus $k<m$. Now, following the same argument as in the first part of the proof, we see that the linear-dual of $(b_1+3,b_2,\dots,b_m)$ is of the

form $(c_1, \dots, c'_m)$, where $c'_m \le c_m$ and $m \le l$. Thus $b_{m+1} = c_{m+1} = b_1 + 3 \ge 5$. But, by Lemma 3.6, $(b_1+3, b_2, \dots, b_m, b_{m+1}) = (b_1+3, b_2, \dots, b_m, b_1+3)$ has linear-dual $(c_1, \dots, c_m, 2^{[b_1+1]})$, implying that $c_{m+1} \ge 5$, which is another contradiction. □

Lemma 3.10 *Suppose* $\boldsymbol{a} \in \mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$ *and* $\boldsymbol{a}^p \in \mathcal{S}_{2c}$ *for some* p*. Then* $\boldsymbol{a} \in \mathcal{S}_{2c}$*.*

Proof It suffices to show that, if $\boldsymbol{a} \in \mathcal{S}_{2a}$ or $\boldsymbol{a} \in \mathcal{S}_{2b}$, then $\boldsymbol{a} \in \mathcal{S}_{2c}$. Let $\boldsymbol{a} \in \mathcal{S}_{2a}$, so that $\boldsymbol{a}^p$ is of the form

$$\begin{aligned} \boldsymbol{a}^p = \big(&b_1+3, b_2, \dots, b_k, 2, c_l, \dots, c_1, \\ &\qquad\qquad\vdots_{\,l} \\ &b_1+3, b_2, \dots, b_k, 2, c_l, \dots, c_1, \\ &b_1+3, b_2, \dots, b_k, 2, c_l, \dots, c_1, \\ &b_1+3, b_2, \dots, b_k, 2, c_l, \dots, c_1, \\ &\qquad\qquad\vdots_{\,p-l-1} \\ &b_1+3, b_2, \dots, b_k, 2, c_l, \dots, c_1\big). \end{aligned}$$

Since $\boldsymbol{a}^p \in \mathcal{S}_{2c}$ and $b_1 + 3 > 3$, $\boldsymbol{a}^p = (d_1+1, d_2, \dots, d_{q-1}, d_q+1, e_1, \dots, e_r)$, where $(d_1, \dots, d_q)$ and $(e_1, \dots, e_r)$ are linear-dual strings. Following as in the proof of Lemma 3.8 and appealing to Lemma 3.6, p is odd, $l = \frac{1}{2}(p-1)$ and $q = \frac{1}{2}(p-1)(k+l+1)+k$, which is the length of the blue substring above. Thus, $(e_1, \dots, e_r)$ is the black substring of $\boldsymbol{a}^p$ above. Comparing the end of both strings, it is clear that $c_l = 2$ and $c_i = c_{l-i}$ for all $1 \le i \le l-1$. As mentioned in the first paragraph of the proof of Lemma 3.8, this implies that $(b_1+1, b_2, \dots, b_k)$ is a palindrome. By Lemma 3.8, $\boldsymbol{a} \in \mathcal{S}_{2c}$.

Now assume $\boldsymbol{a} \in \mathcal{S}_{2b}$. Then $\boldsymbol{a}^p$ is of the form

$$\begin{aligned} \boldsymbol{a}^p = \big(&3+x, b_1, \dots, b_{k-1}, b_k+1, 2^{[x]}, c_l+1, c_{l-1}, \dots, c_1, \\ &\qquad\qquad\qquad\vdots_{\,l} \\ &3+x, b_1, \dots, b_{k-1}, b_k+1, 2^{[x]}, c_l+1, c_{l-1}, \dots, c_1, \\ &3+x, b_1, \dots, b_{k-1}, b_k+1, 2^{[x]}, c_l+1, c_{l-1}, \dots, c_1, \\ &3+x, b_1, \dots, b_{k-1}, b_k+1, 2^{[x]}, c_l+1, c_{l-1}, \dots, c_1, \\ &\qquad\qquad\qquad\vdots_{\,p-l-1} \\ &3+x, b_1, \dots, b_{k-1}, b_k+1, 2^{[x]}, c_l+1, c_{l-1}, \dots, c_1\big). \end{aligned}$$

Since $\boldsymbol{a}^p \in \mathcal{S}_{2c}$, $\boldsymbol{a}^p = (d_1+1, d_2, \dots, d_{q-1}, d_q+1, e_1, \dots, e_r)$, where $(d_1, \dots, d_q)$ and $(e_1, \dots, e_r)$ are linear-dual strings. Following as above, we have that p is odd,

$l = \frac{1}{2}(p-1)$ and $q = \frac{1}{2}(p-1)(k+l+x+1)+k+1$, which is the length of the blue substring above. Thus, on the one hand, $(e_1,\dots,e_r)$ is the black substring of $\boldsymbol{a}^p$ above. On the other hand, by computing the linear-dual of $(d_1,\dots,d_q)$ from the blue string above, $(e_1,\dots,e_r)$ ends in the substring $(c_1+1,\dots,c_l)$. Comparing the end of both strings, it is clear that $(c_1,\dots,c_l) = (c_l,\dots,c_1)$ and thus $(b_1,\dots,b_k)$ is also a palindrome. By Lemma 3.8, $\boldsymbol{a} \in \mathcal{S}_{2c}$. □

Corollary 3.11 *If $\boldsymbol{a}, \boldsymbol{a}^p \in \mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$, where $p \geq 4$, then $\boldsymbol{a} \in \mathcal{S}_{2c}$.*

Proof It follows from Lemma 3.9 that $\boldsymbol{a}^p \in \mathcal{S}_{2c}$; thus, $\boldsymbol{a}^p \in \mathcal{S}_{2c}$. By Lemma 3.10, $\boldsymbol{a} \in \mathcal{S}_{2c}$. □

The final technical lemma shows that the cyclic-duals of strings in $\mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$ are also in $\mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$. Although this result is implicit in the proof of Theorem 1.7, it is also relatively simple to prove directly, with the help of Lemma 3.6.

Lemma 3.12 *Let $\boldsymbol{d}$ be the cyclic-dual of $\boldsymbol{a}$. If $\boldsymbol{a} \in \mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$, then $\boldsymbol{d} \in \mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$.*

Proof Let $\boldsymbol{a} \in \mathcal{S}_{2c}$. Using the description of $\boldsymbol{a}$ as in (1) on page 2465, it is easy to see that $\boldsymbol{d} \in \mathcal{S}_{2c}$. Next let $\boldsymbol{a} = (3+x, b_1,\dots,b_k+1, 2^{[x]}, c_l+1, c_{l-1},\dots,c_1) \in \mathcal{S}_{2b}$. Notice that $(3+x, b_1,\dots,b_k+1)$ has linear-dual $(2^{[x+1]}, c_1+1,\dots,c_l,2)$ and $(2^{[x]}, c_l+1, c_{l-1},\dots,c_1)$ has linear-dual $(2+x, b_k,\dots,b_1)$. Thus, by Lemma 3.6, $\boldsymbol{d} = (2^{[x]}, c_1+1,\dots,c_l, 3+x, b_k,\dots,b_1+1) \in \mathcal{S}_{2b}$.

Finally, let $\boldsymbol{a} = (b_1+3, b_2,\dots,b_k, 2, c_l,\dots,c_1) \in \mathcal{S}_{2a}$. If $k+l=1$, then $\boldsymbol{a} = (4,2)$ and $\boldsymbol{d} = (2,4) \in \mathcal{S}_{2a}$. If $k+l=2$, then $\boldsymbol{a} = (5,2,2)$ and $\boldsymbol{d} = (2,2,5) \in \mathcal{S}_{2a}$. Now let $k+l \geq 3$. Then either $b_k \geq 3$ and $c_l = 2$ or vice versa. Assume the former. Since $(b_1+3, b_2,\dots,b_k)$ has linear-dual $(2,2,2,c_1,\dots,c_l)$ and $(2,c_l,\dots,c_1)$ has linear-dual $(b_k+1, b_{k-1},\dots,b_1)$, by Lemma 3.6,

$$\boldsymbol{d} = (2,2,c_1,\dots,c_{l-1}, c_l+b_k, b_{k-1},\dots,b_2,b_1+1).$$

Let $d_1 = c_l+b_k-3$, $d_k = b_1+1$, and $d_i = b_{k-i+1}$ for all $2 \leq i \leq k-1$. Also let $e_1 = c_{l-1}$, $e_l = 2$, and $e_i = c_{l-i}$ for all $2 \leq i \leq l-1$. Then

$$\boldsymbol{d} = (2, e_l,\dots,e_1, d_1+3, d_2,\dots,d_k)$$

and $(d_1,\dots,d_k) = (b_k-1, b_{k-1},\dots,b_2,b_1+1)$ and $(e_1,\dots,e_l) = (c_{l-1},\dots,c_1,2)$ are linear-dual; thus $\boldsymbol{d} \in \mathcal{S}_{2a}$. Now assume $b_k = 2$ and $c_l \geq 3$. Set $d_1 = c_l+b_k-3$, $d_{l+1} = 2$, $d_i = c_{l-i+1}$ for all $2 \leq i \leq l$, $e_1 = b_{k-1}$, $e_{k-1} = b_1+1$, and $e_i = b_{k-i}$ for all $2 \leq i \leq k-2$. Proceeding as above, we see that $\boldsymbol{d} \in \mathcal{S}_{2a}$. □

3.2 Cyclic covers and proving Theorem 1.1

We are now ready to finish the proof of Theorem 1.1. The next two results explore cyclic covers of $\mathbb{Q}S^1 \times B^3$'s and cyclic covers of hyperbolic torus bundles over S^1. Coupling these results with the results in Section 3.1, we complete the proof of Theorem 1.1 in the subsequent corollaries.

Lemma 3.13 *Let W be a $\mathbb{Q}S^1 \times B^3$ and let $\widetilde{W}$ be a p–fold cyclic cover of W, where p is prime and not a divisor of $|\mathrm{Tor}(H_2(W;\mathbb{Z}))|$. If $\partial\widetilde{W}$ is a $\mathbb{Q}S^1 \times S^2$, then $\widetilde{W}$ is a $\mathbb{Q}S^1 \times B^3$.*

Proof Let $Y = \partial W$ and $\widetilde{Y} = \partial\widetilde{W}$. Since W is a $\mathbb{Q}S^1 \times B^3$ and $H_3(W;\mathbb{Z})$ has no torsion, it follows that $H_3(W;\mathbb{Z}) = 0$. Thus, by Poincaré duality and the universal coefficient theorem, we have the isomorphisms

$$H_1(W, Y;\mathbb{Z}_p) \cong H^3(W;\mathbb{Z}_p) \cong \mathrm{Ext}(H_2(W;\mathbb{Z}), \mathbb{Z}_p).$$

Since p is relatively prime to $|\mathrm{Tor}(H_2(W;\mathbb{Z}))|$, we have

$$H_1(W, Y;\mathbb{Z}_p) \cong \mathrm{Ext}(H_2(W;\mathbb{Z}), \mathbb{Z}_p) = 0.$$

By the proof of [7, Theorem 1.2], since p is prime, it follows that $H_1(\widetilde{W}, \widetilde{Y};\mathbb{Z}_p) = 0$. Once again applying Poincaré duality and the universal coefficient theorem, we have the isomorphisms

$$0 = H_1(\widetilde{W}, \widetilde{Y};\mathbb{Z}_p) \cong H^3(\widetilde{W};\mathbb{Z}_p) \cong \mathrm{Hom}(H_3(\widetilde{W};\mathbb{Z}), \mathbb{Z}_p) \oplus \mathrm{Ext}(H_2(\widetilde{W};\mathbb{Z}), \mathbb{Z}_p).$$

Thus $H_3(\widetilde{W};\mathbb{Z})$ is a torsion group. Thus, if we apply Poincaré duality and the universal coefficient theorem as above, but with $\mathbb{Q}$–coefficients, we obtain

$$H_1(\widetilde{W}, \widetilde{Y};\mathbb{Q}) \cong H^3(\widetilde{W};\mathbb{Q}) \cong \mathrm{Hom}(H_3(\widetilde{W};\mathbb{Z}), \mathbb{Q}) \oplus \mathrm{Ext}(H_2(\widetilde{W};\mathbb{Z}), \mathbb{Q}) = 0.$$

Thus the map $H_1(\widetilde{Y};\mathbb{Q}) \to H_1(\widetilde{W};\mathbb{Q})$ induced by inclusion is surjective. Since $\widetilde{Y}$ is a $\mathbb{Q}S^1 \times S^2$, it follows that $\mathrm{rank}(H_1(\widetilde{W};\mathbb{Q})) \le 1$. Finally, since $\chi(\widetilde{W}) = p\chi(W) = 0$ and $H_3(\widetilde{W};\mathbb{Q}) = 0$, we necessarily have that $H_1(\widetilde{W};\mathbb{Q}) = \mathbb{Q}$ and $H_2(\widetilde{W};\mathbb{Q}) = 0$, proving that $\widetilde{W}$ is indeed a $\mathbb{Q}S^1 \times B^3$. □

Proposition 3.14 *Let $T_{\pm A(\boldsymbol{a})}$ be a hyperbolic torus bundle that bounds a $\mathbb{Q}S^1 \times B^3$, say W. If p is an odd prime that does not divide $|\mathrm{Tor}(H_2(W;\mathbb{Z}))|$, then $T_{\pm A(\boldsymbol{a}^p)}$ bounds a $\mathbb{Q}S^1 \times B^3$.*

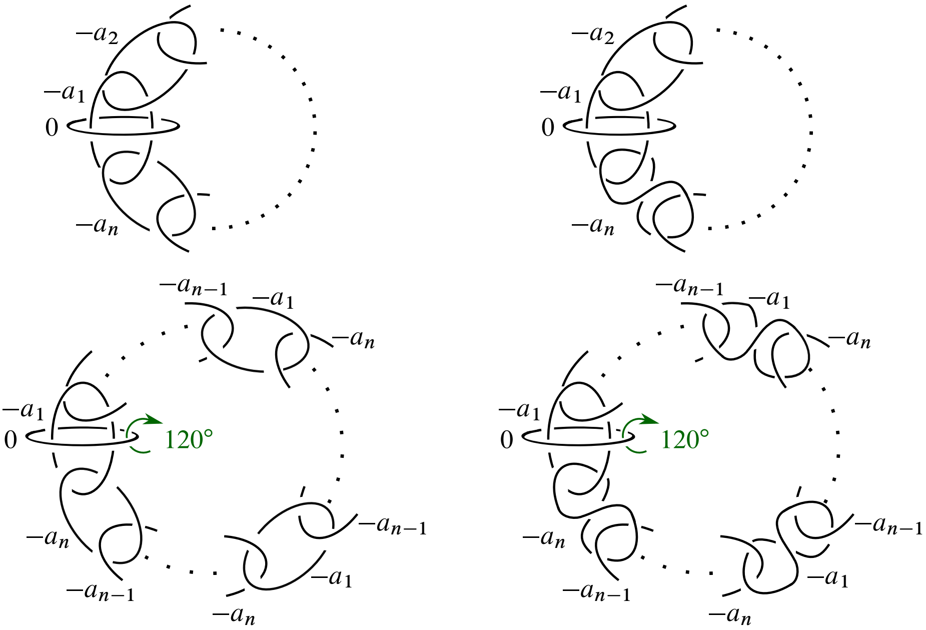


Figure 8: Surgery diagrams for $\boldsymbol{T}_{A(\boldsymbol{a})}$ (top left), $\boldsymbol{T}_{-A(\boldsymbol{a})}$ (top right), $\boldsymbol{T}_{A(\boldsymbol{a}^3)}$ (bottom left) and $\boldsymbol{T}_{-A(\boldsymbol{a}^3)}$ (bottom right). $\boldsymbol{T}_{\pm A(\boldsymbol{a}^3)}$ is a 3–fold cyclic cover of $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$. There is an obvious $\mathbb{Z}_3$–action on $\boldsymbol{T}_{\pm A(\boldsymbol{a}^3)}$ given by a rotation of 120° through the 0–framed unknot. The quotient of $\boldsymbol{T}_{\pm A(\boldsymbol{a}^3)}$ by this action is $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$.

Proof Let W be a $\mathbb{Q}S^1 \times B^3$ bounded by some negative hyperbolic torus bundle $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$, where $\boldsymbol{a} = (a_1, \dots, a_n)$. Let p be an odd prime number that is not a factor of $|\mathrm{Tor}(H_2(W;\mathbb{Z}))|$. Consider the obvious surgery diagrams of $\boldsymbol{T}_{A(\boldsymbol{a})}$ and $\boldsymbol{T}_{-A(\boldsymbol{a})}$ as in Figure 8, top. In both diagrams, let μ_i denote the homology class of the meridian of the $-a_i$–framed surgery curve and let μ_0 denote the homology class of the meridian of the 0–framed surgery curve. Then $H_1(\boldsymbol{T}_{\pm A(\boldsymbol{a})};\mathbb{Z})$ is generated by $\mu_0, \dots, \mu_n$.

Consider the torus bundle $\boldsymbol{T}_{-A(\boldsymbol{a}^p)}$, which has monodromy $-(T^{-a_1}S \cdots T^{-a_n}S)^p$. The standard surgery diagram of this torus bundle includes a -1–half-twisted chain link (as in Table 1). Note that, by sliding the chain link over the 0–framed unknot $\frac{1}{2}(p-1)$ times, we may arrange that the chain link has $-p$ half-twists, as in Figure 8, bottom right (for the case $p=3$). For the torus bundle $\boldsymbol{T}_{A(\boldsymbol{a}^p)}$, which has monodromy $(T^{-a_1}S \cdots T^{-a_n}S)^p$, consider the standard surgery diagram shown in Figure 8, bottom left (for the case $p=3$). There is an obvious $\mathbb{Z}_p$–action on $\boldsymbol{T}_{\pm A(\boldsymbol{a}^p)}$ obtained by rotating the chain link through the 0–framed unknot by an angle of $2\pi/p$, as indicated in Figure 8, bottom. The quotient of $\boldsymbol{T}_{\pm A(\boldsymbol{a}^p)}$ by this action is clearly $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$ and the induced map $f\colon H_1(\boldsymbol{T}_{\pm A(\boldsymbol{a})};\mathbb{Z}) \to \mathbb{Z}_p$ satisfies $f(\mu_0)=1$ and $f(\mu_i)=0$ for all $1\le i\le n$. Consider the long exact sequence of the pair $(W,\boldsymbol{T}_{\pm A(\boldsymbol{a})})$,

$$H_1(\boldsymbol{T}_{\pm A(\boldsymbol{a})};\mathbb{Z}) \xrightarrow{i_*} H_1(W;\mathbb{Z}) \to H_1(W,\boldsymbol{T}_{\pm A(\boldsymbol{a})};\mathbb{Z}) \to 0.$$

Choose a basis $\{m_0,m_1,\dots,m_k\}$ for $H_1(W;\mathbb{Z})$ such that m_0 has infinite order and m_i is a torsion element for all $1\le i\le k$. Since $H_1(W,\boldsymbol{T}_{\pm A(\boldsymbol{a})};\mathbb{Z})$ is a torsion group, $i_*(\mu_0)=\alpha m_0+\sum_{i=1}^k \beta_i m_i$ for some $\alpha,\beta_i\in\mathbb{Z}$, where $\alpha\ne 0$. Since p is not relatively prime to $|\mathrm{Tor}(H_2(W;\mathbb{Z}))| = |H_1(W,\boldsymbol{T}_{\pm A(\boldsymbol{a})};\mathbb{Z})|$ and α divides $|H_1(W,\boldsymbol{T}_{\pm A(\boldsymbol{a})};\mathbb{Z})|$, it follows that α and p are relatively prime; thus there exists an integer t such that $t\alpha\equiv 1 \bmod p$. Define a map $g\colon H_1(W;\mathbb{Z})\to\mathbb{Z}_p$ by $g(m_0)=t$ and $g(m_i)=0$ for all $1\le i\le k$. Then g is a surjective homomorphism satisfying $f=g\circ i_*$. Let $\widetilde{W}$ be the p–fold cyclic cover of W induced by g. Then $\partial\widetilde{W}=\boldsymbol{T}_{\pm A(\boldsymbol{a}^p)}$ and, by Lemma 3.13, $\widetilde{W}$ is a $\mathbb{Q}S^1\times B^3$. □

The two following corollaries conclude the proof of Theorem 1.1.

Corollary 3.15 *No negative hyperbolic torus bundle bounds a $\mathbb{Q}S^1\times B^3$.*

Proof Let $\boldsymbol{T}_{-A(\boldsymbol{a})}$ be a negative hyperbolic torus bundle that bounds a $\mathbb{Q}S^1\times B^3$, say W. Let $p>3$ be an odd prime number that is not a factor of $|\mathrm{Tor}(H_2(W;\mathbb{Z}))|$. By Proposition 3.14, $\boldsymbol{T}_{-A(\boldsymbol{a}^p)}$ also bounds a $\mathbb{Q}S^1\times B^3$. Let $\boldsymbol{d}$ be the cyclic-dual of $\boldsymbol{a}$; by Lemma 3.6, $\boldsymbol{d}^p$ is the linear-dual of $\boldsymbol{a}^p$. By Lemma 1.2, $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}^p}^{-1}$ bound $\mathbb{Q}B^4$'s and so, by Theorem 1.7, $\boldsymbol{a}$ or $\boldsymbol{d}$ belongs to $\mathcal{S}_1\cup\mathcal{O}$ and $\boldsymbol{a}^p$ or $\boldsymbol{d}^p$ belongs to $\mathcal{S}_1\cup\mathcal{O}$.

First assume $\boldsymbol{a},\boldsymbol{a}^p\in\mathcal{S}_1\cup\mathcal{O}$. By Remark 1.6, $-4\le I(\boldsymbol{a}),I(\boldsymbol{a}^p)\le 0$. Moreover, $I(\boldsymbol{a}^p)=pI(\boldsymbol{a})$. If $I(\boldsymbol{a})<0$, then, since $p>3$, we have $I(\boldsymbol{a}^p)<-4$, which is a contradiction. Thus $I(\boldsymbol{a}^p)=I(\boldsymbol{a})=0$. By Remark 1.6, $\boldsymbol{a},\boldsymbol{a}^p\in\mathcal{S}_{2a}\cup\mathcal{S}_{2b}\cup\mathcal{S}_{2c}\cup\mathcal{O}$. Since $\mathcal{S}_1\cap\mathcal{S}_2=\varnothing$, by Lemma 3.4, we necessarily have that $\boldsymbol{a},\boldsymbol{a}^p\in\mathcal{O}$, which is not possible since $p\ne 1$.

Next assume $\boldsymbol{a},\boldsymbol{d}^p\in\mathcal{S}_1\cup\mathcal{O}$. By Remark 1.6, $-4\le I(\boldsymbol{a}),I(\boldsymbol{d}^p)\le 0$. Since $I(\boldsymbol{d}^p)=pI(\boldsymbol{d})=-pI(\boldsymbol{a})$, we necessarily have that $I(\boldsymbol{a})=I(\boldsymbol{d}^p)=0$. As above, this implies that $\boldsymbol{a},\boldsymbol{d}^p\in\mathcal{O}$. But, since $\boldsymbol{a}\in\mathcal{O}$, it is clear that $\boldsymbol{a}=\boldsymbol{d}$ and thus $\boldsymbol{d}\in\mathcal{O}$. As above, it is clear that $\boldsymbol{d}$ and $\boldsymbol{d}^p$ cannot both be contained in $\mathcal{O}$.

Finally, if $\boldsymbol{d},\boldsymbol{d}^p\in\mathcal{S}_1\cup\mathcal{O}$ or $\boldsymbol{d},\boldsymbol{a}^p\in\mathcal{S}_1\cup\mathcal{O}$, similar arguments provide similar contradictions. Therefore, ∂W cannot be a negative hyperbolic torus bundle. □

Corollary 3.16 *If a positive hyperbolic torus bundle $T_{A(\boldsymbol{a})}$ bounds a $\mathbb{Q}S^1 \times B^3$, then $\boldsymbol{a} \in \mathcal{S}_{2c}$.*

Proof Let $T_{A(\boldsymbol{a})}$ be a positive hyperbolic torus bundle that bounds a $\mathbb{Q}S^1 \times B^3$, say W, and let $p > 3$ be an odd prime number that is not a factor of $|\mathrm{Tor}(H_2(W;\mathbb{Z}))|$. Following as in the proof of Corollary 3.15, $\boldsymbol{a}$ or $\boldsymbol{d}$ belongs to $\mathcal{S}_2$ and $\boldsymbol{a}^p$ or $\boldsymbol{d}^p$ belongs to $\mathcal{S}_2$, where $\boldsymbol{d}$ is the cyclic-dual of $\boldsymbol{a}$. Suppose $\boldsymbol{a}, \boldsymbol{a}^p \in \mathcal{S}_2$. As in the proof of Corollary 3.15, $I(\boldsymbol{a}) = I(\boldsymbol{a}^p) = 0$ and so, by Remark 1.6, $\boldsymbol{a}, \boldsymbol{a}^p \in \mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$. By Corollary 3.11, $\boldsymbol{a} \in \mathcal{S}_{2c}$. Next suppose $\boldsymbol{a}, \boldsymbol{d}^p \in \mathcal{S}_2$. Once again, following the argument in Corollary 3.15, $I(\boldsymbol{a}) = I(\boldsymbol{d}^p) = 0$ and so, by Remark 1.6, $\boldsymbol{a}, \boldsymbol{d}^p \in \mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$. By Lemma 3.12, we necessarily have that $\boldsymbol{a}^p \in \mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c}$; proceeding as in the previous case, we find $\boldsymbol{a} \in \mathcal{S}_{2c}$. Finally, if $\boldsymbol{d}, \boldsymbol{a}^p \in \mathcal{S}_2$ or $\boldsymbol{d}, \boldsymbol{d}^p \in \mathcal{S}_2$, we can similarly deduce that $\boldsymbol{a} \in \mathcal{S}_{2c}$. □

4 Surgeries on chain links bounding rational homology 4–balls

In this section, we will prove the necessary conditions of Theorem 1.7. Namely, we will show that the $\mathbb{Q}S^3$'s of Theorem 1.7 bound $\mathbb{Q}B^4$'s by explicitly constructing such $\mathbb{Q}B^4$'s via Kirby calculus. Notice that the necessary condition of Theorem 1.7(2) follows from the necessary condition of Theorem 1.7(1) in light of Lemma 2.3. Therefore, we need only show the following three cases (where $\boldsymbol{a}$ and $\boldsymbol{d}$ are cyclic-duals):

- If $\boldsymbol{a} \in \mathcal{S}_{1a}$, then $Y_{\boldsymbol{a}}^{-1}$ bounds a $\mathbb{Q}B^4$.
- If $\boldsymbol{a} \in \mathcal{S}_{1b} \cup \mathcal{S}_{1c} \cup \mathcal{S}_{1d} \cup \mathcal{S}_{1e}$, then $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{d}}^{-1}$ bound $\mathbb{Q}B^4$'s.
- If $\boldsymbol{a} \in \mathcal{S}_2$, then $Y_{\boldsymbol{a}}^0$ and $Y_{\boldsymbol{d}}^0$ bound $\mathbb{Q}B^4$'s.

Figures 9–15 exhibit the Kirby calculus needed to produce these $\mathbb{Q}B^4$'s. We will describe in detail the $\mathbb{Q}B^4$ constructed in Figure 9, top. The constructions in the other cases are similar. Notice that the top figure of Figure 9, top (without the -1–framed blue unknot) is a surgery diagram for $Y_{\boldsymbol{a}}^{-1}$, where $\boldsymbol{a} = (b_1, \dots, b_k, 2, c_l, \dots, c_1, 2) \in \mathcal{S}_{1a}$. Thicken $Y_{\boldsymbol{a}}^{-1}$ to the 4–manifold $Y_{\boldsymbol{a}}^{-1} \times [0,1]$. By attaching a -1–framed 2–handle to $Y_{\boldsymbol{a}}^{-1} \times \{1\}$ along the blue unknot in Figure 9, top, we obtain a 2–handle cobordism from $Y_{\boldsymbol{a}}^{-1}$ to a new 3–manifold, which we will show is $S^1 \times S^2$. By performing a blowdown, we obtain the middle surgery diagram. Blowing down a second time, the surgery curves with framings $-b_1$ and $-c_1$ link each other once and have framings $-(b_1 - 1)$

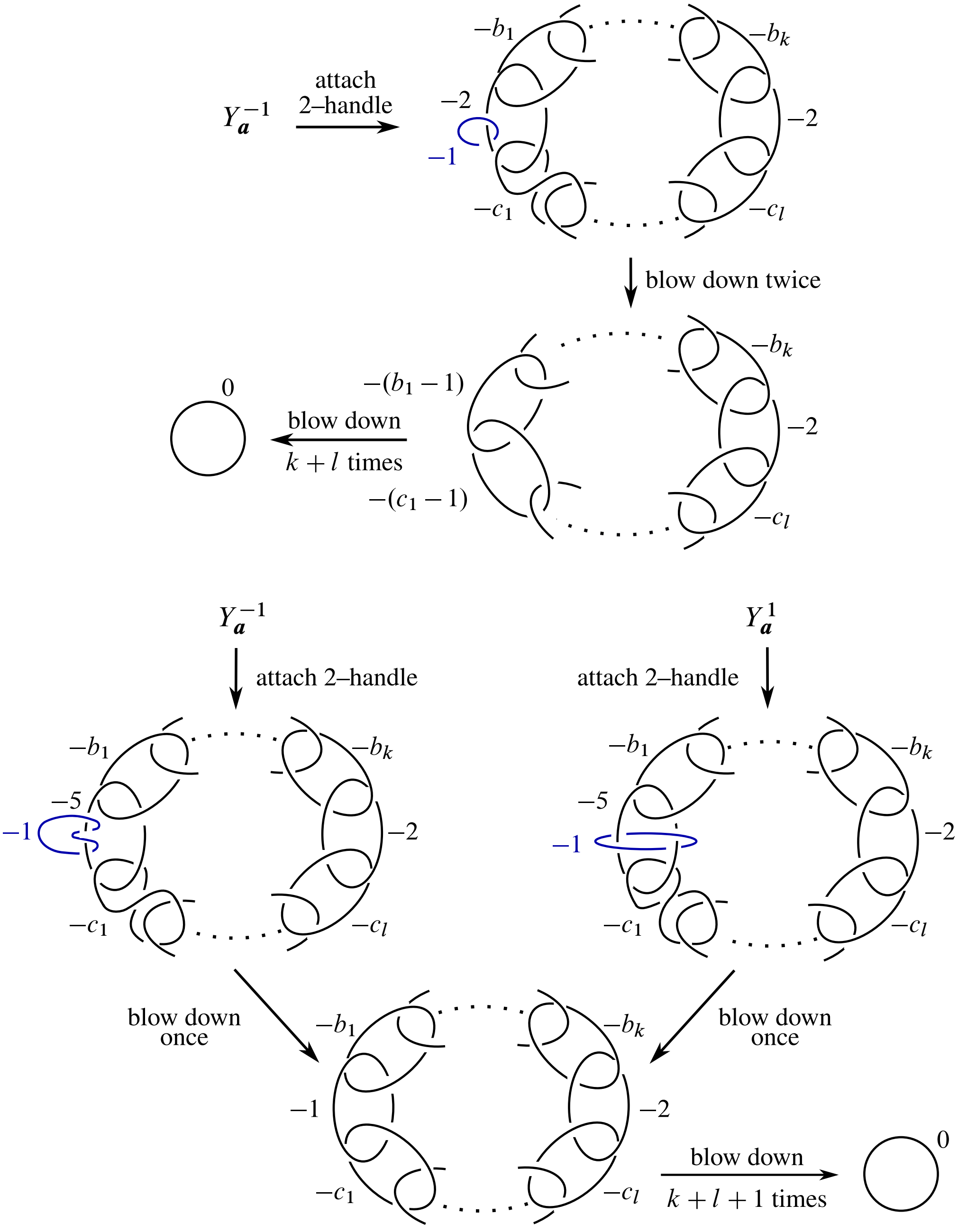


Figure 9: With Figures 10–12, we show the 3–manifolds in Theorem 1.7(1)–(2) bound rational balls. Top: if $\boldsymbol{a} \in \mathcal{S}_{1a}$, then $Y_{\boldsymbol{a}}^{-1}$ bounds a $\mathbb{Q}B^4$. Bottom: if $\boldsymbol{a} \in \mathcal{S}_{1b}$, then $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^{1}$ bound $\mathbb{Q}B^4$'s.

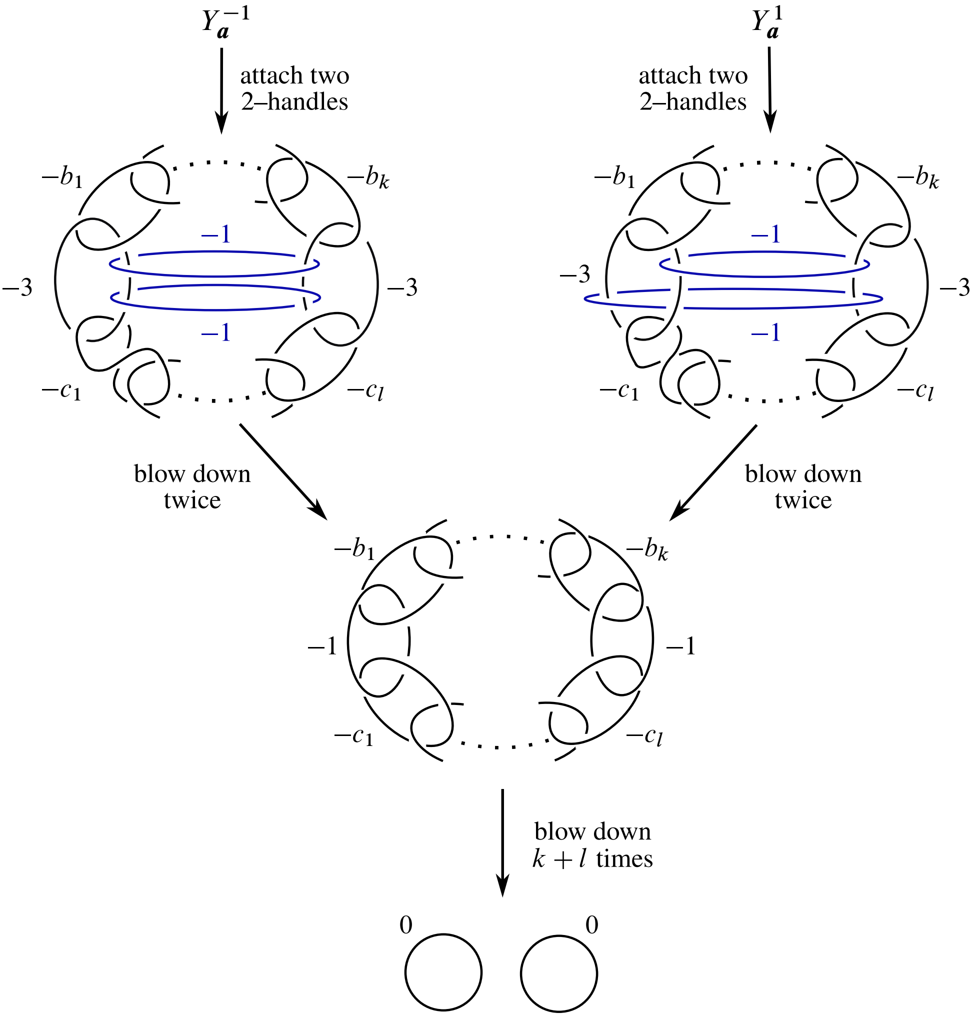


Figure 10: If $\boldsymbol{a} \in \mathcal{S}_{1c}$, then $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^{1}$ bound $\mathbb{Q}B^4$'s.

and $-(c_1-1)$, respectively. Since $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$ are linear-dual, either $-(b_1-1)$ or $-(c_1-1)$ is equal to -1. We can thus blow down again. Continuing in this way, we can continue to blow down -1–framed unknots until we obtain 0–surgery on the unknot, which is shown on the right side of the figure. Thus we have a 2–handle cobordism from $Y_{\boldsymbol{a}}^{-1}$ to $S^1\times S^2$. By gluing this cobordism to $S^1\times B^3$, we obtain the desired $\mathbb{Q}B^4$ bounded by $Y_{\boldsymbol{a}}^{-1}$.

Suppose $\boldsymbol{a} \in \mathcal{S}_{1b}\cup\mathcal{S}_{1c}\cup\mathcal{S}_{1d}\cup\mathcal{S}_{1e}$ and let $\boldsymbol{d}$ be its cyclic-dual. Then, by Lemma 2.3, $\overline{Y}_{\boldsymbol{d}}^{-1} = Y_{\boldsymbol{a}}^{1}$. To show that $Y_{\boldsymbol{d}}^{-1}$ bounds a $\mathbb{Q}B^4$, we will show that $Y_{\boldsymbol{a}}^{1}$ bounds a $\mathbb{Q}B^4$.

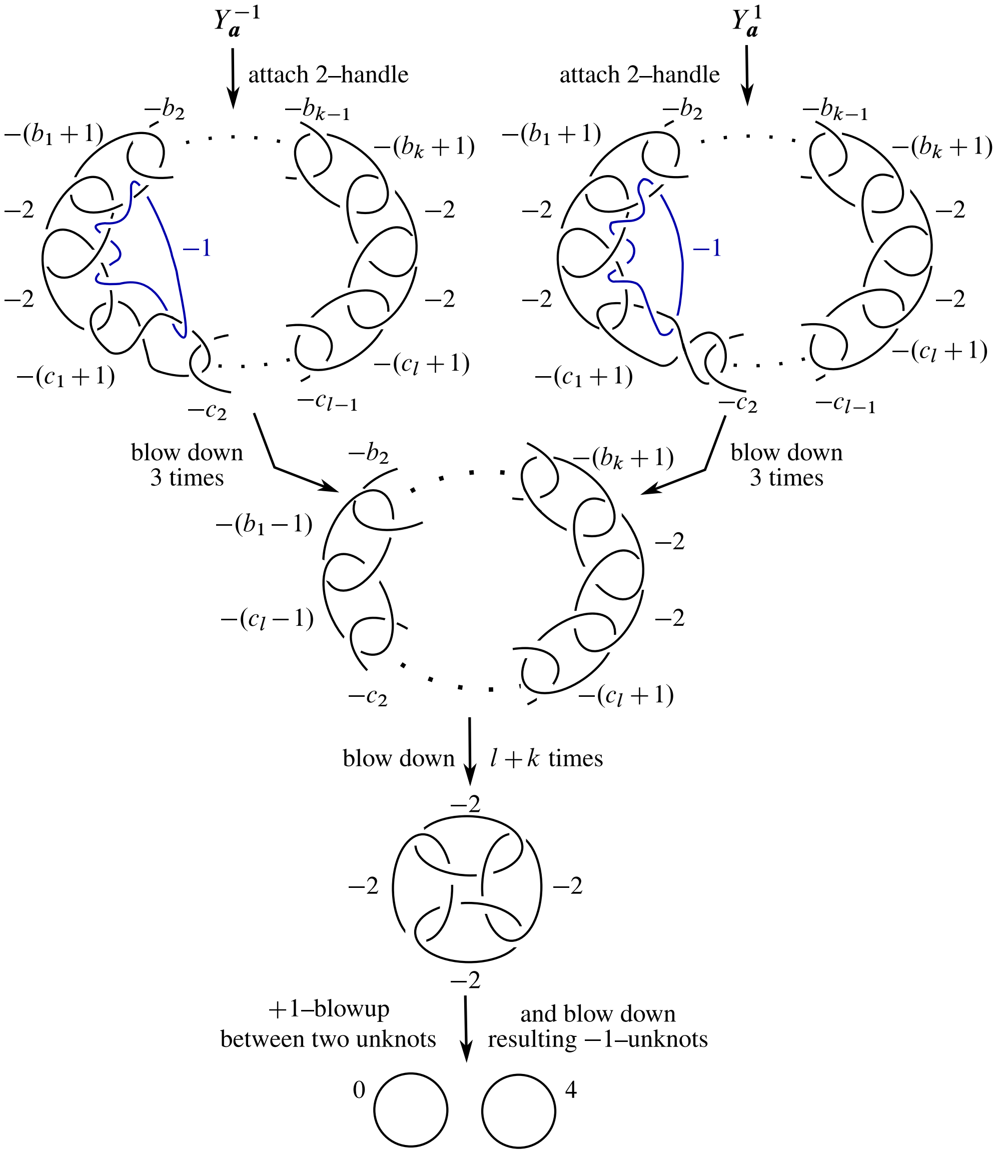


Figure 11: If $\boldsymbol{a} \in \mathcal{S}_{1d}$, then $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^{1}$ bound $\mathbb{Q}B^4$'s.

Figures 9–12 show that, if $\boldsymbol{a} \in \mathcal{S}_{1b} \cup \mathcal{S}_{1c} \cup \mathcal{S}_{1d} \cup \mathcal{S}_{2e}$, then $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^{1}$ bound $\mathbb{Q}B^4$'s. Note that Figure 9, bottom, depicts a cobordism similar to the one constructed in Figure 9, top, which was described in the previous paragraph. However, the cobordisms constructed in Figures 10–12 are slightly different. In Figure 11, we have a 2–handle cobordism from $Y_{\boldsymbol{a}}^{\pm 1}$ to $S^1 \times S^2 \# L(-4,1)$, which bounds a $\mathbb{Q}S^1 \times B^3$,

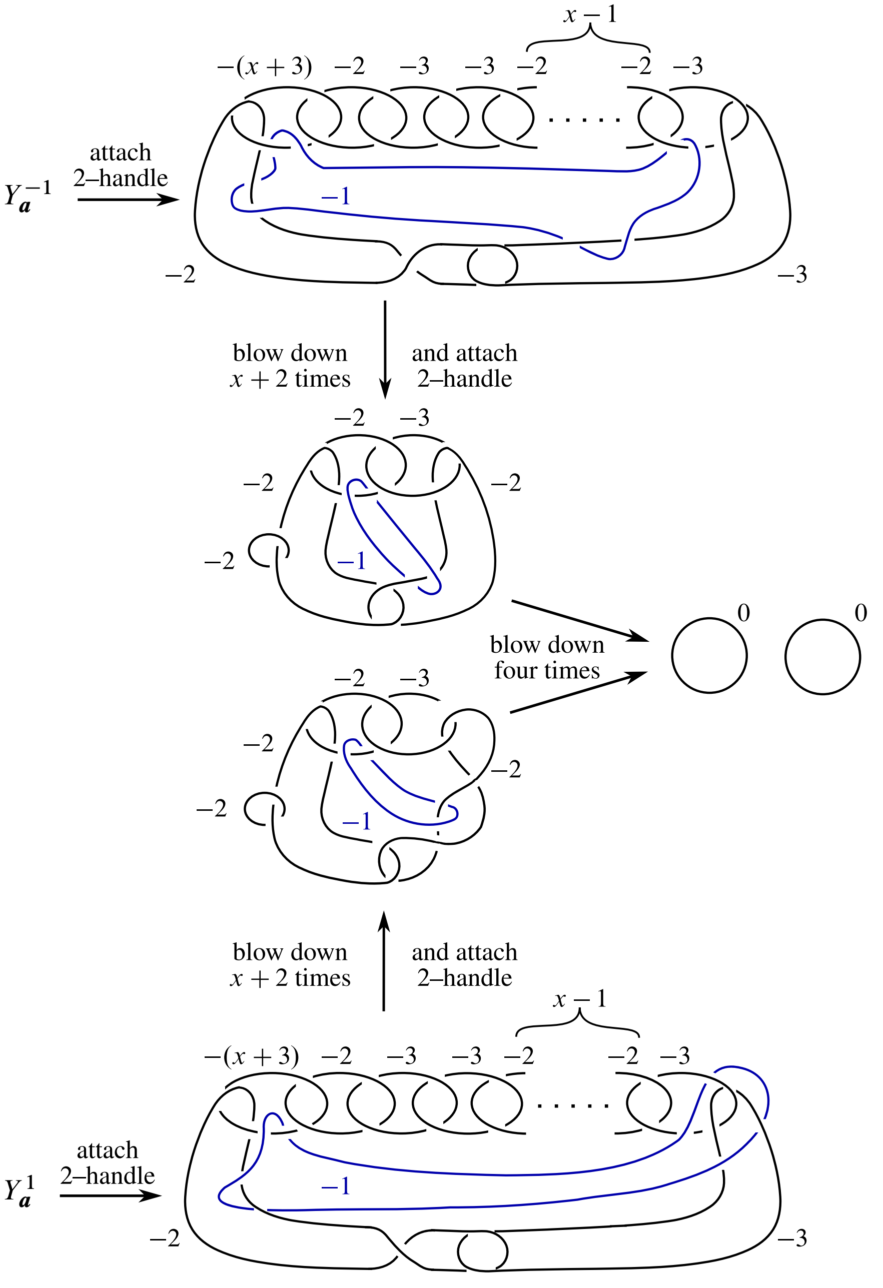


Figure 12: If $\boldsymbol{a} \in \mathcal{S}_{1e}$, then $Y_{\boldsymbol{a}}^{-1}$ and $Y_{\boldsymbol{a}}^{1}$ bound $\mathbb{Q}B^4$'s.

since $L(-4,1)$ bounds a $\mathbb{Q}B^4$ [8]. Gluing this $\mathbb{Q}S^1 \times B^3$ to the cobordism yields the desired $\mathbb{Q}B^4$. The cobordisms depicted in Figures 10 and 12 are built out of two

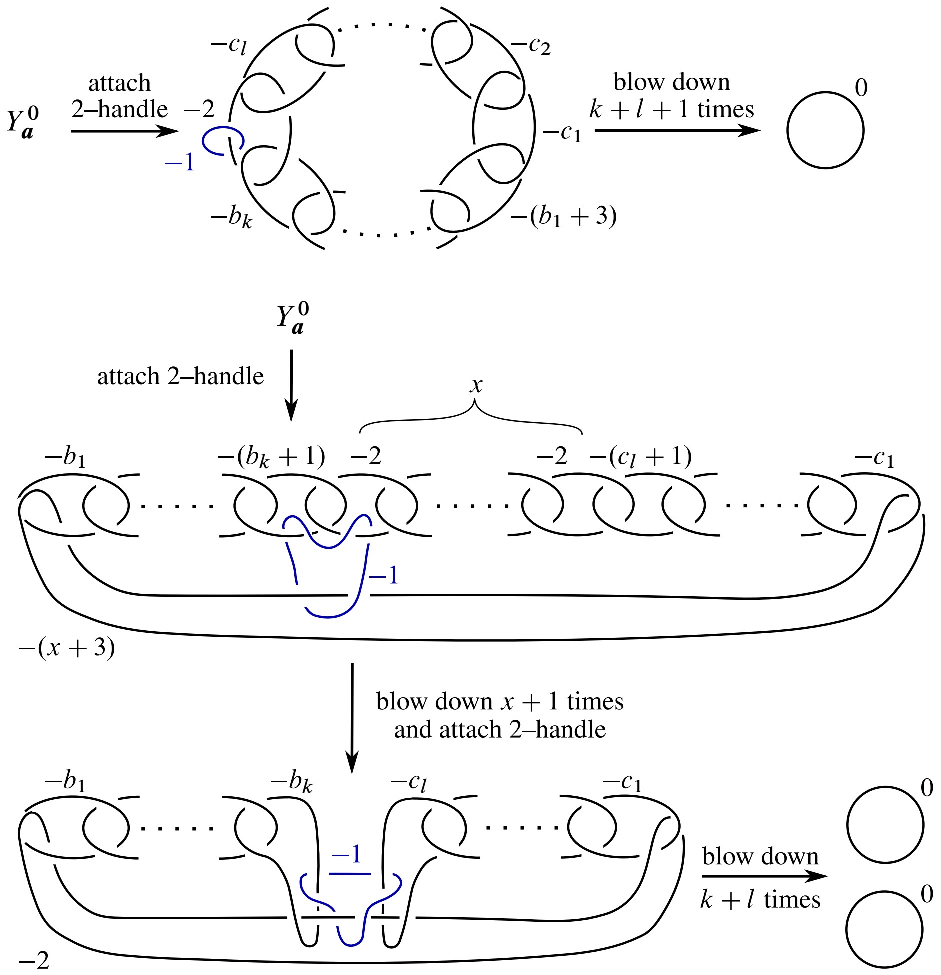


Figure 13: With Figures 14–15, we show the 3–manifolds in Theorem 1.7(3) bound rational balls Top: if $\boldsymbol{a} \in \mathcal{S}_{2a}$, then $Y^0_{\boldsymbol{a}}$ bounds a $\mathbb{Q}B^4$. Bottom: if $\boldsymbol{a} \in \mathcal{S}_{2b}$, then $Y^0_{\boldsymbol{a}}$ bounds a $\mathbb{Q}B^4$.

2–handles. These cobordisms are from $Y^{\pm 1}_{\boldsymbol{a}}$ to $S^1 \times S^2 \# S^1 \times S^2$. Gluing these cobordisms to $S^1 \times B^3 \natural S^1 \times B^3$ yields the desired $\mathbb{Q}B^4$'s.

Lastly, suppose $\boldsymbol{a} \in \mathcal{S}_2$. By Lemma 2.3, $\overline{Y}^0_{\boldsymbol{a}} = Y^0_{\boldsymbol{d}}$. Thus, once we show that $Y^0_{\boldsymbol{a}}$ bounds a $\mathbb{Q}B^4$, it will follow that $Y^0_{\boldsymbol{d}}$ also bounds a $\mathbb{Q}B^4$. Figures 13–15 show that, if $\boldsymbol{a} \in \mathcal{S}_2$, then $Y^0_{\boldsymbol{a}}$ bounds a $\mathbb{Q}B^4$. The $\mathbb{Q}B^4$'s in almost all of the cases are constructed in very similar ways as in the negative cases. The last case, $Y^0_{(2,2,2,3)}$, is much simpler; Figure 15, bottom, shows that $Y^0_{(2,2,2,3)} = L(-4,1)$, which bounds a $\mathbb{Q}B^4$.

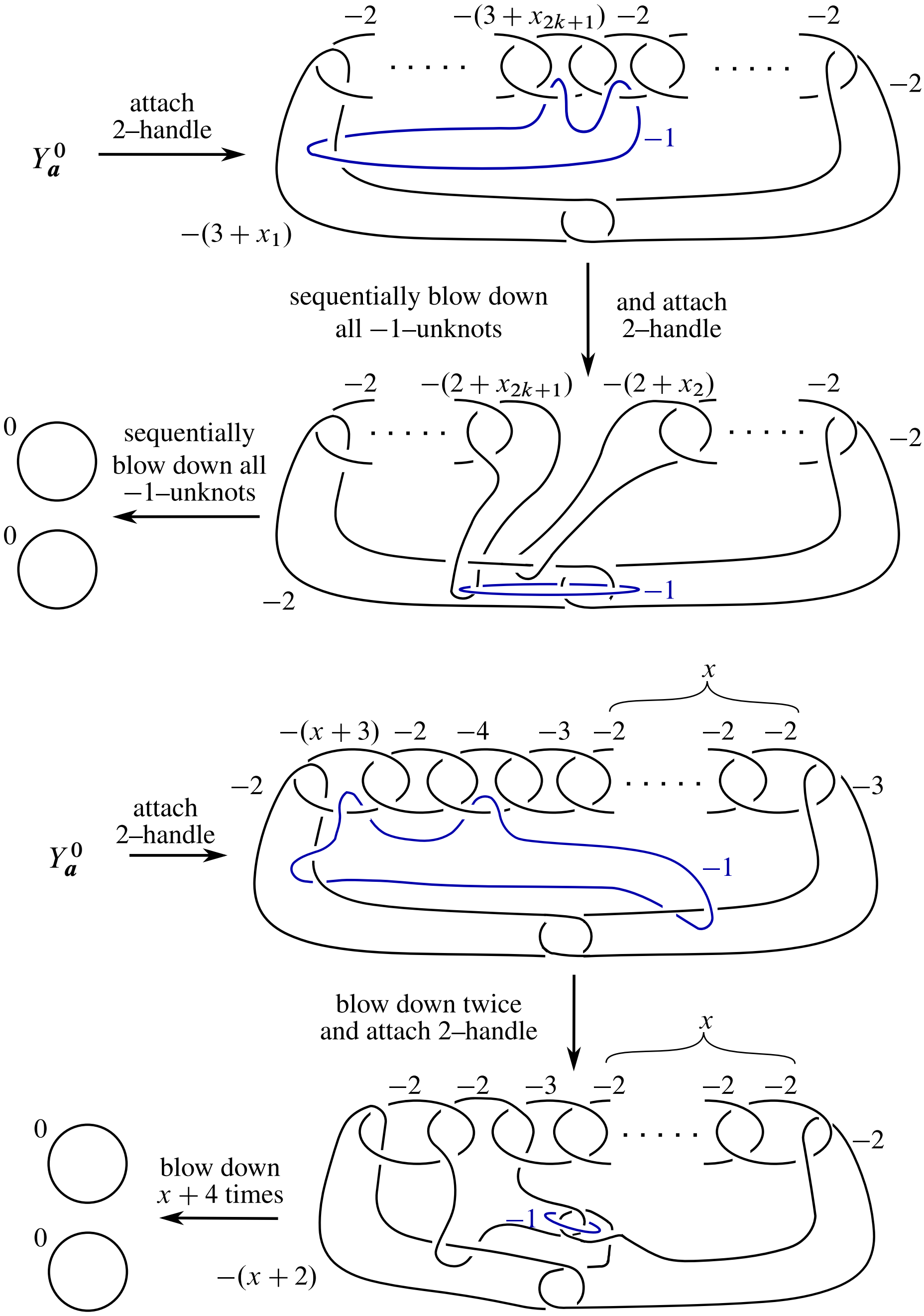


Figure 14: Top: if $\boldsymbol{a} \in \mathcal{S}_{2c}$, then $Y_{\boldsymbol{a}}^0$ bounds a $\mathbb{Q}B^4$. Bottom: if $\boldsymbol{a} \in \mathcal{S}_{2d}$, then $Y_{\boldsymbol{a}}^0$ bounds a $\mathbb{Q}B^4$.

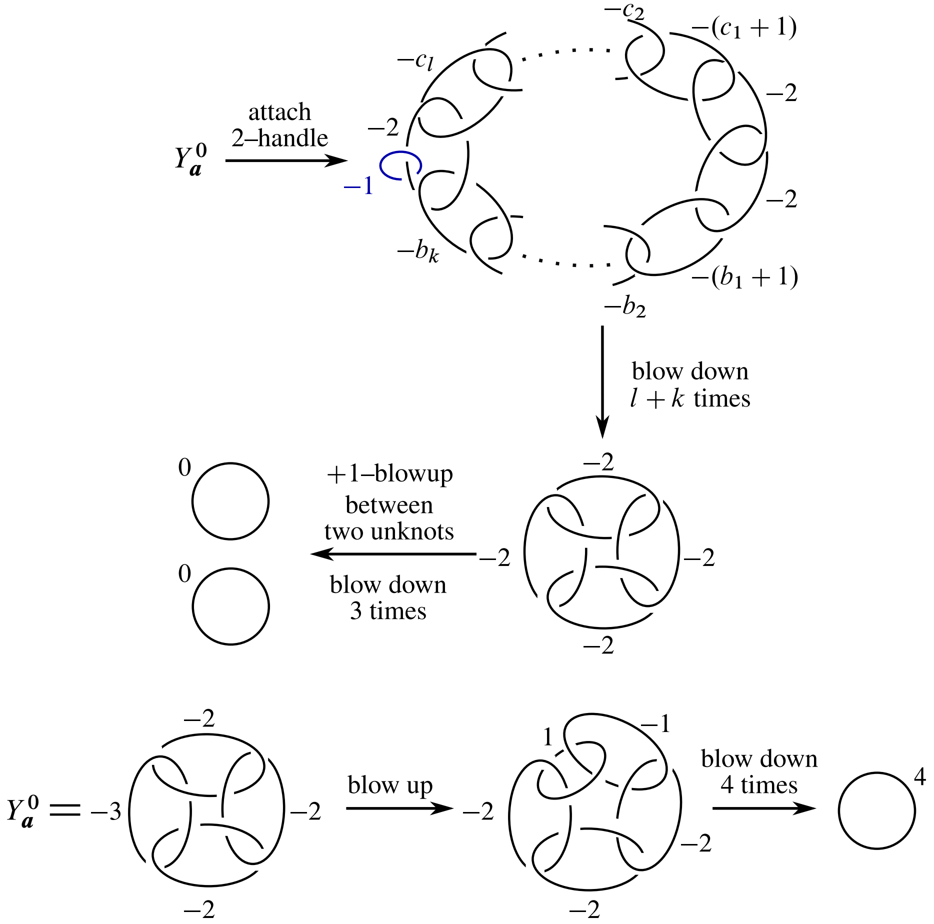


Figure 15: Top: if $\boldsymbol{a} \neq (3,2,2,2) \in \mathcal{S}_{2e}$, then $Y_{\boldsymbol{a}}^0$ bounds a $\mathbb{Q}B^4$. Bottom: if $\boldsymbol{a} = (3,2,2,2) \in \mathcal{S}_{2e}$, then $Y_{\boldsymbol{a}}^0$ bounds a $\mathbb{Q}B^4$.

As shown above, if $\boldsymbol{a} \in \mathcal{S}_{1b} \cup \mathcal{S}_{1c} \cup \mathcal{S}_{1d} \cup \mathcal{S}_{1e}$, then $Y_{\boldsymbol{d}}^{-1}$ bounds a $\mathbb{Q}B^4$. However, as the next results will show, if $\boldsymbol{a} \in \mathcal{S}_{1a}$, then $Y_{\boldsymbol{d}}^{-1}$ does not necessarily bound a $\mathbb{Q}B^4$. The key is that $|H_1(Y_{\boldsymbol{a}}^{-1})|$ can be either even or odd when $\boldsymbol{a} \in \mathcal{S}_{1a}$, but, in all other cases, $H_1(Y_{\boldsymbol{a}}^{-1})$ has even order. Recall that $[b_1,\dots,b_k]$ represents the Hirzebruch–Jung continued fraction (see the appendix for details).

Proposition 4.1 *Let* $\boldsymbol{a} = (b_1,\dots,b_k,2,c_l,\dots,c_1,2) \in \mathcal{S}_{1a}$, *where* $[b_1,\dots,b_k] = p/q$. *Then* $|H_1(Y_{\boldsymbol{a}}^{-1})| = |\mathrm{Tor}(H_1(\boldsymbol{T}_{-A(\boldsymbol{a})}))| = p^2$.

Proof See Proposition A.3. □

Lemma 4.2 *Let $\boldsymbol{a}=(2,b_1,\dots,b_k,2,c_l,\dots,c_1)\in\mathcal{S}_{1a}$, where $[b_1,\dots,b_k]=p/q$, and let $\boldsymbol{d}=(d_1,\dots,d_m)$ be the cyclic-dual of $\boldsymbol{a}$. If p is odd, then $Y_{\boldsymbol{d}}^{-1}$ and $Y_{\boldsymbol{a}}^{1}$ do not bound $\mathbb{Q}B^4$'s.*

Proof By Lemma 2.3, $\overline{Y_{\boldsymbol{d}}^{-1}}=Y_{\boldsymbol{a}}^{1}$, so it suffices to show that $Y_{\boldsymbol{a}}^{1}$ does not bound a $\mathbb{Q}B^4$. Since $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$ are linear-dual strings, it is clear that $\frac{1}{4}I(\boldsymbol{a})=-1$ (see Remark 1.6). By the calculations in Section 2.1, $d(Y_{\boldsymbol{a}}^{1},\mathfrak{s}_0)=1-\frac{1}{4}I(\boldsymbol{a})=2$. Since p is odd, by Proposition 4.1, $|H_1(Y_{\boldsymbol{a}}^{1})|=|H_1(Y_{\boldsymbol{a}}^{-1})|$ has odd order and so $\mathfrak{s}_0$ extends over any $\mathbb{Q}B^4$ bounded by $Y_{\boldsymbol{a}}^{1}$. Thus, if $Y_{\boldsymbol{a}}^{1}$ bounds a $\mathbb{Q}B^4$, then $d(Y_{\boldsymbol{a}}^{1},\mathfrak{s}_0)=0$, which is not possible. □

Remark 4.3 By Lemma 1.2 and Theorem 1.1, we already know that, if $\boldsymbol{a}\in\mathcal{S}_{2c}$, then $Y_{\boldsymbol{a}}^{0}$ bounds a $\mathbb{Q}B^4$. However, by [13], the $\mathbb{Q}B^4$'s constructed via Theorem 1.1 necessarily admit handlebody decompositions with 3–handles. On the other hand, the $\mathbb{Q}B^4$'s constructed in this section do not contain 3–handles. Thus $Y_{\boldsymbol{a}}^{0}$ bounds a $\mathbb{Q}B^4$ without 3–handles, even though $T_{A(\boldsymbol{a})}$ only bounds $\mathbb{Q}S^1\times B^3$'s containing 3–handles.

5 Cyclic subsets

The remainder of the sections are dedicated to proving the sufficient conditions of Theorem 1.7. In fact, we will prove something more general. We will show that if t is odd and $Y_{\boldsymbol{a}}^{t}$ bounds a $\mathbb{Q}B^4$, then $\boldsymbol{a}\in\mathcal{S}_1\cup\mathcal{O}$ or $\boldsymbol{d}\in\mathcal{S}_1\cup\mathcal{O}$, and if t is even and $Y_{\boldsymbol{a}}^{t}$ bounds a $\mathbb{Q}B^4$, then $\boldsymbol{a}\in\mathcal{S}_2$ or $\boldsymbol{d}\in\mathcal{S}_2$. For convenience, we recall the definition of these sets.

Definition 1.4 Two strings are considered to be equivalent if one is a cyclic reordering and/or reverse of the other. Each string in the following sets is defined up to this equivalence. Moreover, strings of the form $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$ are assumed to be linear-dual. We define

$$\mathcal{S}_{1a}=\{(b_1,\dots,b_k,2,c_l,\dots,c_1,2)\mid k+l\geq 3\},$$

$$\mathcal{S}_{1b}=\{(b_1,\dots,b_k,2,c_l,\dots,c_1,5)\mid k+l\geq 2\},$$

$$\mathcal{S}_{1c}=\{(b_1,\dots,b_k,3,c_l,\dots,c_1,3)\mid k+l\geq 2\},$$

$$\mathcal{S}_{1d}=\{(2,b_1+1,b_2,\dots,b_{k-1},b_k+1,2,2,c_l+1,c_{l-1},\dots,c_2,c_1+1,2)\mid k+l\geq 2\},$$

$$\mathcal{S}_{1e}=\{(2,3+x,2,3,3,2^{[x-1]},3,3)\mid x\geq 0\text{ and }(3,2^{[-1]},3):=(4)\},$$

$$\mathcal{S}_{2a}=\{(b_1+3,b_2,\dots,b_k,2,c_l,\dots,c_1)\},$$

$$\mathcal{S}_{2b}=\{(3+x,b_1,\dots,b_{k-1},b_k+1,2^{[x]},c_l+1,c_{l-1},\dots,c_1)\mid x\geq 0\text{ and }k+l\geq 2\},$$

$$\begin{aligned}
\mathcal{S}_{2c} &= \{(b_1+1, b_2, \dots, b_{k-1}, b_k+1, c_1, \dots, c_l) \mid k+l \geq 2\},\\
\mathcal{S}_{2d} &= \{(2, 2+x, 2, 3, 2^{[x-1]}, 3, 4) \mid x \geq 0 \text{ and } (3, 2^{[-1]}, 3) := (4)\},\\
\mathcal{S}_{2e} &= \{(2, b_1+1, b_2, \dots, b_k, 2, c_l, \dots, c_2, c_1+1, 2), (2,2,2,3) \mid k+l \geq 2\},\\
\mathcal{O} &= \{(6,2,2,2,6,2,2,2), (4,2,4,2,4,2,4,2), (3,3,3,3,3,3)\},\\
\mathcal{S}_1 &= \mathcal{S}_{1a} \cup \mathcal{S}_{1b} \cup \mathcal{S}_{1c} \cup \mathcal{S}_{1d} \cup \mathcal{S}_{1e},\\
\mathcal{S}_2 &= \mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2c} \cup \mathcal{S}_{2d} \cup \mathcal{S}_{2e},\\
\mathcal{S} &= \mathcal{S}_1 \cup \mathcal{S}_2,
\end{aligned}$$

Also recall, to remove the necessity of different cases, if $\boldsymbol{a} \in \mathcal{S}_{1d} \cup \mathcal{S}_{2c}$ and $k = 1$, then the substring $(b_1 + 1, b_2, \dots, b_{k-1}, b_k + 1)$ is understood to be the substring $(b_1 + 2)$.

First suppose $n = 1$ and let $\boldsymbol{a} = (a_1)$, where $a_1 \geq 3$. Then L_1^0 and L_1^{-1} are both the unknot and so $Y_{(a_1)}^0 = L(a_1 - 2, 1)$ and $Y_{(a_1)}^{-1} = L(a_1 + 2, 1)$ (see Figure 2). By Lisca's classification of lens spaces that bound $\mathbb{Q}B^4$'s [8], the only such lens spaces that bound $\mathbb{Q}B^4$'s are $L(1,1) = S^3$ and $L(4,1)$. Thus $Y_{(a_1)}^{-1}$ does not bound a $\mathbb{Q}B^4$ for all $a_1 \geq 3$ and $Y_{(a_1)}^0$ bounds a $\mathbb{Q}B^4$'s if and only if $a_1 = 3$ or $a_1 = 6$. In the former case, $\boldsymbol{a} = (3) \in \mathcal{S}_{2c}$, and in the latter case, $\boldsymbol{d} = (2,2,2,3) \in \mathcal{S}_{2e}$.

We now assume the length of $\boldsymbol{a}$ is at least 2. Throughout, we will consider the standard negative definite intersection lattice $(\mathbb{Z}^n, -I_n)$. Let $\{e_1, \dots, e_n\}$ be the standard basis of $\mathbb{Z}^n$. Then, with respect to the product $\cdot$ given by $-I_n$, we have $e_i \cdot e_j = -\delta_{ij}$ for all i and j. We begin by recalling definitions and results from [8] and introducing new terminology for our purposes.

We consider two subsets $S_1, S_2 \subset \mathbb{Z}^n$ to be the same if S_2 can be obtained by applying an element of $\mathrm{Aut}(\mathbb{Z}^n)$ to S_1. Let $S = \{v_1, \dots, v_n\} \subset \mathbb{Z}^n$ be a subset. We call each element $v_i \in S$ a *vector* and we call the string of integers $(a_1, \dots, a_n)$ defined by $a_i = -v_i \cdot v_i$ the *string associated* to S. Two vectors $z, w \in S$ are called *linked* if there exists $e \in \mathbb{Z}^n$ such that $e \cdot e = -1$ and $z \cdot e, w \cdot e \neq 0$. A subset S is called *irreducible* if, for every pair of vectors $v, w \in S$, there exists a finite sequence of vectors $v_1 = v, v_2, \dots, v_k = w \in S$ such that v_i and v_{i+1} are linked for all $1 \leq i \leq k-1$.

Definition 5.1 A subset $S = \{v_1, \dots, v_n\} \in \mathbb{Z}^n$ is

- *good* if it is irreducible and

$$v_i \cdot v_j = \begin{cases} -a_i \leq -2 & \text{if } i = j,\\ 0 \text{ or } 1 & \text{if } |i-j| = 1,\\ 0 & \text{otherwise;} \end{cases}$$

- *standard* if
$$v_i \cdot v_j = \begin{cases} -a_i \le -2 & \text{if } i = j, \\ 1 & \text{if } |i-j| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that, by definition, standard subsets are good. If S is a good subset, then a vertex $v \in S$ is called *isolated* if $v \cdot w = 0$ for all $w \in S \setminus \{v\}$, *final* if there exists exactly one vertex $w \in S \setminus \{v\}$ such that $v \cdot w = 1$, and *internal* otherwise. A *component* of a good subset G is a subset of G corresponding to a connected component of the intersection graph of G (which is the graph consisting of vertices $v_1, \dots, v_n$ and an edge between two vertices v_i and v_j if and only if $v_i \cdot v_j = 1$).

Definition 5.2 A subset $S = \{v_1, \dots, v_n\} \in \mathbb{Z}^n$ is

- *negative cyclic* if either
  (1) $n = 2$ and
  $$v_i \cdot v_j = \begin{cases} -a_i \le -2 & \text{if } i = j, \\ 0 & \text{if } i \ne j, \end{cases}$$
  or
  (2) $n \ge 3$ and there is a cyclic reordering of S such that
  $$v_i \cdot v_j = \begin{cases} -a_i \le -2 & \text{if } i = j, \\ 1 & \text{if } |i-j| = 1, \\ -1 & \text{if } i \ne j \in \{1, n\}, \\ 0 & \text{otherwise;} \end{cases}$$
- *positive cyclic* if $-a_i \le -3$ for some i and either
  (1) $n = 2$ and
  $$v_i \cdot v_j = \begin{cases} -a_i \le -2 & \text{if } i = j, \\ 2 & \text{if } i \ne j, \end{cases}$$
  or
  (2) $n \ge 3$ and there is a cyclic reordering of S such that
  $$v_i \cdot v_j = \begin{cases} -a_i \le -2 & \text{if } i = j, \\ 1 & \text{if } |i-j| = 1, \\ 1 & \text{if } i \ne j \in \{1, n\}, \\ 0 & \text{otherwise;} \end{cases}$$
- *cyclic* if S is negative or positive cyclic.

If S is cyclic, then the indices of each vertex are understood to be defined modulo n (eg $v_{n+1} = v_1$). If $v_i \cdot v_{i+1} = \pm 1$, then we say that v_i and v_j have a *positive/negative* *intersection*. Moreover, if S is cyclic and S' is obtained from S by reversal and/or cyclic reordering, then we consider S and S' to be the same subset. In this way, associated strings of cyclic subsets are well defined up to reversal and cyclic-reordering.

Remark 5.3 By standard linear algebra, it is easy to see that, if S is good, cyclic, or the union of a good subset and a cyclic subset, then S forms a linearly independent set in $\mathbb{Z}^n$ (see [8, Remark 2.1]).

Remark 5.4 Suppose $S = \{v_1, \dots, v_n\}$ is a cyclic subset. Then, by replacing v_k with $v'_k = -v_k$, we obtain a new subset $\widehat{S} = \{v_1, \dots, v_{k-1}, v'_k, v_{k+1}, \dots, v_n\}$ such that $v_{k-1} \cdot v'_k = -v_{k-1} \cdot v_k$ and $v'_k \cdot v_{k+1} = -v_k \cdot v_{k+1}$. Notice that S and $\widehat{S}$ have the same associated strings. Thus we can change the number of positive and negative intersections of S without changing the associated string. Conversely, any subset of the form $S = \{v_1, \dots, v_n\}$, where $n \geq 3$ and

$$v_i \cdot v_j = \begin{cases} -a_i \leq -2 & \text{if } i = j, \\ \pm 1 & \text{if } |i - j| = 1, \\ \pm 1 & \text{if } i \neq j \in \{1, n\}, \\ 0 & \text{otherwise,} \end{cases}$$

can modified into a positive or negative cyclic subset by changing the signs of select vertices. In particular, for any negative cyclic subset, the negative intersection can be moved at will by negating select vertices.

Similarly, any irreducible subset of the form $G = \{v_1, \dots, v_n\}$, where

$$v_i \cdot v_j = \begin{cases} -a_i \leq -2 & \text{if } i = j, \\ \pm 1 & \text{if } |i - j| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

can be modified into a good subset by changing the signs of select vertices. In Section 7, we will often create such subsets and assume that they are good, without specifying the need to possibly negate select vertices first.

Definition 5.5 Let $S = \{v_1, \dots, v_n\} \subset \mathbb{Z}^n$ be a subset with $v_i \cdot v_i = -a_i$. We define

$$I(S) := \sum_{i=1}^{n} (a_i - 3), \qquad E_i^S := \{j : v_j \cdot e_i \neq 0\},$$

$$p_i(S) := \bigl|\{j : |E_j^S| = i\}\bigr|, \quad V_i^S := \{j : v_i \cdot e_j \neq 0\}.$$

In some cases we will drop the superscript S from the above notation if the subset being considered is understood.

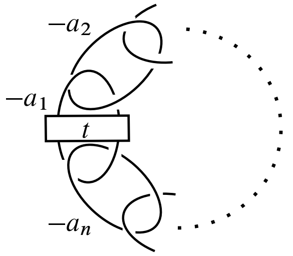


Figure 16: A 4–manifold P^t with boundary $Y^t_{\boldsymbol{a}}$.

Remark 5.6 Lisca [8] classified all standard subsets of $\mathbb{Z}^n$ with $I(S) < 0$. The results in the next three sections rely in part on his classification of standard subsets. We will review his classification in Section 5.1.

Example 5.7 The subset $S = \{e_1 - e_2, e_2 - e_3, \dots, e_{n-1} - e_n, e_n + e_1\} \subset \mathbb{Z}^n$ for $n \geq 2$ is a negative cyclic subset with associated string $(2^{[n]})$. Moreover, $I(S) = -n$, $p_2(S) = n$, and $p_j(S) = 0$ for all $j \neq 2$. When $n = 4$, there is an alternative subset with associated string $(2, 2, 2, 2)$, namely $S' = \{e_1 - e_2, e_2 - e_3, -e_2 - e_1, e_1 + e_4\}$, which satisfies $p_1(S') = p_3(S') = 2$. This latter subset will be used to construct the family strings in $\mathcal{S}_{1a}$.

Let $\boldsymbol{a} = (a_1, \dots, a_n)$. The rational sphere $Y^t_{\boldsymbol{a}}$ is the boundary of the negative definite 2–handlebody P^t whose handlebody diagram is given in Figure 16. Let Q_{P^t} denote the intersection form of P^t. Note that Q_{P^t} depends only on the parity of t. Further suppose $Y^t_{\boldsymbol{a}}$ bounds a rational homology ball B. Then the closed 4–manifold $X^t = P^t \cup B$ is negative definite. By Donaldson's diagonalization theorem [6], the intersection lattice $(H_2(X^t), Q_{X^t})$ is isomorphic to the standard negative definite lattice $(\mathbb{Z}^n, -I_n)$. Thus the intersection lattice $(H_2(P^t), Q_{P^t})$ must embed in $(\mathbb{Z}^n, -I_n)$. The existence of such an embedding implies the existence of a cyclic subset $S \subset \mathbb{Z}^n$ with associated string $(a_1, \dots, a_n)$. Thus our goal is to classify all cyclic subsets of $\mathbb{Z}^n$, where $n \geq 2$.

Recall that, by reversing the orientation of $Y^t_{\boldsymbol{a}}$, we obtain the $\overline{Y}^t_{\boldsymbol{a}} = Y^{-t}_{\boldsymbol{d}}$, where $\boldsymbol{d} = (d_1, \dots, d_m)$ is the cyclic-dual of $(a_1, \dots, a_n)$ (Section 2.2). In particular, $(a_1, \dots, a_n)$ is of the form $(2^{[m_1]}, 3 + n_1, \dots, 2^{[m_k]}, 3 + n_k)$ if and only if $(d_1, \dots, d_m)$ is of the form $(3 + m_1, 2^{[n_1]}, \dots, 3 + m_k, 2^{[n_k]})$. If S and $\overline{S}$ denote the cyclic subsets associated to $(a_1, \dots, a_n)$ and $(d_1, \dots, d_m)$, respectively, then $I(S) + I(\overline{S}) = 0$. Now, since $Y^t_{\boldsymbol{a}}$ bounds a $\mathbb{Q}B^4$ if and only if $Y^{-t}_{\boldsymbol{d}}$ bounds a $\mathbb{Q}B^4$, we will focus our attention on subsets satisfying $I(S) \leq 0$. The following theorem is the main result of our lattice analysis:

Theorem 5.8 *Let S be a cyclic subset such that $I(S) \leq 0$. Then S is either negative with associated string in $\mathcal{S}_1 \cup \mathcal{O} \cup \{(2^{[n]}) \mid n \geq 2\}$ or positive with associated string in $\mathcal{S}_2$.*

Proof The theorem follows from Example 5.7 and Propositions 6.5, 7.5 and 7.14, which will be proven in Sections 6 and 7. □

We can now prove Theorem 1.7, which we recall here for convenience.

Theorem 1.7 *Let $\boldsymbol{a} = (a_1, \dots, a_n)$, where $n \geq 1$, $a_i \geq 2$ for all i, and $a_j \geq 3$ for some j, and let $\boldsymbol{d}$ be the cyclic-dual of $\boldsymbol{a}$.*

1. *Suppose $\boldsymbol{d} \notin \mathcal{S}_{1a} \cup \mathcal{O}$. Then $Y_{\boldsymbol{a}}^{-1}$ bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{a} \in \mathcal{S}_1$ or $\boldsymbol{d} \in \mathcal{S}_{1b} \cup \mathcal{S}_{1c} \cup \mathcal{S}_{1d} \cup \mathcal{S}_{1e}$.*
2. *Suppose $\boldsymbol{a} \notin \mathcal{S}_{1a} \cup \mathcal{O}$. Then $Y_{\boldsymbol{a}}^{1}$ bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{d} \in \mathcal{S}_1$ or $\boldsymbol{a} \in \mathcal{S}_{1b} \cup \mathcal{S}_{1c} \cup \mathcal{S}_{1d} \cup \mathcal{S}_{1e}$.*
3. *$Y_{\boldsymbol{a}}^{0}$ bounds a $\mathbb{Q}B^4$ if and only if $\boldsymbol{a} \in \mathcal{S}_2$ or $\boldsymbol{d} \in \mathcal{S}_2$.*

Proof The sufficient conditions of Theorem 1.7 follow from the calculations in Section 4. The necessary conditions of Theorem 1.7 follow from Theorem 5.8 and the fact that $Y_{\boldsymbol{a}}^{t}$ bounds a $\mathbb{Q}B^4$ if and only if $Y_{\boldsymbol{d}}^{-t}$ bounds a $\mathbb{Q}B^4$. □

The proof of Theorem 5.8 will span the next three sections. The proof will begin in earnest in Section 6. The proof applies two strategies. The first will be to reduce certain cyclic subsets to good subsets and standard subsets and appeal to Lisca's work [8; 9]. The second will be to reduce certain cyclic subsets (via *contractions*) to a small list of base cases. In the upcoming subsection, we will recall Lisca's classification of standard subsets. In the subsequent subsection, we will describe how to perform contractions and list the relevant base cases. In the final subsection, we will prove a few preliminary lemmas that will be useful going forward.

5.1 Lisca's standard and good subsets

In Section 7, we will construct good subsets and standard subsets satisfying $I < 0$ from cyclic subsets, thus reducing the problem of classifying certain cyclic subsets to Lisca's work [8; 9]. In this section, we collect relevant results proved by Lisca. The first two propositions can be found in [8, Sections 3–7]. In particular, the "moreover" statements in Proposition 5.10 are obtained by examining the proofs of [8, Lemmas 7.1–7.3].

Proposition 5.9 *Let $T = \{v_1, \dots, v_n\}$ be a standard subset with $I(T) < 0$. Then*:

1. $I(T) \in \{-1, -2, -3\}$.
2. $|v_i \cdot e_j| \le 1$ *for all i and j.*
3. $p_1(T) = 1$ *if and only if* $I(T) = -3$ *and, if* $p_1(T) = 0$, *then* $p_2(T) > 0$.
4. *If* $I(T) = -3$, *then* $p_1(T) = p_2(T) = 1$ *and* $p_3(T) = n - 2$.
5. *If* $I(T) = -2$, *then* $p_2(T) = 3$, $p_4(T) = 1$, *and* $p_3(T) = n - 4$.
6. *If* $I(T) = -1$, *then* $p_2(T) = 2$, $p_4(T) = 1$ *and* $p_3(T) = n - 3$.

Proposition 5.10 *Let T be standard with $I(T) < 0$. Let $x, y \ge 0$.*

(1) *If* $I(T) = -3$, *then, if* $E_i = \{s\}$, *then* v_s *is internal* (*ie* $1 < s < n$) *and* $v_s \cdot v_s = -2$; *if* $|E_j| = 2$, *then* $E_j = \{1, n\}$; *either* $v_1 \cdot v_1 = -2$ *or* $v_n \cdot v_n = -2$; *and* $v_1 \cdot e_j = -v_n \cdot e_j$. *Moreover, T has associated string of the form* $(b_1, \dots, b_k, 2, c_l, \dots, c_1)$, *where* $(b_1, \dots, b_k)$ *and* $(c_1, \dots, c_l)$ *are linear-dual strings.*

(2) *If* $I(T) = -2$, *then* (*up to reversal*) *T has associated string of the form*

(a) $(2^{[x]}, 3, 2 + y, 2 + x, 3, 2^{[y]})$,

(b) $(2^{[x]}, 3 + y, 2, 2 + x, 3, 2^{[y]})$, *or*

(c) $(b_1, \dots, b_{k-1}, b_k + 1, 2, 2, c_l + 1, c_{l-1}, \dots, c_1)$, *where the strings* $(b_1, \dots, b_k)$ *and* $(c_1, \dots, c_l)$ *are linear-dual.*

Moreover, up to the action of $\mathrm{Aut}(\mathbb{Z}^n)$, *the corresponding embeddings are of the form*

(a) $$\Big\{e_{x+4} - e_{x+3}, e_{x+3} - e_{x+2}, \dots, e_5 - e_4, e_4 - e_2 - e_3, e_2 + e_1 + \sum_{\alpha=x+5}^{x+y+4} e_i, -e_2 - e_4 - \sum_{\alpha=5}^{x+4} e_i, e_2 - e_1 - e_3, e_1 - e_{x+5}, e_{x+5} - e_{x+6}, \dots, e_{x+y+3} - e_{x+y+4}\Big\},$$

(b) $$\Big\{e_{x+4} - e_{x+3}, e_{x+3} - e_{x+2}, \dots, e_5 - e_4, e_4 - e_2 - e_3 - \sum_{\alpha=x+5}^{x+y+4} e_i, e_2 + e_1, -e_2 - e_4 - \sum_{\alpha=5}^{x+4} e_i, e_2 - e_1 - e_3, e_3 - e_{x+5}, e_{x+5} - e_{x+6}, \dots, e_{x+y+3} - e_{x+y+4}\Big\},$$

(c) $\{u_1, \dots, u_{k-1}, u_k + e_4 - e_2 - e_3, e_2 + e_1, -e_2 - e_4, e_2 - e_1 - e_3 + w_1, w_2, \dots, w_l\}$, *where $k+l \geq 3$, $u_k = 0$ or $w_1 = 0$, $|E_1| = |E_4| = 2$. Furthermore (up to reversal), in* (c) *we may assume that $u_1^2 = -2$; consequently, there exist integers j_1 and j_2 such that $|E_{j_1}| = 2$, $|E_{j_2}| = 3$, $u_1 \cdot e_{j_2} = -u_2 \cdot e_{j_2} = -w_l \cdot e_{j_2} = 1$, and $|u_1 \cdot e_{j_2}| = |w_l \cdot e_{j_2}| = 1$.*

(3) *If $I(T) = -1$, then (up to reversal) T has associated string of the form*

- (a) $(2+x, 2+y, 3, 2^{[x]}, 4, 2^{[y]})$,
- (b) $(2+x, 2, 3+y, 2^{[x]}, 4, 2^{[y]})$, *or*
- (c) $(3+x, 2, 3+y, 3, 2^{[x]}, 3, 2^{[y]})$.

Moreover, up to the action of $\mathrm{Aut}(\mathbb{Z}^n)$, *the corresponding embeddings are of the form*

$$\text{(a)}\quad \Bigl\{e_2 + e_4 + \sum_{\alpha=5}^{x+4} e_\alpha, e_1 - e_2 + \sum_{\alpha=x+5}^{x+y+4} e_\alpha, e_2 - e_3 - e_4, e_4 - e_5, \\ e_5 - e_6, \dots, e_{x+3} - e_{x+4}, e_{x+4} - e_1 - e_2 - e_3, e_1 - e_{x+5}, \\ e_{x+5} - e_{x+6}, \dots, e_{x+y+3} - e_{x+y+4}\Bigr\},$$

$$\text{(b)}\quad \Bigl\{e_2 + e_4 + \sum_{\alpha=5}^{x+4} e_\alpha, e_1 - e_2, e_2 - e_3 - e_4 - \sum_{\alpha=x+5}^{x+y+4} e_\alpha, e_4 - e_5, \dots, e_{x+3} - e_{x+4}, \\ e_{x+4} - e_1 - e_2 - e_3, e_3 - e_{x+5}, e_{x+5} - e_{x+6}, \dots, e_{x+y+3} - e_{x+y+4}\Bigr\},$$

$$\text{(c)}\quad \Bigl\{e_1 - e_2 - e_5 - \sum_{\alpha=6}^{x+5} e_\alpha, e_2 + e_3, -e_2 - e_1 - e_4 - \sum_{\alpha=x+6}^{x+y+5} e_\alpha, -e_5 + e_2 - e_3, \\ e_5 - e_6, e_6 - e_7, \dots, e_{x+4} - e_{x+5}, e_{x+5} + e_1 - e_4, e_4 - e_{x+6}, \\ e_{x+6} - e_{x+7}, \dots, e_{x+y+4} - e_{x+y+5}\Bigr\}.$$

The next proposition follows from the first case (S irreducible) of the proof of the main theorem in [9, page 2160ff] and [8, Lemma 6.2] (see also [1, Lemma 6.6]). See [8, Definition 4.1] for the definition of *bad component*.

Proposition 5.11 [9] *Let $G \subset \mathbb{Z}^n$ be a good subset with two components and $I(G) \leq -2$. If G has no bad components, then $I(G) = -2$ and G has associated string of the form $(b_1, \dots, b_k) \cup (c_1, \dots, c_l)$, where $(b_1, \dots, b_k)$ and $(c_1, \dots, c_l)$ are linear-dual strings. Moreover, if $G = \{v_1, \dots, v_k, v_{k+1}, \dots, v_{k+l}\}$, where $-v_i^2 = b_i$ for $1 \leq i \leq k$ and $-v_{k+j}^2 = c_j$ for all $1 \leq j \leq l$, then there exist integers α and β such that $E_\alpha = \{1, k+1\}$ and $E_\beta = \{k, k+l\}$.*

5.2 Contractions, expansions and base cases

In this section, we discuss how to reduce the length of certain cyclic subsets via contractions.

Definition 5.12 Suppose $S=\{v_1,\dots,v_n\}$ with $n\ge 3$ is a cyclic subset and suppose there exist integers i, s and t such that $E_i=\{s,\tilde{s},t\}$, where $\tilde{s}\in\{s\pm 1\}$, $V_{\tilde{s}}\cap V_s=\{i\}$, $|v_u\cdot e_i|=1$ for all $u\in E_i$, and $a_t\ge 3$. After possibly cyclically reordering and reindexing S, we may assume that $s\notin\{1,n\}$. Let $S'\subset\mathbb{Z}^{n-1}=\langle e_1,\dots,e_{i-1},e_{i+1},\dots,e_n\rangle$ be the subset defined by

$$S'=(S\setminus\{v_s,v_{\tilde{s}},v_t\})\cup\{v_s+v_{\tilde{s}},\pi_{e_i}(v_t)\},$$

where $\pi_{e_i}(v_t)=v_t+(v_t\cdot e_i)e_i$. We say that S' is obtained from S by a *contraction* and S is obtained from S' by an *expansion*.

Since $s\notin\{1,n\}$ and $|v_{\tilde{s}}\cdot e_i|=|v_s\cdot e_i|=1$, we have $v_{s-1}\cdot e_i=-v_s\cdot e_i$. Thus

$$(v_s+v_{\tilde{s}})\cdot v_u=\begin{cases}1 & \text{if } \tilde{s}=s+1 \text{ and } u\in\{s-1,s+2\},\\ 1 & \text{if } \tilde{s}=s-1 \text{ and } u\in\{s-2,s+1\},\\ 0 & \text{otherwise.}\end{cases}$$

Moreover, $(\pi_{e_i}(v_t))^2=v_t^2+1\le -2$ and

$$\pi_{e_i}(v_t)\cdot v_u=\begin{cases}1 & \text{if } u=t\pm 1,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore, S' is a positive/negative cyclic subset if and only if S is positive/negative cyclic. Moreover, $I(S')=I(S)$, $p_j(S')=p_j(S)$ for all $j\ne 3$, and $p_3(S')=p_3(S)-1$.

Definition 5.13 Using the notation above, if $v_t\cdot v_s=1$ (so that $t=s\pm 1$ if $\tilde{s}=s\mp 1$) and $a_{\tilde{s}}=2$, then we say

- v_s is the *center of S relative to e_i*,
- S' is obtained by a *contraction of S centered at v_s*, and
- S is obtained by a *−2–expansion of S*.

Note that a subset obtained by a contraction of S centered at v_s is unique. Indeed, if $E_i=\{s-1,s,s+1\}$, $a_{s-1}=2$ and $a_{s+1}\ge 3$, then $V_{s-1}\cap V_s=\{i\}$ and the only contraction centered at v_s is $S\setminus\{v_s,v_{s-1},v_{s+1}\}\cup\{v_{s-1}+v_s,\pi_{e_i}(v_{s+1})\}$. Similarly, if $E_i=\{s-1,s,s+1\}$, $a_{s-1}=2$ and $a_{s+1}\ge 3$, then $V_{s-1}\cap V_s=\{i\}$ and the only contraction centered at v_s is $S \setminus \{v_s, v_{s-1}, v_{s+1}\} \cup \{v_s + v_{s+1}, \pi_{e_i}(v_{s-1})\}$. Now let S have associated string $(a_1, \dots, a_n)$. Then, under the contraction centered at v_s, the associated string changes via

$$(a_1,\dots,a_{s-2},2,\boldsymbol{a}_s,a_{s+1},a_{s+2},\dots,a_n) \to (a_1,\dots,a_{s-2},\boldsymbol{a}_s,a_{s+1}-1,a_{s+2},\dots,a_n)$$

or

$$(a_1,\dots,a_{s-2},a_{s-1},\boldsymbol{a}_s,2,a_{s+2},\dots,a_n) \to (a_1,\dots,a_{s-2},a_{s-1}-1,\boldsymbol{a}_s,a_{s+1},\dots,a_n).$$

Notice that two strings $(b_1, \dots, b_k)$ and $(c_l, \dots, c_1)$ are reverse linear-dual if and only if $(b_1, \dots, b_{k-1})$ and $(c_l - 1, \dots, c_1)$ or $(b_1, \dots, b_k - 1)$ and $(c_{l-1}, \dots, c_1)$ are reverse linear-dual. Thus the substrings on either side of a_s in the associated string of S are reverse linear-dual if and only if the substrings on either side of a_s in the associated string of the contraction of S centered at v_s are reverse linear-dual.

More generally, let $S = \{v_1, \dots, v_n\}$ and consider a sequence of contractions $S^0 = S$, S^1, $S^2, \dots, S^m$ such that S^k is obtained from S^{k-1} by performing a contraction centered at $v_s^{(k-1)} \in S^{k-1}$, where $v_s^{(0)} = v_s$. We call such a sequence of contractions *the sequence of contractions centered at v_s* and call the reverse sequence of expansions *a sequence of -2–expansions centered at $v_s^{(m)}$*. Notice that, for all $1 \le k \le m$, $v_s^{(k)} = v_s^{(k-1)} + v_{\tilde{s}}^{(k-1)}$, where $v_{\tilde{s}}^{(k-1)}$ is the unique vertex of S^{k-1} adjacent to $v_s^{(k-1)}$ with square -2. We have proven the following:

Lemma 5.14 *Let S' be obtained from S by a sequence of contractions centered at v and let $v^2 = -a$. Then S has associated string of the form $(b_1, \dots, b_k, a, c_l, \dots, c_1)$, where $(b_1, \dots, b_k)$ and $(c_l, \dots, c_1)$ are reverse linear-dual, if and only if S' has associated string of the form $(b'_1, \dots, b'_{k'}, a, c'_{l'}, \dots, c'_1)$, where $(b'_1, \dots, b'_{k'})$ and $(c'_{l'}, \dots, c'_1)$ are reverse linear-dual.*

When $I(S) \le 0$ and either $p_1(S) > 0$ or $p_1(S) = p_2(S) = 0$, we will be able to sequentially perform contractions until we arrive at certain base cases. In light of Example 5.7, we will restrict our attention to cyclic subsets containing at least one vector with square at most -3. We will now list all such cyclic subsets of length 2 and 3 with $I(S) \le 0$. It can be concretely checked case by case that the only such cyclic subsets are positive and (up to the action of $\mathrm{Aut}(\mathbb{Z}^2)$) are of the form

- $\{2e_1, -e_1 + e_2\}$, which has associated string $(4, 2) \in \mathcal{S}_{2a}$;
- $\{2e_1 - e_3, e_3 + e_2, -e_1 - e_3\}$, which has associated string $(5, 2, 2) \in \mathcal{S}_{2a}$; and
- $\{e_1 - e_2 - e_3, e_3 - e_1 - e_2, e_2 - e_3 - e_1\}$, which has associated string $(3, 3, 3) \in \mathcal{S}_{2c}$.

Notice that the second and third vertices of the subset with associated string $(5,2,2)$ are both centers relative to e_3. If we perform a contraction centered at either vertex relative to e_3, we obtain the subset with associated string $(4,2)$. Note that, when $n=3$, centers are not unique, but when $n \geq 4$, centers are necessarily unique.

Remark 5.15 We will usually denote cyclic subsets by S, standard subsets by T, and good subsets by G. Moreover, S' will be reserved for contractions of S.

5.3 Preliminary lemmas

The following lemmas will be important in future sections. The first follows from the proof of [8, Lemma 2.5].

Lemma 5.16 [8, Lemma 2.5] *If $S=\{v_1,\dots,v_n\} \subset \mathbb{Z}^n = \langle e_1,\dots,e_n\rangle$ is any subset, then*

$$2p_1(S)+p_2(S)+I(S) \geq \sum_{j=4}^{n}(j-3)p_j(S),$$

with equality if and only if $|v_\alpha \cdot e_\beta| \leq 1$ for all $1 \leq \alpha,\beta \leq n$.

Lemma 5.17 *Let S be cyclic and such that $p_2(S)>0$ and $|v_\alpha \cdot e_\beta| \leq 1$ for all $1 \leq \alpha,\beta \leq n$. Then $\sum_i p_{2i}(S) \equiv -I(S) \bmod 4$.*

Proof First notice that, since $I(S)=\sum_{i=1}^n (a_i-3)$, we have $\sum_{i=1}^n a_i = 3n+I(S)$. Now

$$-\Bigl(\sum_{i=1}^n v_i\Bigr)^2 = \sum_{i=1}^n a_i - \sum_{i=1}^{n-1} 2v_i \cdot v_{i+1} - 2v_1 \cdot v_n = \begin{cases} n+I(S) & \text{if } S \text{ is positive,} \\ n+4+I(S) & \text{if } S \text{ is negative.} \end{cases}$$

On the other hand, set $\sum_{i=1}^n v_i = \sum_{i=1}^n \lambda_i e_i$ and let $k_\alpha = \bigl|\{i : |\lambda_i| = 2\alpha+1\}\bigr|$ and $x_\beta = \bigl|\{i : |\lambda_i| = 2\beta\}\bigr|$. Finally, let $m \in \mathbb{Z}$ be the largest integer such that $k_m \neq 0$ and $k_t = 0$ for all $t > m$, and let $y \in \mathbb{Z}$ be the largest integer such that $x_y \neq 0$ and $x_t = 0$ for all $t > y$. Since $|v_\alpha \cdot e_\beta| \leq 1$ for all α and β, we have $\sum_i p_{2i}(S) = x_0 + \dots + x_y$. Hence,

$$-\Bigl(\sum_{i=1}^n v_i\Bigr)^2 = -\sum_{i=1}^n \lambda_i^2 = \biggl(n - \Bigl(\sum_{\alpha=1}^m k_\alpha\Bigr) - \Bigl(\sum_{\beta=0}^y x_\beta\Bigr)\biggr) + \sum_{\alpha=1}^m (2\alpha+1)^2 k_\alpha + \sum_{\beta=0}^y (2\beta)^2 x_\beta$$

$$
\begin{aligned}
&= n + \sum_{\alpha=1}^{m}(4\alpha^2+4\alpha)k_\alpha + \sum_{\beta=0}^{y}(4\beta^2-1)x_\beta \\
&= n + \sum_{\alpha=1}^{m}(4\alpha^2+4\alpha)k_\alpha + \sum_{\beta=0}^{y}(4\beta^2)x_\beta - \Bigl(\sum_i p_{2i}(S)\Bigr).
\end{aligned}
$$

Thus,

$$\sum_{\alpha=1}^{m}(4\alpha^2+4\alpha)k_\alpha + \sum_{\beta=1}^{y}(4\beta^2)x_\beta = \begin{cases} \sum_i p_{2i}(S) + I(S) & \text{if } S \text{ is positive,} \\ \sum_i p_{2i}(S) + 4 + I(S) & \text{if } S \text{ is negative.} \end{cases}$$

It follows that $\sum_i p_{2i}(S) \equiv -I(S) \bmod 4$. □

Lemma 5.18 *If $G = \{v_1, \dots, v_n\} \subset \mathbb{Z}^n$ is a good subset with $I(G) = 0$, $p_3(G) = n$, and n components, then, up to the action of* $\operatorname{Aut}\mathbb{Z}^n$, *negating vertices, and permuting vertices,*

- *$G = \{e_1 - e_2 + e_3 - e_4, e_1 + e_2, -e_1 + e_2 + e_3 - e_4, e_3 + e_4\}$ with associated string $(4, 2, 4, 2)$, or*
- *$G = \{e_1 - e_2 - e_3, e_1 + e_2 - e_4, e_2 - e_3 + e_4, e_1 + e_3 + e_4\}$ with associated string $(3, 3, 3, 3)$.*

Proof First notice that, by Lemma 5.16, $|v_\alpha \cdot e_\beta| \le 1$ for all α and β. Let i, s, t and u be integers such that $E_i = \{s, t, u\}$. Since every vertex of G is isolated, up to negating vertices we may assume that $v_s \cdot e_i = v_t \cdot e_i = v_u \cdot e_i = -1$.

First suppose $a_s = 2$ and let $v_s = e_i + e_j$. Then, since $v_s \cdot v_t = v_s \cdot v_u = 0$, we have $v_t = e_i - e_j + a$ and $v_u = e_i - e_j + b$. Since $v_t \cdot v_u = 0$, there are integers $k, l \in V_t \cap V_u$ such that $v_t = e_i - e_j + e_k - e_l + a'$ and $v_u = e_i - e_j - e_k + e_l + b'$. If $(a')^2 \ne 0$, then let $R = \{v'_1, \dots, v'_{s-1}, v'_{s+1}, \dots, v'_n\} \subset \mathbb{Z}^{n-2} = \langle e_1, \dots, e_n \rangle / \langle e_i, e_j \rangle$, where $v'_t = \pi_{e_j}(\pi_{e_i}(v_t))$, $v'_u = \pi_{e_j}(\pi_{e_i}(v_u))$, and $v'_x := v_x$ for all $x \notin \{t, u\}$. Then $(v'_t)^2 \le -3$, $v'_t \cdot v'_u = 2$, and $v'_t \cdot v_x = v'_u \cdot v'_x = 0$ for all $x \notin \{t, u\}$. Consequently, R is the union of a positive cyclic subset $\{v'_t, v'_u\}$ and a good subset $R \setminus \{v'_t, v'_u\}$. Thus, by Remark 5.3, R is a linearly independent set of $n-1$ vectors in $\mathbb{Z}^{n-2}$, which is impossible. Thus $(a')^2 = 0$ and, similarly, $(b')^2 = 0$; hence, $v_t = e_i - e_j + e_k - e_l$ and $v_u = e_i - e_j - e_k + e_l$. Now, since $|E_k| = |E_l| = 3$, there exists an integer z such that $k, l \in V_z$ and, since $v_z \cdot v_t = 0$, we may assume that $v_z = e_k + e_l + c$. By a similar argument as above, $c^2 = 0$ and so $v_z = e_k + e_l$. Since G is irreducible, it follows that

$n = 4$ and so G has associated string of the form $(4, 2, 4, 2)$. Setting $i = 3$, $j = 4$, $k = 1$ and $l = 2$, we have the subset listed in the statement of the lemma.

Next suppose $a_s, a_t, a_u \geq 3$. Assume $a_s > 3$. Let $R = \{v'_1, \dots, v'_{s-1}, v'_{s+1}, \dots, v'_n\} \subset \mathbb{Z}^{n-1} = \langle e_1, \dots, e_n\rangle / \langle e_i\rangle$, where $v'_s = \pi_{e_i}(v_s)$, $v'_t = \pi_{e_i}(v_t)$, $v'_u = \pi_{e_i}(v_u)$, and $v'_x := v_x$ for all $x \notin \{s, t, u\}$. Then $(v'_s)^2 < -2$ and $v'_s \cdot v'_t = v'_s \cdot v'_u = v'_t \cdot v'_u = 1$; hence, $\{v'_s, v'_t, v'_u\}$ is a positive cyclic subset. Moreover, $v'_s \cdot v'_x = v'_t \cdot v'_x = v'_u \cdot v'_x = 0$ for all $x \notin \{s, t, u\}$. Thus R is the union of a positive cyclic subset and a good subset and so, by Remark 5.3, R is a linearly independent set of $n-1$ vectors in $\mathbb{Z}^{n-2}$, which is impossible. Thus $a_s = 3$; similarly, $a_t = a_u = 3$. Without loss of generality, $v_s = e_i - e_j - e_k$, $v_t = e_i + e_j - e_l$ and $v_u = e_i + e_k + e_l$ for some integers j, k and l. Since $|E_j| = 3$, there exists an integer z such that $j \in V_z$. Since $v_z \cdot v_s = v_z \cdot v_t = v_z \cdot v_u = 0$, we have $v_z = e_j - e_k + e_l + a$. If $a^2 \neq 0$, then we can define a subset R as above and arrive at a similar contradiction. Thus $v_z = e_j - e_k - e_l$. Since G is irreducible, it follows that $n = 4$ and so G has associated string of the form $(3, 3, 3, 3)$. Setting $i = 1$, $j = 2$, $k = 3$ and $l = 4$, we have the subset listed in the statement of the lemma. □

6 Lattice analysis, case I: $p_1(S) > 0$

Throughout this section, we will assume that $S = \{v_1, \dots, v_n\}$ is a cyclic subset with $I(S) \leq 0$ and $p_1(S) > 0$. Thus there exist integers i and s such that $E_i = \{s\}$. Lemmas 6.1–6.3 will ensure that we can contract such subsets.

Lemma 6.1 *Let S be a cyclic subset of length 4 such that $I(S) \leq 0$ and $E_i = \{s\}$ for some integers i and s. If $a_{s+1} \geq 3$ or $a_{s-1} \geq 3$, then S is positive and has associated string of the form $(6, 2, 2, 2)$ or $(5, 2, 2, 3)$. If $a_{s\pm1} = 2$, then S is either negative and has associated string of the form $(2, 2, 2, 2)$ or $(2, 2, 2, 5)$, or positive and has associated string of the form $(2, 2, 2, 3)$ or $(2, 2, 2, 6)$.*

Proof If $|V_s| = 1$, then, since $E_i = \{s\}$, we obtain $v_s \cdot v_{s+1} = 0$, which is a contradiction. Thus $|V_s| \geq 2$.

Suppose $a_{s-1} \geq 3$. If $|V_s| \geq 3$, then let $R \subset \mathbb{Z}^3$ be the subset obtained by replacing v_s by $v_s + (v_s \cdot e_i)e_i$. Then R is a cyclic subset and, by Remark 5.3, R is made of four linearly independent vectors in $\mathbb{Z}^3$, which is not possible. Thus $|V_s| = 2$. Let $V_s = \{i, j\}$. Then $E_j = \{s-1, s, s+1\}$, since otherwise we would necessarily have that $|E_i| > 1$. Moreover, since $V_{s-1} \cap V_s = V_{s+1} \cap V_s = \{j\}$, we necessarily have that $|v_{s-1}\cdot e_j| = |v_s\cdot e_j| = |v_{s+1}\cdot e_j| = 1$. If S is positive cyclic, then it is clear that $v_{s-1}\cdot e_j = v_{s+1}\cdot e_j = -v_s\cdot e_j$. If S is negative cyclic, then, by possibly moving the negative intersection (see Remark 5.4), we may assume that $v_{s-1}\cdot e_j = v_{s+1}\cdot e_j = -v_s\cdot e_j$. Thus we may perform a contraction of S centered at v_s relative to e_j to obtain a length 3 cyclic subset S' with $I(S') = I(S) \le 0$ and $p_1(S') > 0$. By considering the base cases in Section 5.2, it is clear that $S' = \{2e_1-e_3, e_3+e_2, -e_1-e_3\}$ (up to the action of $\mathrm{Aut}(\mathbb{Z}^3)$), which has associated string $(5,2,2)$. Thus $i=2$, $j=4$, and either $S=\{2e_1-e_3-e_4, e_2+e_4, -e_4+e_3, -e_1-e_3\}$ or $S=\{2e_1-e_3, e_3-e_4, e_4+e_2, -e_4-e_1-e_3\}$. Therefore, S is positive and has associated string $(6,2,2,2)$ or $(5,2,2,3)$.

Now suppose $a_{s-1} = a_{s+1} = 2$. Without loss of generality, assume $s=j=4$. Let $T=\{v_1,v_2,v_3\}\subset\mathbb{Z}^3=\langle e_1,e_2,e_3\rangle$ be the length 3 standard subset obtained by removing v_s from S. Then T has associated string of the form $(2,a_2,2)$. Since $I(S)\le 0$, we must have $a_2\le 6$. It is easy to see that $a_2\ne 6$, since otherwise $v_2 = 2e_1-e_2-e_3$ (up to the action of $\mathrm{Aut}(\mathbb{Z}^3)$), implying that $v_1\cdot v_2\ne\pm1$, which is a contradiction. If $a_2=5$, then T is of the form $\{e_1-e_2, e_2+2e_3, -e_2-e_1\}$ and therefore S must be of the form $\{e_1-e_2, e_2+2e_3, -e_2-e_1, e_1+e_4\}$ (up to the action of $\mathrm{Aut}(\mathbb{Z}^3)$). Thus S is negative with associated string $(2,5,2,2)$ (equivalently $(2,2,2,5)$). If $a_2\le 4$, then $I(T)<0$. By Proposition 5.10, the only such length 3 standard subset has associated string $(2,2,2)$. Moreover, T is of the form $T=\{e_1-e_2, e_2-e_3, -e_2+e_1\}$ (see [8, Lemma 2.4]). Since $v_3\cdot v_4=\pm1$, either $1\in V_4^S$, $2\in V_4^S$, or both. If $1,2\in V_4^S$, then since $v_2\cdot v_4=0$, we must have $3\in V_4^S$; thus $|V_4^S|=4$. Moreover, since $v_1\cdot v_4=\pm1$, we must have that $v_4\cdot e_1 = v_4\cdot e_2\pm 1$, implying that $a_4\ge 7$, which is not possible. Thus either $1\in V_4^S$ or $2\in V_4^S$, but not both. If $1\in V_4^S$, then S is negative and of the form $\{e_1-e_2, e_2-e_3, -e_2-e_1, e_1+e_4\}$ or $\{e_1-e_2, e_2-e_3, -e_2-e_1, e_1+2e_4\}$, which have associated strings $(2,2,2,2)$ and $(2,2,2,5)$ (note that we found the latter subset above). If $2\in V_4^S$, then $3\in V_4^S$ and S is positive and of the form $\{e_1-e_2, e_2-e_3, -e_2-e_1, e_2+e_3+e_4\}$ or $\{e_1-e_2, e_2-e_3, -e_2-e_1, e_2+e_3+2e_4\}$, which have associated strings $(2,2,2,3)$ and $(2,2,2,6)$. □

Lemma 6.2 *Let S be a cyclic subset of length at least 5 such that $E_i=\{s\}$ for some i and s. Then $|V_s|=2$. Moreover, if $V_s=\{i,j\}$, then $E_j=\{s-1,s,s+1\}$ and $v_{s-1}\cdot e_j = v_{s+1}\cdot e_j = -v_s\cdot e_j = \pm1$.*

Proof First note that, if $|V_s|=1$, then, since $E_i=\{s\}$, we obtain $v_s\cdot v_{s+1}=0$, which is a contradiction. Now suppose $|V_s|\ge 3$. Then, by replacing v_s with $v_s' = v_s+(v_s\cdot e_i)e_i$

and relabeling $v'_u = v_u$ for all $u \neq s$, we obtain a subset

$$R = \{v'_1, \dots, v'_{s-1}, v'_s, v'_{s+1}, \dots, v'_n\} \subset \mathbb{Z}^{n-1} = \langle e_1, \dots, e_{i-1}, e_{i+1}, \dots, e_n \rangle.$$

Let $(a'_1, \dots, a'_n)$ be the string associated to R, where $-a'_s := v'_s \cdot v'_s \leq -2$ and $a'_j = a_j$ for all $j \neq i$. If S is negative cyclic, then so is R and thus, by Remark 5.3, R is made of n linearly independent vectors in $\mathbb{Z}^{n-1}$, which is not possible. If S is positive cyclic and either $a'_s \geq 3$ or $a_i \geq 3$ for some $i \neq s$, then R is also positive cyclic, and we obtain a similar contradiction. Now suppose S is positive cyclic, $a'_s = 2$ and $a'_t = a_t = 2$ for all $t \neq s$. Let T be the subset obtained by removing v_s from S. Then T has associated string $(2^{[n-1]})$ and so $I(T) = -(n-1) \leq -4$. If $|E^S_k| \geq 2$ for all $k \in V^S_s$, where $k \neq i$, then T is a standard subset of $\mathbb{Z}^{n-1}$ with $I(T) \leq -4$, which contradicts Proposition 5.9. If $|E^S_k| = 1$ for some $k \in V^S_s$ such that $k \neq i$, then, by Remark 5.3, T consists of $n-1$ linearly independent vectors in $\mathbb{Z}^m$, where $m < n-1$, which is not possible. Thus $|V_s| = 2$. Let $V^S_s = \{i, j\}$. Then, as in the proof of Lemma 6.1, $E_j = \{s-1, s, s+1\}$ and $v_{s-1} \cdot e_j = v_{s+1} \cdot e_j = -v_s \cdot e_j = \pm 1$. □

Lemma 6.3 *Let S be a cyclic subset of length at least 5 such that $I(S) \leq 0$ and $E_i = \{s\}$ for some i and s. Then either $a_{s-1} \geq 3$ or $a_{s+1} \geq 3$. Moreover, if $a_{s\pm 1} \geq 3$, then S is positive with associated string* $(2,3,2,3,2)$ *or* $(2,3,5,3,2)$.

Proof By Lemma 6.2, $V_s = \{i, j\}$ and $E_j = \{s-1, s, s+1\}$. Assume that $a_{s-1} = a_{s+1} = 2$. Then $V_{s-1} = \{j, k\}$ for some k, $V_{s+1} = \{j, k'\}$ for some k', and $|v_{s\pm 1} \cdot e_j| = |v_{s-1} \cdot e_k| = |v_{s+1} \cdot e_{k'}| = 1$. Since $v_{s-1} \cdot v_{s+1} = 0$, we must have $k = k'$. Since $|v_{s-2} \cdot v_{s-1}| = 1$ and $j \notin V_{s-2}$, we must have $k \in V_{s-2}$. But then $v_{s-2} \cdot v_{s+1} \neq 0$, which is a contradiction.

Now suppose $a_{s-1}, a_{s+1} \geq 3$ and let R be the subset obtained by removing v_s and replacing $v_{s\pm 1}$ with $v'_{s\pm 1} = v_{s\pm 1} + (v_{s\pm 1} \cdot e_j) \cdot e_j$. Note that $v'_{s-1} \cdot v'_{s+1} = \pm 1$. As in the proof of Lemma 6.2, either R is cyclic or S is positive cyclic and R has associated string of the form $(2^{[n-1]})$. In the former case, by Remark 5.3, $R \subset \mathbb{Z}^{n-2}$ contains $n-1$ linearly independent vectors, which is not possible. In the latter case, let $T \subset \mathbb{Z}^{n-1}$ be the standard subset obtained from S by only removing v_s. Then T has associated string $(3, 2, \dots, 2, 3)$. By Proposition 5.10, the only such standard subset is $\{e_4 + e_3 - e_2, e_2 + e_1, -e_2 - e_4, e_2 + e_3 - e_1\}$ (up to the action of $\mathrm{Aut}(\mathbb{Z}^4)$), which has associated string $(3, 2, 2, 3)$. Thus $j = 3$, $|v_s \cdot e_3| = 1$. Since $I(S) \leq 0$, S is of the form $\{-e_2 - e_4, e_2 + e_3 - e_1, e_5 - e_3, e_4 + e_3 - e_2, e_2 + e_1\}$ or $\{-e_2 - e_4, e_2 + e_3 - e_1, 2e_5 - e_3, e_4 + e_3 - e_2, e_2 + e_1\}$, which are positive and have associated strings $(2, 3, 2, 3, 2)$ and $(2, 3, 5, 3, 2)$, respectively. □

Let $S=\{v_1,\dots,v_n\}$ be a cyclic subset such that $n\geq 6$, $I(S)\leq 0$ and $E_i^S=\{s\}$ for some integers i and s. By Lemma 6.2, we may assume that $V_s^S=\{i,j\}$ and $E_j^S=\{s-1,s,s+1\}$ for some integer j. Thus v_s is the center vertex of S relative to e_j. By Lemma 6.3, we may further assume that $a_{s+1}\geq 3$ and $a_{s-1}=2$ and so $V_{s-1}^S=\{j,j_1\}$ for some integer j_1. Let $S'=\{v_1',\dots,v_{s-2}',v_s',v_{s+1}',\dots,v_n'\}$ be the contraction of S centered at v_s, where $v_x'=v_x$ for all $x\notin\{s-1,s,s+1\}$, $v_s'=v_{s-1}+v_s$, and $v_{s+1}'=\pi_{e_j}(v_t)$. Since $V_s^{S'}=\{i,j_1\}$ and $E_{j_1}^{S'}=\{s-2,s,s+1\}$, v_s' is the center vertex of S' relative to e_{j_1} and, by Lemma 6.3, either $(v_{s-2}')^2\leq -3$ or $(v_{s+1}')^2\leq -3$. If $(v_{s-2}')^2\leq -3$ and $(v_{s+1}')^2\leq -3$, then, by Lemma 6.3, S' is positive and has associated string of the form $(2,3,2,3,2)$ or $(2,3,5,3,2)$. If $(v_{s-2}')^2=-2$ or $(v_{s+1}')^2=-2$, then we can perform the contraction centered at v_s' relative to e_{j_1}, as above. Continuing in this way, we have a sequence of contractions centered at v_s, which ends in a subset $\widehat{S}$ either of length 4 or of length 5 with associated string $(2,3,2,3,2)$ or $(2,3,5,3,2)$. Let $\hat{v}_s$ denote the resulting center vertex of $\widehat{S}$. Then $V_s^{\widehat{S}}=\{i,k\}$ for some integer k and $|E_k^{\widehat{S}}|=3$.

Suppose that $\widehat{S}$ has length 4. By considering the length 4 cyclic subsets in the proof of Lemma 6.1, it is clear that $\widehat{S}$ is either negative and of the form

- $\{e_1-e_2,e_2-e_3,-e_2-e_1,e_1+e_4\}$ with associated string $(2,\mathbf{2},2,\mathbf{2})$, or
- $\{e_1-e_2,e_2-e_3,-e_2-e_1,e_1+2e_4\}$ with associated string $(2,\mathbf{2},2,\mathbf{5})$;

or positive and of the form

- $S=\{2e_1-e_3-e_4,e_2+e_4,-e_4+e_3,-e_1-e_3\}$ with associated string $(6,\mathbf{2},2,2)$, or
- $S=\{2e_1-e_3,e_3-e_4,e_4+e_2,-e_4-e_1-e_3\}$ with associated string $(5,2,\mathbf{2},3)$.

Each bold number in the above strings corresponds to a vertex $\hat{v}_m$ satisfying $E_\alpha^{\widehat{S}}=\{m\}$ for some integers α and m. In particular, one of the bold numbers in each of the above strings corresponds to $\hat{v}_s$. In the first two cases, notice that the substrings between the bold numbers (ie (2) and (2)) are reverse linear-dual. Thus, by Lemma 5.14, S has associated string of the form $(b_1,\dots,b_k,2,c_l,\dots,c_1,2)$ or $(b_1,\dots,b_k,2,c_l,\dots,c_1,5)$, where $(b_1,\dots,b_k)$ and $(c_l,\dots,c_1)$ are reverse linear-dual. Similarly, the third and fourth strings are of the form $(b_1+3,b_2,\dots,b_k,2,c_l,\dots,c_1)$, where $(b_1,\dots,b_k)$ and $(c_l,\dots,c_1)$ are reverse linear-dual, and so S has associated string of the same form. Note that the strings $(5,2,2)$ and $(4,2)$ also fall under this family (recall that the linear-dual of (1) is the empty string).

Now suppose $\widehat{S}$ has length 5. Then, by the proof of Lemma 6.3, $\widehat{S}$ is positive and of the form

- $\{-e_2-e_4, e_2+e_3-e_1, e_5-e_3, e_4+e_3-e_2, e_2+e_1\}$ with associated string $(2,3,\mathbf{2},3,2)$, or
- $\{-e_2-e_4, e_2+e_3-e_1, 2e_5-e_3, e_4+e_3-e_2, e_2+e_1\}$ with associated string $(2,3,\mathbf{5},3,2)$.

As above, the bold numbers in these two strings correspond to the vertex $\hat{v}_s$. Notice that, after performing a -2–expansion centered at $\hat{v}_s$, the first and last entries in each string remain unchanged. Moreover, the substrings adjacent to the bold numbers are (3) and (3); notice $(3-1)=(2)$ and $(3-1)=(2)$ are reverse linear-dual strings. Thus, as above, S has associated string of the form $(2, b_1+1, b_2, \dots, b_k, 2, c_l, \dots, c_2, c_1+1, 2)$ or $(2, b_1+1, b_2, \dots, b_k, 5, c_l, \dots, c_2, c_1+1, 2)$, where $(b_1, \dots, b_k)$ and $(c_l, \dots, c_1)$ are reverse linear-dual strings.

Remark 6.4 Consider the length 5 subsets above. We can perform contractions to obtain the cyclic subsets of Lemma 6.1 with associated strings $(2,2,2,3)$ and $(2,2,2,6)$. However, these do not fall under the general formulas listed above. Moreover, the string $(2,2,2,6)$ is also the associated string of a different subset, as seen in Lemma 6.1. This string already appeared in first set of cases we considered and so we will not count this string again.

Combining all of these cases, we have proven the following:

Proposition 6.5 *Let S be a cyclic subset with $I(S) \le 0$ and $p_1(S) > 0$. Then S is either negative with associated string in $\mathcal{S}_{1a} \cup \mathcal{S}_{1b}$, or positive with associated string in $\mathcal{S}_{2a} \cup \mathcal{S}_{2b} \cup \mathcal{S}_{2e}$.*

7 Lattice analysis, case II: $p_1(S) = 0$

In this section, we will assume that $S = \{v_1, \dots, v_n\}$ is cyclic with $I(S) \le 0$ and $p_1(S) = 0$. By Lemma 5.16, $p_2(S) \ge \sum_{j=4}^{n} (j-3) p_j(S)$. If $p_2(S) = 0$, then the inequality is necessarily an equality and so $p_j(S) = 0$ for all $4 \le j \le n$. Thus, in this case, $I(S) = 0$ and $p_3(S) = n$. Therefore, we have two cases to consider: $p_2(S) = 0$ and $p_2(S) > 0$.

7.1 Case IIa

Let S be cyclic and $p_1(S) = p_2(S) = 0$. Then, as shown above, $I(S) = 0$ and $p_3(S) = n$. The next two lemmas provide some general properties of S.

Lemma 7.1 *If S is cyclic and $p_1(S) = p_2(S) = 0$, then $|v_\alpha \cdot e_\beta| \le 1$ for all $1 \le \alpha, \beta \le n$.*

Proof Let $v_i = \sum_{j=1}^n m_{ij} e_j$ for each i, where $m_{ij} = v_i \cdot e_j$. Then, since $I(S) = 0$, we have $3n = -\sum_{i=1}^n v_i^2 = \sum_{i,j} m_{ij}^2 \ge \sum_{i,j} |m_{ij}| \ge 3n$. Thus $m_{ij}^2 = |m_{ij}|$ for all i and j and so $|v_i \cdot e_j| = |m_{ij}| \le 1$ for all i and j. □

Lemma 7.2 *If S is cyclic and $p_1(S) = p_2(S) = 0$, then S is positive cyclic.*

Proof Again, let $v_i = \sum_{j=1}^n m_{ij} e_j$. By Lemma 7.1, $|m_{ij}| \le 1$ for all i and j. Let $\sum_{i=1}^n v_i = \sum_{i=1}^n \lambda_i e_i$. Then, since $p_3(S) = n$, $\lambda_i \in \{\pm 1, \pm 3\}$ for all i. Now, if S is negative, then $-3n = \sum_{i=1}^n v_i^2 = \left(\sum_{i=1}^n v_i\right)^2 - 2\sum_{i<j} v_i \cdot v_j = \left(-\sum_{i=1}^n \lambda_i^2\right) - 2(n-2)$ or $\sum_{i=1}^n \lambda_i^2 = n+4$. Thus there must exist j such that $\lambda_j = \pm 3$. But then $n - 1 \le \sum_{i \ne j} \lambda_i^2 = n - 5$, which is impossible. Thus S must be positive. □

If $p_3(S) = n$, then it is clear that $n \ge 3$. If $n = 3$, then S is the subset with associated string $(3, 3, 3) \in \mathcal{S}_{2b} \cap \mathcal{S}_{2c}$ found in Section 5.2. From now on, we will assume that $n \ge 4$.

Lemma 7.3 *Let S be cyclic with $p_1(S) = p_2(S) = I(S) = 0$. Suppose there exist integers i and s such that $E_i = \{s-1, s, s+1\}$. Then S is positive and has associated string in $\mathcal{S}_{2b}$.*

Proof By Lemma 7.2, we know that S is necessarily positive. Now, since $E_i = \{s-1, s, s+1\}$, we necessarily have that $a_s \ge 3$; otherwise, if $a_s = 2$ and $V_s = \{i, i'\}$, then $|E_{i'}| = 1$, which is a contradiction. We further claim that $a_{s-1} \ge 3$ or $a_{s+1} \ge 3$. Suppose otherwise: $a_{s-1} = a_{s+1} = 2$. Then $V_{s-1} = V_{s+1} = \{i, j\}$ for some integer j and, since $|E_i| = 3$, we necessarily have that $j \in V_{s-2} \cap V_{s+2}$. Since $|E_j| = 3$, we necessarily have that $n = 4$. But then there exists an integer $k \in V_s$ such that either $|E_k| = 1$ or $|E_k| = 2$, which is a contradiction. Without loss of generality, assume that $a_{s-1} \ge 3$.

First assume that $v_{s-1} \cdot e_i = v_s \cdot e_i$ (or similarly $v_{s+1} \cdot e_i = v_s \cdot e_i$). Let $x \ge 0$ be the smallest integer such that $a_{s+x+1} \ge 3$. Since $a_{s+1} = \cdots = a_{s+x} = 2$, we have $V_{s+\alpha} = \{i_{\alpha-1}, i_\alpha\}$ for all $1 \le \alpha \le x$, where $i_0 := i$ and $\{i_0, \dots, i_x\}$ contains $x+1$ distinct integers.

Moreover, $E_{i_\alpha} = \{s-1, s+\alpha, s+\alpha+1\}$ for all $1 \le \alpha \le x$. Since $v_{s-1} \cdot e_i = v_s \cdot e_i$, by Lemmas 7.1 and 7.2, there exist integers $m, k \in V_{s-1} \cap V_s$ such that $v_{s-1} \cdot e_m = -v_s \cdot e_m$ and $v_{s-1} \cdot e_k = -v_s \cdot e_k$. Thus $a_{s-1} \ge x+3$. Let $R = \{v'_1, \dots, v'_{s-1}, v'_{s+x+1}, \dots, v_n\} \subset \mathbb{Z}^{n-x-1} = \langle e_1, \dots, e_n\rangle / \langle e_{i_0}, \dots, e_{i_x}\rangle$, where $v'_{s-1} = \pi_{e_{i_0}}\big(\pi_{e_{i_1}}(\cdots(\pi_{e_{i_x}}(v_{s-1}))\cdots)\big)$, $v'_{s+x+1} = \pi_{e_{i_x}}(v_{s+x+1})$, and $v'_y = v_y$ for all $y \notin \{s-1, \dots, s+x+1\}$. Then R is negative cyclic with $I(R) = 1 - a_s \le -2$. By Proposition 7.14 in Section 7.2, R must have associated string in $\mathcal{S}_{1c} \cup \mathcal{S}_{1d} \cup \mathcal{S}_{1e} \cup \mathcal{O} \cup \{(2^{[n]}) \mid n \ge 2\}$ and hence either $I(R) = -(n-x-1)$ or $I(R) = -2$. In the former case, we necessarily have that $a_{s-1} = 3+x$, $a_s = n+x$, and $a_{s+x+1} = 3$; hence S has associated string of the form $(3+x, n+x, 2^{[x]}, 3, 2^{[n-x-3]}) \in \mathcal{S}_{2b}$. In the latter case, $a_s = 3$ and so $V_s^S \cap V_{s-1}^S = \{i, m, k\}$. Since $v_s^2 = -3$, it follows that $V_m^S = V_k^S = \{s-1, s, z\}$ for some integer $z \notin \{s-1, s\}$. It is easy to see that $v_{s-1}^2 \le -(4+x)$ and $\tilde{v}_z^2 \le -3$. Let $T = (S \setminus \{v_z, v_s, v_{s-1}\}) \cup \{\pi_{e_k}(v_s), \pi_{e_m}(\pi_{e_k}(v_{s-1}))\}$. Then T is standard with $I(T) \le -3$ and $E_m^T = \{s\}$. By Proposition 5.9, $I(T) = -3$ and so $v_z^2 = -3$; by Proposition 5.10(1), T has associated string of the form $(b_1, \dots, b_k, 2, c_l, \dots, c_1)$, where $(b_1, \dots, b_k)$ and $(c_1, \dots, c_l)$ are linear-dual strings and the middle vertex with square -2 is $\pi_{e_k}(v_s)$. Thus S has associated string $(3, b_1, \dots, b_k+2, 3, c_l, \dots, c_1)$. Since $(\beta_1, \dots, \beta_\kappa) = (b_1, \dots, b_k+1)$ has linear-dual $(\gamma_1, \dots, \gamma_\lambda) = (2, c_1, \dots, c_l)$ (see Lemma 3.6), we have

$$(3, b_1, \dots, b_k+2, 3, c_l, \dots, c_1) = (3, \beta_1, \dots, \beta_{\kappa-1}, \beta_\kappa+1, \gamma_\lambda+1, \gamma_{l-1}, \dots, \gamma_1) \in \mathcal{S}_{2b}.$$

Now assume that $v_{s-1} \cdot e_i = -v_s \cdot e_i = v_{s+1} \cdot e_i$. Suppose $a_{s+1} = 2$ and set $V_{s+1} = \{i, j\}$. Note that $E_j = \{s-1, s+1, s+2\}$ and $V_s \cap V_{s+1} = \{i\}$. Thus v_s is the center of S relative to e_i. Perform the contraction of S centered at v_s to obtain the positive cyclic subset $S' = \{v'_1, \dots, v'_s, v'_{s+2}, \dots, v'_n\}$, where $v'_x = v_x$ for all $x \notin \{s-1, s, s+1\}$, $v'_s = v_s + v_{s+1}$, and $v'_{s-1} = \pi_{e_i}(v_{s-1})$. Then $I(S') = 0$ and $p_3(S') = n-1$. Now the vertices v'_{s-1}, v'_s, and v'_{s+2} are adjacent in S', $E_j^{S'} = \{s-1, s, s+2\}$, and $(v'_s)^2 = v_s^2 \le -3$. Thus v'_s is the center of S' relative to e_j. Moreover, $v'_{s-2} \cdot e_j = -v'_s \cdot e_j = v'_{s+2} \cdot e_j$. If $(v'_{s-2})^2 = -2$ or $(v'_{s+1})^2 = -2$, then we can contract S' centered at v'_s. Continuing in this way, we have a sequence of contractions centered at v_s which terminates in a positive subset $\widetilde{S}$ such that the resulting center vertex $\tilde{v}_s$ has adjacent vertices whose squares are both at most -3. Reindex $\widetilde{S}$ chronologically and let $u = s$ under the new indexing. Then $\tilde{v}_u^2 = v_s^2 \le -3$, $\tilde{v}_{u\pm1}^2 \le -3$, and there is an integer l such that $E_l^{\widetilde{S}} = \{u-1, u, u+1\}$ and $\tilde{v}_{u-1} \cdot e_l = -\tilde{v}_u \cdot e_l = \tilde{v}_{u+1} \cdot e_l$. Note that, if $a_{s+1} \ge 3$, then $\widetilde{S} = S$. Let C be the subset obtained from $\widetilde{S}$ by removing $\tilde{v}_u$, replacing $\tilde{v}_{u\pm1}$ with $\tilde{v}'_{u\pm1} = \pi_{e_l}(\tilde{v}_{u\pm1})$, and setting $\tilde{v}'_x = \tilde{v}_x$ for all $x \notin \{u-1, u, u+1\}$. Then $I(C) \le -2$, $p_1(C) = 0$, $p_2(C) > 0$, and

$\tilde{v}_{u-1}\cdot\tilde{v}_{u+1}=1$. If there exists a vertex of C with square at most -3, then C is a positive cyclic subset. However, by Proposition 7.14 in Section 7.2, positive cyclic subsets with $p_1=0$ and $p_2>0$ have associated strings in $\mathcal{S}_{2c}\cup\mathcal{S}_{2d}$ and thus have $I\in\{-1,0\}$. Since $I(C)\le -2$, every vertex of C must have square equal to -2 and so $\widetilde{S}$ has associated string of the form $(3+x,3,2^{[x]},3)$, where $-(\tilde{v}_u)^2=3+x$. Notice that $(3-1)=(2)$ and $(3-1)=(2)$ are reverse linear-dual strings. Thus, by Lemma 5.14, S has associated string of the form $(3+x,b_1,\dots,b_{k-1},b_k+1,2^{[x]},c_l+1,c_{l-1},\dots,c_1)\in\mathcal{S}_{2b}$, where $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$ are linear-dual strings. □

Lemma 7.4 *Let S be a cyclic subset with $p_1(S)=p_2(S)=I(S)=0$. Suppose that, for all $1\le i\le n$, $E_i\ne\{s-1,s,s+1\}$ for some integer s. Then S is positive with associated string in $\mathcal{S}_{2c}$.*

Proof Let s be an integer such that $a_s\ge 3$. Let i be an integer such that $v_s\cdot e_i=-v_{s+1}\cdot e_i$, which exists by Lemmas 7.1 and 7.2. Finally, let $E_i=\{s-1,s,t\}$. By assumption, $t\notin\{s-2,s+1\}$. Let $x\ge 0$ be the smallest integer such that $a_{s+x+1}\ge 3$. Since $a_{s+1}=\dots=a_{s+x}=2$, we have $V_{s+\alpha}=\{i_{\alpha-1},i_\alpha\}$ for all $1\le\alpha\le x$, where $i_0:=i$ and $\{i_0,\dots,i_x\}$ contains $x+1$ distinct integers. Since $i\in V_t$ and $v_t\cdot v_{s+\alpha}=0$ for all $1\le\alpha\le x-1$, we have $i_0,\dots,i_{x-1}\in V_t$. If $t=s+x+1$, then it is clear that $i_x\notin V_t$ and so $|E_{i_x}|=1$, which is a contradiction. Thus $v_t\cdot v_{s+x}=0$ and so $i_x\in V_t\cap V_{s+x+1}$, and $a_t\ge x+1$. Moreover, since $E_{i_x}=\{s+x,s+x+1,t\}$, by assumption, $t\ne s+x+2$. Now, since $v_t\cdot v_{s-1}=v_t\cdot v_{s+x+1}=0$, there exist integers $m_1\in(V_t\setminus\{i_0,\dots,i_x\})\cap V_{s-1}$ and $m_2\in(V_t\setminus\{i_0,\dots,i_x\})\cap V_{s+x+1}$, implying that $a_t\ge 2+x$. If $a_t=2+x$, then $m_1=m_2$; set $m:=m_1=m_2$. But then $m\in V_{t\pm1}$, implying that $|E_m|\ge 5$, which is a contradiction. Thus $a_t\ge 3+x$. Let $G=\{v'_1,\dots,v'_{s-1},v'_{s+x+1},\dots,v'_{t-1},v'_{t+1},\dots,v_n\}\subset\mathbb{Z}^{n-x-1}=\langle e_1,\dots,e_n\rangle/\langle e_{i_0},\dots,e_{i_x}\rangle$, where $v'_{s-1}=\pi_{e_i}(v_{s-1})$, $v'_{s+x+1}=\pi_{e_{i_x}}(v_{s+x})$, and $v'_\alpha=v_\alpha$ for all $\alpha\notin\{s-1,\dots,s+x+1,t\}$. Then G has two components and $p_1(G)=p_4(G)=0$ and $I(G)\le -2$.

We first claim that G is irreducible and thus a good subset. Suppose otherwise. Then G is the union of two standard subsets G_1 and G_2. By Proposition 5.9, $I(G_1),I(G_2)\ge -3$. Since $p_1(G)=p_4(G)=0$, Proposition 5.9 tells us that $I(G_1),I(G_2)\ge 0$. Consequently, $-2=I(G)=I(G_1)+I(G_2)\ge 0$, a contradiction. Thus G is a good subset. Moreover, by the hypothesis, there do not integers l and z such that $E_l^G=\{z-1,z,z+1\}$, implying that neither component of G is bad (see [8, Definition 4.1]). By Proposition 5.11, $I(G)=-2$ (so $a_t=3+x$) and G_1 and G_2 have associated strings of the form $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$, where $(b_1,\dots,b_k)$ and $(c_1,\dots,c_l)$

are linear-dual strings. Thus G has associated string of the form $(b_1,\dots,b_k,c_1,\dots,c_l)$ or $(b_1,\dots,b_k,c_l,\dots,c_1)$.

To determine which string is correct, we first claim that $m_1\neq m_2$. Assume otherwise, and set $m:=m_1=m_2$. Since $a_t=3+x$, we have $V_t^S=\{i_0,\dots,i_x,m,z\}$ for some integer z. Since $E_m^S=\{s-1,s+x+1,t\}$, we necessarily have that $E_z^S=\{t-1,t,t+1\}$, contradicting the hypothesis of the lemma. Thus $m_1\neq m_2$ and $V_t^S=\{i_0,\dots,i_x,m_1,m_2\}$. Once again by the hypothesis, we may assume that $m_1\in V_{t-1}^S$ and $m_2\in V_{t+1}^S$. Thus $E_{m_1}^G=\{s-1,t-1\}$ and $E_{m_2}^G=\{s+x+1,t+1\}$. By Proposition 5.11, G must have associated string $(b_1,\dots,b_k,c_1,\dots,c_l)$. Consequently, S has associated string of the form $(3+x,b_1,\dots,b_{k-1},b_k+1,2^{[x]},c_1+1,c_2,\dots,c_l)$. Note that, by Lemma 3.6, $(\beta_1,\dots,\beta_\kappa)=(2+x,b_1,\dots,b_k)$ has linear-dual $(\gamma_1,\dots,\gamma_\lambda)=(2^{[x]},c_1+1,c_2,\dots,c_1)$; hence S has associated string

$$(\beta_1+1,\beta_2,\dots,\beta_{\kappa-1},\beta_\kappa+1,\gamma_1,\dots,\gamma_\lambda)\in\mathcal{S}_{2c}. \qquad \square$$

Combining the above two lemmas, we have proven the following:

Proposition 7.5 *Let S be a cyclic subset with $I(S)\le 0$ and $p_1(S)=p_2(S)=0$. Then S is positive with associated string in $\mathcal{S}_{2b}\cup\mathcal{S}_{2c}$.*

7.2 Case IIb: $p_2(S)>0$

Throughout this section, we will consider cyclic subsets satisfying $p_1(S)=0$ and $p_2(S)>0$. In light of Example 5.7, we will further restrict ourselves to cyclic subsets containing at least one vertex with square at most -3. By the discussion in Section 5.2, there are no such cyclic subsets of length 2 or 3. Thus we assume that $n\ge 4$. We start with some useful notation and some preliminary lemmas.

Definition 7.6 Let $S=\{v_1,\dots,v_n\}\subset\mathbb{Z}^n$ be any subset. We define the sets

$$\mathcal{I}^S=\{i\mid E_i=\{s,t\}\text{ and }a_s=2\text{ or }a_t=2\},\quad \mathcal{J}^S=\{i\mid E_i=\{s,t\}\text{ and }a_s,a_t\ge 3\}.$$

In some cases, we will drop the superscript S from the notation if the subset being considered is understood. Notice that $p_2(S)=|\mathcal{I}^S|\cup|\mathcal{J}^S|$. For each $i\in\mathcal{I}^S\cup\mathcal{J}^S$, let $E_i=\{s(i),t(i)\}$. For each $i\in\mathcal{I}^S$, assume $a_{s(i)}=2$.

Lemma 7.7 *Let S be cyclic, $I(S)\le 0$, $p_1(S)=0$, $p_2(S)>0$, and $n\ge 4$. If $i\in\mathcal{I}$, then $a_{t(i)}\ge 3$.*

Proof Set $s := s(i)$ and $t := t(i)$. Assume $a_t = 2$. Suppose $v_s \cdot v_t = 0$. Then $V_s = V_t = \{i, j\}$ for some j, and $E_j \supseteq \{s-1, s, s+1, t-1, t, t+1\}$. If $n \geq 5$, then either $v_{s-1} \cdot v_t = 0$ or $v_{s+1} \cdot v_t = 0$, and so $i \in V_{s-1}$ or $i \in V_{s+1}$, which is a contradiction. If $n = 4$, then $t \pm 1 = s \mp 1$. Since $v_{t-1} \cdot v_{t+1} = 0$, there exists an integer k such that $k \in V_{t\pm1}$. Moreover, there exists a fourth integer m such that $m \in V_{t+1}$ or V_{t-1}, but not both, since $v_{t-1} \cdot v_{t+1} = 0$. Thus $p_1(S) > 0$, contradicting the hypothesis.

Now suppose $|v_s \cdot v_t| = 1$ and, without loss of generality, let $t = s+1$. Since $a_s = a_{s+1} = 2$, we have $V_s = \{i, j\}$ and $V_{s+1} = \{i, i_1\}$, where $i_1 \neq j$. Let $l \geq 2$ be the smallest integer such that $a_{s+l} \geq 3$. Then it is easy to see that $V_{s+\alpha} = \{i_{\alpha-1}, i_\alpha\}$ for all $1 \leq \alpha \leq l-1$, where $i_0 := i$, $i_\alpha \notin \{i, j\}$ for all $1 \leq \alpha \leq l-1$ and the i_α are all distinct. Similarly, let $m \geq 1$ be the smallest integer such that $a_{s-m} \geq 3$. Then, as above, $V_{s-\beta} = \{j_{\beta-1}, j_\beta\}$ for all $1 \leq \beta \leq m-1$, where $j_0 := j$ and the set $\{j_\beta, i, i_\alpha\}$ has $m+l$ distinct elements. Now, since $|v_{s+l-1} \cdot v_{s+l}| = 1$, we must have that $V_{s+l-1} \cap V_{s+l} = \{i_{l-1}\}$ and $|v_{s+l} \cdot e_{i_{l-1}}| = 1$. Similarly, $V_{s-m+1} \cap V_{s-m} = \{j_{m-1}\}$ and $|v_{s-m} \cdot e_{j_{m-1}}| = 1$. Moreover, $E_{i_\alpha} = \{s+\alpha, s+\alpha+1\}$ and $E_{j_\beta} = \{s-\beta, s-\beta-1\}$ for all α and β.

If $v_{s-m} = v_{s+l} = v_u$, then $\{i_{l-1}, j_{m-1}\} \subset V_u$. Since $|v_u \cdot e_{i_{l-1}}| = |v_u \cdot e_{j_{m-1}}| = 1$ and $a_u \geq 3$, we have $|V_u| \geq 3$. Thus there is an integer p such that $E_p = \{u\}$, which contradicts $p_1(S) = 0$. Now suppose $v_{s-m} \neq v_{s+l}$. Let $T = \{v'_1, \dots, v'_{s-1}, v'_{s+l}, \dots, v'_n\} \subset \mathbb{Z}^{n-(m+l)} = \langle e_1, \dots, e_n\rangle / \langle e_{i_0}, \dots, e_{i_{l-1}}, e_{j_0}, \dots, e_{j_{m-1}}\rangle$, where $v'_{s-m} = \pi_{e_{j_{m-1}}}(v_{s-m})$ and $v'_{s+l} = \pi_{e_{i_{l-1}}}(v_{s+l})$. Since $|v_{s+l} \cdot e_{i_{l-1}}| = |v_{s-m} \cdot e_{j_{m-1}}| = 1$ and $a_{s-m}, a_{s+l} \geq 3$, we have $(v'_{s-m})^2, (v'_{s+l})^2 \leq -2$. Thus T is a standard subset made of $n - (l+m-1)$ vectors. However, by Remark 5.3, these vectors are linearly independent in $\mathbb{Z}^{n-(l+m)}$, which is not possible. □

Lemma 7.8 *Let S be cyclic, $I(S) \leq 0$, $p_1(S) = 0$, $p_2(S) > 0$, and $n \geq 4$. If $i \in \mathcal{I}$, then $v_{s(i)} \cdot v_{t(i)} = 0$.*

Proof Set $s := s(i)$ and $t := t(i)$. Let $V_s = \{i, j\}$. Then, by Lemma 7.7, $a_t \geq 3$. Assume $|v_s \cdot v_t| = 1$ and, without loss of generality, assume $t = s+1$. Then $\{s-1, s\} \subseteq E_j$. If there exists an integer $u \notin \{s-1, s, s+1\}$ such that $u \in E_j$, then we necessarily have that $i \in V_u$, implying that $|E_i| \geq 3$, which is not possible. Thus either $E_j = \{s-1, s\}$ or $E_j = \{s-1, s, s+1\}$.

If $E_j = \{s-1, s\}$, then, by Lemma 7.7, $a_{s-1} \geq 3$. Moreover, since $|v_s \cdot e_i| = |v_s \cdot e_j| = 1$, $V_s \cap V_{s-1} = \{j\}$ and $V_s \cap V_{s+1} = \{i\}$, we have $|v_{s+1} \cdot e_i| = |v_{s-1} \cdot e_j| = 1$. Let $T = \{v'_1, \dots, v'_{s-1}, v'_{s+1}, \dots, v'_n\} \subset \mathbb{Z}^{n-2} = \langle e_1, \dots, e_n\rangle / \langle e_i, e_j\rangle$, where $v'_{s+1} = \pi_{e_j}(v_{s+1})$,

$v'_{s-1}=\pi_{e_j}(v_{s-1})$, and $v'_x=v_x$ for all $x\notin\{s-1,s,s+1\}$. Then $(v'_{s\pm1})^2\le-2$ and $v'_{s-1}\cdot v'_{s+1}=0$. Thus T is standard with final vertices v'_{s-1} and v'_{s+1}. By Remark 5.3, $T\subset\mathbb{Z}^{n-2}$ contains $n-1$ linearly independent vectors, which is impossible.

If $E_j=\{s-1,s,s+1\}$, then, since $v_{s-1}\cdot v_{s+1}=0$, there exists an integer $k\notin\{i,j\}$ such that $k\in V_{s-1}\cap V_{s+1}$. Moreover, $|v_{s-1}\cdot e_j|=1$ and, since $V_{s+1}\cap V_s=\{i,j\}$ and $|v_{s+1}\cdot v_s|=1$, we have $|v_{s+1}\cdot e_i|=x$ and $|v_{s+1}\cdot e_j|=x\pm1$, where $x,x\pm1\ne0$. Thus $a_{s+1}\ge x^2+(x\pm1)^2+1\ge6$. If $|v_{s+1}\cdot e_i|=x\ge2$, let $T=\{v'_1,\dots,v'_{s-1},v'_{s+1},\dots,v'_n\}\subset\mathbb{Z}^{n-1}=\langle e_1,\dots,e_n\rangle/\langle e_i\rangle$, where $v'_{s+1}=\pi_{e_i}(v_{s+1})$ and $v'_x=v_x$ for all $x\notin\{s,s+1\}$. Then T is standard and $0\ge I(S)=I(T)+x^2+(a_s-3)=I(T)+x^2-1$. Thus $I(T)\le1-x^2<0$ and so, by Proposition 5.9, we necessarily have that $I(T)=-3$ and $p_1(T)=1$. But then $p_1(S)=p_1(T)=1$, which contradicts our assumption that $p_1(S)=0$. Now suppose $|v_{s+1}\cdot e_i|=1$, so that $|v_{s+1}\cdot e_j|=2$. Since $|v_{s-1}\cdot e_j|=1$ and $|v_{s-1}\cdot v_{s+1}|=0$, either $a_{s-1}\ge3$ or $a_{s-1}=2$ and $|v_{s+1}\cdot e_k|=2$. In the latter case, note that $E_k=\{s-2,s-1,s+1\}$ and $E_j=\{s-1,s,s+1\}$. In this case, let $T'=\{v'_1,\dots,v'_{s-2},v'_{s+1},\dots,v'_n\}=\subset\mathbb{Z}^{n-2}=\langle e_1,\dots,e_n\rangle/\langle e_i,e_j\rangle$, where $v'_{s+1}=\pi_{e_i}(\pi_{e_j}(v_{s+1}))$ and $v'_x=v_x$ for all $x\notin\{s-1,s,s+1\}$. Then T' is standard with $p_1(T')=0$ and $0\ge I(S)=I(T')+5+(a_{s-1}-3)+(a_s-3)=I(T')+3$, implying that $I(T')\le-3$. But, by Proposition 5.9, no such subset exists. In the former case ($a_{s-1}\ge3$), let $T''=\{v'_1,\dots,v'_{s-1},v'_{s+2},\dots,v'_n\}\subset\mathbb{Z}^{n-2}=\langle e_1,\dots,e_n\rangle/\langle e_i,e_j\rangle$, where $v'_{s-1}=\pi_{e_j}(v_{s-1})$ and $v'_x=v_x$ for all $x\notin\{s-1,s,s+1\}$. Then T'' is a standard subset such that $0\ge I(S)=I(T'')+1+(a_s-3)+(a_{s+1}-3)\ge I(T'')+3$. By Proposition 5.9, we necessarily have that $I(T'')=-3$ and $p_1(T'')=1$. Thus $a_{s+1}=6$ and $V^S_{s+1}=\{i,j,k\}$. This implies that $|E^{T''}_k|=1$. But $k\in V^{T''}_{s-1}$ and v_{s-1} is a final vertex of T''. By Proposition 5.10(a), no such standard subset exists. □

Lemma 7.9 *Let S be cyclic, $I(S)\le0$, $p_1(S)=0$, $|\mathcal{I}|>0$, and $n\ge4$.*

(a) *If there exist integers $i,i'\in\mathcal{I}$ such that $|v_{s(i)}\cdot v_{s(i')}|=1$, then S is negative and has associated string in $\mathcal{S}_{1d}$, $|\mathcal{J}|=0$, and $|v_\alpha\cdot v_\beta|\le1$ for all $1\le\alpha,\beta\le n$.*

(b) *If $v_{s(i)}\cdot v_{s(i')}=0$ for all $i,i'\in\mathcal{I}$, then $p_4(S)\ge|\mathcal{I}|$.*

Proof Suppose $|v_{s(i)}\cdot v_{s(i')}|=1$ and, without loss of generality, assume $s(i')=s(i)+1$. Then $t(i)=s(i)+2$, $t(i')=s(i)-1$, and there exists an integer j such that $E_j=\{s(i)-1,s(i),s(i)+1,s(i)+2\}$. Set $s:=s(i)$. By Lemma 7.7, $a_{s-1},a_{s+2}\ge3$; consequently, $n\ge5$. Without loss of generality, assume $v_{s-1}\cdot v_s=v_s\cdot v_{s+1}=1$, so that $v_{s-1}\cdot e_j=-v_s\cdot e_j=v_{s+1}\cdot e_j\in\{\pm1\}$. Let $S'=\{v'_1,\dots,v'_{s-1},v'_{s+1},\dots,v'_n\}\subset\mathbb{Z}^{n-1}=\langle e_1,\dots,e_n\rangle/\langle e_i\rangle$, where $v'_{s+2}=\pi_{e_i}(v_{s+2})$, $v'_{s-1}=\pi_{e_{i'}}(v_{s-1})$, and $v'_x:=v_x$

for all $x \notin \{s-1, s, s+2\}$. Then S' is cyclic with $I(S') = I(S) - 1 < 0$ and $p_1(S') = 1$ (since $E_{i'}^{S'} = \{s+1\}$). Moreover, $v'_{s-1} \cdot e_j = v'_{s+1} \cdot e_j$ and so S' is positive if and only if S is negative. By the proof of Proposition 6.5, the only cyclic subset with $p_1 = 1$ and $I < 0$ is positive and has associated string of the form $(2, b_1+1, b_2, \dots, b_k, 2, c_l, \dots, c_2, c_1+1, 2) \in \mathcal{S}_{2e}$. Moreover, the vertex with square 2 in the middle of the string is v'_{s+1}. Thus S is negative and has associated string of the form $(2, b_1+1, b_2, \dots, b_k+1, 2, 2, c_l+1, \dots, c_2, c_1+1, 2) \in \mathcal{S}_{1d}$. Furthermore, by the proof of Proposition 6.5, it is easy to see that $|v'_\alpha \cdot v'_\beta| \leq 1$ for all α and β and $|\mathcal{J}^{S'}| = 0$; hence $|v_\alpha \cdot v_\beta| \leq 1$ for all $1 \leq \alpha, \beta \leq n$ and $|\mathcal{J}^S| = 0$.

By Lemma 7.8, for all $i \in \mathcal{I}^S$, there exists an integer $j(i)$ such that

$$E_{j(i)} = \{s(i)-1, s(i), s(i)+1, t(i)\}.$$

If $v_{s(i)} \cdot v_{s(i')} = 0$ for some $i, i' \in \mathcal{I}^S$, it follows that $j(i) \neq j(i')$; hence, if $v_{s(i)} \cdot v_{s(i')} = 0$ for all $i, i' \in \mathcal{I}^S$, then $p_4(S) \geq |\mathcal{I}^S|$. □

Lemma 7.10 *Let S be cyclic, $I(S) \leq 0$, $p_1(S) = 0$, $p_2(S) > 0$, and $n \geq 4$. Then $|v_\alpha \cdot e_\beta| \leq 1$ for all integers α and β.*

Proof By Lemma 7.9, we may assume that $v_{s(i)} \cdot v_{s(i')} = 0$ for all $i, i' \in \mathcal{I}$, so that $p_4(S) \geq |\mathcal{I}|$. First suppose that $|\mathcal{J}| \neq 0$. Let $i \in \mathcal{J}$ and set $s := s(i)$ and $t := t(i)$. Notice that we cannot have $|V_s| = |V_t| = 2$. Without loss of generality, assume that $|V_s| \geq 3$. Let $T = \{v'_1, \dots, v'_{t-1}, v'_{t+1}, \dots, v'_n\} \subset \mathbb{Z}^{n-1} = \langle e_1, \dots, e_n \rangle / \langle e_i \rangle$, where $v'_s = \pi_{e_i}(v_s)$ and $v'_x = v_x$ for all $x \notin \{s, t\}$. Then $(v'_s)^2 \leq -2$ and $v'_{t-1} \cdot v'_{t+1} = 0$, and so T is standard. Let $|v_s \cdot e_i| = x \geq 1$. Then

$$0 \geq I(S) = I(T) + x^2 + (a_t - 3) \geq I(T) + x^2 \geq I(T) + 1.$$

Thus $I(T) \leq -1$ and so, by Proposition 5.9, $I(T) \in \{-1, -2, -3\}$. Thus $a_t \leq 5$ and $|v_s \cdot e_i| = x = 1$. Moreover, by Proposition 5.9, $|v'_\alpha \cdot e_\beta| \leq 1$ for all α and β. Thus $|v_\alpha \cdot e_\beta| \leq 1$ for all $\alpha \neq t$ and all β. If $|v_t \cdot e_j| \geq 2$ for some j, then, since $a_t \leq 5$, we necessarily have $V_t = \{i, j\}$ and $a_t = 5$; consequently, $I(T) = -3$ and, by Proposition 5.9, $p_1(T) = 1$. In particular, $|E_j^T| = 1$ and $E_j^S = \{s, t\}$. If $v_s \cdot v_t = 0$, then $v_t \cdot v_{t\pm1} = 0$, which is a contradiction. If $|v_s \cdot v_t| = 1$ and, say, $t = s+1$, then $v_{s+1} \cdot v_{s+2} = 0$, which is a contradiction. Thus $|v_\alpha \cdot e_\beta| \leq 1$ for all α and β.

Now suppose $|\mathcal{J}| = 0$. Then $p_4(S) \geq p_2(S)$ and so, by Lemma 5.16, $I(S) = 0$, $p_2(S) = p_4(S)$ and $p_j(S) = 0$ for all $j = 5, \dots, n$. Thus $p_3(S) = n - 2p_2(S)$. Let

$m_{ij} := v_i \cdot e_j$. Then

$$3n = \sum a_i = \sum_{i,j} m_{ij}^2 \geq \sum_{i,j} |m_{ij}| \geq \sum i p_i(S) = 2p_2(S) + 4p_2(S) + 3(n - 2p_2(S)) = 3n.$$

Thus $|v_i \cdot e_j| = |m_{ij}| \leq 1$ for all i and j. □

In light of Lemma 7.10, it will now be a standing assumption that $|v_\alpha \cdot e_\beta| \leq 1$ for all integers α and β.

Lemma 7.11 *Suppose S is cyclic with $n \geq 4$ and $|\mathcal{J}| \neq 0$. If there exists $i \in \mathcal{J}$ with $a_{s(i)}, a_{t(i)} \geq 4$, then S is positive with associated string $(4,4,2,2,2) \in \mathcal{S}_{2d}$.*

Proof By cyclically reordering and negating vertices, we may assume $s(i) = 1$ and $t(i) = k$ for some integer k. Let $R = \{v'_1, \dots, v'_n\} \subset \mathbb{Z}^{n-1} = \langle e_1, \dots, e_n \rangle / \langle e_i \rangle$, where $v'_1 = \pi_{e_i}(v_1)$, $v'_k = \pi_{e_i}(v_k)$, and $v'_i := v_i$ for all $i \neq 1, k$.

Case 1 ($v_1 \cdot v_k = 0$, so $k \notin \{2, n\}$) By Lemma 7.10, $-(v'_1)^2 = a_1 - 1$, $-(v'_k)^2 = a_k - 1$, and $v'_1 \cdot v'_k = \pm 1$. Let A be the intersection matrix $A = (v'_i \cdot v'_j)$. Assume $a_1, a_k \geq 4$. By Lemma A.4, if S is negative cyclic or S is positive cyclic with $v_1 \cdot e_i = -v_k \cdot e_i$, then A is negative definite; in these cases R is a linearly independent set of n vectors in $\mathbb{Z}^{n-1}$, which is not possible. Thus we may assume that S is positive and $v_1 \cdot e_i = v_k \cdot e_i$. Again by Lemma A.4, we arrive at another linear independence contradiction unless $a_1 = a_k = 4$ and $a_x = 2$ for all $x \notin \{1, k\}$. Thus $I(S) = -(n-4)$. Let $T = \{v'_2, \dots, v'_n\} \subset \mathbb{Z}^{n-1} = \langle e_1, \dots, e_n \rangle / \langle e_i \rangle$, where $v'_k = \pi_{e_i}(v_k)$ and $v'_x = v_x$ for all $x \notin \{1, k\}$. Then T is a standard subset and $I(T) = I(S) - 2 = -(n-2)$. Since $I(T) \geq -3$ by Proposition 5.9, it follows that $n \leq 5$. If $n = 5$, then $I(S) = -1$, $I(T) = -3$, and T has length 4. By Proposition 5.10(1), up to reversal, T has associated string of the form $(3,2,2,2)$. Since $a_t = 4$, this implies that $k = 2$, a contradiction. If $n = 4$, then $I(S) = 0$, $I(T) = -2$, and T has length 3. But, by Proposition 5.10(2), no such standard subset exists.

Case 2 ($|v_1 \cdot v_k| = 1$) Without loss of generality, assume $k = 2$. If $v_1 \cdot e_i = -v_2 \cdot e_i$, then $v'_1 \cdot v'_2 = 0$; hence R is standard and so, by Remark 5.3, R is a linearly independent set of n vectors in $\mathbb{Z}^{n-1}$, a contradiction. If $v_1 \cdot e_i = v_2 \cdot e_i$, then $v'_1 \cdot v'_2 = 2$; by applying Lemma A.5 as in Case 1, we obtain a contradiction unless S is positive, $a_1 = a_2 = 4$, and $a_3 = \cdots = a_n = 2$. In this case, $I(S) = -(n-4)$. As in Case 1, we necessarily have that $n \leq 5$. If $n = 4$, then $I(T) = -2$ and T has length 3; by Proposition 5.10(2), no such subset exists. Suppose $n = 5$, so that $I(T) = -3$ and T has length 4. By Proposition 5.10(1), up to reversal, T has associated string of the form $(3, 2, 2, 2)$. Hence S is positive and has associated string of the form $(4, 4, 2, 2, 2) \in \mathcal{S}_{2d}$. □

We are now ready to finish the classification of cyclic subsets with $I(S) \leq 0$, $p_1(S) = 0$, and $p_2(S) > 0$. We will consider two cases: $|\mathcal{J}| \neq 0$ and $|\mathcal{J}| = 0$. These cases are handled respectively in the next two propositions.

Proposition 7.12 *Let S be cyclic, $I(S) \leq 0$, $p_1(S) = 0$, $p_2(S) > 0$, and $n \geq 4$. If $|\mathcal{J}| \neq 0$, then S is positive with associated string in $\mathcal{S}_{2c} \cup \mathcal{S}_{2d}$ or negative with associated string in $\mathcal{S}_{1c} \cup \mathcal{S}_{1e} \cup \mathcal{O}$.*

Proof Let $i \in \mathcal{J}$ and set $s := s(i)$ and $t := t(i)$. If $a_s, a_t \geq 4$, then, by Lemma 7.11, S is positive with associated string in $\mathcal{S}_{2d}$. Without loss of generality, we may now assume that $a_s = 3$. Moreover, by Lemma 7.9, $v_{s(i_1)} \cdot v_{s(i_2)} = 0$ for all $i_1, i_2 \in \mathcal{I}$, implying that $p_4(S) \geq |\mathcal{I}|$. Let $T = \{v'_1, \dots, v'_{s-1}, v'_{s+1}, \dots, v'_n\} \subset \mathbb{Z}^{n-1} = \langle e_1, \dots, e_n \rangle / \langle e_i \rangle$, where $v'_t = \pi_{e_i}(v_t)$ and $v'_x = v_x$ for all $x \notin \{s, t\}$. By Lemma 7.10, $(v'_t)^2 = v_t^2 + 1$ and so T is a standard subset and $I(T) = I(S) - 1 \leq -1$. By Proposition 5.9, $I(T) \in \{-3, -2, -1\}$. We will work case by case, considering each of the standard subsets listed in Proposition 5.10.

Case 1 $(I(T) = -1)$ By Proposition 5.9, $p_1(T) = 0$, $p_2(T) = 2$, $p_4(T) = 1$, and $p_j(T) = 0$ for all $j \geq 5$. Thus $p_2(S) \leq 3$, $p_4(S) \leq 3$, $p_5(S) \leq 1$, and $p_j(S) = 0$ for all $j \geq 6$. Note that, since $a_s = 3$, if $p_5(S) = 1$, then $p_4(S) = p_2(S) - 2$, and if $p_5(S) = 0$, then $p_2(S) = p_4(S)$. By Lemma 5.17, $p_2(S) + p_4(S) \equiv 0 \mod 4$, implying that either $p_5(S) = 1$, $p_2(S) = 3$ and $p_4(S) = 1$, or $p_5(S) = 0$ and $p_2(S) = p_4(S) = 2$. By Proposition 5.10(3), T is of one of the forms (a)–(c) given there.

Case 1(a) Without loss of generality, we may assume that the listed vertices are $v'_{s+1}, \dots, v'_n, v'_1, \dots, v'_{s-1}$. First assume $p_5(S) = 1$, $p_2(S) = 3$, and $p_4(S) = 1$. Then $2 \in V_s^S$ and $3, x + y + 4 \notin V_s^S$ (where $x + y + 4 = 1$ if $y = 0$). If $y = 0$, then, since $v_{s+2} \cdot v_s = 0$ and $1 \notin V_s^S$, we have $i \in V_{s+2}^S$. Since $v_{s+3} \cdot v_s = 0$ and $2 \in V_s^S$, we have $4 \in V_s^S$ and $v_s \cdot e_2 = v_s \cdot e_4$. Since $V_s^S = \{i, 2, 4\}$, if $x \geq 1$, then $v_{s+4} \cdot v_s \neq 0$, which is a contradiction, and if $x = 0$, then $v_{s-1} \cdot v_s = 0$, which is a contradiction. Thus we may assume $y \geq 1$. Since $v_s \cdot v_{s+2} = v_s \cdot v_{s+3} = 0$ and $a_s = 3$, either $i \in V_{s+2}^S$ and $4 \in V_s^S$, or $i \in V_{s+3}^S$ and $|\{1, x+5, \dots, x+y+3\} \cap V_s^S| = 1$. In the former case, $V_s^S = \{i, 2, 4\}$ and so $|v_s \cdot v_{s+1}| \neq 1$, which is a contradiction. In the latter case, if $1 \in V_s^S$, then $V_s^S = \{i, 1, 2\}$ and $v_s \cdot e_1 = v_s \cdot e_2$ (since $v_s \cdot v_{s+2} = 0$); but then $|v_{s+x+4} \cdot v_s| = 2$, which is a contradiction. On the other hand, if $|\{x+5, \dots, x+y+3\} \cap V_s^S| = 1$, then, since $v_s \cdot v_{s-\alpha} = 0$ for all $2 \le \alpha \le y$, $\{x+5, \dots, x+y+3\} \subset V_s^S$, implying that $y = 1$ and $1 \in V_s^S$, which is again a contradiction.

Now assume $p_5(S) = 0$ and $p_2(S) = p_4(S) = 2$. Then $2 \notin V_s^S$ and either $x+y+4 \in V_s^S$ or $3 \in V_s^S$, but not both (where $x+y+4 = 1$ if $y = 0$). First assume $x+y+4 \in V_s^S$. Since $x+y+4 \in V_{s+2}^S$ and $v_{s+2} \cdot v_s = 0$, either $|\{1, x+5, \dots, x+y+3\} \cap V_s^S| = 1$ or $i \in V_{s+2}^S$. In the former case, $y \ge 1$ and, since $v_{s-\alpha} \cdot v_s = 0$ for all $2 \le \alpha \le y$, it follows that $\{1, x+5, \dots, x+y+3\} \subset V_s^S$, implying that $|v_s \cdot v_{s-1}| \neq 1$, which is a contradiction. In the latter case, since $|v_s \cdot v_{s+1}| = 1$, we have $|\{4, 5, \dots, x+4\} \cap V_s^S| = 1$. Since $v_{s+\alpha} \cdot v_s = 0$ for all $4 \le \alpha \le x+4$, we have $\{4, 5, \dots, x+4\} \subset V_s^S$, which implies that $x = 0$ and $V_s^S = \{i, 4, x+y+4\}$; but then $|v_{s+3} \cdot v_s| = 1$, which is a contradiction.

Now suppose $3 \in V_s^S$. Since $v_s \cdot v_{s+3} = 0$ and $3 \in V_{s+3}^S$, either $i \in V_{s+3}^S$ or $4 \in V_s^S$. In the former case, since $|v_s \cdot v_{s+1}| = 1$, we have $|\{4, 5, \dots, x+4\} \cap V_s^S| = 1$. As in the previous case, we see that $x = 0$ and $V_s^S = \{i, 3, 4\}$ and so $v_{s+3} \cdot v_s \neq 0$, which is a contradiction. In the latter case, since $4 \in V_{s+4}^S$, we have $i \in V_{s+4}^S$ and, since $|v_s \cdot v_{s-1}| = 1$, we necessarily have that $y = 0$. Consequently, S is of the form

$$\Bigl\{e_i - e_4 + e_3, e_2 + e_4 + \sum_{\alpha=5}^{x+4} e_\alpha, e_1 - e_2, e_2 - e_3 - e_4, e_i + e_4 - e_5, \\ e_5 - e_6, \dots, e_{x+3} - e_{x+4}, e_{x+4} - e_1 - e_2 - e_3\Bigr\},$$

which is positive and has associated string $(3, 2+x, 2, 3, 3, 2^{[x-1]}, 4) \in \mathcal{S}_{2c}$.

Case 1(b) As in the previous case, we may label the vertices $v'_{s+1}, \dots, v'_n, v'_1, \dots, v'_{s-1}$. Note that, if $y = 0$, then S is also of the form in Case 1(a), which we already covered. Thus we may assume $y \ge 1$. Consequently, $|\mathcal{I}^T| = 2$. If $p_5(S) = 1$, then $p_2(S) = 3$ and so $|\mathcal{I}^S| = 2$; but we also have that $p_4(S) = 1 \ge |\mathcal{I}^S|$, which is a contradiction. Thus $p_5(S) = 0$ and $p_2(S) = p_4(S) = 2$; hence $2 \notin V_s^S$ and either $1 \in V_s^S$ or $x+y+4 \in V_s^S$, but not both. Assume $x+y+4 \in V_s^S$. Since $x+y+4 \in V_{s+3}^S$, either $i \in V_{s+3}^S$ or $|\{3, 4, x+5, \dots, x+y+3\} \cap V_s^S| = 1$. In the former case, since $|v_s \cdot v_{s+1}| = 1$, following as in Case 1(a) we see that $x = 0$ and $V_s^S = \{i, x+y+4, 4\}$, which implies that $|v_s \cdot v_{s+3}| = 1$, which is a contradiction. In the latter case, since $v_{s-\alpha} \cdot v_s = 0$ for all $2 \le \alpha \le y$, it is clear that $3, x+5, \dots, x+y+3 \notin V_s^S$ and so $4 \in V_s^S$. Since

$4, x+y+4 \in V^S_{s+3}$ and $4 \in V^S_{s+4}$, we have $i \in V^S_{s+4}$. Hence, if $x \geq 1$, S is of the form

$$\Bigl\{e_i - e_4 + e_{x+y+4}, e_2 + e_4 + \sum_{\alpha=5}^{x+4} e_\alpha, e_1 - e_2, e_2 - e_3 - e_4 - \sum_{\alpha=x+5}^{x+y+4} e_\alpha, \\ e_i + e_4 - e_5, \dots, e_{x+3} - e_{x+4}, e_{x+4} - e_1 - e_2 - e_3, e_3 - e_{x+5}, \\ e_{x+5} - e_{x+6}, \dots, e_{x+y+3} - e_{x+y+4}\Bigr\},$$

which is positive and has associated string $(3, 2+x, 2, 3+y, 3, 2^{[x-1]}, 4, 2^{[y]}) \in \mathcal{S}_{2c}$, and if $x = 0$, then S is of the form

$$\Bigl\{e_i - e_4 + e_{y+4}, e_2 + e_4, e_1 - e_2, e_2 - e_3 - e_4 - \sum_{\alpha=5}^{y+4} e_\alpha, e_i + e_4 - e_1 - e_2 - e_3, \\ e_3 - e_5, e_6 - e_7, \dots, e_{y+3} - e_{y+4}\Bigr\},$$

which is positive and has associated string $(3, 2, 2, 3+y, 5, 2^{[y]}) \in \mathcal{S}_{2c}$.

Next assume $3 \in V^S_s$. Since $v_s \cdot v_{s+3} = v_s \cdot v_{s+x+4} = 0$ and $3 \in V^S_{s+3} \cap V^S_{s+x+4}$, either $i \in V^S_{s+3}$ or $i \in V^S_{s+x+4}$. Since $y \geq 1$ and $|v_{s-1} \cdot v_s| = 1$, it follows that $x+y+3 \in V^S_s$ (where $x+y+3 = 3$ if $y = 1$). But then $v_s \cdot v_{s+1} = 0$, which is a contradiction.

Case 1(c) Label the vertices $v'_{s+1}, \dots, v'_n, v'_1, \dots, v'_{s-1}$. Notice that $|\mathcal{I}^T| = 2$ if $y \geq 1$. By the same argument as in Case 1(b), if $y \geq 1$, then $p_5(S) \neq 0$. Suppose $y = 0$, $p_5(S) = 1$, and $p_2(S) = 3$. Then $2 \in V^S_s$ and $3, 4 \notin V^S_s$. Since $2, 3 \in V^S_{s+2}$ and $v_s \cdot v_{s+2} = 0$, we necessarily have that $i \in V^S_{s+2}$. Now, since $V^S_{s+3} \cap V^S_{s+4} = \{2\}$, it follows that either $v_s \cdot v_{s+3} \neq 0$ or $v_s \cdot v_{s+4} \neq 0$, which is a contradiction. Thus we may assume that $p_5(S) = 0$ and $p_2(S) = p_4(S) = 2$. Thus $2 \notin V^S_s$ and either $3 \in V^S_s$ or $x+y+5 \in V^S_s$, but not both (where $x+y+5 = 4$ if $y = 0$). If $x+y+5 \in V^S_{s+3}$, then either $i \in V^S_{s+3}$ or $|\{1, 4, x+6, \dots, x+y+3\} \cap V^S_s| = 1$. In the former case, we obtain a contradiction as in Cases 1(a) and 1(b). In the latter case, we obtain similar contradictions unless $1 \in V^S_s$. In this case, since $1, x+y+5 \in V^S_{s+3}$ and $1 \in V^S_{s+x+5}$, we have $i \in V^S_{s+x+4}$. Thus S is of the form

$$\Bigl\{e_i - e_1 + e_{x+y+5}, e_1 - e_2 - e_5 - \sum_{\alpha=6}^{x+5} e_\alpha, e_2 + e_3, -e_2 - e_1 - e_4 - \sum_{\alpha=x+6}^{x+y+5} e_\alpha, \\ -e_5 + e_2 - e_3, e_5 - e_6, \dots, e_{x+4} - e_{x+5}, -e_i + e_{x+5} + e_1 - e_4, \\ e_4 - e_{x+6}, e_{x+6} - e_{x+7} \dots, e_{x+y+4} - e_{x+y+5}\Bigr\},$$

which is positive and has associated string $(3, 3+x, 2, 3+y, 3, 2^{[x]}, 4, 2^{[y]}) \in \mathcal{S}_{2c}$.

Next suppose $3 \in V_s^S$. Since $2 \notin V_{s+2}^S$ and $v_s \cdot v_{s+2} = 0$, we necessarily have that $i \in V_{s+2}^S$. Since $v_s \cdot v_{s+4} = 0$, we have $5 \in V_s^S$ and so $V_s^S = \{i, 3, 5\}$. Moreover, since $5 \in V_{s+5}^S$, $v_s \cdot v_{s+5} = 0$, and $|v_s \cdot v_{s-1}| = 1$, we must have that $x = y = 0$. Hence S is of the form

$$\{e_i - e_3 + e_5, e_1 - e_2 - e_5, e_2 + e_3 + e_i, -e_2 - e_1 - e_4, -e_5 + e_2 - e_3, e_5 + e_1 - e_4\},$$

which is negative cyclic with associated string $(3, 3, 3, 3, 3, 3) \in \mathcal{O}$.

Case 2 ($I(T) = -2$, so that $I(S) = -1$) By Proposition 5.10(2), $p_1(T) = 0$, $p_2(T) = 3$, $p_4(T) = 1$, $p_j(T) = 0$ for all $j \geq 5$, and $|\mathcal{I}^T| = 2$. Then, since $a_s = 3$, $p_2(S) \leq 4$, $p_4(S) \leq 3$, and $p_5(S) \leq 1$. By Lemma 5.17, $p_2(S) + p_4(S) = 1 \bmod 4$. By a similar argument as in Case 1(b), $p_5(S) = 0$ and so $p_2(S) = 3$ and $p_4(S) = 2$. By Proposition 5.10(2), T is of one of the forms (a)–(c) given there.

Case 2(a) Label the vertices $v'_{s+1}, \dots, v'_n, v'_1, \dots, v'_{s-1}$. Notice that, if $y = 0$, then T is also of the form given in Case 2(b). Moreover, if $x = 0$, then the reverse of T is of the form given in Case 2(b). We will assume that $x, y \geq 1$ and handle the cases $x = 0$ and $y = 0$ in Case 2(b). Since $p_5(S) = 0$ and $p_2(S) = 3$, we have $2 \notin V_s^S$ and $|\{x+4, x+y+4, 3\} \cap V_s^S| = 1$. If $x + 4 \in V_s^S$ or $x + y + 4 \in V_s^S$, then, arguing as in Case 1, we arrive at contradictions. Assume $3 \in V_s^S$. Since $3 \in V_{s+x+4}$ and $v_s \cdot v_{s+x+4} = 0$, either $i \in V_{s+x+4}^S$ or $1 \in V_s^S$, but not both. In the former case, since $|v_s \cdot v_{s\pm 1}| = 1$, we have $x + 3, x + y + 3 \in V_s^S$, implying that $a_s \geq 4$, which is a contradiction. In the latter case, $V_s^S = \{i, 1, 3\}$, implying that $v_s \cdot v_{s+1} = 0$, which is a contradiction.

Case 2(b) Label the vertices $v'_{s+1}, \dots, v'_n, v'_1, \dots, v'_{s-1}$. Notice that, if $x = 0$, then T is of the form in Case 2(c). We will assume that $x \geq 1$ and handle $x = 0$ in Case 2(c). Since $p_5(S) = 0$ and $p_2(S) = 3$, we have $2 \notin V_s^S$ and $|\{x+4, x+y+4, 1\} \cap V_s^S| = 1$ (where $x+y+4 = 3$ if $y = 0$). If $1 \in V_s^S$, then, since $v_{s+x+2} \cdot v_s = 0$, we necessarily have that $i \in V_{s+x+2}^S$. Now, since $|v_{s+1} \cdot v_s| = 1$, we have $x + 3 \in V_s^S$ and so $V_s^S = \{i, 1, x+3\}$; but then $|v_s \cdot v_{s+2}| = 1$, which is a contradiction. If $x + 4 \in V_s^S$, then, since $v_s \cdot v_{s+\alpha} = 0$ for all $2 \leq \alpha \leq x$, it follows that $4, \dots, x+3 \notin V_s^S$. Since $x + 4 \in V_{s+x+3}^S$, we must have that $i \in V_{s+x+3}^S$; consequently, since $|v_{s-1} \cdot v_s| = 1$, we necessarily have that $y \geq 1$ and $x + y + 3 \in V_s^S$. But then $v_{s-2} \cdot v_s \neq 0$, which is a contradiction. Thus $x + y + 4 \in V_s^S$. As above, it is easy to see that $3, x+5, \dots, x+y+3 \notin V_s^S$. Since $x + y + 4 \in V_{s+x+1}^S$, it follows that either $i \in V_{s+x+1}^S$ or $4 \in V_s^S$. In the former case, since $|v_s \cdot v_{s+1}| = 1$, we have $x + 3 \in V_s^S$, which leads to a contradiction. In the latter case, since $4 \in V^S_{s+x+3}$, we see that $i \in V^S_{s+x+4}$. Since $|v_s \cdot v_{s-1}| = 1$, it follows that $x = 1$. Thus S is of the form

$$\Bigl\{e_i + e_4 + e_{x+y+4}, e_5 - e_4, e_4 - e_2 - e_3 - \sum_{i=x+5}^{x+y+4} e_i, e_2 + e_1, e_i - e_2 - e_4 - e_5, \\ e_2 - e_1 - e_3, e_3 - e_{x+5}, e_{x+5} - e_{x+6}, \dots, e_{x+y+3} - e_{x+y+4}\Bigr\},$$

which is positive cyclic with associated string $(3, 2, 3 + y, 2, 4, 3, 2^{[y]}) \in \mathcal{S}_{2d}$.

Case 2(c) Label the vertices $v'_{s+1}, \dots, v'_n, v'_1, \dots, v'_{s-1}$. As usual, since $p_5(S) = 0$, $2 \notin V^S_s$. Notice $2 \in V^S_{s+k+1} \cap V^S_{s+k+2}$. By our standing assumption that $v_{s(i)} \cdot v_{s(i')} = 0$ for all $i, i' \in \mathcal{I}^S$, we necessarily have that either $1 \in V^S_s$ or $4 \in V^S_s$, but not both. Consequently, since $v_s \cdot v_{s+k+1} = v_s \cdot v_{s+k+2} = 0$, either $i \in V^S_{s+k+1}$ or $i \in V^S_{s+k+2}$. Moreover, since $p_2(S) = 3$, $j_1 \notin V^S_s$ and so $j_2 \in V^S_s$. Now, since $j_2 \in V^S_{s+2}$ and $v_s \cdot v_{s+2} = 0$, we necessarily have that $k = 2$ and $4 \in V^S_s$. Hence $V^S_s = \{4, i, j_2\}$, $i \in V^S_{s+k+2}$, and T has associated string of the form $(2, 3 + x, 2, 2, 3, 2^{[x-1]}, 3)$. Moreover, $v_s \cdot e_{j_2} = \pm v_{s-1} \cdot e_{j_2} = \mp v_{s+1} \cdot e_{j_2}$. Thus S is negative and has associated string of the form $(3, 2, 3 + x, 2, 3, 3, 2^{[x-1]}, 3) \in \mathcal{S}_{1e}$.

Case 3 ($I(T) = -3$, so that $I(S) = -2$) By Proposition 5.9, $p_1(T) = 1$, $p_2(T) = 1$, and $p_j(T) = 0$ for all $j \geq 4$. Thus $p_j(S) = 0$ for all $j \geq 5$. Let l be the unique integer such that $|E^T_l| = 1$ and let u be the integer such that $E^T_l = \{u\}$, where $u \neq s \pm 1$. Then, since $p_1(S) = 0$, $l \in V^S_s$. Since $a_s = 3$, we have $p_2(S) \in \{2, 3\}$ and $p_4(S) = p_2(S) - 2$. By Lemma 5.17, $p_2(S) + p_4(S) = 2p_2(S) - 2 \equiv 2 \bmod 4$, implying that $p_2(S) = 2$ and $p_4(S) = 0$. By Proposition 5.10(1), there is an integer k such that $E^T_k = \{s - 1, s + 1\}$ and $v_{s-1} \cdot e_k = -v_{s+1} \cdot e_k$. Since $p_2(S) = 2$, $k \in V^S_s$, and so $V^S_s = \{i, l, k\}$. Since $k \notin V^S_u$, we must have that $i \in V^S_u$. Thus $a_u = 3$. Now, by Proposition 5.10(1), T has associated string $(b_1, \dots, b_k, 2, c_l, \dots, c_1)$, where the middle entry "2" corresponds to the square of v'_u. Now, since $v_{s-1} \cdot e_k = -v_{s+1} \cdot e_k$, we have $v_s \cdot e_k = \pm v_{s-1} \cdot e_k = \mp v_{s+1} \cdot e_k$ and so S is negative and has associated string of the form $(3, b_1, \dots, b_k, 3, c_l, \dots, c_1) \in \mathcal{S}_{1c}$. □

Proposition 7.13 *Let S be cyclic, $I(S) \leq 0$, $p_1(S) = 0$, $p_2(S) > 0$, and $n \geq 4$. If $|\mathcal{J}| = 0$, then S is negative and has associated string in $\mathcal{S}_{1d} \cup \mathcal{O}$.*

Proof Note that $|\mathcal{I}| = p_2(S)$. By Lemma 7.7, $a_{t(i)} \geq 3$ for all $i \in \mathcal{I}$. If there exist $i_1, i_2 \in \mathcal{I}$ such that $v_{s(i_1)} \cdot v_{s(i_2)} \neq 0$, then, by Lemma 7.9, S is negative with associated string in $\mathcal{S}_{1d}$. Now assume that $v_{s(i_1)} \cdot v_{s(i_2)} = 0$ for all $i_1, i_2 \in \mathcal{I}$. Then, by Lemmas 5.16 and 7.9, $p_4(S) = p_2(S)$, $I(S) = 0$, and $p_j(S) = 0$ for all $j \notin \{2,3,4\}$. Let $G = (S \setminus \{v_{s(i)}, v_{t(i)} \mid i \in \mathcal{I}\}) \cup \{\pi_{e_i}(v_{t(i)}) \mid i \in \mathcal{I}\}$ and set $v'_{t(i)} = \pi_{e_i}(v_{t(i)})$ for all $i \in \mathcal{I}$, $v'_x := v_x$ for all $x \notin \{s(i), t(i) \mid i \in \mathcal{I}\}$, and $a'_x = -(v'_x)^2$ for all x. Then $p_2(G) = p_4(S) = 0$, $I(G) = 0$, $p_3(G) = n - p_2(G)$, and, by Lemma 7.9, G has $|\mathcal{I}|$ components. Finally, since, for each $i \in \mathcal{I}$, there exists an integer $j(i)$ such that $E^S_{j(i)} = \{s(i)-1, s(i), s(i)+1, t(i)\}$, G is irreducible and hence a good subset.

Assume C is a component of G of length at least 2. After possibly relabeling, let $C = \{v'_1, \dots, v'_m\}$. Since $v'_1 \cdot v'_2 = 1$, by Lemma 7.10, there is an integer $k \in V^G_1 \cap V^G_2$ such that $v'_1 \cdot e_k = -v'_2 \cdot e_k$. Since $|E^G_k| = 3$, there exists an integer z such that $k \in V^G_z$. Since v'_1 is a final vertex, $v'_z \cdot v'_1 = 0$ and so there exists an integer $l \in V^G_1 \cap V^G_z$. Moreover, since $|E^G_l| = 3$, we necessarily have that $a'_1 \geq 3$. We claim that, if $a'_z = 2$, then $v'_z = v'_3$. If $v'_z \neq v'_3$, then it is clear that v'_z must be isolated. In this case, since $v'_z \cdot v'_2 = 0$, we have $l \in V^G_2$ and $v'_1 \cdot e_l = -v'_2 \cdot e_l$. Since $v'_1 \cdot v'_2 = 1$, there exists another integer $m \in V^G_1 \cap V^G_2$ and so $a'_1, a'_2 \geq 3$. Let $L = (G \setminus \{v'_1, v'_2\}) \cup \{\pi_{e_k}(v'_1), \pi_{e_k}(v'_2)\}$; then L is good and $p_1(L) = 1$. By [8, Corollary 3.5], $I(L) = -3$; but it is clear that $I(L) = I(G) - 2 = -2$, which is a contradiction.

Thus, if $a'_z = 2$, then $v'_z = v'_3$ and we can perform a contraction yielding the subset $G' = G \setminus \{v'_1, v'_2, v'_3\} \cup \{\pi_{e_k}(v'_1), v'_2 + v'_3\}$. Notice that G' is a good subset with $I(G') = 0$ and $p_j(G') = 0$ for all $j \neq 3$; moreover, the component $C' = \{\pi_{e_k}(v'_1), v'_2 + v_3, v'_4, \dots, v'_m\}$ has length one less than the length of C. On the other hand, if $a'_z \geq 3$, then we can perform a contraction yielding the subset $G'' = G \setminus \{v'_1, v'_2, v'_z\} \cup \{v'_1 + v'_2, \pi_{e_k}(v'_z)\}$. As above, G'' is a good subset with $I(G'') = 0$ and $p_j(G'') = 0$ for all $j \neq 3$, and the component C'' resulting from C has length one less than the length of C. We may continue performing contractions in this way until the component C is reduced to an isolated vertex. We can similarly perform contractions on all remaining components until they are all isolated vertices. We obtain a good subset K that contains only isolated vertices. By Lemma 5.18, K is of the form

- $\{e_1 - e_2 + e_3 - e_4, e_1 + e_2, -e_1 + e_2 + e_3 - e_4, e_3 + e_4\}$, or
- $\{e_1 - e_2 - e_3, e_1 + e_2 - e_4, e_2 - e_3 + e_4, e_1 + e_3 + e_4\}$.

It is easy to see that no expansion of either subset exists. Thus $K = G$. Moreover, by construction, $|\mathcal{I}| = 4$ and we may assume that $1 = j(i_1)$, $2 = j(i_2)$, $3 = j(i_3)$, and $4 = j(i_4)$, where $\mathcal{I} = \{i_1, i_2, i_3, i_4\}$. Thus (up to the action of $\operatorname{Aut} \mathbb{Z}^8$), S is of the form either

$$\{e_1-e_2+e_3-e_4-e_{i_2}+e_{i_3}, e_{i_1}-e_1, e_1+e_2,\\ e_{i_2}-e_2, -e_1+e_2+e_3-e_4-e_{i_1}-e_{i_4}, e_{i_3}-e_3, e_3+e_4, e_{i_4}-e_4\}$$

or

$$\{e_1-e_2-e_3-e_{i_2}, e_{i_1}-e_1, e_1+e_2-e_4-e_{i_4}, e_{i_2}-e_2, e_2+e_3+e_4+e_{i_3},\\ e_{i_4}-e_4, e_1+e_3+e_4+e_{i_1}, e_{i_3}-e_3\},$$

So S is negative cyclic with associated string $(6,2,2,2,6,2,2,2)$ or $(4,2,4,2,4,2,4,2)$, both of which are in $\mathcal{O}$. □

To summarize, we have proven the following:

Proposition 7.14 *Let S be a cyclic subset with $p_1(S)=0$, $p_2(S)>0$ and $I(S)\le 0$. Then S is positive with associated string in $\mathcal{S}_{2c}\cup\mathcal{S}_{2d}$ or negative with associated string in $\mathcal{S}_{1c}\cup\mathcal{S}_{1d}\cup\mathcal{S}_{1e}\cup\mathcal{O}\cup\{(2^{[n]}\mid n\ge 2)\}$.*

Appendix

Given a sequence of integers $(a_1,\dots,a_n)$ the (Hirzebruch–Jung) continued fraction expansion is given by

$$[a_1,\dots,a_n]=a_1-\cfrac{1}{a_2-\cfrac{1}{\ddots-\cfrac{1}{a_n}}}.$$

If $a_i\ge 2$ for all $1\le i\le n$, then this fraction is well defined and the numerator is greater than the denominator. In fact, for coprime $p>q>0\in\mathbb{Z}$, there exists a unique continued fraction expansion $[a_1,\dots,a_n]=p/q$, where $a_i\ge 2$ for all $1\le i\le n$. Moreover, by reversing the order of the continued fraction, $[a_n,\dots,a_1]=p/q'$, where q' is the unique integer such that $1\le q'<p$ and $qq'\equiv 1 \bmod p$.

Lemma A.1 *Let $p/q=[a_1,\dots,a_n]$, $s/r=[a_1,\dots,a_{n-1}]$, and $\boldsymbol{a}=(a_1,\dots,a_n)$. Then $|\mathrm{Tor}(H_1(\boldsymbol{T}_{\pm A(\boldsymbol{a})}))|=p-(r\pm 2)$.*

Proof Let $\boldsymbol{a}=(a_1,\dots,a_n)$. By [11, Theorem 6.1], hyperbolic torus bundles are of the form $\boldsymbol{T}_{\pm A(\boldsymbol{a})}=T^2\times[0,1]/(\boldsymbol{x},1)\sim(\pm A\boldsymbol{x},0)$, where

$$A=A(\boldsymbol{a})=\begin{pmatrix} p & q\\ -s & -r\end{pmatrix},\quad \frac{p}{q}=[a_1,\dots,a_n]\quad\text{and}\quad \frac{s}{r}=[a_1,\dots,a_{n-1}].$$

Note that, since $A \in \mathrm{SL}_2(\mathbb{Z})$, we have $qs - pr = 1$. Moreover, since $\boldsymbol{T}_{\pm A(\boldsymbol{a})}$ is hyperbolic, $\operatorname{tr} A(\boldsymbol{a}) = p - r > 2$. Now, by [12, Lemma 10], $|\mathrm{Tor}(H_1(\boldsymbol{T}_{\pm A(\boldsymbol{a})}))| = |\mathrm{tr}(\pm A(\boldsymbol{a})) - 2| = |\pm(p-r) - 2| = |\pm(p-(r \pm 2))| = p - (r \pm 2)$. □

Lemma A.2 *Let $(b_1, \dots, b_k)$ and $(c_1, \dots, c_l)$ be linear-dual strings, where $l + k \geq 2$, $x \geq 1$ be an integer, and $[b_1, \dots, b_k] = p/q$. Then $[b_1, \dots, b_k, x+1, c_l, \dots, c_1] = xp^2/(xpq+1)$ and $[c_1, \dots, c_l, x+1, b_k, \dots, b_1] = xp^2/(xp^2 - xpq + 1)$.*

Proof Given the first conclusion, the second follows since $(xpq+1)(xp^2-xpq+1) = xp^2(xpq - q^2 + 1) + 1$. We will now prove that $[b_1, \dots, b_k, x+1, c_l, \dots, c_1] = xp^2/(xpq+1)$.

Let $n = k + l + 1$ be the length of $(b_1, \dots, b_k, x+1, c_l, \dots, c_1)$. We proceed by induction on n. If $n = 3$, then $k = 1$, $l = 1$, $(b_1) = \frac{2}{1}$, and $[2, x+1, 2] = 4x/(2x+1) = x2^2/(x(2)(1)+1)$. Suppose the lemma is true for all length $n-1$ continued fractions and consider $[b_1, \dots, b_k, x+1, c_l, \dots, c_1]$. By definition of linear-dual strings, either $b_1 = 2$ and $c_1 \geq 3$ or vice versa.

Assume that $b_1 = 2$. Then the strings $(b_2, \dots, b_k)$ and $(c_1 - 1, \dots, c_l)$ are linear-dual and, by the inductive hypothesis,

$$[b_2, \dots, b_k, x+1, c_l, \dots, c_1 - 1] = \frac{xm^2}{xmn+1},$$
$$[c_1 - 1, c_2, \dots, c_l, x+1, b_k, \dots, b_2] = \frac{xm^2}{xm^2 - xmn + 1},$$

where $[b_2, \dots, b_k] = m/n$. Thus,

$$[c_1, c_2, \dots, c_l, x+1, b_k, \dots, b_2] = 1 + \frac{xm^2}{xm^2 - xmn + 1} = \frac{2xm^2 - xmn + 1}{xm^2 - xmn + 1}.$$

Since $(2xmn - xn^2 + 2)(xm^2 - xmn + 1) = (2xm^2 - xmn + 1)(xmn - xn^2 + 1) + 1$,

$$[b_2, \dots, b_k, x+1, c_l, \dots, c_1] = \frac{2xm^2 - xmn + 1}{2xmn - xn^2 + 2}.$$

Thus,

$$[b_1, \dots, b_k, x+1, c_l, \dots, c_1] = 2 - \frac{2xmn - xn^2 + 2}{2xm^2 - xmn + 1} = \frac{x(2m-n)^2}{x(2m-n)m + 1},$$
$$[b_1, \dots, b_k] = 2 - \frac{n}{m} = \frac{2m-n}{m}.$$

Setting $p = 2m - n$ and $q = m$ yields the result.

Now suppose $c_1 = 2$. Then $(b_1 - 1, \dots, b_k)$ and $(c_2, \dots, c_l)$ are linear-dual and

$$[b_1 - 1, \dots, b_k, x+1, c_l, \dots, c_2] = \frac{xm^2}{xmn+1},$$
$$[c_2, \dots, c_l, x+1, b_k, \dots, b_1 - 1] = \frac{xm^2}{xm^2 - xmn + 1},$$

where $[b_1 - 1, \dots, b_k] = m/n$. Thus,

$$[c_1, \dots, c_l, x+1, b_k, \dots, b_1 - 1] = 2 - \frac{xm^2 - xmn + 1}{xm^2} = \frac{xm^2 + xmn - 1}{xm^2}.$$

Since $(xmn + xn^2 + 1)xm^2 = (xm^2 + xmn - 1)(xmn + 1) + 1$,

$$[b_1 - 1, \dots, b_k, x+1, c_l, \dots, c_2, c_1] = \frac{xm^2 + xmn - 1}{xmn + xn^2 + 1}.$$

Thus,

$$[b_1, \dots, b_k, x+1, c_l, \dots, c_2, c_1] = 1 + \frac{xm^2 + xmn - 1}{xmn + xn^2 + 1} = \frac{x(m+n)^2}{x(m+n)n + 1},$$
$$[b_1, \dots, b_k] = 1 + \frac{m}{n} = \frac{m+n}{n}.$$

Setting $p = m + n$ and $q = n$ yields the result. □

Proposition A.3 *Let* $[b_1, \dots, b_k] = p/q$ *and let* $\boldsymbol{a} = (a_1, \dots, a_n) \in \mathcal{S}_{1a}$. *Then* $|\mathrm{Tor}(H_1(\boldsymbol{T}_{-A(\boldsymbol{a})}))| = p^2$.

Proof Let $\boldsymbol{a} = (2, b_1, \dots, b_k, 2, c_l, \dots, c_1)$, where $(b_1, \dots, b_k)$ and $(c_1, \dots, c_l)$ are linear-dual (up to cyclic reordering). By Lemma A.2, $[b_1, \dots, b_k, 2, c_l, \dots, c_1] = p^2/(pq+1)$ and so

$$[2, b_1, \dots, b_k, 2, c_l, \dots, c_1] = 2 - \frac{pq+1}{p^2} = \frac{2p^2 - pq - 1}{p^2}.$$

By Lemma A.1, $|\mathrm{Tor}(H_1(\boldsymbol{T}_{-A(\boldsymbol{a})}))| = |2p^2 - pq - 1 - (\alpha - 2)|$, where α is the denominator of $[2, b_1, \dots, b_k, 2, c_l, \dots, c_2]$. By Lemma A.2,

$$[c_1, \dots, c_l, 2, b_k, \dots, b_1] = \frac{p^2}{p^2 - pq + 1}$$

and so

$$[c_2, \dots, c_l, 2, b_k, \dots, b_1] = \frac{p^2 - pq + 1}{(c_1 - 1)p^2 - c_1 pq + c_1}.$$

Thus,

$$[b_1, \dots, b_k, 2, c_l, \dots, c_2] = \frac{p^2 - pq + 1}{s} \quad \text{for some } s.$$

Now it is clear that $\alpha = p^2 - pq + 1$ and so

$$|\mathrm{Tor}(H_1(\boldsymbol{T}_{-A(\boldsymbol{a})}))| = |2p^2 - pq - 1 - (p^2 - pq + 1 - 2)| = p^2. \qquad \square$$

Lemma A.4 *Let*

$$A=(a_{ij})=\begin{bmatrix} -a_1 & 1 & (-1)^t & & & & (-1)^r \\ 1 & -a_2 & & & & & \\ & & \ddots & 1 & & & \\ (-1)^t & & 1 & -a_k & 1 & & \\ & & & 1 & \ddots & & \\ & & & & & -a_{n-1} & 1 \\ (-1)^r & & & & & 1 & -a_n \end{bmatrix}.$$

Suppose $a_i \geq 2$ for all $1 \leq i \leq n$, $a_1 \geq 3$, $a_k \geq 3$, and $r,t \in \{0,1\}$.

(1) *If $r=1$ or $t=1$, then A is negative definite.*

(2) *If $r=t=0$ and either $a_1 \geq 4$, $a_k \geq 4$, or there exists an integer $i \notin \{1,k\}$ such that $a_i \geq 3$, then A is negative definite.*

Proof Let $s_i=\sum_{j=1}^n a_{ij}$ be the i^{th} row sum of A. Then $s_i \leq 0$ for all i. Moreover, since either $a_1 \geq 4$, $a_k \geq 4$, or there exists an integer $i \notin \{1,s\}$ such that $a_i \geq 3$, there exists a row sum that is strictly less than 0. Let $w \in \mathbb{Z}^n$. Then

$$\begin{aligned} w^TAw &= \sum_{i,j} a_{ij}w_iw_j = \frac{1}{2}\sum_{i,j} a_{ij}(w_i^2+w_j^2-(w_i-w_j)^2) \\ &= \sum_{i,j} a_{ij}w_i^2 - \sum_{i<j} a_{ij}(w_i-w_j)^2 = \sum_i s_iw_i^2 - \sum_{i<j} a_{ij}(w_i-w_j)^2. \end{aligned}$$

First suppose $r=t=0$. Then every term in the above expression is at most zero and so $w^TAw \leq 0$. Moreover, if either $a_1 \geq 4$, $a_k \geq 4$ or there exists an integer $i \notin \{1,k\}$ such that $a_i \geq 3$, then one of the row sums s_i is strictly less than 0. In this case, $w^TAw=0$ if and only if $w=0$. Thus A is negative definite. Next suppose $r=1$ and $t=0$. Then $s_1, s_n \leq -2$ and so

$$\begin{aligned} w^TAw &= s_1w_1^2+s_nw_n^2+(w_1-w_n)^2+\sum_{i\neq 1,n} s_iw_i^2 - \sum_{\substack{i<j \\ (i,j)\neq(1,n)}} (w_i-w_j)^2 \\ &\leq -2w_1^2-2w_n^2+(w_1-w_n)^2+\sum_{i\neq 1,n} s_iw_i^2 - \sum_{\substack{i<j \\ (i,j)\neq(1,n)}} (w_i-w_j)^2 \\ &= -(w_1+w_n)^2+\sum_{i\neq 1,n} s_iw_i^2 - \sum_{\substack{i<j \\ (i,j)\neq(1,n)}} (w_i-w_j)^2. \end{aligned}$$

Each term in this expression is clearly negative. If $w^TAw=0$, then, from the first term, $w_1=-w_n$. From the terms in the last summand, $w_1=w_2=\cdots=w_n$. Hence

$w_n = -w_n$, implying that $w_1 = \cdots = w_n = 0$. Therefore, A is negative definite. We obtain a similar result if $r = 0$ and $t = 1$. Finally assume $r = t = 1$. Then $s_1 \leq -4$ and $s_k, s_n \leq -2$. Arguing as above,

$$\begin{aligned} w^T A w &= s_1 w_1^2 + s_k w_k^2 + s_n w_n^2 + (w_1 - w_n)^2 + (w_1 - w_k)^2 \\ &\qquad + \sum_{i \neq 1,k,n} s_i w_i^2 - \sum_{\substack{i<j \\ (i,j)\neq(1,n),(1,k)}} (w_i - w_j)^2 \\ &\leq -(w_1 + w_n)^2 - (w_1 + w_k)^2 + \sum_{i \neq 1,n} s_i w_i^2 - \sum_{\substack{i<j \\ (i,j)\neq(1,n),(1,k)}} (w_i - w_j)^2. \end{aligned}$$

Once again, we can see that A is necessarily negative definite. □

Lemma A.5 *Let*

$$A = \begin{bmatrix} -a_1 & 2 & & & & (-1)^r \\ 2 & -a_2 & 1 & & & \\ & 1 & -a_3 & & & \\ & & & \ddots & & \\ & & & & -a_{n-1} & 1 \\ (-1)^r & & & & 1 & -a_n \end{bmatrix}.$$

Suppose $a_i \geq 2$ *for all* $1 \leq i \leq n$, $a_1 \geq 3$, $a_2 \geq 3$, *and* $r \in \{0, 1\}$.

(a) *If* $r = 1$, *then* A *is negative definite.*

(b) *If* $r = 0$ *and either* $a_1 \geq 4$, $a_2 \geq 4$ *or there exists an integer* $i \notin \{1, k\}$ *such that* $a_i \geq 3$, *then* A *is negative definite.*

Proof The proof is similar to the proof of Lemma A.4. □

References

[1] **P Aceto**, **M Golla**, **K Larson**, **A G Lecuona**, *Surgeries on torus knots, rational balls, and cabling*, preprint (2020) arXiv 2008.06760

[2] **S Akbulut**, **R Kirby**, *Branched covers of surfaces in 4–manifolds*, Math. Ann. 252 (1979/80) 111–131 MR Zbl

[3] **J A Baldwin**, *Heegaard Floer homology and genus one, one-boundary component open books*, J. Topol. 1 (2008) 963–992 MR Zbl

[4] **A J Casson**, **C M Gordon**, *Cobordism of classical knots*, from "À la recherche de la topologie perdue" (L Guillou, A Marin, editors), Progr. Math. 62, Birkhäuser, Boston, MA (1986) 181–199 MR Zbl

[5] **T D Cochran**, **B D Franklin**, **M Hedden**, **P D Horn**, *Knot concordance and homology cobordism*, Proc. Amer. Math. Soc. 141 (2013) 2193–2208 MR Zbl

[6] **S K Donaldson**, *The orientation of Yang–Mills moduli spaces and* 4*–manifold topology*, J. Differential Geom. 26 (1987) 397–428 MR Zbl

[7] **D L Goldsmith**, *A linking invariant of classical link concordance*, from "Knot theory" (J-C Hausmann, editor), Lecture Notes in Math. 685, Springer (1978) 135–170 MR Zbl

[8] **P Lisca**, *Lens spaces, rational balls and the ribbon conjecture*, Geom. Topol. 11 (2007) 429–472 MR Zbl

[9] **P Lisca**, *Sums of lens spaces bounding rational balls*, Algebr. Geom. Topol. 7 (2007) 2141–2164 MR Zbl

[10] **P Lisca**, *On* 3*–braid knots of finite concordance order*, Trans. Amer. Math. Soc. 369 (2017) 5087–5112 MR Zbl

[11] **W D Neumann**, *A calculus for plumbing applied to the topology of complex surface singularities and degenerating complex curves*, Trans. Amer. Math. Soc. 268 (1981) 299–344 MR Zbl

[12] **M Sakuma**, *Surface bundles over S^1 which are* 2*–fold branched cyclic coverings of* S^3, Math. Sem. Notes Kobe Univ. 9 (1981) 159–180 MR Zbl

[13] **J Simone**, *Using rational homology circles to construct rational homology balls*, Topology Appl. 291 (2021) art. id. 107626 MR Zbl

School of Mathematics, Georgia Institute of Technology
Atlanta, GA, United States

jsimone7@gatech.edu

Received: 10 July 2020 Revised: 1 March 2022

Algebraic & Geometric Topology **23**:6 (2023) 2519–2543
DOI: 10.2140/agt.2023.23.2519
Published: 7 September 2023

A mnemonic for the Lipshitz–Ozsváth–Thurston correspondence

Artem Kotelskiy
Liam Watson
Claudius Zibrowius

When $\Bbbk$ is a field, type D structures over the algebra $\Bbbk[u,v]/(uv)$ are equivalent to immersed curves decorated with local systems in the twice-punctured disk. Consequently, knot Floer homology, as a type D structure over $\Bbbk[u,v]/(uv)$, can be viewed as a set of immersed curves. With this observation as a starting point, given a knot K in S^3, we realize the immersed curve invariant $\widehat{HF}(S^3 \smallsetminus \mathring{\nu}(K))$ of Hanselman, Rasmussen and Watson by converting the twice-punctured disk to a once-punctured torus via a handle attachment. This recovers a result of Lipshitz, Ozsváth and Thurston calculating the bordered invariant of $S^3 \smallsetminus \mathring{\nu}(K)$ in terms of the knot Floer homology of K.

57K18, 57K31; 57R58

Recent work interprets relative versions of homological invariants in terms of immersed curves, including Heegaard Floer homology for manifolds with torus boundary (see Hanselman, Rasmussen and Watson [4]) as well as link Floer homology (see Zibrowius [23]), singular instanton knot homology (see Hedden, Herald and Kirk [7]), and Khovanov homology (see Kotelskiy, Watson and Zibrowius [12]) for 4–ended tangles. In particular, Section 5 of [12] classifies type D structures over a quiver algebra associated with a surface with boundary in terms of immersed curves on this surface; compare Haiden, Katzarkov and Kontsevich [2] and Hanselman, Rasmussen and Watson [4]. Denoting a field by $\Bbbk$, perhaps the simplest algebra to illustrate these classification results is $\mathcal{R} = \Bbbk[u,v]/(uv)$. This algebra arises as the path algebra of a quiver that is associated with the decorated surface shown in Figure 1. Work of Lekili and Polishchuk [13; 14] describes the role of $\mathcal{R}$, and its relationship with the twice-punctured disk, in the context of homological mirror symmetry; see in particular [14, Figures 1 and 2]. The algebra $\mathcal{R}$ equipped with the Alexander and δ gradings $\mathrm{gr}(u) = (-1,1)$ and $\mathrm{gr}(v) = (1,1)$ plays a central role in knot Floer homology; see Dai, Hom, Stoffregen and Truong [1], for instance.

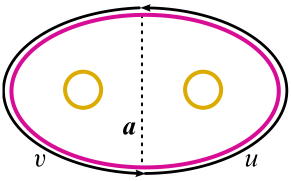


Figure 1: An arc system associated with the algebra $\mathcal{R}$.

Theorem 1 *Every bigraded type D structure over $\mathcal{R}$ is equivalent to an immersed curve* (*decorated with local systems*) *in the twice-punctured disk*, *which is unique up to regular homotopy* (*and equivalence of local systems*).

As stated, this is a special case of a theorem proved in [12, Section 5] appealing to techniques from [4] (see also [23]). The observation could alternatively be extracted from [4, Section 3.4] (see the aside starting on page 2527 below accompanying Figure 8), and also follows from work of Haiden, Katzarkov and Kontsevich [2]; see Section 1.8 of [12] for more discussion. We will review the algebraic objects in Section 1 and, without reproducing the proof in full, explain some key steps in this special case in Section 2. Theorem 1 gives rise to a graphical interpretation $\boldsymbol{\gamma}$ for (a variant of) knot Floer homology ${}^{\mathcal{R}}CFK(Y, K)$, which is a bigraded type D structure over $\mathcal{R}$. Our proof is constructive and, in particular, foregrounds the role of vertically and horizontally simplified bases that arise in knot Floer homology. An explicit example of a curve $\boldsymbol{\gamma}$ in the twice-punctured disk is shown in Figure 2, left. This particular curve corresponds to the type D structure associated with the right-hand trefoil $T_{2,3}$ in S^3:

$$[\diamond_1 \xleftarrow{u} \diamond_2 \xrightarrow{v} \diamond_3] = {}^{\mathcal{R}}CFK(S^3, T_{2,3}).$$

Note that, while the local system in this example is trivial, these are easy to add to the picture in general, being equivalent to isomorphism classes of flat vector bundles over the curves in question. There is an obvious handle attachment, identifying the two punctures in the disk, which yields a once-punctured torus. Denote this handle attachment by ∩ and consider the curve ∩$(\boldsymbol{\gamma})$. Note that, given a choice of meridian on the torus, this operation has an inverse, which we will denote by ◖◗.

Denote by $\widehat{HF}(M)$ the immersed curve in the once-punctured torus associated with a manifold M with torus boundary [4]. This is equivalent to the bordered Heegaard Floer invariant of M; see Lipshitz, Ozsváth and Thurston [17]. Here is the mnemonic we propose:

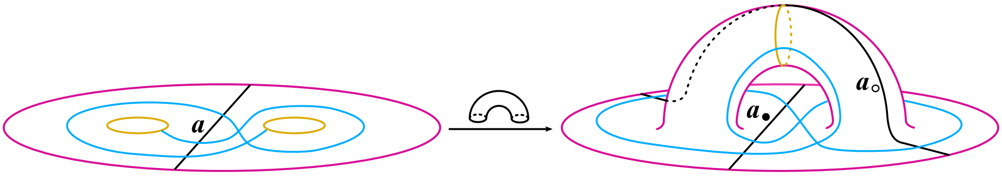


Figure 2: Adding a handle to the twice-punctured disk results in the once-punctured torus. This carries immersed curves to immersed curves; the immersed curve on the left corresponds to the type D structure ${}^{\mathcal{R}}CFK(S^3, T_{2,3}) = [\diamond_1 \xleftarrow{u} \diamond_2 \xrightarrow{v} \diamond_3]$, which is carried to the curve $\widehat{HF}(M)$, where M is the trefoil exterior $S^3 \smallsetminus \mathring{\nu}(K)$.

Theorem 2 *If $\boldsymbol{\gamma}$ is a curve representing the knot Floer invariant ${}^{\mathcal{R}}CFK(S^3, K)$ over the two-element field, then ⌒$(\boldsymbol{\gamma})$ is equivalent to $\widehat{HF}(M)$, where M is the exterior of the knot K. Conversely, given a meridian for $M = S^3 \smallsetminus \mathring{\nu}(K)$, the curve ◜◝$(\widehat{HF}(M))$ represents the knot Floer type D structure for K.*

Figure 2 illustrates this theorem for the right-hand trefoil knot; the proof is given in Section 4.

Remark There is an apparent ambiguity in the statement of Theorem 2, namely the number of twists (along the belt of the handle ⌒) one adds to the noncompact component of the curve $\boldsymbol{\gamma}$. However, recall that the curve $\widehat{HF}(M) \subset \partial M$ is null-homologous in M [4, Sections 5 and 6]; to resolve the ambiguity it is enough to identify the once-punctured torus obtained after adding the handle with the boundary of the knot exterior (minus a small disk). We identify the arc $\boldsymbol{a}_\bullet$ from Figure 2 with the meridian μ, and the second arc $\boldsymbol{a}_\circ$ with a longitude λ of K. This pair provides a bordered structure, in the sense of Lipshitz, Ozsváth and Thurston [17]. Concerning the framing λ: On one hand, there is a preferred choice given by the Seifert longitude λ_0, and the corresponding identification is depicted in Figure 3, right. On the other hand, it is often easiest to work with the "blackboard framing", which simply joins the endpoints of $\boldsymbol{\gamma}$ without new twisting as they run over the handle, as in Figure 2. In general, the latter gives the $2\tau(K)$–framed longitude $\lambda_{2\tau} = 2\tau \cdot \mu + \lambda_0$, where the value $\tau(K)$ is the Ozsváth–Szabó concordance invariant (we describe how to extract this value below). This choice of longitude is illustrated in Figure 3, left. These choices differ by Dehn twists along μ; note that in both cases $[$⌒$(\boldsymbol{\gamma})] = [\lambda_0]$ in homology. Different choices of twisting precisely correspond to different unstable chains appearing in [17, Theorem A.11], due to Lipshitz, Ozsváth and Thurston, which Theorem 2 recasts.

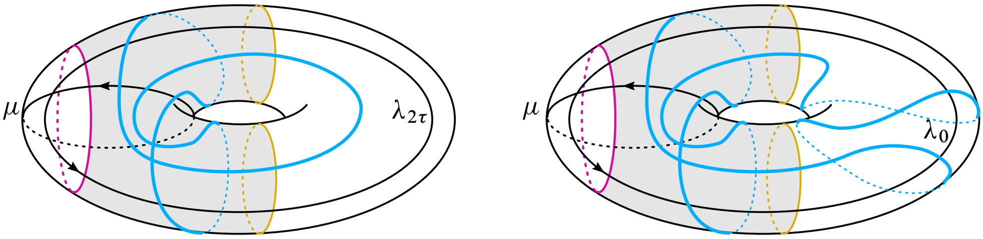


Figure 3: Choices of framing on the right-hand trefoil invariant: $\lambda_{2\tau} = 2\mu + \lambda_0$ (left) and the Seifert longitude λ_0 (right). The resulting curve γ on the boundary of the trefoil exterior coincides with Figure 9 of Hanselman, Rasmussen and Watson [5].

This result generalizes to knots in arbitrary three-manifolds; see Section 5 for further discussion.

A graphical interpretation of the family of concordance homomorphism $\{\phi_i\}$ due to Dai, Hom, Stoffregen and Truong [1] is given by Hanselman and the second author [6]. This can be read off the current picture: Denote by $\boldsymbol{\gamma}_0(K) \subset \boldsymbol{\gamma}(K)$ the noncompact curve in the twice-punctured disk associated with ${}^{\mathcal{R}}CFK(S^3, K)$. (The curve $\boldsymbol{\gamma}_0(K)$ is a concordance invariant [6, Proposition 2].) Orient $\boldsymbol{\gamma}_0(K)$ so that it leaves from the v–puncture; this is the left-hand puncture in Figure 1, which records the v^i coefficient maps. Contracting the arc $\boldsymbol{a}$ to a point gives a wedge of annuli $A_v \vee A_u$, and the oriented segments of $\boldsymbol{\gamma}_0(K)$ around the v–puncture give a collection of homotopy classes in $\pi_1 A_v \cong \langle t \rangle$, where the generator t winds counterclockwise. As a result, given $\boldsymbol{\gamma}_0(K)$, with our choice of orientation we obtain $t^{n_1} t^{n_2} \cdots t^{n_k}$ for the k oriented segments winding around the v–puncture, and

$$\phi_i(K) = \sum_{n_j = \pm i} \operatorname{sign}(n_j), \qquad \tau(K) = \sum_{j=1}^{k} n_j,$$

so that $\tau(K)$ is simply the winding number of $\boldsymbol{\gamma}$ around the v–puncture. One can check that this gives $\tau(T_{2,3}) = \phi_1(T_{2,3}) = 1$. A more complicated example is shown in Figure 4. The same construction works with the u–puncture instead of the v–puncture, due to a symmetry interchanging u and v in knot Floer homology; see Ozsváth and Szabó [18].

Relevant to concordance is the behaviour under connect sum. Denote by ${}_{\mathcal{R}}HFK(S^3, K)$ the knot Floer invariant obtained as the homology of a complex $CFK(S^3, K)$ freely generated over $\mathcal{R}$. In Section 6 we prove:

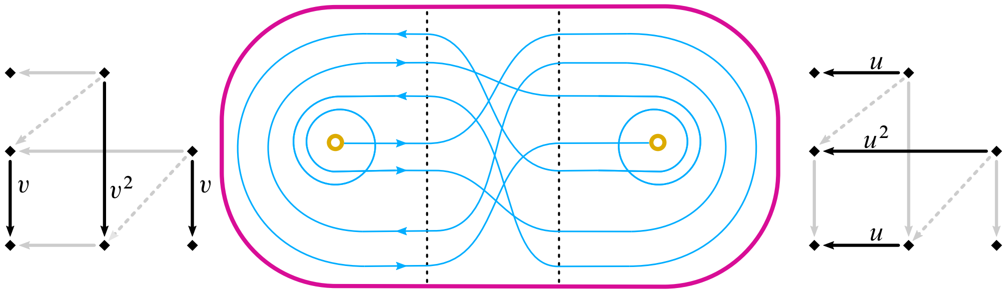


Figure 4: The curve associated with ${}^{\mathcal{R}}CFK(S^3, K)$ when K is the $(2,1)$–cable of the right-hand trefoil. The vertical and horizontal complexes are shown beside the relevant annuli; including the diagonal arrows describes the invariant over $\Bbbk[u,v]$. Applying Theorem 2 results in the curve invariant in the torus, which can be compared with [6, Figure 1]. Orientating the curve as shown, we calculate $\phi_1(K) = 0$, $\phi_2(K) = 1$ and $\tau(K) = 2$.

Theorem 3 *The knot Floer homology over $\mathcal{R}$ of a connected sum of two knots is equal to the wrapped Lagrangian Floer homology of the corresponding curves*:

$$_{\mathcal{R}}HFK(S^3, \mathrm{m}K \# K') \cong HF(\boldsymbol{\gamma}(K), \boldsymbol{\gamma}(K')).$$

A proof is given in Section 6. As is the case with Theorem 1, the proof appeals to the techniques in [12, Section 5].

1 Algebraic objects

Let $\mathcal{B}$ be a bigraded unital algebra over a field $\Bbbk$, with a subring of idempotents $\mathcal{I}$ being equal to $\Bbbk^n$. The object of interest is a bigraded chain complex over $\mathcal{B}$: Let V be a finite-dimensional bigraded left $\mathcal{I}$–module, and suppose further that we have a morphism of $\mathcal{I}$–modules

$$d\colon V \to \mathcal{B} \otimes_{\mathcal{I}} V$$

satisfying the compatibility condition

$$(\mu \otimes \mathrm{id}_V) \circ (\mathrm{id}_{\mathcal{B}} \otimes d) \circ d = 0,$$

where μ denotes multiplication in $\mathcal{B}$. In our setting the morphism d has bidegree $(a, \delta) = (0, 1)$, and the pair (V, d) is a bigraded type D structure over $\mathcal{B}$.

A couple of remarks: We work with left actions for consistency with [17], and our type D structures will always be reduced, which means that $d(x) = \sum_i b_i \otimes y_i$, where none of the $b_i \in \mathcal{B}$ are invertible. This is justified by the fact that any bigraded type D structure is homotopy equivalent to a reduced one [12, Lemma 2.16].

Such algebraic structures appear naturally in a variety of settings. For example, given a knot K in S^3, the knot Floer invariant $HFK(S^3, K)$, due to Ozsváth and Szabó [18] and to Rasmussen [21], can be viewed as a $\Bbbk[u,v]$–module obtained as the homology of a chain complex $CFK(S^3, K)$ over the ring $\Bbbk[u,v]$ [22, Section 3]. This complex is freely generated as a module over this ring. As such, it is natural to view $CFK(S^3, K)$ as a type D structure over $\Bbbk[u,v]$, which we denote by ${}^{\Bbbk[u,v]}CFK(S^3, K)$.

Given a type D structure over $\mathcal{B}$, a homomorphism of $\mathcal{I}$–algebras $\mathcal{B} \to \mathcal{B}'$ gives rise to an induced type D structure over $\mathcal{B}'$. In particular, the quotient $\Bbbk[u,v] \to \Bbbk[u,v]/(uv)$ defines a truncated version of the knot Floer type D structure,

$$ {}^{\mathcal{R}}CFK(S^3, K) = {}^{\Bbbk[u,v]}CFK(S^3, K)|_{uv=0}. $$

The associated module object ${}_{\mathcal{R}}CFK(S^3, K)$ (see [17, Lemma 2.20]) is the knot Floer complex freely generated over $\mathcal{R}$, which is studied in depth by Dai, Hom, Stoffregen and Truong [1] and Ozsváth and Szabó [19]. A concise formula connecting the type D structure and the associated module object uses the box tensor product (see [16, Section 2.3.2 and Proposition 2.3.18], and also the beginning of Section 4 for a similar construction):

$$ {}_{\mathcal{R}}CFK(S^3, K) = {}_{\mathcal{R}}\mathcal{R}_{\mathcal{R}} \boxtimes {}^{\mathcal{R}}CFK(S^3, K). $$

We note that there are two further type D structures obtained from ${}^{\mathcal{R}}CFK(S^3, K)$ by setting the appropriate variables equal to zero: the horizontal type D structure $C^{\boldsymbol{h}}$ and the vertical type D structure $C^{\boldsymbol{v}}$. For instance, in the case of the type D structure ${}^{\mathcal{R}}CFK(S^3, T_{2,3})$ (see Figure 2), we have

$$ C^{\boldsymbol{h}} = [\diamond_1 \xleftarrow{u} \diamond_2 \quad \diamond_3], \qquad C^{\boldsymbol{v}} = [\diamond_1 \quad \diamond_2 \xrightarrow{v} \diamond_3]. $$

As the type D structures are reduced, the isomorphisms of vector spaces $C^{\boldsymbol{h}}|_{u=0} \cong \widehat{HFK}(S^3, K) \cong C^{\boldsymbol{v}}|_{v=0}$ induce an isomorphism

$$ \varphi \colon C^{\boldsymbol{h}}|_{u=0} \to C^{\boldsymbol{v}}|_{v=0}. $$

We have:

Proposition 4 *The data specified by the triple $(C^{\boldsymbol{h}}, C^{\boldsymbol{v}}, \varphi)$ is equivalent to the type D structure ${}^{\mathcal{R}}CFK(S^3, K)$.*

Proof This is immediate from the definitions, but also follows from the discussion in Section 2 outlining the proof of Theorem 1. □

2 Geometric objects

Often, when an invariant of a topological object is a type D structure over an algebra $\mathcal{B}$, the invariant is only well defined up to homotopy equivalence. As such, it is of general interest to be able to classify homotopy equivalence classes of type D structures. Such classification turns out to be possible when the algebra $\mathcal{B}$ is isomorphic to an endomorphism algebra of certain objects in the (wrapped) Fukaya category of a surface Σ. In this case, homotopy equivalence classes of type D structures over $\mathcal{B}$ correspond to certain curves (decorated with local systems) immersed in Σ. This is a powerful structural result allowing us to translate algebra into geometry, something not so often encountered in mathematics. The classification result is established in [2] using representations of nets; an alternative, more geometric approach is given in [4], which appeals to train tracks in a surface. The simplification algorithm proved in [4] that is central to the classification is further developed and leveraged in [12; 23], where train tracks reappear as precurves. We focus on this latter approach.

To provide a useful toy model for the classification result, we restrict to type D structures over $\mathcal{R}$. The algebra $\mathcal{R}$ indeed arises as the endomorphism algebra of an object in the (wrapped) Fukaya category of a surface. The surface is the oriented, twice-punctured disk D and the object is an arc connecting the two punctures; see Figure 5. More explicitly, from this figure we can extract a quiver with a single vertex corresponding to the object in the Fukaya category, and arrows labelled u and v corresponding to the two paths around the punctures in D:

$$\begin{tikzcd} \diamond \arrow[loop left, "v"] \arrow[loop right, "u"] \end{tikzcd}$$

It is useful to view this quiver as a deformation retract of the twice-punctured disk. The algebra $\mathcal{R}$ is the path algebra of this quiver modulo the relations $uv = 0 = vu$. In terms of Figure 5, these relations have the effect that paths that run along the dashed arc are zero in $\mathcal{R}$, while paths that only wind around a single puncture are nonzero.

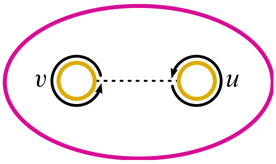


Figure 5: A dual arc system associated with the algebra $\mathcal{R}$.

To match the setup in [12], a different viewpoint, which is in some sense dual to the previous one, will be more useful. Namely, choose an arc $\boldsymbol{a}$ that is properly embedded in $(D, \partial D)$ and that divides D into a pair of annuli, as illustrated in Figure 1. From this, we can also recover the quiver: the vertex corresponds to the arc $\boldsymbol{a}$ and the arrows correspond to paths on the boundary of D. Again, it is useful to consider the quiver as a deformation retract that contracts the arc to the quiver vertex. The relations that we impose on the quiver algebra to obtain $\mathcal{R}$ now have a different geometric interpretation: paths that at an endpoint of the dashed arc continue along the boundary of D are zero in $\mathcal{R}$, while paths that, at such a point, always choose to follow the dashed arc are nonzero; see also [12, Section 5.1].

The choice of arc $\boldsymbol{a}$ is an example of an arc system on D, in the sense of Section 5.1 of [12]. In general, an arc system, giving rise to an algebra $\mathcal{B}$, allows for a graphical representation of type D structures over $\mathcal{B}$ as subobjects of the surface. These show up as train tracks in [4] and precurves in [12]; we describe them explicitly in the case of $\mathcal{R}$ and the twice-punctured disk D. It will be convenient to specify the annuli $D \smallsetminus \boldsymbol{a} = A_v \sqcup A_u$; these annuli are called faces.

Let (V, d) be a type D structure over $\mathcal{R}$. Given a homogeneous basis $\{x_1, \dots, x_n\}$ for V (as a vector space over $\Bbbk$, say), we can pick n distinct points on $\boldsymbol{a}$ and label these with the x_i. To describe the morphism d, suppose $b \otimes x_j$ is a summand of $d(x_i)$. Then, since b is a sum of polynomials, we may assume without loss of generality that b is λu^k or λv^k for some $\lambda \in \Bbbk$ and $k > 0$. (The assumption that this power is nonzero comes from our restriction to reduced type D structures.) There are two cases: if $b = \lambda u^k$ then we connect x_i to x_j by an oriented arc immersed in A_u that winds algebraically k times in the positive direction; and if $b = \lambda v^k$ then we connect x_i to x_j by an oriented curve immersed in A_v that winds algebraically k times in the positive direction. In both cases the arc is decorated by the field coefficient λ, noting that when $\lambda = 1$ our convention is to drop the label. In particular, when $\Bbbk$ is the two-element field, only the arcs are needed. Lastly, if an intersection point x_i does not have outgoing arcs in the annulus A_u, we connect x_i straight to the u–puncture; we do the same for the A_v annulus and the v–puncture. To see that this information, having added all of the arcs described, can be viewed as an immersed train track in D, we simply require that every curve is perpendicular to $\boldsymbol{a}$ in a neighbourhood of each x_i. An explicit example is given in Figure 6. Note that in this example there are no arcs going to interior punctures.

These train tracks can be put into a simple form that makes them easier to manage: we require that they are simply faced in the sense of [12, Definition 5.9]. In the present

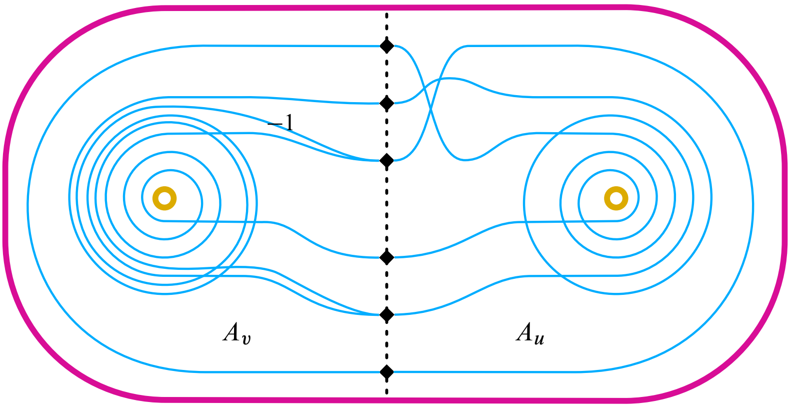


Figure 6: A sample train track representation of a type D structure over $\mathcal{R}$. Note that every curve segment is oriented so that it runs counterclockwise around a puncture, so this orientation is omitted. Similarly, unlabelled edges (of which there are all but one in this example) carry the decoration $\lambda = 1$.

setting, this amounts to expressing

$$D = A_v \cup_{\boldsymbol{a}\times\{1\}} (\boldsymbol{a} \times [-1,1]) \cup_{\boldsymbol{a}\times\{-1\}} A_u$$

and requiring that the train track restricted to A_u and to A_v describes a type D structure over $\Bbbk[u]$ and $\Bbbk[v]$, respectively, with the property that each x_i connects to at most one x_j. For an illustration, see Figure 7. All of the interesting switching is confined to the strip $\boldsymbol{a} \times [-1,1]$, which amounts to a graphical interpretation (reading from right to left) of an isomorphism $\varphi\colon V_u \to V_v$, where V_v and V_u are the underlying vector spaces associated with the type D structure in each face. As such, the general fact that we can restrict to simply faced train tracks (see [12, Proposition 5.10]) boils down to the fact that type D structures over $\mathcal{R}$ admit vertically and horizontally simplified bases [17, Definition 11.23] — though not necessarily one that is simultaneously vertically and horizontally simplified, whence the choice of isomorphism. This last assertion explains the presence of φ; compare Proposition 4. We remark that this is one step in which the grading plays a key role.

Aside We make a digression to describe that, in order to classify type D structures in terms of immersed curves, other choices of surface decomposition are possible. Namely, another option would be to

(1) cut the annuli A_u and A_v further, as described in Figure 8;

(2) associate with this new geometric picture a different algebra $\mathcal{E}$;

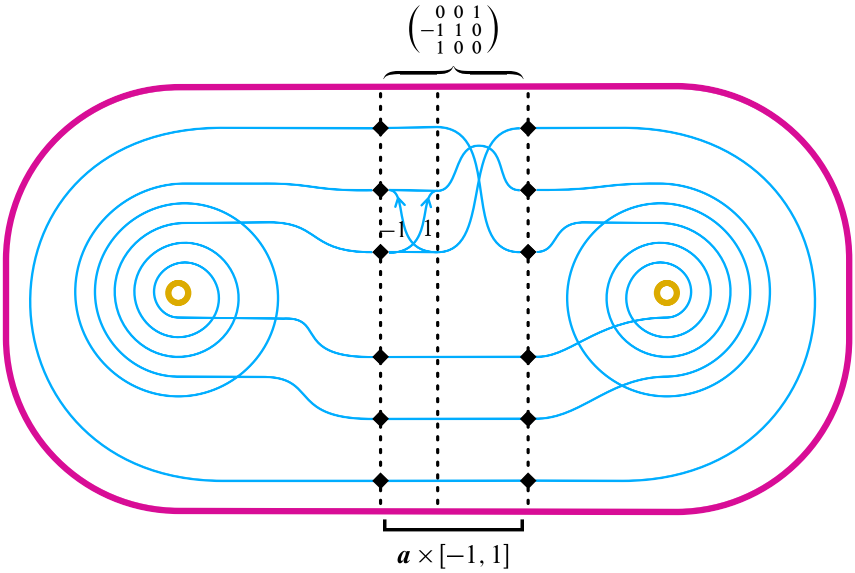

$\begin{pmatrix} 0 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$
-1
1
$\boldsymbol{a} \times [-1,1]$

Figure 7: Expressing the train track from Figure 6 as a simply faced precurve. The isomorphism described can be read off the tracks in $\boldsymbol{a} \times [-1,1]$ from right to left; in the present setting the resulting matrix block-decomposes into two 3×3 parts, of which one is shown and the other is the identity matrix.

(3) interpret the type D structure ${}^{\mathcal{R}}V$ as a type D structure ${}^{\mathcal{E}}W$ over the algebra $\mathcal{E}$;

(4) apply the methods from [4] to interpret ${}^{\mathcal{E}}W$ as an immersed curve.

To describe this in more detail, let us focus first on the annulus A_v in step (2).

Consider Figure 8. Any type D structure ${}^{\Bbbk[v]}V_\diamond$ may be regarded as a type D structure $V_\diamond \oplus V_\bullet \oplus V_\circ$ over the quiver algebra $\Bbbk[\bullet \xrightarrow{a} \diamond \xrightarrow{b} \circ]$ together with an isomorphism between the vector spaces $V_\bullet$ and $V_\circ$. To repackage the latter into a type D structure without extra data, we consider a subalgebra generated by idempotents $\iota_\bullet + \iota_\circ$ and $\iota_\diamond$

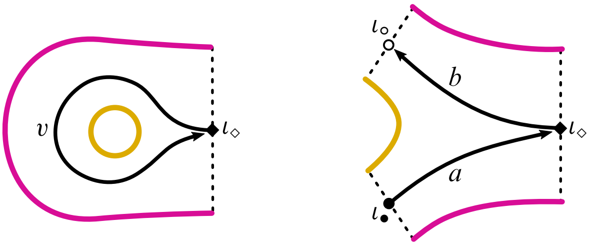


Figure 8: A quiver associated with the annulus describing the algebra $\Bbbk[v]$, and a quiver for an algebra associated with an additional cut.

(because eventually the idempotents $\iota_\circ$ and $\iota_\bullet$ are identified). Writing $\iota_{\text{⌖}} = \iota_\bullet + \iota_\circ$, the subalgebra is equal to

$$\mathcal{C} = \Bbbk[\text{⌖} \underset{b}{\overset{a}{\rightleftarrows}} \diamond]/(ba).$$

The type D structure ${}^{\Bbbk[v]}V_\diamond$ can now be interpreted as a type D structure ${}^{\mathcal{C}}(V_\diamond \oplus V_{\text{⌖}})$: generators $\diamond$ in ${}^{\Bbbk[v]}V_\diamond$ and ${}^{\mathcal{C}}(V_\diamond \oplus V_{\text{⌖}})$ are in one-to-one correspondence, while a differential $\diamond \xrightarrow{v^n} \diamond$ in ${}^{\Bbbk[v]}V_\diamond$ corresponds to the sequence of differentials

$$\diamond \xrightarrow{b} \underbrace{\text{⌖} \xrightarrow{ab} \text{⌖} \xrightarrow{ab} \cdots \xrightarrow{ab} \text{⌖}}_{n \text{ generators}} \xrightarrow{a} \diamond$$

in ${}^{\mathcal{C}}(V_\diamond \oplus V_{\text{⌖}})$. To add the second annulus A_u to the picture, given a type D structure ${}^{\mathcal{R}}V_\diamond$ one translates it into a type D structure ${}^{\mathcal{E}}W$ over the algebra

$$\mathcal{E} = \Bbbk[\text{⌖}_1 \underset{b_1}{\overset{a_1}{\rightleftarrows}} \diamond \underset{a_2}{\overset{b_2}{\rightleftarrows}} \text{⌖}_2]/(b_1a_1, b_2a_2, a_1b_2, a_2b_2)$$

via the dictionary

$$\diamond \xrightarrow{v^n} \diamond \longmapsto \diamond \xrightarrow{b_1} \underbrace{\text{⌖}_1 \xrightarrow{a_1b_1} \text{⌖}_1 \xrightarrow{a_1b_1} \cdots \xrightarrow{a_1b_1} \text{⌖}_1}_{n \text{ generators}} \xrightarrow{a_1} \diamond, \tag{1}$$

$$\diamond \xrightarrow{u^n} \diamond \longmapsto \diamond \xrightarrow{b_2} \underbrace{\text{⌖}_2 \xrightarrow{a_2b_2} \text{⌖}_2 \xrightarrow{a_2b_2} \cdots \xrightarrow{a_2b_2} \text{⌖}_2}_{n \text{ generators}} \xrightarrow{a_2} \diamond. \tag{2}$$

With this type D structure ${}^{\mathcal{E}}W$ in hand, the methods from [4] allow us to interpret ${}^{\mathcal{E}}W$ as an immersed curve.

A possible difficulty might arise from the following. The passage from ${}^{\mathcal{R}}V_\diamond$ to ${}^{\mathcal{E}}W$ does not respect homotopy equivalences: there exist homotopy equivalent type D structures ${}^{\mathcal{R}}V_\diamond \simeq {}^{\mathcal{R}}V'_\diamond$ such that the corresponding type D structures ${}^{\mathcal{E}}W$ and ${}^{\mathcal{E}}W'$ are not homotopy equivalent (take for example ${}^{\mathcal{R}}V_\diamond = [\diamond \xleftarrow{v} \diamond \xrightarrow{v} \diamond]$ and ${}^{\mathcal{R}}V'_\diamond = [\diamond \xleftarrow{v} \diamond] \oplus [\diamond]$). This problem is mitigated by the fact that the curves associated with ${}^{\mathcal{E}}W$ and ${}^{\mathcal{E}}W'$ will differ only by how many times their ends wrap around the two punctures, and initially we regard such curves as the same. Another way to mitigate this problem is to find vertically and horizontally simplified bases $\{\xi_i\}$ and $\{\eta_j\}$ for ${}^{\mathcal{R}}V_\diamond$ at the outset, and apply the operation (1) to the basis $\{\xi_i\}$ and the operation (2) to the basis $\{\eta_i\}$. This will ensure that the curve associated with ${}^{\mathcal{E}}W$ will not have extra wrapping around the punctures (and, of course, there may be nontrivial train tracks in the middle as in Figure 7).

We now return to the main text and make some comments about our conventions, reviewing [12, Section 5.6]. The object appearing in the strip $\boldsymbol{a} \times [-1, 1]$ represents an invertible matrix, where the i^{th} column records the edges leaving the point labelled x_i on $\boldsymbol{a} \times \{-1\}$ ($\boldsymbol{a}$ is oriented from top to bottom in our figures, so that $\{-1\}$ is the right-most

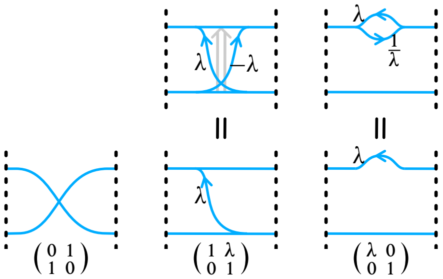

λ
−λ
λ
1/λ
=
=
λ
λ
$\begin{pmatrix}0&1\\1&0\end{pmatrix}$
$\begin{pmatrix}1&\lambda\\0&1\end{pmatrix}$
$\begin{pmatrix}\lambda&0\\0&1\end{pmatrix}$

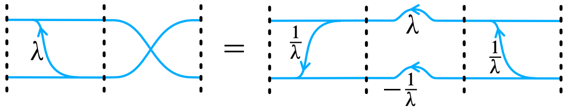
Figure 9: A crossing, a crossover switch, and a passing loop, each with the elementary matrix they represent by reading paths right to left. To declutter pictures, we will be using the pictures in the lower row, where the arrows pointing left to right are dropped.

edge of the strip). Using the row-reduction algorithm, this matrix can be factorized into elementary matrices corresponding to three geometric subobjects, as shown in Figure 9. These subobjects differ from the ones in [4], where the coefficients are restricted to the two-element field. New in the context of general fields are the nonzero coefficients $\lambda \in \Bbbk$, recorded on the crossover switches (these correspond to crossover arrows from [4]), as well as the passing loops, which introduce coefficients at various points. The main point is that, when two coefficients appear consecutively on one edge connecting the source and the target, the coefficients multiply, while if two edges share a common source and a common target, the coefficients on those edges add. We note that the geometric objects contain not only the information encoding φ (reading right to left) but also the information about the inverse φ^{-1} (reading left to right). As such, some of the data in the crossover switches and in the passing loops is superfluous. In particular, to simplify pictures here, we will record only the arrows running right to left.

It is convenient to put the matrix representing φ into a normal form, namely the LPU normal form: any invertible matrix can be written as a product of a lower triangular matrix, a permutation matrix (which may be multiplied, additionally, by a diagonal matrix to change coefficients), and an upper triangular matrix. For example, the matrix $\left(\begin{smallmatrix}\lambda&1\\1&0\end{smallmatrix}\right)$ may be expressed as

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

and this identity has the geometric interpretation



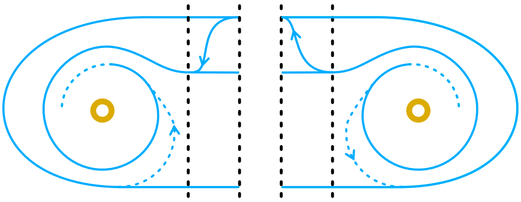


Figure 10: Arrows running counterclockwise can be removed.

More generally, writing the matrix for φ in LPU normal form corresponds to modifying the train track in the region $\boldsymbol{a} \times [-1, 1]$ so that the downward arrows are on the left, the upward arrows are on the right, and there is a permutation in the middle. A complete list of geometric moves corresponding to different factorizations into elementary matrices is given in [12, Figure 23]. As an example, the reader should compare Figures 7 and 11.

The reason this form is useful is that it allows us to remove arrows and simplify. This is possible in general, by appealing to an algorithm given in [4], and ultimately gives rise to the proof of Theorem 1; see [12, Section 5] for details. The main point is that arrows winding counterclockwise around a puncture can be removed. Namely, suppose there is an arrow near an edge of the strip $\boldsymbol{a} \times [-1, 1]$ that, when pushed into the relevant annulus, runs counterclockwise between curve segments with different amounts of wrapping. Then there is a homotopy equivalence that produces a new train track — with the counterclockwise arrow removed — representing the same type D structure; see Figure 10. This is described in detail in [12, Lemma 5.11]. The result of this procedure, applied to the example described in Figure 11, is shown in Figure 12.

Recall that a local system over an immersed curve is a vector bundle over the curve. In general, all of our curves carry local systems, but when the associated bundle is one-dimensional and trivial we drop it from the notation. When working with signs, one-dimensional local systems are quite common as the coefficients along any given curve component multiply. Of course, noncompact curves do not carry interesting local systems since all vector bundles are trivial in this case. On the other hand, for compact curves it should be clear from the construction described above where a local system can arise: if two compact curves run parallel, then a crossover switch running between them cannot be removed by a chain isomorphism of type D structures. In general, local systems provide a clean way of presenting the relevant invariants, while the formalism expressing curves with local systems in terms of train tracks gives a concrete means of working with these objects. An example is shown in Figure 13; notice that, by replacing φ with φ^{-1} in this example, one can obtain a vertically simplified basis or

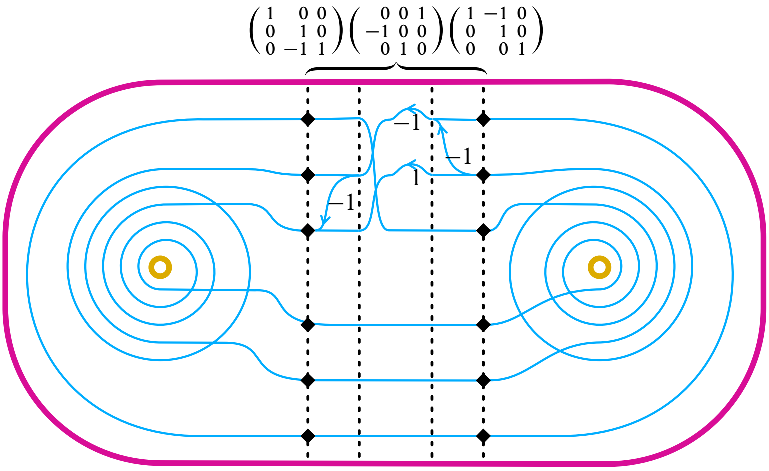

$\begin{pmatrix}1&0&0\\0&1&0\\0&-1&1\end{pmatrix}\begin{pmatrix}0&0&1\\-1&0&0\\0&1&0\end{pmatrix}\begin{pmatrix}1&-1&0\\0&1&0\\0&0&1\end{pmatrix}$
−1
−1
1
−1

Figure 11: Modifying the train track from Figure 7 according to an LPU decomposition of the matrix.

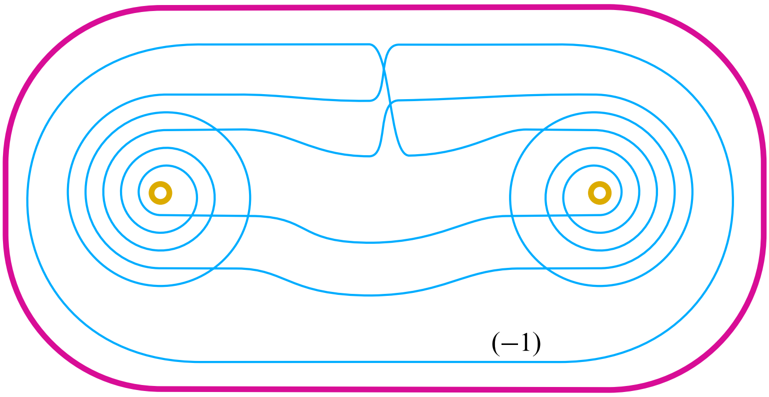


Figure 12: Modifying the train track from Figure 11 by removing the counterclockwise arrows. This produces an immersed curve — an object that is equivalent to the train track from Figure 6, and which carries a one-dimensional local system with automorphism that multiplies by −1.

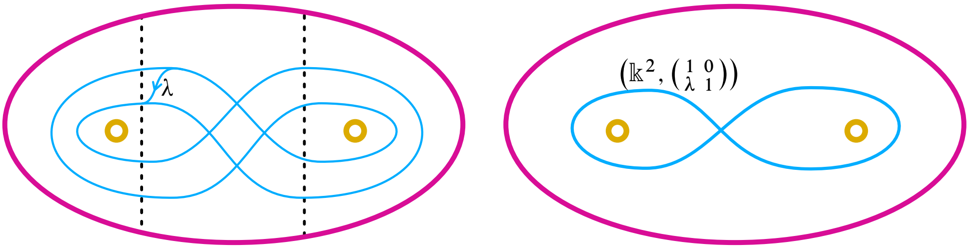

λ
$(\mathbb{k}^2, \begin{pmatrix}1&0\\\lambda&1\end{pmatrix})$

Figure 13: An arrow that cannot be cancelled gives rise to a nontrivial local system.

a horizontally simplified basis, but not both simultaneously. It appears to still be an open question if such phenomena arise for invariants associated with knots; see [11, Remark 2.9].

3 Adding a handle

We now introduce the second algebra: the extended torus algebra $\widetilde{\mathcal{A}}$. This algebra is introduced in [4], and is also the algebra arising naturally in our setting. By construction, the map takes the twice-punctured disk to the once-punctured torus T. An arc system for the latter is shown in Figure 14, from which the associated quiver

$$\begin{tikzcd}\bullet \arrow[r, bend left=40, "\rho_1"] \arrow[r, bend right=40, "\rho_3"'] & \circ \arrow[l, bend right=70, "\rho_0"'] \arrow[l, bend left=70, "\rho_2"]\end{tikzcd}$$

can be extracted — as before, we contract the arcs to the quiver vertices. Consulting Figure 16, note that $\boldsymbol{a}_\bullet$ is identified with the meridian μ and $\boldsymbol{a}_\circ$ is identified with the choice of longitude λ. With this arc system we associate an algebra $\widetilde{\mathcal{A}}$. Analogous to the relation $uv = 0$ from Figure 1, the algebra $\widetilde{\mathcal{A}}$ has relations

$$\rho_{i+1}\rho_i = 0$$

(indices interpreted modulo 4), as explained in Section 2. Note that the products $\rho_i\rho_{i+1} = \rho_{i(i+1)}$ are nonzero. For consistency with [4, Section 3.1] we would need to add an additional relation $\rho_0\rho_1\rho_2\rho_3\rho_0 = 0$, but this is not necessary in the present setting.

The arc system associated with $\widetilde{\mathcal{A}}$ decomposes the torus into a single disk, so type D structures associated with compact train tracks will be curved. We fix the curvature

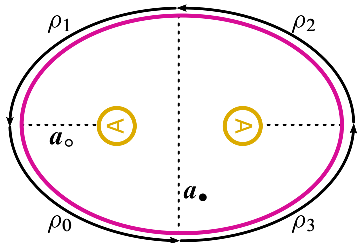


Figure 14: An arc system for the extended algebra $\widetilde{\mathcal{A}}$. The two discs are identified, producing a handle.

term $c=\rho_{0123}+\rho_{1230}+\rho_{2301}+\rho_{3012}$. Recall that a curved type D structure over $\widetilde{\mathcal{A}}$ satisfies the compatibility condition

$$(\mu\otimes\mathrm{id}_V)\circ(\mathrm{id}_{\widetilde{\mathcal{A}}}\otimes d)\circ d=c\cdot\mathrm{id}_{\widetilde{\mathcal{A}}}$$

and that, in this setting, the underlying $\Bbbk$–vector space decomposes so that $V=V_\bullet\oplus V_\circ$ as an $\mathcal{I}$–module.

The torus algebra is the quotient $\mathcal{A}=\widetilde{\mathcal{A}}/(\rho_0)$. Notice that in this quotient the curvature vanishes and the compatibility condition for type D structures given in Section 1 is recovered. This algebra is explored in depth in [17, Section 11; 4].

4 The proof of Theorem 2

To set the stage, we first describe three general constructions. First, given a type D structure ${}^{\Bbbk[u]}N$ over the polynomial ring $\Bbbk[u]$, there is a natural way to produce a dg module/chain complex over $\Bbbk[u]$: substitute each generator $\diamond$ in ${}^{\Bbbk[u]}N$ with a copy of the ring $\Bbbk[u]$, producing a free $\Bbbk[u]$–module, and then endow this module with a differential by substituting every arrow $\diamond\xrightarrow{\ell u^n}\diamond$ in ${}^{\Bbbk[u]}N$ with a map $\Bbbk[u]\xrightarrow{\cdot(\ell u^n)}\Bbbk[u]$ (where $\ell\in\Bbbk$). We denote the resulting dg module by $\Bbbk[u]\boxtimes{}^{\Bbbk[u]}N$, because it coincides with the result of box tensoring the type D structure with the module $\Bbbk[u]$ viewed as a bimodule over itself [16, Section 2.3.2]. Note that this operation respects homotopy equivalences and also can be reversed [16, Proposition 2.3.18], albeit in a less than straightforward way.

For the second construction, let $\Bbbk[v]$ be the graded polynomial ring in one variable with grading $a(v)=1$. (Below, a will be the Alexander grading.) Suppose ${}^{\Bbbk[v]}N=\bigoplus_{a\in\mathbb{Z}}{}^{\Bbbk[v]}N^a$ is a graded type D structure over $\Bbbk[v]$ such that the differential preserves the grading a. We can then produce a complex ${}^{\Bbbk[v]}N|_{v=1}$ by substituting arrows $\diamond\xrightarrow{\ell v^n}\diamond$ in ${}^{\Bbbk[v]}N$ by arrows $\diamond\xrightarrow{\ell}\diamond$. Clearly, this amounts to passing to the quotient $\Bbbk=\Bbbk[v]/(v-1)$. However, since $a(v)=1$, all the differentials in ${}^{\Bbbk[v]}N$ that involved v^n for $n\neq 0$ now change the grading in ${}^{\Bbbk[v]}N|_{v=1}$ by n. Thus, we can consider ${}^{\Bbbk[v]}N|_{v=1}$ as a filtered chain complex, where the filtration levels are $\mathcal{F}_j=\bigoplus_{a\leq j}N^a$. As a category, type D structures over $\Bbbk[v]$ are equivalent to filtered chain complexes via the construction above. In particular, type D structure homomorphisms and homotopies between them precisely correspond to filtered chain maps and filtered homotopies between them.

The third construction is similar to the second. Given, a graded type D structure ${}^{\Bbbk[v]}N$ over $\Bbbk[v]$ whose differential preserves the grading a, we define a complex ${}^{\Bbbk[v]}N|_{v=0}$ by removing all arrows $\diamond \xrightarrow{\ell v^n} \diamond$ for $n > 0$ in ${}^{\Bbbk[v]}N$. This amounts to passing to the quotient $\Bbbk = \Bbbk[v]/(v)$ or, equivalently, to passing to the associated graded complex of the filtered complex ${}^{\Bbbk[v]}N|_{v=1}$.

We can now provide a dictionary between the knot Floer structures used here and those in [17]. In this paper, the most general knot Floer invariant is the type D structure ${}^{\Bbbk[u,v]}CFK(S^3, K)$. In [17], two kinds of invariants appear. The first is the filtered chain complex $CFK^-(S^3, K)$ over $\Bbbk[u]$, which is a dg module over $\Bbbk[u]$ filtered with respect to the Alexander grading. It is obtained from ${}^{\Bbbk[u,v]}CFK(S^3, K)$ by applying the first construction to the variable u and the second construction to the variable v:

$$CFK^-(S^3, K) = \Bbbk[u] \boxtimes {}^{\Bbbk[u]}\big({}^{\Bbbk[u,v]}CFK(S^3, K)|_{v=1}\big).$$

The second invariant used in [17] is $gCFK^-(S^3, K)$, the associated graded complex of $CFK^-(S^3, K)$. It is obtained from ${}^{\Bbbk[u,v]}CFK(S^3, K)$ by applying the first construction to the variable u and the third construction to the variable v:

$$gCFK^-(S^3, K) = \Bbbk[u] \boxtimes {}^{\Bbbk[u]}\big({}^{\Bbbk[u,v]}CFK(S^3, K)|_{v=0}\big).$$

Example Consider the right-hand trefoil and its knot Floer invariants. The type D structure invariant is

$${}^{\Bbbk[u,v]}CFK(S^3, T_{2,3}) = [\diamond_1^1 \xleftarrow{u} \diamond_1^0 \xrightarrow{v} \diamond_1^{-1}],$$

where the superscripts and subscripts indicate the Alexander and δ gradings, respectively. Recall that the Alexander and δ gradings are $\mathrm{gr}(u) = (-1, 1)$ and $\mathrm{gr}(v) = (1, 1)$, so that the differential in the type D structure is of bidegree $(a, \delta) = (0, 1)$. The filtered chain complex over $\Bbbk[u]$ now becomes

$$CFK^-(S^3, K) = \Bbbk[u] \boxtimes {}^{\Bbbk[u]}\big({}^{\Bbbk[u,v]}CFK(S^3, K)|_{v=1}\big) = \big[\Bbbk[u]_1^1 \xleftarrow{\cdot u} \Bbbk[u]_1^0 \xrightarrow{1} \Bbbk[u]_1^{-1}\big],$$

while the associated graded chain complex over $\Bbbk[u]$ is equal to

$$gCFK^-(S^3, K) = \Bbbk[u] \boxtimes {}^{\Bbbk[u]}\big({}^{\Bbbk[u,v]}CFK(S^3, K)|_{v=0}\big) = \big[\Bbbk[u]_1^1 \xleftarrow{\cdot u} \Bbbk[u]_1^0\big] \oplus [\Bbbk[u]_1^{-1}].$$

We now proceed to the proof. We start with the knot Floer type D structure

$${}^{\mathcal{R}}CFK(S^3, K) = {}^{\mathbb{F}[u,v]}CFK(S^3, K)|_{vu=0},$$

and then homotope it to a representative (following the steps from Section 2) from which the curve invariant $\boldsymbol{\gamma}$ can be extracted. With the dictionary above in mind, Theorem A.11 of [17] describes in detail how to pass from ${}^{\mathcal{R}}CFK(S^3, K)$ to the type D structure ${}^{\mathcal{A}}\widehat{CFD}(M)$, which then produces a curve $\widehat{HF}(M)$ in the punctured torus ∂M. Our task is to prove that the resulting curve coincides with $\frown(\boldsymbol{\gamma})$.

We focus on segments of the curve $\boldsymbol{\gamma}$ in each of the annuli A_v and A_u, and consider their images under the map $\frown$. Starting with an illustrative example, the image of a curve segment corresponding to the arrow $\diamond \xrightarrow{v^3} \diamond$ is drawn in thick in Figure 15, left, relative to the arc system of the algebra $\widetilde{\mathcal{A}}$. Focussing on the first part of this segment, shown are the two ways in can be retracted to the boundary of the torus union the two arcs: in one case the homotoped path runs along ρ_1, and in the other case it runs along ρ_2 then ρ_3 then ρ_0. In the type D structure language, then, according to [4] and the discussion in Section 2, this part of the curve results in $\bullet \underset{\rho_{230}}{\overset{\rho_1}{\rightleftarrows}} \circ$. Similarly, the whole thick curve segment depicted in Figure 15, left, corresponds to the following part of a type D structure over $\widetilde{\mathcal{A}}$:

$$\begin{tikzcd}
\bullet \arrow[r, bend left, "\rho_1"] & \circ \arrow[l, bend left, "\rho_{230}"] \arrow[r, bend left, "\rho_{01}"] & \circ \arrow[l, bend left, "\rho_{23}"] \arrow[r, bend left, "\rho_{01}"] & \circ \arrow[l, bend left, "\rho_{23}"] \arrow[r, bend left, "\rho_0"] & \bullet \arrow[l, bend left, "\rho_{123}"]
\end{tikzcd}$$

More generally, the image of a curve segment corresponding to the arrow $\diamond \xrightarrow{v^i} \diamond$ is

$$\overbrace{\hphantom{\circ \qquad \circ \qquad \circ \qquad \cdots \qquad \circ}}^{\dim V_\circ = i}$$

$$\begin{tikzcd}
\bullet \arrow[r, bend left, "\rho_1"] & \circ \arrow[l, bend left, "\rho_{230}"] \arrow[r, bend left, "\rho_{01}"] & \circ \arrow[l, bend left, "\rho_{23}"] \arrow[r, bend left, "\rho_{01}"] & \circ \arrow[l, bend left, "\rho_{23}"] & \cdots & \circ \arrow[r, bend left, "\rho_0"] & \bullet \arrow[l, bend left, "\rho_{123}"]
\end{tikzcd}$$

Analogously, the image of a curve segment corresponding to the arrow $\diamond \xrightarrow{u^i} \diamond$ is

$$\overbrace{\hphantom{\circ \qquad \circ \qquad \circ \qquad \cdots \qquad \circ}}^{\dim V_\circ = i}$$

$$\begin{tikzcd}
\bullet \arrow[r, bend left, "\rho_3"] & \circ \arrow[l, bend left, "\rho_{012}"] \arrow[r, bend left, "\rho_{23}"] & \circ \arrow[l, bend left, "\rho_{01}"] \arrow[r, bend left, "\rho_{23}"] & \circ \arrow[l, bend left, "\rho_{01}"] & \cdots & \circ \arrow[r, bend left, "\rho_2"] & \bullet \arrow[l, bend left, "\rho_{301}"]
\end{tikzcd}$$

Passing to the quotient algebra $\mathcal{A}$ by setting $\rho_0 = 0$ simplifies the above two images to

$$\bullet \xrightarrow[\rho_1]{} \overbrace{\circ \xleftarrow[\rho_{23}]{} \circ \xleftarrow[\rho_{23}]{} \circ \quad \cdots \quad \circ}^{\dim V_\circ = i} \xleftarrow[\rho_{123}]{} \bullet$$

and

$$\bullet \xrightarrow[\rho_3]{} \overbrace{\circ \xrightarrow[\rho_{23}]{} \circ \xrightarrow[\rho_{23}]{} \circ \quad \cdots \quad \circ}^{\dim V_\circ = i} \xrightarrow[\rho_2]{} \bullet$$

These are precisely the two *stable* chains appearing in the statement of Theorem A.11 of [17]; according to their result, these are the parts of ${}^{\mathcal{A}}\widehat{CFD}(M)$ that correspond to the differentials $\diamond \xrightarrow{v^i} \diamond$ and $\diamond \xrightarrow{u^i} \diamond$ in ${}^{\mathcal{R}}CFK(S^3, K)$.

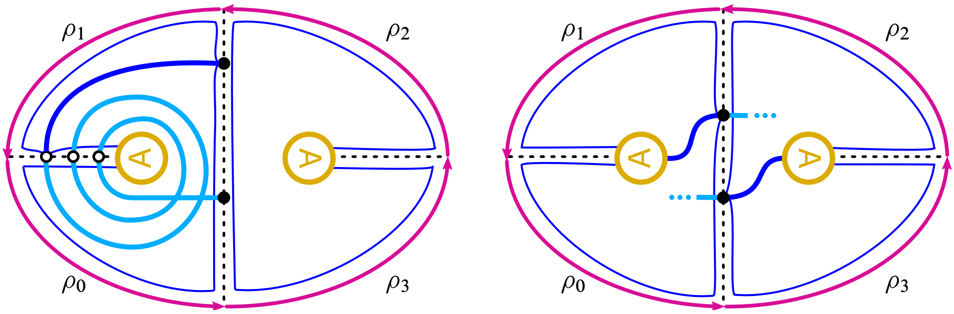


Figure 15: Sample parts of the curve corresponding to a stable chain (left) and an unstable chain (right) from [17, Theorem A.11].

The main subtlety is the appearance of the *unstable* chain, which we have already touched on. Defining in such a way that there is no extra twisting introduced (see Figure 3, left), the straight segment running over the handle in Figure 15, right, retracts in the two ways shown, producing the final part of the type D structure, $\bullet \underset{\rho_{30}}{\overset{\rho_{12}}{\rightleftarrows}} \bullet$. Setting $\rho_0 = 0$ results in $\bullet \xrightarrow{\rho_{12}} \bullet$, which is precisely the unstable chain from [17, Theorem A.11]; according to their result, this is the final piece (in addition to the stable chains) in ${}^{\mathcal{A}}\widehat{CFD}(M)$ (computed relative to the parametrization $(\mu, 2\tau)$ of the torus $T^2 = \partial M$). In [17, Theorem A.11], this final piece connects the distinguished generators ξ_0 and η_0 in the vertically and horizontally simplified bases of $CFK^-(K)$. It is left to note that the two generators in Figure 15, right, are precisely ξ_0 and η_0, because each is incident to only one arrow $\xrightarrow{v^i}$ or $\xrightarrow{u^j}$ in the complex ${}^{\mathcal{R}}CFK(S^3, K)$. We also remark that, while the unstable chain $\bullet \xrightarrow{\rho_{12}} \bullet$ corresponds to the 2τ–framing of the knot K, there are other type D structure presentations of the unstable chain in [17, Theorem A.11], and those would correspond to other choices of twisting in .

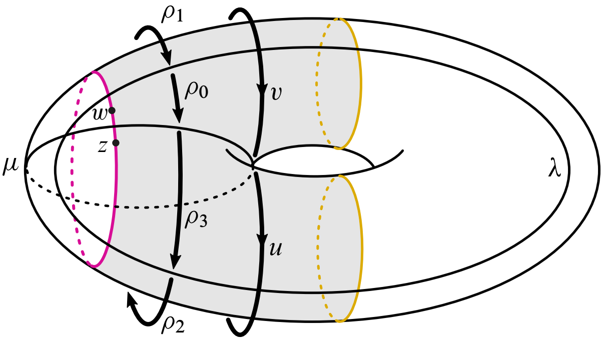


Figure 16: Both algebras $\mathcal{R}$ and $\widetilde{\mathcal{A}}$ in context.

The statement about the reverse operation follows from the discussion above. Namely, its clear that $\smallfrown\smallfrown(\frown(\boldsymbol{\gamma})) = \boldsymbol{\gamma}$, and, since we proved $\frown(\boldsymbol{\gamma}) = \widehat{HF}(M)$, we obtain $\smallfrown\smallfrown(\widehat{HF}(M)) = \boldsymbol{\gamma}$.

5 Comments on generalizations and related work

Perhaps the most interesting step in this constructive review of the Lipshitz–Ozsváth–Thurston correspondence comes about when the endpoints of the noncompact component $\boldsymbol{\gamma}_0 \subset \boldsymbol{\gamma}$ are identified to give a new compact component in the once-punctured torus. Note that the output of $\frown$ is always a compact curve, and this is consistent with the observation that $\widehat{HF}(M)$ is a compact curve. The latter, in turn, follows from the fact that $\widehat{CFD}(M)$ is an extendable type D structure [4, Appendix].

Joining the endpoints of the immersed curve $\boldsymbol{\gamma}_0$ associated with a knot K requires a choice of automorphism of $\Bbbk^n$ where n is the number of components in $\boldsymbol{\gamma}_0$. Denote the horizontal homology by $H^{\boldsymbol{h}} = H_*(C^{\boldsymbol{h}}|_{u=1})$ and the vertical homology by $H^{\boldsymbol{v}} = H_*(C^{\boldsymbol{v}}|_{v=1})$. Then, in Theorem 2, because the knot is in S^3, it follows that $n = 1$ and the automorphism is given, tautologically, by

$$H^{\boldsymbol{v}}(CFK^-(S^3, K)) \cong H^{\boldsymbol{h}}(CFK^-(S^3, K)) \cong \widehat{HF}(S^3) \cong \Bbbk,$$

as explained in [17, Section 11.5]. Thus, the operation $\frown$ is defined over any field provided that we choose a coefficient $a \in \Bbbk$ when we identify the ends of $\boldsymbol{\gamma}_0$ along a handle. We choose this coefficient to be $+1$ so that the bordered invariant for the solid torus is a circle with the trivial local system. We note that bordered Floer homology is only defined over the two-element field $\mathbb{F}$. As such, the map $\frown$ and Theorem 2 gives a *candidate* bordered invariant for the knot exterior when $\Bbbk \neq \mathbb{F}$.

We now consider the general case of a knot K in Y. Decomposing along spin^c–structures, the same strategy as above works if Y is an L–space [8]. More generally, however, one needs to know the isomorphism

$$H^{\boldsymbol{v}}(CFK^-(Y, K)) \cong H^{\boldsymbol{h}}(CFK^-(Y, K)) \cong \widehat{HF}(Y)$$

(which may be block-decomposed according to spin^c–structures). This recovers a generalization of [17, Theorem A.11], which may be found in forthcoming work of Hockenhull [9] building on his invariant $\mathrm{Poly}(L, \Lambda)$ [10]. From our perspective, the passage from the knot Floer homology of a knot K in Y to the bordered invariants of $Y \smallsetminus \mathring{\nu}(K)$ requires the isomorphism shown above. As there is a decomposition

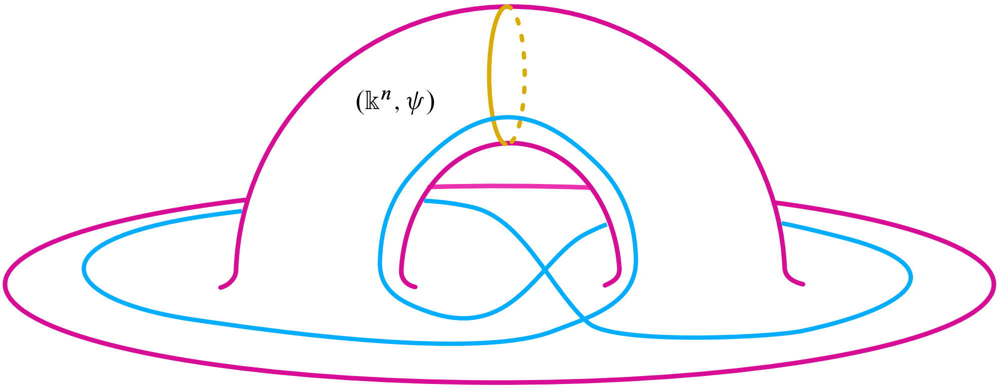


Figure 17: A sample hypothetical local system, where $n = \dim \widehat{HF}(Y)$.

according to spin^c–structures, there is no loss of generality in considering the case where Y is an integer homology sphere. When such a Y is not an L–space, we have that $\dim \widehat{HF}(Y) > 1$ and, in principle, the automorphism ψ induced by the isomorphism between the homologies $H^{\boldsymbol{h}}$ and $H^{\boldsymbol{v}}$ can be interesting. In particular, while all components of $\boldsymbol{\gamma}_0(K)$ carry trivial local systems, the new compact object $(\boldsymbol{\gamma}_0(K))$ obtains an additional local system $(\Bbbk^{\boldsymbol{n}}, \psi)$; see Figure 17. The key point of difference is that the output will be equivalent to a simply faced precurve (in the torus) in general, and a further application of the arrow sliding algorithm may be required to obtain immersed curves. The algebraic side of this story is laid out carefully by Hockenhull [9; 10].

Finally, Hanselman gives another approach [3]: his construction takes the complex $CFK^-(K)$ and outputs an immersed curve in the strip covering the twice-punctured disk D, containing a countable set of pairs of punctures. This cover of the disk is useful for recording the Alexander grading, and also works with general fields (hence producing candidate bordered invariants). We advertise that Hanselman's construction has a different aim in mind, namely a candidate bordered-minus invariant obtained by promoting the curves to describe type D structures over $\Bbbk[u, v]$.

6 The proof of Theorem 3

For simplicity we first focus on the case of the two-element field $\Bbbk = \mathbb{F}$. A few properties of the invariant ${}^{\mathcal{R}}CFK(S^3, K)$ are needed for the proof. First, given two type D structures over the polynomial algebra $\Bbbk[u, v]$ or its quotient $\mathcal{R}$, their tensor product is another type D structure

$$(V,d)\otimes(V',d')=(V\otimes_{\Bbbk}V',d\otimes\mathrm{id}+\mathrm{id}\otimes d').$$

Now, reformulating [18, Theorem 7.1], the behaviour of knot Floer homology under taking the connected sum can be described as

$$^{\mathcal{R}}CFK(S^3,K\mathbin{\#}K')\simeq{}^{\mathcal{R}}({}^{\mathcal{R}}CFK(S^3,K)\otimes{}^{\mathcal{R}}CFK(S^3,K')).$$

The mirroring operation is also well understood — see [18, Proposition 3.7] —

$$^{\mathcal{R}}CFK(S^3,\mathrm{m}K)\simeq{}^{\mathcal{R}}\overline{CFK(S^3,K)},$$

where the latter is the *dual type D structure*, equal to the original one but with all differentials reversed [15, Definition 2.5] (since $\mathcal{R}$ is commutative, the fact that dualizing turns left type D structure to right ones is not a problem). Finally, we need an algebraic relationship between morphism spaces of type D structures [16, Section 2.2.3] and their tensor products. Given any two type D structures, the definitions imply the isomorphism of chain complexes

$$\mathcal{R}\boxtimes{}^{\mathcal{R}}({}^{\mathcal{R}}\overline{N}\otimes{}^{\mathcal{R}}N')\cong\mathrm{Mor}({}^{\mathcal{R}}N,{}^{\mathcal{R}}N').$$

With the properties above in place, the proof of Theorem 3 is the sequence of isomorphisms

$$\begin{aligned}
{}_{\mathcal{R}}HFK(S^3,\mathrm{m}K\mathbin{\#}K')&\cong H_*(\mathcal{R}\boxtimes{}^{\mathcal{R}}CFK(S^3,\mathrm{m}K\mathbin{\#}K'))\\
&\cong H_*\big(\mathcal{R}\boxtimes{}^{\mathcal{R}}[{}^{\mathcal{R}}CFK(S^3,\mathrm{m}K)\otimes{}^{\mathcal{R}}CFK(S^3,K')]\big)\\
&\cong H_*\big(\mathcal{R}\boxtimes{}^{\mathcal{R}}[{}^{\mathcal{R}}\overline{CFK(S^3,K)}\otimes{}^{\mathcal{R}}CFK(S^3,K')]\big)\\
&\cong H_*\big(\mathrm{Mor}({}^{\mathcal{R}}CFK(S^3,K),{}^{\mathcal{R}}CFK(S^3,K'))\big)\\
&\cong HF(\boldsymbol{\gamma}(K),\boldsymbol{\gamma}(K')),
\end{aligned}$$

where the final isomorphism follows from the general description of morphism spaces between type D structures over surface algebras [12, Theorem 1.5].

The recipe for adding signs follows the Koszul sign rule, which is discussed in Section 12 of [20] in detail. We find that the resulting signs are a bit more natural if one considers right type D structures [12, Example 2.10], rather than left ones [20, Section 12.3], as then there are no extra signs when box tensoring with $-\boxtimes{}_{\mathcal{R}}\mathcal{R}_{\mathcal{R}}$; this is explained in [12, page 19]. Now, since the algebra $\mathcal{R}$ is commutative, our left type D structures can be viewed as right type D structures, and after that filling in the signs becomes straightforward. We refer the reader to [12, Sections 2 and 5]. □

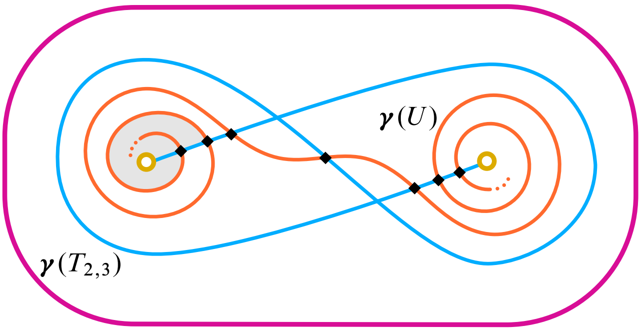


Figure 18: Illustrating Theorem 3 in the case of the unknot $K = U$ and the right-hand trefoil $K' = T_{2,3}$. The curve $\boldsymbol{\gamma}(U)$ is a horizontal arc connecting the punctures, but because we are in the wrapped setting, one needs to wrap $\boldsymbol{\gamma}(U)$ infinitely many times around the punctures when pairing with another curve.

To illustrate this gluing result, suppose $K = U$ and $K' = T_{2,3}$. Then the knot Floer homology of the connected sum is equal to

$$\begin{aligned} {}_{\mathcal{R}}HFK(S^3, T_{2,3}) &= H_*(\mathcal{R} \xleftarrow{\cdot v} \mathcal{R} \xrightarrow{\cdot u} \mathcal{R}) \\ &= [\cdots \xleftarrow{v} \diamond \xleftarrow{v} \diamond \xleftarrow{v} \diamond \xrightarrow{u} \diamond \xleftarrow{v} \diamond \xrightarrow{u} \diamond \xrightarrow{u} \diamond \xrightarrow{u} \cdots], \end{aligned}$$

where the arrows indicate the $\mathcal{R}$–action. The corresponding wrapped Lagrangian Floer homology $HF(\boldsymbol{\gamma}(U), \boldsymbol{\gamma}(T_{2,3}))$ is illustrated in Figure 18. Note that in this example the $\mathcal{R}$–action can be seen geometrically by counting Maslov index 2 disks covering the punctures; one of these is shaded in the picture. The same is true for the $\Bbbk[H]$–action on Bar-Natan homology, viewed as wrapped Lagrangian Floer homology of immersed curves in [12, Example 7.7]. In general, to recover these module-structures, only some of the Maslov index 2 disks should be counted — we will investigate this in future work.

Acknowledgements

This short paper benefited from stimulating conversations with Jonathan Hanselman, Thomas Hockenhull and Matthew Stoffregen. The work was initiated during the CRM50 thematic program *Low-dimensional topology* and we thank the CRM and CIRGET for hosting a great event.

Kotelskiy is supported by an AMS–Simons travel grant. Watson is supported by an NSERC discovery/accelerator grant and was partially supported by funding from the Simons Foundation and the Centre de Recherches Mathématiques, through the

Simons–CRM scholar-in-residence program. Zibrowius is supported by the Emmy Noether Programme of the DFG, project number 412851057, and the SFB 1085 *Higher invariants* in Regensburg.

References

[1] **I Dai**, **J Hom**, **M Stoffregen**, **L Truong**, *More concordance homomorphisms from knot Floer homology*, Geom. Topol. 25 (2021) 275–338 MR Zbl

[2] **F Haiden**, **L Katzarkov**, **M Kontsevich**, *Flat surfaces and stability structures*, Publ. Math. Inst. Hautes Études Sci. 126 (2017) 247–318 MR Zbl

[3] **J Hanselman**, *Knot Floer homology as immersed curves*, preprint (2023) arXiv 2305.16271

[4] **J Hanselman**, **J Rasmussen**, **L Watson**, *Bordered Floer homology for manifolds with torus boundary via immersed curves*, preprint (2016) arXiv 1604.03466 To appear in J. Amer. Math. Soc.

[5] **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

[6] **J Hanselman**, **L Watson**, *Cabling in terms of immersed curves*, Geom. Topol. 27 (2023) 925–952 MR Zbl

[7] **M Hedden**, **C M Herald**, **P Kirk**, *The pillowcase and perturbations of traceless representations of knot groups*, Geom. Topol. 18 (2014) 211–287 MR Zbl

[8] **M Hedden**, **A S Levine**, *Splicing knot complements and bordered Floer homology*, J. Reine Angew. Math. 720 (2016) 129–154 MR Zbl

[9] **T Hockenhull**, *Duality patterns in knot Floer homology*, in preparation

[10] **T Hockenhull**, *Holomorphic polygons and the bordered Heegaard Floer homology of link complements*, preprint (2018) arXiv 1802.02443

[11] **J Hom**, *Heegaard Floer invariants and cabling*, PhD thesis, University of Pennsylvania (2011) Available at `https://repository.upenn.edu/edissertations/329/`

[12] **A Kotelskiy**, **L Watson**, **C Zibrowius**, *Immersed curves in Khovanov homology*, preprint (2019) arXiv 1910.14584

[13] **Y Lekili**, **A Polishchuk**, *Auslander orders over nodal stacky curves and partially wrapped Fukaya categories*, J. Topol. 11 (2018) 615–644 MR Zbl

[14] **Y Lekili**, **A Polishchuk**, *Homological mirror symmetry for higher-dimensional pairs of pants*, Compos. Math. 156 (2020) 1310–1347 MR Zbl

[15] **R Lipshitz**, **P S Ozsváth**, **D P Thurston**, *Heegaard Floer homology as morphism spaces*, Quantum Topol. 2 (2011) 381–449 MR Zbl

[16] **R Lipshitz**, **P S Ozsváth**, **D P Thurston**, *Bimodules in bordered Heegaard Floer homology*, Geom. Topol. 19 (2015) 525–724 MR Zbl

[17] **R Lipshitz**, **P S Ozsvath**, **D P Thurston**, *Bordered Heegaard Floer homology*, Mem. Amer. Math. Soc. 1216, Amer. Math. Soc., Providence, RI (2018) MR Zbl

[18] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and knot invariants*, Adv. Math. 186 (2004) 58–116 MR Zbl

[19] **P Ozsváth**, **Z Szabó**, *Algebras with matchings and knot Floer homology*, preprint (2019) arXiv 1912.01657

[20] **P Ozsváth**, **Z Szabó**, *Bordered knot algebras with matchings*, Quantum Topol. 10 (2019) 481–592 MR Zbl

[21] **J A Rasmussen**, *Floer homology and knot complements*, PhD thesis, Harvard University (2003) MR arXiv math/0306378

[22] **I Zemke**, *Connected sums and involutive knot Floer homology*, Proc. Lond. Math. Soc. 119 (2019) 214–265 MR Zbl

[23] **C Zibrowius**, *Peculiar modules for* 4*–ended tangles*, J. Topol. 13 (2020) 77–158 MR Zbl

Department of Mathematics, Indiana University
Bloomington, IN, United States

Current address: *Department of Mathematics, Stony Brook University*
Stony Brook, NY, United States

Department of Mathematics, University of British Columbia
Vancouver, BC, Canada

Faculty of Mathematics, University of Regensburg
Regensburg, Germany

Current address: *Department of Mathematical Sciences, Durham University*
Durham, United Kingdom

artofkot@gmail.com, liam@math.ubc.ca, claudius.zibrowius@posteo.net

https://artofkot.github.io/, https://personal.math.ubc.ca/~liam/,
https://cbz20.raspberryip.com/

Received: 30 July 2020 Revised: 8 October 2021

Algebraic & Geometric Topology **23**:6 (2023) 2545–2559
DOI: 10.2140/agt.2023.23.2545
Published: 7 September 2023

New bounds on maximal linkless graphs

Ramin Naimi
Andrei Pavelescu
Elena Pavelescu

We construct a family of maximal linklessly embeddable graphs on n vertices and $3n-5$ edges for all $n \geq 10$, and another family on n vertices and $m < \frac{25}{12}n - \frac{1}{4}$ edges for all $n \geq 13$. The latter significantly improves the lowest edge-to-vertex ratio for any previously known infinite family. We construct a family of graphs showing that the class of maximal linklessly embeddable graphs differs from the class of graphs that are maximal without a K_6 minor studied by L Jørgensen. We give necessary and sufficient conditions for when the clique sum of two maximal linklessly embeddable graphs over K_2, K_3 or K_4 is a maximal linklessly embeddable graph, and use these results to prove our constructions yield maximal linklessly embeddable graphs.

57M15; 05C10

1 Introduction

All graphs in this paper are finite and simple. A graph is *intrinsically linked* (IL) if every embedding of it in $\mathbb{R}^3$ (or, equivalently, S^3) contains a nontrivial 2–component link. A graph is *linklessly embeddable* if it is not intrinsically linked (nIL). A nIL graph G is *maxnil* if it is not a proper subgraph of a nIL graph of the same order. The combined work of Conway and Gordon [2], Sachs [11] and Robertson, Seymour and Thomas [9] fully characterized IL graphs: a graph is IL if and only if it contains a graph in the Petersen family as a minor. The Petersen family consists of seven graphs obtained from K_6 by ∇Y–moves and $Y\nabla$–moves, as described in Figure 1. The ∇Y–move and the $Y\nabla$–move preserve the IL property.

The property of being maxnil is, in a way, analogous to the property of being maximal planar. While it is well known that every maximal planar graph with n vertices has $3n-6$ edges, an analogous statement for maxnil graphs does not exist. For example,

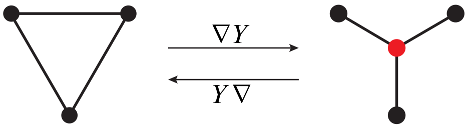


Figure 1: ∇Y– and $Y\nabla$–moves.

start with a maximal planar graph G and add one vertex v together with all the edges from v to the vertices of G. Such a graph is maxnil by [11], and if it has n vertices, then it has $4n-10$ edges. In fact, $4n-10$ is an upper bound on the number of edges of a maxnil graph on n vertices. This follows from work of Mader [7], who proved that having more than $4n-10$ edges implies the existence of a K_6 minor, which implies the graph is IL.

On the other hand, Jørgensen [5] and Dehkordi and Farr [3] constructed maxnil graphs with n vertices and $3n-3$ edges. Jørgensen's maxnil graphs are obtained from the Jørgensen graph in Figure 2, left, by subdividing the highlighted edge incident to the vertex y and then adding edges that connect every new vertex to u and v. We denote the graph obtained this way through i subdivisions by J_i for $i \geq 1$. See Figure 2, right.

Recently, Aires [1] found a family of graphs with fewer than $3n-3$ edges. For each value $n \geq 13$ with $n \equiv 3 \pmod{10}$, he constructed a maxnil graph with $\frac{14}{5}n - \frac{27}{5}$ edges. He also proved that, if G is a maxnil graph with $n \geq 5$ vertices and m edges, then $m \geq 2n$. This bound is sharp: the maxnil graph $Q(13,3)$ described by Maharry [8] has 26 edges and 13 vertices.

In Section 2, we present two constructions of maxnil graphs. The first one is a family of maxnil graphs with $n \geq 10$ vertices and $3n-5$ edges. This construction builds upon a maxnil graph on 10 vertices and 25 edges and uses edge subdivisions. The second

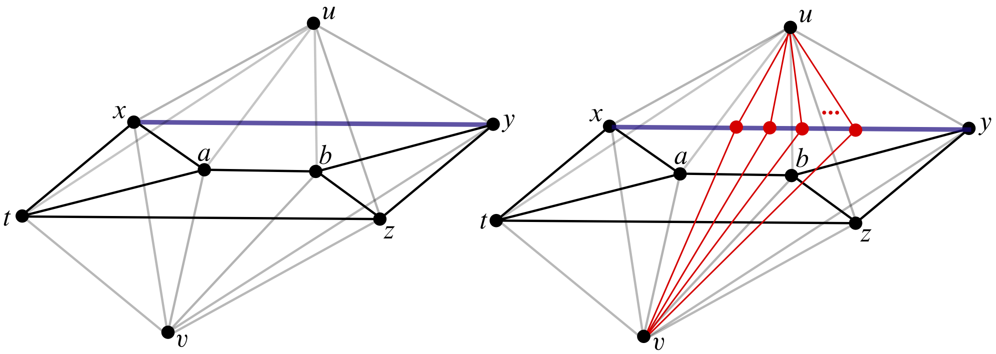


Figure 2: Left: the Jørgensen graph. Right: the graph J_i in Jørgensen's $3n-3$ family.

construction significantly improves on Aires' result on the number of edges. Using clique sums of copies of $Q(13,3)$, we construct examples with a smaller "edge-to-vertex ratio", as in the following theorem:

Theorem *For each $n \geq 13$, there exists a maxnil graph G with n vertices and $m < \frac{25}{12}n - \frac{1}{4}$ edges.*

In Section 3, we study the properties of maxnil graphs under clique sums. Some of these results are used in the constructions of Section 2. We give sufficient and necessary conditions for when the clique sum of two maxnil graphs over K_2, K_3 or K_4 is maxnil. Jørgensen [5] studied clique sums of graphs that are maximal without a K_6 minor. We give examples showing that the class of maxnil graphs and the class of graphs that are maximal without a K_6 minor are distinct.

2 Two families of maxnil graphs

We note that the Jørgensen graph is 2–apex, ie removing the vertices u and v leaves a planar graph P. Furthermore, the embedding of P in $\mathbb{R}^2$ shown in Figure 2, left, has no separating cycles, ie for every cycle C in P, one of the components of $\mathbb{R}^2 \setminus C$ contains no vertices of P. These properties are generalized in the next lemma, which we use to prove the graphs in the $3n-5$ family are nIL.

Lemma 1 *Let G be a graph with two nonadjacent vertices u, v such that there exists an embedding Σ of $G - \{u, v\}$ in $\mathbb{R}^2$, where, for every cycle C in Σ, $\mathbb{R}^2 \setminus C$ has a component X such that $X \cup C$ separates u and v (ie every path in G from u to v contains a vertex in $X \cup C$). Then embedding u as $(0,0,1)$ and v as $(0,0,-1)$ and connecting each of them to its neighbors in Σ with straight edges yields a linkless embedding of G in $\mathbb{R}^3$.*

Proof Let Γ denote the embedding of G as described in the lemma, and let $K \cup K'$ be a 2–component link in Γ. We consider two cases.

Case 1 (neither K nor K' contains both u and v) Then we have three subcases: zero, one or both of K and K' are in Σ. In each of these three subcases it is easy to see that $K \cup K'$ is a trivial link. We prove this for one of the three subcases here; the other two are similar and easier. Suppose K contains u but not v, and $K' \subset \Sigma$. Then K consists of two edges incident to u and a path $P \subset \Sigma$. Connecting u with straight line segments to every point in P gives us a Γ–panel for K. On the other hand, K' bounds a disk D in $\mathbb{R}^2$. We isotop D, while keeping its boundary fixed, by pushing its interior slightly below $\mathbb{R}^2$, to make it disjoint from K (since K contains no points below $\mathbb{R}^2$). It follows that $K \cup K'$ is a trivial link.

Case 2 (one of the link's components, say K, contains both u and v) Then $K' \subset \Sigma$. So $\mathbb{R}^2 \setminus K'$ has two components such that one of them, X, separates u and v. Therefore all vertices of K except u and v lie in X. Now, K has exactly two vertices, call them a and b, that are adjacent to u, and two vertices, c and d, adjacent to v. Note that $\{a, b\}$ is not necessarily disjoint, or even distinct, from $\{c, d\}$. Furthermore, $K \cap X$ consists of two components, P_1 and P_2, each of which is a path of length zero or greater. We can assume $a, c \in P_1$ and $b, d \in P_2$. We consider three subcases.

Case 2.1 ($a = c$ and $b = d$) Join a to b by an arc $\beta \subset X$ (not necessarily in Σ), and then connect each of u and v by straight line segments to every point in β. See Figure 3, left. This gives us a disk bounded by K and disjoint from K'. Similarly to Case 1 above, K' also bounds a disk disjoint from K. Hence $K \cup K'$ is a trivial link.

Case 2.2 ($a = c$ and $b \neq d$) Join a to each of b and d by disjoint arcs β and δ respectively, both in X, such that $\beta \cup \delta \cup P_2$ is a simple closed curve. See Figure 4, right. Connect each of u and v by straight line segments to every point in β and δ respectively. This gives us two disks whose union with the disk bounded by $\beta \cup \delta \cup P_2$ in X is a disk bounded by K and disjoint from K'. As before, K' bounds a disk disjoint from K. Hence, $K \cup K'$ is a trivial link.

Case 2.3 ($a \neq c$ and $b \neq d$) This case is similar to Case 2.2, except that we join a to b and c to d by disjoint arcs β and δ in X such that $\beta \cup \delta \cup P_1 \cup P_2$ is a simple closed curve. □

2.1 The $3n - 5$ family

We construct a family of graphs with n vertices and $3n - 5$ edges for $n \geq 10$. This family is obtained from the graph G pictured in Figure 4, left, through a sequence of subdivisions and edge additions. The graph G is obtained from the Jørgensen graph by splitting (the opposite of contracting edges) the vertices a and b into the edges ad and bc. See Figures 2, left, and 4, left. With the notation in Figure 4, left, construct the graph G_1 by subdividing the edge xy with a new vertex z_1, then adding edges z_1u and z_1v. Construct graphs G_i for $i \geq 2$ as follows: subdivide the edge $z_{i-1}y$ of G_{i-1} with a new vertex z_i, then add edges z_iu and z_iv to G_{i-1}. Notice that G_i has one more vertex and three more edges than G_{i-1}. The graph G_i has $10 + i$ vertices and $25 + 3i = 3(10 + i) - 5$ edges. We note that the graphs G_i can also be obtained by successive splittings of the vertex y into the edge yz_i.

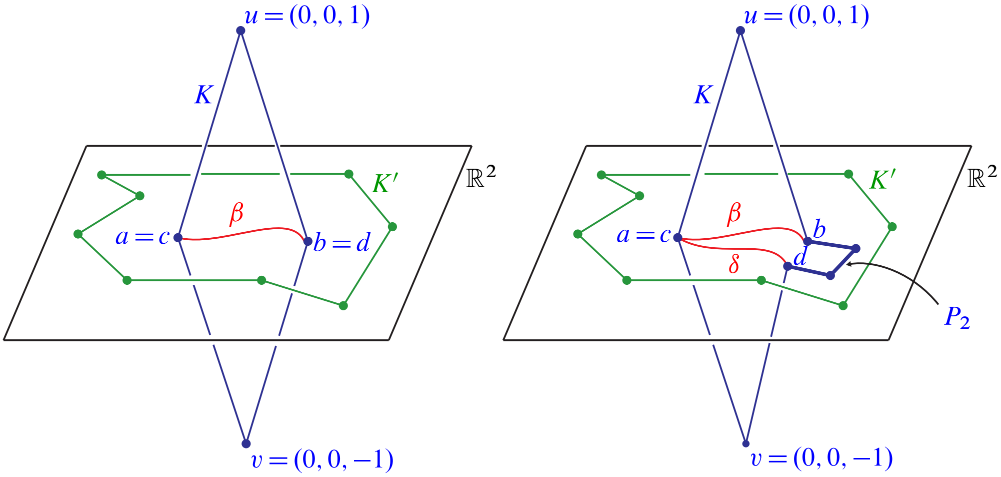


Figure 3: Left: configuration for Case 2.1. Right: configuration for Case 2.2.

Proposition 2 *The graphs G and G_i in Figure 4 are linklessly embeddable.*

Proof It is straightforward to check that these graphs satisfy the hypotheses of Lemma 1 and hence are nIL. □

Proposition 3 *The graph G in Figure 4, left, is maxnil.*

Proof Since G is linklessly embeddable, it remains to show that adding any edge to G gives an IL graph.

Note that both of the minors $G/(ab \cup cd)$ and $G/(ad \cup bc)$ are isomorphic to the Jørgensen graph. If an edge e other than bd is added to $G - \{u, v\}$, then e is an edge

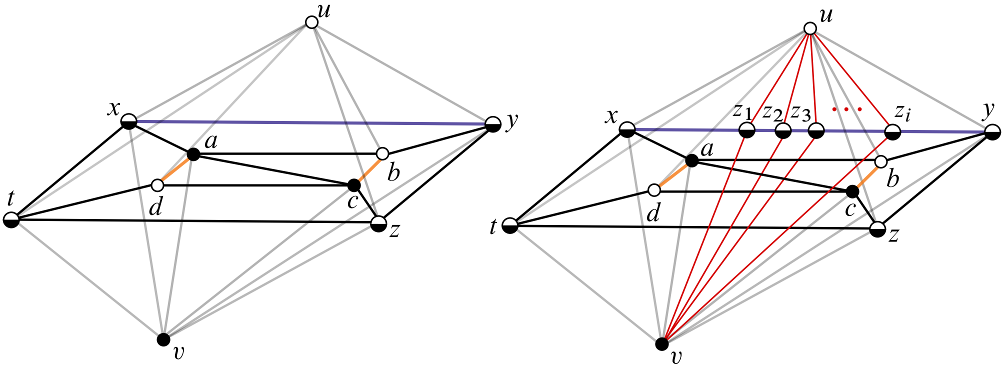


Figure 4: Left: the graph G is maxnil with 10 vertices and 25 edges. Right: the graph G_i is obtained through i edge subdivisions and edge additions.

in $(G+e)/(ab \cup cd)$ or $(G+e)/(ad \cup bc)$. Thus $G+e$ contains a minor that itself contains the Jørgensen graph plus an edge.

Since the Jørgensen graph is maxnil, $G+e$ is IL. The same holds if $e = uv$ is added to G. If the edge bd is added, then contracting the edges dt, cz, ux and vy creates a K_6 minor of $G+bd$.

Lastly, suppose an edge e from u or v to $G - \{u, v\}$ is added; by symmetry, we can assume that $e = ua$ or $e = vb$. If $e = ua$, then contracting the edges cd, dt, by and uz creates a K_6 minor of $G+ua$. If $e = vb$, then contracting the edges ax, cz, du and dt creates a K_6 minor of $G+vb$. □

Proposition 4 *All graphs G_i for $i \geq 1$ are maxnil.*

Proof Since G_i is linklessly embeddable, it remains to show that adding any edge to G_i gives an IL graph. Adding any edge e different from xy and disjoint from $\{z_1, z_2, \dots, z_i\}$ to G_i gives a graph $G_i + e$ that contains $G+e$ as a minor (obtained by contracting the path $xz_1z_2 \dots z_i$). Since G is maxnil, $G+e$ is IL and so is $G_i + e$. Adding an edge e that is either xy or has at least one endpoint in $\{z_1, z_2, \dots, z_i\}$ to G_i gives a graph $G_i + e$ that contains $J_i + e$ as a minor (obtained by contracting the edges ad and bc). Since J_i is maxnil, $J_i + e$ is IL and so is $G_i + e$. □

2.2 The $Q(13,3)$ family

A graph G is called *triangular* if each edge of G belongs to at least one triangle. In a nontriangular graph, an edge that is not part of a triangle is a *nontriangular edge*. In Section 3, we study the properties of maxnil graphs under the operation of clique sum (defined in Section 3). For the construction presented in the next theorem we use the result of Lemma 10 about clique sums of maxnil graphs over K_2.

Theorem 5 *For each $n \geq 13$, there exists a maxnil graph G with n vertices and $m < \frac{25}{12}n - \frac{1}{4}$ edges.*

Proof The construction is based on the maxnil graph $Q_{13,3}$ described by Maharry [8]. See Figure 5, left. This graph has 13 vertices and 26 edges, and it is triangle free.

For each n with $13 \leq n \leq 39$, we construct a set of maxnil graphs with n vertices and $2n$ edges by adding $n-13$ new vertices, and then choosing $n-13$ edges in $Q_{13,3}$ and

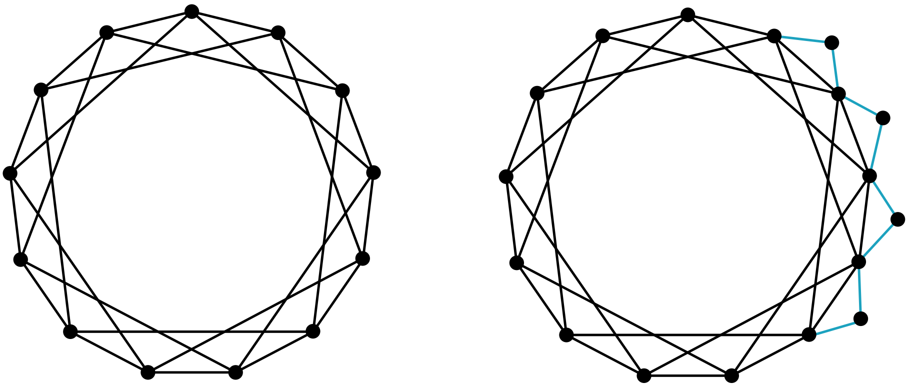


Figure 5: Left: $Q_{13,3}$ is a maxnil graph with 13 vertices and 26 edges. Right: a maxnil graph with 17 vertices and 34 edges obtained from $Q_{13,3}$ by adding four vertices of degree 2 and eight edges.

connecting the two endpoints of each of them to one of the new vertices. Equivalently, we are taking the clique sum of $Q_{13,3}$ with $n-13$ disjoint triangles over $n-13$ copies of K_2. See Figure 5, right. By Lemma 10, the resulting graph is maxnil.

The graph on 39 vertices obtained this way is triangular, so the construction cannot proceed further. To build graphs with a larger number of vertices, we use multiple copies of $Q_{13,3}$ joined along an edge (clique sum over K_2). Consider $k \geq 1$ copies of $Q_{13,3}$ and choose one edge in each copy. Then join the k graphs together by identifying the k chosen edges into one edge. This graph, which we denote by H_k, is maxnil (by repeated application of Lemma 10) and has $11k+2$ vertices and $25k+1$ edges. All edges of H_k are nontriangular and adding vertices of degree 2 (as above) along any subset of the edges of H_k gives a maxnil graph.

For $n \geq 13$, let $k = \lceil \frac{1}{36}(n-3) \rceil$ and add $n-(11k+2)$ vertices of degree 2 along any $n-(11k+2)$ edges of H_k. With every added vertex of degree 2, the number of edges is increased by 2. This gives a maxnil graph with n vertices and $m = (25k+1)+2[n-(11k+2)] = 2n+3k-3$ edges. Moreover,

$$m = 2n + 3\lceil \tfrac{1}{36}(n-3) \rceil - 3 < 2n + 3\big(\tfrac{1}{36}(n-3)+1\big) - 3 = \tfrac{25}{12}n - \tfrac{1}{4}. \qquad \square$$

Remark 6 The above shows there exist maxnil graphs of arbitrarily large order n with an edge-to-vertex ratio of less than $\frac{25}{12} - 1/(4n)$. Whether this edge-to-vertex ratio can be lowered further is an open question.

3 Clique sums of maxnil graphs

In this section we study the properties of maxnil graphs under taking clique sums. A set $S \subset V(G)$ is a *vertex cut set* of a connected graph G if $G - S$ is disconnected. We say a vertex cut set $S \subset V(G)$ is *minimal* if no proper subset of S is a vertex cut set of G. A graph G is the *clique sum* of G_1 and G_2 over K_t if $V(G) = V(G_1) \cup V(G_2)$, $E(G) = E(G_1) \cup E(G_2)$ and the subgraphs induced by $V(G_1) \cap V(G_2)$ in both G_1 and G_2 are complete of order t. Since the vertices of the clique over which a clique sum is taken form a vertex cut set in the resulting graph, the vertex connectivity of a clique sum over K_t is at most t. For a set of vertices $\{v_1, v_2, \dots, v_k\} \subset V(G)$, $\langle v_1, v_2, \dots, v_k \rangle_G$ denotes the subgraph of G induced by this set of vertices. By abuse of notation, the subgraph induced in G by the union of the vertices of subgraphs $H_1, H_2, \dots, H_k$ is denoted by $\langle H_1, H_2, \dots, H_k \rangle_G$.

Holst, Lovász and Schrijver [4, Theorem 2.10] studied the behavior of the Colin de Verdière μ–invariant for graphs under clique sums. Since a graph G is nIL if and only if $\mu(G) \leq 4$ [6; 10], their theorem implies the following:

Theorem 7 (Holst, Lovász and Schrijver [4]) *If G is the clique sum over S of two nIL graphs, then G is IL if and only if one can contract two or three components of $G - S$ so that the contracted nodes together with S form a K_7 minus a triangle.*

Theorem 7 implies that, for $t \leq 3$, the clique sum over K_t of nIL graphs is nIL. While Theorem 7 shows when a clique sum is nIL, it does not establish when a clique sum of maxnil graphs is maxnil.

Lemma 8 *Any maxnil graph is 2–connected.*

Proof Let G be a maxnil graph. If G is disconnected, let A and B denote two of its connected components. Let $a \in V(A)$ and $b \in V(B)$. Then $G + ab$ is a nIL graph, as it can be obtained by performing two consecutive clique sums over K_1 of nIL summands, namely

$$G + ab = A \cup_{\{a\}} ab \cup_{\{b\}} (G - A).$$

But this contradicts the maximality of G.

If the vertex connectivity of G is one, assume $x \in V(G)$ is a cut vertex; that is, $G - \{x\} = A \sqcup B$, with A and B nonempty, and no edges between vertices of A and vertices of B. Let $a \in V(A)$ and $b \in V(B)$ be neighbors of x in G. Then $G + ab$ is nIL, as it can be obtained by performing two consecutive clique sums over K_2 of nIL summands. If Δ denotes the triangle axb,

$$G + ab = \langle A, x\rangle_G \cup_{ax} \Delta \cup_{xb} \langle B, x\rangle_G.$$

But this contradicts the maximality of G. □

Lemma 9 *Let G be a maxnil graph with a vertex cut set $S = \{x, y\}$, and let $G_1, G_2, \dots, G_r$ denote the connected components of $G - S$. Then $xy \in E(G)$ and $\langle G_i, S\rangle_G$ is maxnil for all $1 \le i \le r$.*

Proof By Lemma 8, x and y are distinct and each of them has at least one neighbor in each G_i. Suppose $xy \notin E(G)$. Let $G' = G + xy$ and $G'_i = \langle G_i, S\rangle_{G'}$. Then, for every i, G'_i is a minor of G since, if we pick a $j \ne i$ and in $\langle G_i, G_j, S\rangle_G$ contract G_j to x, we get a graph isomorphic to G'_i. So G'_i is nIL. Then, by Theorem 7, $G' = G'_1 \cup_{xy} \cdots \cup_{xy} G'_r$ is nIL, contradicting the assumption that G is maxnil. So $xy \in E(G)$.

For each i, we repeatedly add new edges to $\langle G_i, S\rangle_G$, if necessary, to get a maxnil graph H_i. Then $H := H_1 \cup_{xy} \cdots \cup_{xy} H_r$ is nIL and contains G as a subgraph, so $H = G$ and every $\langle G_i, S\rangle_G$ is maxnil. □

Lemma 10 *Let G_1 and G_2 be maxnil graphs. Pick an edge in each G_i and label it e. Then $G = G_1 \cup_e G_2$ is maxnil if and only if e is nontriangular in at least one G_i.*

Proof The graph G is nIL by Theorem 7. Suppose e is nontriangular in at least one G_i, say G_2. Denote the endpoints of e in G by x and y. To prove G is maxnil, it is enough to show that $G + b_1b_2$ is IL for all $b_i \in V(G_i) \setminus \{x, y\}$. By Lemma 8, G_1 is 2–connected, so each of x and y has at least one neighbor in G_1. So, if we contract G_1 to b_1 and then contract b_1b_2 to b_2, we obtain a graph G'_2 that contains G_2 as a proper subgraph, since b_2x and b_2y are both in G'_2, while e is nontriangular in G_2. So G'_2 is IL since G_2 is maxnil. But G'_2 is a minor of G, which is nIL, so we have a contradiction.

To prove the converse, suppose e is triangular in G_1 and G_2. Let $t_i \in V(G_i)$ be adjacent to both endpoints of e. Let K be a complete graph on four vertices, with vertices labeled x, y, t_1 and t_2. Denote the triangles induced by x, y and t_i in K and in G_i by Δ_i. Then, by Theorem 7, $G' := G_1 \cup_{\Delta_1} K \cup_{\Delta_2} G_2$ is nIL. But G' is isomorphic to $G + t_1t_2$, so G is not maxnil. □

Lemma 11 *Let G be a maxnil graph with vertex connectivity 3 and a vertex cut set $S = \{x, y, z\}$. Let $G_1, G_2, \dots, G_r$ denote the connected components of $G - S$. Then $\langle S\rangle_G \simeq K_3$ and $\langle G_i, S\rangle_G$ is maxnil for all $1 \le i \le r$.*

Proof Suppose $\langle S\rangle_G \not\simeq K_3$. Let G' be the graph obtained from G by adding one or more edges to $\langle S\rangle_G$ so that S induces a triangle T in G'. For $1 \le i \le r$, let $G'_i = \langle G_i, T\rangle_{G'}$. We see that G'_i is nIL as follows. Pick any $j \ne i$ and, in the graph $\langle G_i, G_j, S\rangle_G$, contract G_j to an arbitrary vertex v in G_j. Then v is connected to each of x, y and z since G is 3–connected and hence each of x, y and z has at least one neighbor in G_j. The graph M_i obtained this way is a minor of G, and hence is nIL. Performing a ∇Y–move on $T \subset G'_i$ we obtain a subgraph of M_i. Since M_i is nIL, so is G'_i. By Theorem 7, $G' = G'_1 \cup_T \cdots \cup_T G'_r$ is nIL, which contradicts the maximality of G. So $T = \langle S\rangle_G \simeq K_3$.

To show $\langle G_i, S\rangle_G$ is maxnil, repeatedly add new edges to it, if necessary, to get a maxnil graph H_i. Then $H := H_1 \cup_T \cdots \cup_T H_r$ is nIL by Theorem 7 and contains G as a subgraph, so $H = G$ and every $\langle G_i, S\rangle_G$ is maxnil. □

Let G be a graph and let $T = \langle x, y, z, t\rangle_G$ be an induced K_4 subgraph (*tetrahedral graph*). We say T is *strongly separating* if $G - T$ has at least two connected components C_1 and C_2 such that every vertex of T has a neighbor in each C_i.

Lemma 12 *Let G_1 and G_2 be maxnil graphs and let $G = G_1 \cup_\Delta G_2$ be the clique sum of G_1 and G_2 over a K_3 subgraph $\Delta = \langle x, y, z\rangle_G$. Assume Δ is a minimal vertex cut set in G. Then G is maxnil if and only if, for some $i \in \{1, 2\}$, every induced K_4 subgraph of the form $\langle x, y, z, t\rangle_{G_i}$ is strongly separating.*

Proof By Theorem 7, $G := G_1 \cup_\Delta G_2$ is nIL. Then G is maxnil if and only if, for every $t_1 \in V(G_1) \setminus V(\Delta)$ and $t_2 \in V(G_2) \setminus V(\Delta)$, the graph $G' := G + t_1t_2$ is IL.

First, suppose for some i at least one of x, y and z is not connected to t_i, say $xt_2 \notin E(G_2)$. Contracting $G_1 - \{y, z\}$ to x produces $G_2 + t_2x$ as a minor of G'. Since G_2 is maxnil, this minor is IL, and hence G' is IL, as desired. So we can assume $\langle x, y, z, t_i\rangle_{G_i}$ is a tetrahedral graph for both $i = 1, 2$.

Assume every tetrahedral graph in G_2 that contains Δ is strongly separating. So $G_2 - \langle x, y, z, t_2\rangle_{G_2}$ has at least two connected components each of which, when contracted to a single vertex, is adjacent to all four vertices x, y, z and t_2. In Figure 6,

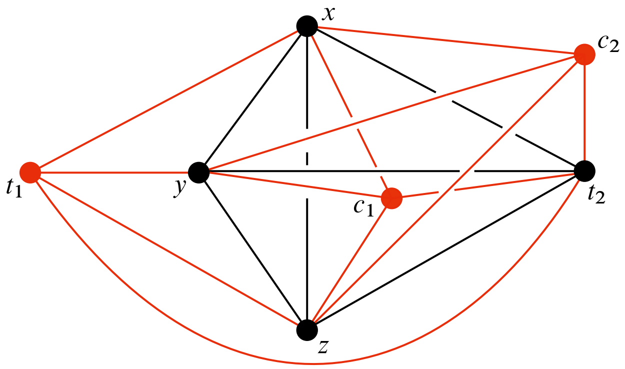


Figure 6: A K_7 minus a triangle minor of the graph G.

these vertices are denoted by c_1 and c_2. Now, if the component of $G_1 - \Delta$ that contains t_1 is contracted to t_1, this vertex too will be adjacent to x, y, z and t_2. So we get a minor of G' isomorphic to K_7 minus a triangle, which is IL since it contains a Petersen family graph (the one obtained by one ∇Y–move on K_6) as a minor. It follows that G' is IL, and therefore G is maxnil.

To prove the converse, for $i = 1, 2$ let t_i be a vertex in G_i such that $T_i := \langle x, y, z, t_i \rangle_{G_i}$ is a tetrahedral graph that is not strongly separating. Let $G' = G + t_1 t_2$. Then $G' = G_1 \cup_{T_1} \langle x, y, z, t_1, t_2 \rangle_{G'} \cup_{T_2} G_2$. Each of these clique sums is over a K_4, each summand is nIL, and each of T_1 and T_2 is nonstrongly separating; so, by Theorem 7, G' is nIL, and hence G is not maxnil. □

Unlike the vertex connectivity 2 and 3 cases, it is not true that a minimal vertex cut set in a 4–connected maxnil graph must be a clique. The four neighbors of b in the graph depicted in Figure 4, left, form a vertex cut set, but the graph induced by its vertices has exactly two edges. The four neighbors of any vertex in the graph $Q_{13,3}$ in Figure 5, left, form a discrete vertex cut set. However, if a maxnil graph G has vertex connectivity 4, the following lemma provides some restrictions on the shape of the subgraph induced by the vertices of any minimal vertex cut set:

Lemma 13 *Let G be a maxnil graph and assume $\{x, y, z, t\}$ is a minimal vertex cut. Let $S = \langle x, y, z, t \rangle_G$. Then S is either a clique or a subgraph of a 4–cycle.*

Proof Assume that S is neither a clique nor a subgraph of a 4–cycle. This implies that, if every vertex of S has degree less than 3, then S contains K_3 as a subgraph; and if S has a vertex of degree at least 3, then it contains $K_{1,3}$ as a subgraph. Below, we consider these two cases separately. In both cases, we use the fact that since $\{x,y,z,t\}$ is a minimal vertex cut set in G, each of x, y, z and t has at least one neighbor in each component of $G-S$.

Case 1 (S has a K_3 subgraph) We can assume that x, y and z induce a triangle in G. If $G-S$ has at least three connected components, contracting each of them to a single node would produce a minor of G which has a subgraph isomorphic to G_7, the graph in the Petersen family obtained by one ∇Y move on K_6. This contradicts the fact that G is nIL.

It follows that $G-S$ has at most two components, G_1 and G_2. For each $i=1,2$, contract $\langle G_i,t\rangle_G$ to t to produce a minor of G, denoted by G'_i, which must be nIL. Then $\{x,y,z,t\}$ induces a 4–clique K in both G'_1 and G'_2. By Theorem 7, the clique sum $G'=G'_1\cup_K G'_2$ is nIL since $G'-K$ has only two components and K has only four vertices. But G' strictly contains G as a subgraph; this implies G is not maxnil, a contradiction.

Case 2 (S has a $K_{1,3}$ subgraph) We can assume that t is adjacent to x, y and z in G. If $G-S$ has at least three connected components, contracting each of them to a single node would produce a minor of G containing a subgraph isomorphic to $K_{3,3,1}$; thus, G is IL. So $G-S=G_1\sqcup G_2$, with G_1 and G_2 connected. For $i=1,2$, contracting each of G_i to a single node t_i, deleting the edge t_it, deleting any existing edges of $\langle x,y,z\rangle_G$, and then performing a $Y\nabla$–move at t_i produces a nIL graph, denoted by G'_i. Let $G'=G'_1\cup_{K_4} G'_2$ be the clique sum over the complete graph with vertices x, y, z and t. By Theorem 7, G' is nIL since $G'-S=G_1\sqcup G_2$; but G' strictly contains G as a subgraph, a contradiction. □

Lemma 14 *Let $G=G_1\cup_S G_2$ be the clique sum of maxnil graphs G_1 and G_2 over $S=\langle x,y,z,t\rangle_G\simeq K_4$. Assume S is a minimal vertex cut set in G. Then G is maxnil if and only if, in both G_1 and G_2, S is not strongly separating.*

Proof If S is strongly separating in G_1 or G_2, then $G-S$ has at least three connected components and contracting each of them to a single node produces a minor isomorphic to K_7 minus a triangle.

If, in both G_1 and G_2, S is not strongly separating, then $G-S$ has only two connected components. Contracting each of the two components to a single node produces K_6

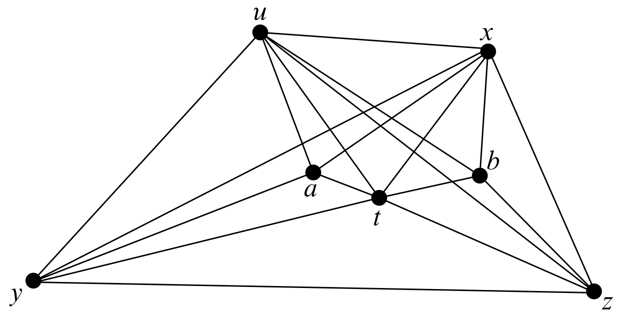


Figure 7: A maxnil graph that is a clique sum over K_5.

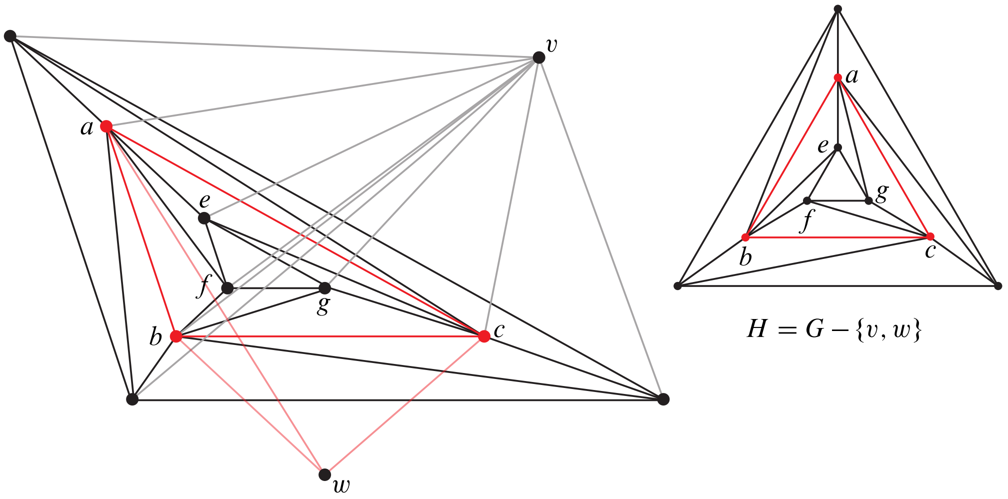
minus an edge as a minor (not K_7 minus a triangle); hence, G is nIL by Theorem 7. Adding an edge between a vertex in $G_1 - S$ and a vertex in $G_2 - S$ and contracting $G_1 - S$ and $G_2 - S$ to single nodes produces a K_6 minor. It follows that G is maxnil in this case. □

The graph G of Figure 7 is maxnil since $G - \{u\}$ is a maximal planar graph. If $S = \langle x, y, z, t, u\rangle$, $G_1 = \langle a, x, y, z, t, u\rangle$ and $G_2 = \langle b, x, y, z, t, u\rangle$, then $S \simeq K_5$, $G_1 \simeq G_2 \simeq K_6^-$ (K_6 minus one edge) and $G = G_1 \cup_S G_2$. This shows it is possible for the clique sum of two maxnil graphs over $S \simeq K_5$ to be nIL (and maxnil). However, no clique S of order 5 can be a minimal vertex cut set in a nIL graph G, since then any connected component of $G - S$ would form a K_6–minor together with S, which would imply G is IL. For $t \geq 6$, any clique sum over K_t is IL since K_6 is IL.

Jørgensen studied clique sums of graphs that are maximal without a K_6 minor [5]. These are graphs that do not contain a K_6 minor and a K_6 minor is created by the addition of any edge. The class of maxnil graphs and the class of graphs that are maximal without a K_6 minor are not the same, as shown in the following proposition:

Proposition 15 *The graph in Figure 8 is maxnil, and it is not maximal without a K_6 minor.*

Proof The graph G in Figure 8 is obtained by adding vertices v and w to the plane triangulation H: the vertex v connects to all nine vertices of H and the vertex w connects to the vertices a, b and c of H. The graph $H + v$ is maxnil since it is a cone over a maximal planar graph [11]. The graph G is the clique sum over $K_3 = \langle a, b, c\rangle_G$ of maxnil graphs $H + v$ and $K_4 = \langle a, b, c, w\rangle_G$. The graph $\langle a, b, c, v\rangle_{H+v}$ is the only induced K_4 subgraph in $H + v$ containing a, b and c and it is strongly separating



$H = G - \{v, w\}$

Figure 8: A maxnil graph G (left) that is not maximal without a K_6 minor is obtained by adding two vertices to a plane triangulation with nine vertices (right).

in $H+v$. So, by Lemma 12, G is maxnil; in particular, it has no K_6 minor. The graph $G+vw$ is a clique sum over $K_4 = \langle a,b,c,v \rangle_G$ of graphs $H+v$ and $K_5 = \langle a,b,c,v,w \rangle$, both of which are K_6 minor free. Hence, by [5], $G+vw$ is K_6 minor free, so G is not maximal without a K_6 minor. The graph $G+vw$ has order 11 and size 34, so it is maximal without a K_6 minor by Mader's result [7], since $34 = 4 \times 11 - 10$. □

Remark 16 Starting with the graph G in Proposition 15, one can construct graphs G_n with $n \geq 11$ vertices that are maxnil but not maximal without a K_6 minor. Take $G_{11} = G$ and construct G_{11+k} from G by triangulating the disk bounded by the triangle efg with k new vertices, and then adding edges between v and these new vertices. The argument used in the proof of Proposition 15 shows that G_n for $n \geq 11$ is maxnil but not maximal without a K_6 minor. Furthermore, $n = 11$ is the minimal order of a graph with this property, ie every maxnil graph with $n \leq 10$, vertices is maximal without a K_6 minor. We used Mathematica to generate all 136 maxnil graphs of orders between 6 and 10 and we confirmed that all of them are maximal without a K_6 minor.

References

[1] **M Aires**, *On the number of edges in maximally linkless graphs*, J. Graph Theory 98 (2021) 383–388 MR

[2] **J H Conway**, **C M Gordon**, *Knots and links in spatial graphs*, J. Graph Theory 7 (1983) 445–453 MR Zbl

[3] **H R Dehkordi**, **G Farr**, *Non-separating planar graphs*, Electron. J. Combin. 28 (2021) art. id. 1.11 MR Zbl

[4] **H van der Holst**, **L Lovász**, **A Schrijver**, *The Colin de Verdière graph parameter*, from "Graph theory and combinatorial biology" (L Lovász, A Gyárfás, G Katona, A Recski, L Székely, editors), Bolyai Soc. Math. Stud. 7, János Bolyai Math. Soc., Budapest (1999) 29–85 MR Zbl

[5] **L K Jørgensen**, *Some maximal graphs that are not contractible to K_6*, Aalborg Universitetscenter Institut for Elektroniske Systemer (1989) art.id. 1989:R89–28

[6] **L Lovász**, **A Schrijver**, *A Borsuk theorem for antipodal links and a spectral characterization of linklessly embeddable graphs*, Proc. Amer. Math. Soc. 126 (1998) 1275–1285 MR Zbl

[7] **W Mader**, *Homomorphiesätze für Graphen*, Math. Ann. 178 (1968) 154–168 MR Zbl

[8] **J Maharry**, *A splitter for graphs with no Petersen family minor*, J. Combin. Theory Ser. B 72 (1998) 136–139 MR Zbl

[9] **N Robertson**, **P D Seymour**, **R Thomas**, *Linkless embeddings of graphs in* 3*–space*, Bull. Amer. Math. Soc. 28 (1993) 84–89 MR Zbl

[10] **N Robertson**, **P D Seymour**, **R Thomas**, *A survey of linkless embeddings*, from "Graph structure theory" (N Robertson, P Seymour, editors), Contemp. Math. 147, Amer. Math. Soc., Providence, RI (1993) 125–136 MR Zbl

[11] **H Sachs**, *On spatial representations of finite graphs*, from "Finite and infinite sets, II" (A Hajnal, L Lovász, V T Sós, editors), Colloq. Math. Soc. János Bolyai 37, North-Holland, Amsterdam (1984) 649–662 MR Zbl

Department of Mathematics, Occidental College
Los Angeles, CA, United States

Mathematics and Statistics Department, University of South Alabama
Mobile, AL, United States

Mathematics and Statistics Department, University of South Alabama
Mobile, AL, United States

`rnaimi@oxy.edu, andreipavelescu@southalabama.edu,`
`elenapavelescu@southalabama.edu`

Received: 19 September 2020 Revised: 28 December 2021

Algebraic & Geometric Topology **23**:6 (2023) 2561–2591
DOI: 10.2140/agt.2023.23.2561
Published: 7 September 2023

Legendrian large cables and new phenomenon for nonuniformly thick knots

ANDREW MCCULLOUGH

We define the notion of a knot type having Legendrian large cables and show that having this property implies that the knot type is not uniformly thick. We then show that there exists an infinite family of ribbon knots that have Legendrian large cables. These knots fail to be uniformly thick in several ways not previously seen. We also give a general construction of ribbon knots, and show when they give similar such examples.

57K10, 57K33, 57R65

1 Introduction

The *contact width* $w(K)$ of a knot $K \subset (S^3, \xi_{\mathrm{std}})$ was defined by Etnyre and Honda in [4] as follows.[1] An embedding $\phi\colon S^1 \times D^2 \hookrightarrow S^3$ is said to represent K if the core curve of $\phi(S^1 \times D^2)$ is isotopic to K. (To simplify notation, we will not distinguish between $S^1 \times D^2$ and its image under ϕ.) Define the *slope* of homotopically nontrivial curves on $\partial(S^1 \times D^2)$ by identifying $\partial(S^1 \times D^2) \simeq \mathbb{R}^2/\mathbb{Z}^2$, where the meridian has slope ∞ and the longitude (which is well defined since K is inside S^3) has slope 0. Now define the *contact width* $w(K)$ as

$$w(K) = \sup \operatorname{slope}(\Gamma_{\partial(S^1 \times D^2)})$$

where the supremum is taken over $S^1 \times D^2 \hookrightarrow S^3$ representing K with $\partial(S^1 \times D^2)$ convex.

Etnyre and Honda [4] also defined K to have the *uniform thickness property* if

(1) any solid torus representing the knot type $\mathcal{K}$ can be thickened to a *standard neighborhood* of a Legendrian representative of K, and

[1] These definitions are slightly different than those originally made in [4] since we are using a different slope convention in this paper; see Remark 1.1.

(2) $w(K)$ is equal to the maximal Thurston–Bennequin invariant $\overline{\mathrm{tb}}(K)$ of Legendrian representatives of K.

Here a *standard neighborhood* $N(L)$ of a Legendrian knot L is an embedded solid torus $N(L)$ representing L with convex boundary ∂N such that $\#\Gamma_{\partial N} = 2$ and $\mathrm{tb}(L) = \mathrm{slope}(\Gamma_{\partial N})$. Standard neighborhoods are contact isotopic to any sufficiently small tubular neighborhood N of L with convex boundary and $\#\Gamma_{\partial N} = 2$.

The usefulness of this property became evident when Etnyre and Honda showed in the same work that if $L \subset S^3$ is Legendrian simple and uniformly thick, then cables of L are Legendrian simple as well. Recall that a knot type is Legendrian simple if Legendrian knots in this knot type are completely determined (up to Legendrian isotopy) by their Thurston–Bennequin invariant and rotation number. They also showed that if the cables are sufficiently negative, then they too satisfy the uniform thickness property. This allows that certain iterated cables of Legendrian simple knots are Legendrian simple, for example.

Uniform thickness has become a key hypothesis in work since then. For example, generalizing the above work on cables, Etnyre and Vértesi [6] showed that given a companion knot $L \subset S^3$ which is both Legendrian simple and uniformly thick, and a pattern $P \subset S^1 \times D^2$ satisfying certain symmetry hypothesis, the knots in the satellite knot type P_K may be understood.

Broadly, if one wants to classify Legendrian knots in a satellite knot type with companion knot $K \subset S^3$, and a pattern $P \subset S^1 \times D^2$, then as a first step one needs to understand

(1) contact structures on the complement of a neighborhood N of K,

(2) contact structures on a neighborhood N of K, and

(3) a classification of Legendrian knots in the knot type of the pattern P in the possible contact structures on N.

If K is uniformly thick, then N can always be taken to be a standard neighborhood of K with dividing curves on the boundary of slope $\overline{\mathrm{tb}}(K)$ (ie maximal Thurston–Bennequin invariant of K), which reduces the problem to items (1) and (3) above. Moreover, if K is Legendrian simple and uniformly thick, then (1) is more or less known as well [4]. If K is not uniformly thick, then understanding satellites is much more complicated.

Similarly, uniform thickness can be useful in understanding contact surgery constructions. A typical way to obtain a new contact 3–manifold is by removing a solid torus in the knot type K, and gluing in some new contact solid torus. To understand the new manifold, one needs to understand items (1) and (2) above, and the gluing map defining the surgery. If K is uniformly thick, then N can always be taken to be a standard neighborhood of K with dividing curves on the boundary of slope $\overline{\mathrm{tb}}(K)$, which simplifies (1) and (2) considerably.

On the other hand, there are knot types that are not uniformly thick. For such knot types, it is important to understand in what ways they can fail to be uniformly thick.

1.1 New phenomenon for nonuniformly thick knots

Given a knot type $\mathcal{K} \subset S^3$, the *contact width* of $\mathcal{K}$ is

$$w(\mathcal{K}) = \sup\{\mathrm{slope}(\Gamma_{\partial N}) \mid N \text{ is a solid torus representing } \mathcal{K} \text{ with convex boundary}\}.$$

We say a solid torus represents $\mathcal{K}$ if its core is in the knot type of $\mathcal{K}$. The contact width satisfies the inequality $\overline{\mathrm{tb}}(\mathcal{K}) \le w(\mathcal{K}) \le \overline{\mathrm{tb}}(\mathcal{K}) + 1$; see [4].

Remark 1.1 A word about slope conventions: If μ and λ are the meridional and longitudinal curves, respectively, on a torus T then $[\lambda]$ and $[\mu]$ form a basis for $H_1(T)$. A (p,q) curve, or a curve of slope q/p, will refer to any simple closed curve in T that is in the homology class of $p[\lambda] + q[\mu]$, where $p, q \in \mathbb{Z}$ are relatively prime. This is the opposite convention to the one used in several of the main references in this paper, which were some of the first works in convex surface theory. However, it is the convention that is standard in low-dimensional topology. We caution however that, when the phrase "integer slope" is used, it would correspond to the phrase "one over integer slope" in Etnyre and Honda [3; 4; 10] among others.

We are now in position to define uniform thickness. We say that a knot type $\mathcal{K}$ has the *uniform thickness property* or is *uniformly thick* if

(1) $\overline{\mathrm{tb}}(\mathcal{K}) = w(\mathcal{K})$, and

(2) every solid torus representing $\mathcal{K}$ can be thickened to a standard neighborhood of a maximal tb representative of $\mathcal{K}$.

By a standard neighborhood of a Legendrian knot L, we mean a solid torus neighborhood N of L with convex boundary, and dividing set $\Gamma_{\partial N}$ consisting of two curves with slope $\mathrm{tb}(L)$.

In past work, it is shown that a knot type $\mathcal{K}$ can fail to have the uniform thickness property in two ways. It can have neighborhoods whose slopes are larger than $\overline{\mathrm{tb}}$, as is the case with the unknot U which has $\overline{\mathrm{tb}}(U) = -1$ and $w(U) = 0$. It can also happen that there are neighborhoods with slope strictly less than $\overline{\mathrm{tb}}$, but that do not thicken.

The first and only such examples are due to [4] and Etnyre, LaFountain and Tosun [5] where it is shown that all positive torus knots $T_{p,q}$ have tori N with slopes satisfying $\mathrm{slope}(\Gamma_{\partial N}) < \overline{\mathrm{tb}}(T_{p,q})$ but that do not thicken. Moreover, the contact structure on all of these N is universally tight.

In what follows we will denote the set of Legendrian knots, up to isotopy, in the same topological knot type as K by $\mathcal{L}(K)$. We also use the convention that for a pair of relatively prime integers p and q, the (p,q) cable of K, that is, the knot type of a curve of slope q/p on the boundary of a torus neighborhood of K, is denoted by $K_{p,q}$. Notice that if $p = \pm 1$, then $K_{p,q}$ is a trivial cable in the sense that it is isotopic to the underlying knot K. The following theorem of Etnyre and Honda motivates us to define some new terminology.

Theorem 1.2 (Etnyre and Honda [4]) *If $K \subset S^3$ satisfies the uniform thickness property, then for $|p| > 1$ and any $L \in \mathcal{L}(K_{p,q})$ we have $\mathrm{tb}(L) \leq pq$.*

We generalize this result in Lemma 3.3 below. Notice that if we have a uniformly thick knot K and we fix a Legendrian representative $L \in \mathcal{L}(K)$ with $\mathrm{tb}(L) = k$, then there is an isotopy of K which arranges that L is a trivial cable $L = K_{1,k-1}$. But then we have that $\mathrm{tb}(K_{1,k-1}) = \mathrm{tb}(L) = k \nleq k-1$, so the inequality in Theorem 1.2 is not satisfied.

Definition 1.3 Given $|p| > 1$, we will say that a Legendrian cable $L \in \mathcal{L}(K_{p,q})$ is *large* if $\mathrm{tb}(L) > pq$, and call $K_{p,q}$ *Legendrian large* if there exists large $L \in \mathcal{L}(K_{p,q})$. We will then say that K has *Legendrian large cables*, or has the *Legendrian large cable* (*LLC*) *property*, if any of its nontrivial cables are Legendrian large.

Notice the example above indicates that if we allowed trivial cables, the LLC property would be vacuous. Our main theorem relates the LLC property to uniform thickness.

Theorem 1.4 *If K has Legendrian large cables, then there exist solid tori $V = S^1 \times D^2$ representing K such that $\xi|_V$ is virtually overtwisted. Moreover, V cannot be thickened to a standard neighborhood of a Legendrian knot, and so K is not uniformly thick.*

Recall that the term *universally tight* refers to a contact structure that is tight, and that, when lifted to the universal cover, remains tight. If the lift becomes overtwisted, then we will refer to the contact structure as *virtually overtwisted.*

Theorem 1.5 *Given K, if there exists a slope $q/p > \overline{\mathrm{tb}}(K)$ with $|p| > 1$ such that $K_{p,q}$ is Legendrian large, then $w(K) > \overline{\mathrm{tb}}(K)$.*

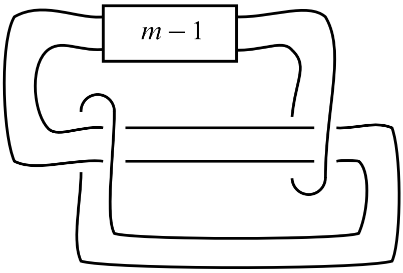


Figure 1: The ribbon knots K^m. There are $m-1$ right-handed full twists.

Question 1 Are there knots K and slopes $q/p < \overline{\mathrm{tb}}(K)$ such that $K_{p,q}$ is Legendrian large?

Question 2 If ξ is a virtually overtwisted contact structure on $S = S^1 \times D^2$, for which p and q is there a Legendrian (p,q) knot L in S with $\mathrm{tw}(L) > pq$?

In [15], Yasui gave some interesting examples of ribbon knots, which we will denote by K^m, shown in Figure 1. Yasui [15] shows that these knots have $\overline{\mathrm{tb}}(K^m) = -1$. In what follows, we will be concerned with integers $m < 0$. Building on his work, we observe that $K^m_{(-n,1)}$ is Legendrian large whenever $m \leq -5$ and $1 < n \leq \lfloor \frac{1}{4}(3-m) \rfloor$. This leads to the following theorem.

Theorem 1.6 *The knots K^m in Figure 1, with $m \leq -5$, are not uniformly thick in $(S^3, \xi_{\mathrm{std}})$. In particular, there are solid tori T representing K^m such that*

$$\mathrm{slope}(\Gamma_{\partial T}) > \overline{\mathrm{tb}}(K^m)$$

and $\xi|_T$ is tight, but virtually overtwisted.

Remark 1.7 Previously, there were no known examples of $\mathcal{K}$ in $(S^3, \xi_{\mathrm{std}})$ with $w(\mathcal{K}) > \overline{\mathrm{tb}}(\mathcal{K})$ except for the unknot. These are also the first examples of solid tori in $(S^3, \xi_{\mathrm{std}})$ with virtually overtwisted contact structures.

It would be interesting to know what $w(K^m)$ is, and what the possible nonthickenable tori in the knot type of K^m are. We have the following partial result, following from Theorem 1.6 and its proof.

Proposition 1.8 *For $m \leq -5$, the knots K^m in Figure 1 have*

$$w(K^m) \geq \frac{-1}{\lfloor \frac{1}{4}(3-m) \rfloor}.$$

The origin of the examples in Theorem 1.6 come from an interesting connection between contact structures and the famous cabling conjecture first observed by Lidman and Sivek in [12] where it is shown that for a knot K with $\overline{\mathrm{tb}}(K) > 0$, Legendrian surgery on K — ie $(\mathrm{tb}(K)-1)$–surgery — never yields a reducible manifold. They conjectured that this might be true with no condition on $\overline{\mathrm{tb}}(K)$. This is equivalent to the following conjecture for any K in S^3.

Conjecture 1.9 *For a Legendrian representative in the knot type $L \in \mathcal{L}(K_{p,q})$, we have $\mathrm{tb}(L) \leq pq$.*

If $\mathrm{tb}(L) > pq$ for such an L, then there exists L' with $\mathrm{tb}(L') = pq + 1$ (we can always stabilize to achieve this). Legendrian surgery on this L' would then yield a reducible manifold; see Moser [14].

Theorem 1.10 (Yasui [15]) *There exist infinitely many Legendrian knots in $(S^3, \xi_{\mathrm{std}})$* (see Figure 1), *each of which yields a reducible 3–manifold by a Legendrian surgery in the standard tight contact structure. Furthermore, K can be chosen so that the surgery coefficient is arbitrarily less than $\overline{\mathrm{tb}}(K)$.*

Yasui shows that for infinitely many pairs of integers $m, n \in \mathbb{Z}$ with $m \leq -5$, Legendrian surgery on the cables $K^m_{n,-1}$ yields a reducible manifold. This shows Lidman and Sivek's conjecture to be false, and stands in contrast with Theorem 1.2 of Etnyre and Honda.

We can now easily see that K^m — see Figure 1 — does not have the uniform thickness property. The interesting features of how K^m fails to be uniformly thick, given in Theorem 1.6, require much more work.

In [15], Yasui shows that for integers $n \leq \frac{1}{4}(3-m)$, the cables $K^m_{n,-1}$ have the property that $\overline{\mathrm{tb}}(K^m_{n,-1}) = -1$. But by Theorem 1.2, if K^m is uniformly thick, then we must have that $\overline{\mathrm{tb}}(K^m_{n,-1}) \leq -n$. So, for any $m \leq -5$ and any $1 < n \leq \lfloor \frac{1}{4}(3-m) \rfloor$, we arrive at a contradiction. This addresses the first assertion of Theorem 1.6.

Theorem 1.6 can be used to address the following conjecture.

Conjecture 1.11 *If $K \subset S^3$ is fibered, then K is uniformly thick if and only if $\xi_K \neq \xi_{\mathrm{std}}$, where ξ_K is the contact structure induced by an open book decomposition of K.*

Building on our above work, Hyunki Min [13] recognized that the K^m are counterexamples. Min showed that the K^m are all fibered. We also know that K^m are slice (since they are ribbon knots) and not strongly quasipositive (since they are obtained

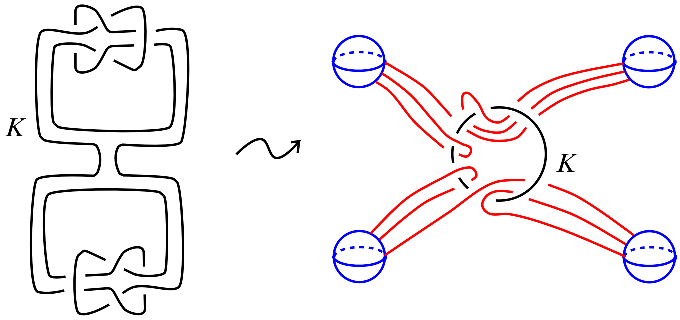


Figure 2: An example ribbon knot before running the algorithm in Theorem 1.12 (left), and after running the algorithm (right).

by attaching negative bands to two parallel disks), which implies that $\xi_K \neq \xi_{\mathrm{std}}$ by a result of Matthew Hedden [9, Proposition 2.1]. Theorem 1.6 tells us that the K^m are not uniformly thick however, so at least one direction of this conjecture is false. The other direction remains an interesting open question.

1.2 Ribbon knots and Legendrian large cable examples

Yasui's examples are all ribbon knots with Legendrian large cables, and can be generalized to other families of ribbon knots. We first observe a folk result that any ribbon knot can be described in a simple way.

Theorem 1.12 *Suppose $K \subset S^3$ is an arbitrary ribbon knot with $n \in \mathbb{N}$ ribbon singularities. Then there is an algorithm to construct a 2–handlebody for D^4 having $n-1$ or fewer 1–2–handle-canceling pairs such that there is an unknot U in the boundary of the 1–subhandlebody which, after attaching the 2–handles, is isotopic to K.*

A representation of a ribbon knot K as in Theorem 1.12 will be called a *handlebody picture for K*. The proof of Theorem 1.12 will be given in Section 3. Figure 2 gives an example ribbon knot and its image after running the algorithm.

Theorem 1.13 *Given an arbitrary ribbon knot K, we can associate to it a handlebody picture. If it is possible to Legendrian realize the attaching circles of the 2–handles so that the handle attachments are Stein (ie framings are all* $\mathrm{tb}-1$*), and also Legendrian realize K so that* $\mathrm{tb}(K) = -1$*, then K is a Legendrian ribbon knot that bounds a Lagrangian disk in $(B^4, \omega_{\mathrm{std}})$.*

Proof Given a handlebody picture for K, there is an unknot U in the boundary of the 1–subhandlebody which, by hypothesis, can be realized with $\mathrm{tb}(U) = -1$. Such

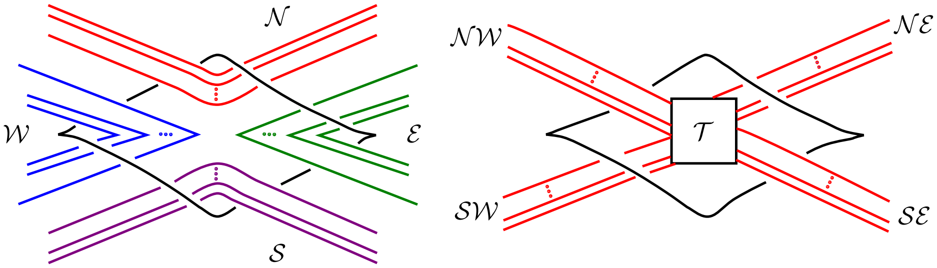


Figure 3: Possible examples of knots with Legendrian large cables. The ellipses are meant to indicate a finite number of strands bundled as shown, while $\mathcal{T}$ is an arbitrary Legendrian tangle.

an unknot bounds a Lagrangian disk in the 1–subhandlebody. Since the 2–handles are attached disjointly from this disk, K bounds a Lagrangian disk after they are attached, that is, K bounds a Lagrangian disk in $(B^4, \omega_{\text{std}})$. □

Conway, Etnyre and Tosun [1] make use of this fact to describe when contact surgery on a knot in (S^3, ξ_{std}) preserves symplectic fillability.

Corollary 1.14 *Given an arbitrary ribbon knot K, we can associate to it a handlebody picture. If it is possible to Legendrian realize the attaching circles of the 2–handles so that the handle attachments are Stein, Legendrian realize K so that $\text{tb}(K) = -1$, and also arrange the local picture of K to be as in Figure 3, left, then K has Legendrian large cables.*

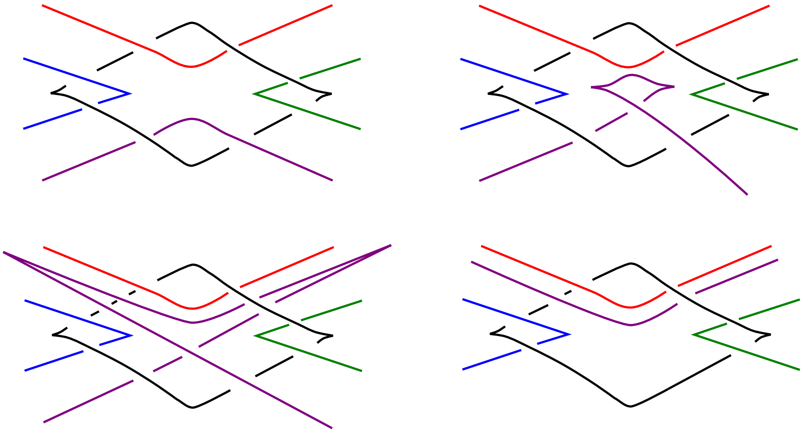


Figure 4: The steps in a Legendrian isotopy to change strands of type $\mathcal{S}$ into strands of type $\mathcal{N}$.

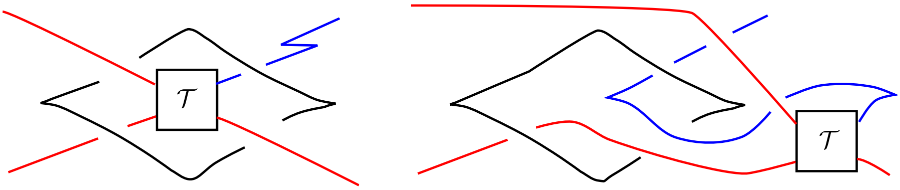


Figure 5: A Legendrian isotopy of the tangle $\mathcal{T}$. In this example, strands of type $\mathcal{NE}$ are assumed to have stabilizations.

Proof The proof is exactly the same as that of Yasui's Theorem 1.3 [15, pages 7–13], when there are only strands of type $\mathcal{N}$, since everything in the arguments can be done locally. The rest of the cases follow by Legendrian isotopy of Figure 3, left. For example, we can change all strands of type $\mathcal{S}$ into strands of type $\mathcal{N}$ by the Legendrian isotopy shown in Figure 4. We can also change all strands of types $\mathcal{E}$ and $\mathcal{W}$ into strands of type $\mathcal{N}$ by even easier isotopies. □

Remark 1.15 If the framings of the 2–handles allow stabilizations, then there are more examples. Given an arbitrary ribbon knot K, we can associate to it a handlebody picture. If it is possible to Legendrian realize the attaching circles of the 2–handles so that the handle attachments are Stein, Legendrian realize K so that $\operatorname{tb}(K) = -1$, arrange the local picture of K to be as in Figure 3, left, and arrange that there is a stabilization on each of the strands of at least one group of strands $\mathcal{NE}$, $\mathcal{NW}$, $\mathcal{SE}$, or $\mathcal{SW}$, then K has Legendrian large cables. This is true since we can isotope the stabilizations to have the form of Figure 5, left, Legendrian isotope the tangle $\mathcal{T}$ off to the side as shown in Figure 5, right, and then apply Corollary 1.14.

Acknowledgements The author would like to express profound gratitude to his advisor John B Etnyre for his patience, encouragement, and many helpful comments and suggestions, without which this paper would not have been possible. He would also like to extend thanks to James Conway for making him aware of Conjecture 1.11, and to Sudipta Kolay, Hyunki Min, Surena Hozoori and Peter Lambert-Cole for many useful and productive conversations.

2 Background

We will assume that the reader is familiar with Legendrian knots and basic convex surface theory. Some excellent sources for this material are [3; 7; 10; 11]. We will need to understand the twisting of a contact structure along a Legendrian curve with respect to two different framings. Suppose we are given a solid torus $S \subset (S^3, \xi)$ with convex boundary which represents the knot K. This just means that $S = D^2 \times S^1$ and $K = \{\mathrm{pt}\} \times S^1$ for some point in $\mathrm{int}(D^2)$. Further suppose that we are given a Legendrian (p,q) curve L in S. Since L is null homologous in S^3, there is a well defined framing on L given by any Seifert surface Σ, and measuring the twisting of ξ along L with respect to this framing gives us $\mathrm{tw}(L;\Sigma) = \mathrm{tb}(L)$, that is, the Thurston–Bennequin invariant of L. We can also find a boundary parallel torus $T^2 \subset S$ containing L, and measure the twisting of ξ along L with respect to the framing coming from T^2. We will denote this twisting by $\mathrm{tw}(L;\partial S)$. The relationship between these twistings is given by the expression

$$\mathrm{tw}(L;\partial S) + pq = \mathrm{tb}(L);$$

see [4].

Consider a contact structure ξ on $T^2 \times I$ with convex boundary, let T_1 and T_2 be its two torus boundary components, and assume without loss of generality that

$$s_1 = \mathrm{slope}(\Gamma_{T_1}) \le \mathrm{slope}(\Gamma_{T_2}) = s_2,$$

where Γ_S denotes the dividing curves on a convex surface S. Then we will say that ξ is *minimally twisting* if every convex, boundary parallel torus $S \subset T^2 \times I$ has $s_1 \le \mathrm{slope}(\Gamma_S) \le s_2$. This is the same notion of minimal twisting that Honda defined in [10]. We will also need to make use of his basic slices to decompose $T^2 \times I$ into layers. Using the same notation as above, we will call $(T^2 \times I, \xi)$ a *basic slice* if

(1) ξ is tight, and minimally twisting;

(2) T_i are convex and $\#\Gamma_{T_i} = 2$;

(3) s_i form an integral basis for $\mathbb{Z}^2$.

Honda showed that, up to isotopy fixing the boundary, there are exactly two tight contact structures on a basic slice, distinguished by their relative Euler classes in $H^2(T^2 \times I, \partial(T^2 \times I); \mathbb{Z})$.

The Farey tessellation, Figure 6, gives a convenient way to describe curves on T^2.

To construct the eastern half of the Farey tessellation, first label the north pole by $0 = \frac{0}{1}$, the south pole by $\infty = \frac{1}{0}$, and connect them by an edge (by edge, we mean a hyperbolic geodesic). Next, label the eastern most point that is midway between 0 and ∞ by $1 = \frac{1}{1}$, as shown in Figure 6. Connect 1 by edges to 0 and ∞. For rational numbers on the tessellation with the same sign, we can define an addition on the Farey tessellation by $a/b + c/d = (a+c)/(b+d)$, locate $(a+c)/(b+d)$ midway between a/b and

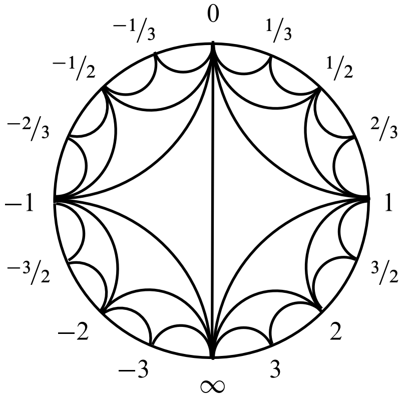


Figure 6: Farey tessellation.

c/d, and connect $(a+c)/(b+d)$ by edges with a/b and c/d respectively. Thus we can fill in the rest of the positive side of the Farey tessellation by iterating this addition. Notice that, if a/b and c/d are assumed to be an integral basis for $\mathbb{Z}^2$, then both

$$\begin{vmatrix} a & a+c \\ b & b+d \end{vmatrix} = ad-bc = \begin{vmatrix} a & c \\ b & d \end{vmatrix} = \pm 1$$

and, similarly,

$$\begin{vmatrix} a+c & c \\ b+d & d \end{vmatrix} = \pm 1,$$

so any two points connected by an edge are an integral basis for $\mathbb{Z}^2$. Also notice that, given two positive rational numbers $a/b > c/d$, there are exactly two other points with edges to both a/b and c/d, namely $(a+c)/(b+d)$ and $(a-c)/(b-d)$.

To construct the western (negative) half of the Farey tessellation, first relabel the north pole by $0 = \frac{0}{-1}$. Next, label the western most point that is midway between 0 and ∞ by $-1 = \frac{1}{-1}$, as shown in Figure 6. Connect -1 by edges to 0 and ∞. Now using the same addition we defined above, we can iteratively build up the negative side of our Farey tessellation. Notice that the only point which was labeled twice was the north pole, which is now given by $\frac{0}{\pm 1}$.

For any two points p_1 and p_2 on the Farey tessellation, we define the interval $[p_1, p_2]$ to be the set of all points encountered starting from p_1 and moving clockwise around the tessellation until reaching p_2. Given a clockwise sequence of three points connected by edges p_1, p_2 and p_3 on the Farey tessellation, we say that a jump from p_2 to p_3 is *half maximal* if p_3 is the half way point of the maximum possible clockwise jump one could make in the interval (p_2, p_1). We will consider only clockwise paths in the

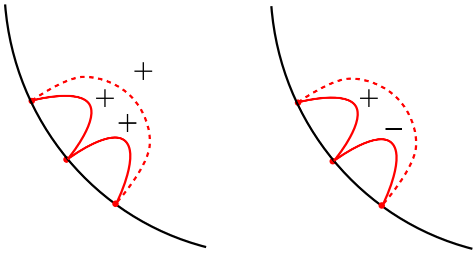


Figure 7: Left: a consistent shortening. Right: a shortening which is not consistent.

Farey tessellation, where a path is a sequence of jumps along edges. We call a path between two points $s_1, s_2 \in \mathbb{Q}$ a *continued fraction block* if, after the first jump, every jump is half maximal. Notice that, by construction, a path that is a continued fraction block cannot be shortened. We will also need to consider decorated paths (ie paths for which each jump gets a "$+$" or "$-$"). We can define an equivalence relation "$\sim$" on decorated paths in the Farey tessellation which says that any two paths with the same endpoints and which differ only by shuffling of signs within continued fraction blocks are in the same class. The following result, due to Honda [10], and in a different terminology Giroux [7], describes a relationship between contact structures on $T^2 \times I$ and minimal decorated paths in the Farey tessellation. Given a manifold M and a multicurve Γ in ∂M, let $\mathrm{Tight}(M, \Gamma)$ denote the set of isotopy classes of tight contact structures on M with convex boundary such that Γ is a set of dividing curves for ∂M. Similarly, given $T^2 \times I$ with boundary $T_1 \sqcup T_2$, and two multicurves Γ_i on T_i, let $\mathrm{Tight}(T^2 \times I, T_1 \cup T_2)$ denote the set of tight, minimally twisting contact structures on $T^2 \times I$ with convex boundary such that Γ_i is a set of dividing curves for T_i.

Theorem 2.1 (Honda [10]) *Given $T^2 \times I$ with boundary $T_1 \sqcup T_2$, and two multicurves Γ_i on T_i with $\#\Gamma_i = 2$ such that $s_1 = \mathrm{slope}(\Gamma_1) \leq \mathrm{slope}(\Gamma_2) = s_2$, there is a one-to-one correspondence*

$$\mathrm{Tight}(T^2 \times I, \Gamma_1 \cup \Gamma_2) \leftrightarrow \{\textit{minimal decorated paths from } s_1 \textit{ to } s_2\}/\sim.$$

Given $T^2 \times I$ with a two-component multicurve on each of its two torus boundary components, and with boundary slopes $s_1, s_2 \in \mathbb{Q}$, then any decorated path starting from s_1 and ending at s_2 describes a contact structure on $T^2 \times I$. Each jump in the path describes a basic slice, and therefore has two possible contact structures distinguished by the relative Euler class. We then get $T^2 \times I$ by concatenating these basic slices. For more details, see [10]. It follows from Theorem 2.1 that within any continued fraction block, shuffling the signs of the jumps results in isotopic contact structures.

Suppose we have a decorated path which can be shortened; see Figure 7. It follows from Honda's gluing theorem that if the two jumps which are being combined into a single jump have different signs, then the contact structure on $T^2 \times I$ described by this path is overtwisted. If the signs agree, then the contact structure will be tight. For this reason, we say that a shortening is *consistent* if the signs of the smaller jumps agree, and make the following theorem owing to Honda.

Theorem 2.2 *Given a decorated path in the Farey tessellation from s_1 to s_2, the contact structure on $T^2 \times I$ with convex boundary $T_1 \sqcup T_2$, $\#\Gamma_{T_i} = 2$, $s_1 = \operatorname{slope}(\Gamma_{T_1})$ and $\operatorname{slope}(\Gamma_{T_2}) = s_2$ described by this path is tight if and only if every shortening is consistent.*

To classify the tight contact structures on solid tori, we will consider a slightly different type of path. Let a *truncated path* be a decorated path, as defined above, with the sign of the first jump omitted from consideration. In other words, the first jump is not decorated. Suppose we have $S^1 \times D^2$ with a two-component multicurve on its torus boundary, and with boundary slope $s_2 \in \mathbb{Q}$. If the meridian of $\partial(S^1 \times D^2)$ has slope $s_1 \in \mathbb{Q}$, then we have the following classification. Given $S^1 \times D^2$ with boundary T, and a multicurve Γ on T, let $\operatorname{Tight}(S^1 \times D^2, \Gamma)$ denote the set of isotopy classes of tight, minimally twisting contact structures on $S^1 \times D^2$ with convex boundary, such that Γ is a set of dividing curves for T.

Theorem 2.3 (Honda [10]) *Given $S^1 \times D^2$ with boundary T, and a multicurve Γ on T with $\#\Gamma = 2$ such that $s_2 = \operatorname{slope}(\Gamma)$ and $s_1 = \operatorname{slope}(\mu)$, where μ is a meridional curve for T,*

$$\operatorname{Tight}(S^1 \times D^2, \Gamma) \leftrightarrow \{\text{minimal truncated paths from } s_1 \text{ to } s_2\}/\sim.$$

Theorem 2.4 (Honda [10]) (1) *Given $T^2 \times I$ with boundary $T_1 \sqcup T_2$, and two multicurves Γ_i on T_i with $\#\Gamma_i = 2$ such that $s_1 = \operatorname{slope}(\Gamma_1) \le \operatorname{slope}(\Gamma_2) = s_2$, there are exactly two tight contact structures on $T^2 \times I$, and these contact structures are universally tight. The paths describing these two structures are the same, one decorated entirely by "$+$", and the other decorated entirely by "$-$".*

(2) *Given $S^1 \times D^2$ with boundary T, and a multicurve Γ on T with $\#\Gamma = 2$ such that $s_2 = \operatorname{slope}(\Gamma)$, and $s_1 = \operatorname{slope}(\mu)$, where μ is a meridional curve for T, then, if $s_1 \cdot s_2 \neq \pm 1$, there are exactly two tight contact structures on $S^1 \times D^2$, and these contact structures are universally tight. The paths describing these two structures are the same, one decorated entirely by "$+$", and the other decorated entirely by "$-$". If*

$s_1 \cdot s_2 = \pm 1$, *then there exists a unique tight contact structure on* $S^1 \times D^2$, *and this contact structure is universally tight.*

It follows from Theorem 2.4 that if we have a path with a mixture of signs, then the contact structure described by this path on either $T^2 \times I$, or on $S^1 \times D^2$, must be virtually overtwisted.

3 Cables in solid tori

In this section, we will give the proof of Theorems 1.4, 1.5 and 1.6 and Proposition 1.8. We would like to record and make use of the following result.

Theorem 3.1 (Etnyre and Honda [4]) *Any cable in a standard neighborhood of a Legendrian knot can be put on a convex torus.*

Proposition 3.2 *If* ξ *is a universally tight contact structure on a solid torus* S *with convex boundary, then any Legendrian* (p,q) *knot* $L \subset S$ *has* $\operatorname{tw}(L;\partial S) \leq 0$.

We delay the proof of Proposition 3.2 to the end of this section, but use it here to give proofs of our main theorems stated in the introduction.

Proof of Theorem 1.4 If K has Legendrian large cables, then there exists $L \in \mathcal{L}(K_{p,q})$ such that $\operatorname{tb}(L) > pq$. Take a solid torus S representing K and containing L as a (p,q) curve. Perturb S to have convex boundary. By hypothesis, $\operatorname{tw}(L;\partial S) > 0$, so by Proposition 3.2, $\xi|_S$ must be virtually overtwisted. Suppose that it were possible to thicken S to a standard neighborhood $\widetilde{S}$ of K. Then $\operatorname{slope}(\Gamma_{\partial \widetilde{S}}) \in \mathbb{Z}$, which implies, by a result of Kanda [11], that $\xi|_{\widetilde{S}}$ is the unique tight contact structure on $\widetilde{S}$, and moreover that $\xi|_{\widetilde{S}}$ is universally tight. But this is a contradiction since $S \subset \widetilde{S}$ and $\xi|_S$ is virtually overtwisted, so no such thickening exists. If K were uniformly thick, then any neighborhood of K would be thickenable to a $\operatorname{slope}(\overline{\operatorname{tb}}(K))$ standard neighborhood of K, which we have just seen is not possible. □

Proof of Theorem 1.5 By assumption, there exists $L \in \mathcal{L}(K_{p,q})$ such that $\operatorname{tb}(L) > pq$. Stabilize L to obtain $\widetilde{L}$ such that $\operatorname{tb}(\widetilde{L}) = pq$. There is a solid torus S representing K for which $\widetilde{L} \subset \partial S$, and as discussed at beginning of Section 2, we see that $\operatorname{tw}(\widetilde{L};\partial S) = 0$. We can therefore C^0 perturb a collar neighborhood N of $\widetilde{L}$ in ∂S to be convex, and then C^∞ perturb $\partial S \setminus N$ to obtain a solid torus $\widetilde{S}$ representing K with convex boundary. Since $\operatorname{tw}(\widetilde{L};\partial \widetilde{S}) = 0$, and since $\operatorname{slope}(\widetilde{L}) = q/p$, we must have that $\operatorname{slope}(\Gamma_{\partial \widetilde{S}}) = q/p$, owing to the fact that $\operatorname{tw}(\widetilde{L}; \widetilde{S}) = -\frac{1}{2}|\widetilde{L} \bullet \varGamma_{\partial\widetilde{S}}|$ where $C_1 \bullet C_2$ denotes the geometric intersection number of two curves on a torus. But $q/p > \overline{\operatorname{tb}}(K)$ by assumption, so $w(K) > \overline{\operatorname{tb}}(K)$. □

Proof of Theorem 1.6 In [15], Yasui shows that for integers $n \le \frac{1}{4}(3-m)$, the cables $K^m_{n,-1}$ have the property that $\overline{\operatorname{tb}}(K^m_{n,-1}) = -1$. So, for any $m \le -5$ and any $1 < n \le \lfloor \frac{1}{4}(3-m) \rfloor$, we see that K^m has Legendrian large cables $L \in \mathcal{L}(K^m_{n,-1})$. Then, by Theorem 1.4, K^m is not uniformly thick and has virtually overtwisted neighborhoods, and by Theorem 1.5 we have that $w(K^m) > \overline{\operatorname{tb}}(K^m)$ □

Proof of Proposition 1.8 The slope of the cable $K^m_{n,-1}$ is $\operatorname{slope}(K^m_{n,-1}) = -1/n$. Whenever $n \le \frac{1}{4}(3-m)$, we know there exist $L \in \mathcal{L}(K^m_{n,-1})$ which are Legendrian large. Stabilize L to obtain $\widetilde{L}$ such that $\operatorname{tb}(\widetilde{L}) = -n$. There is a solid torus S representing K^m for which $\widetilde{L} \subset \partial S$, and we have seen that $\operatorname{tw}(\widetilde{L}; \partial S) = 0$. Using the strategy of the proof of Theorem 1.5, we can C^0 perturb a collar neighborhood N of $\widetilde{L}$ in ∂S to be convex, and then C^∞ perturb $\partial S \setminus N$ to obtain a solid torus $\widetilde{S}$ representing K^m with convex boundary. Since $\operatorname{tw}(\widetilde{L}; \partial\widetilde{S}) = 0$, and since $\operatorname{slope}(\widetilde{L}) = -1/n$, we must have $\operatorname{slope}(\varGamma_{\partial\widetilde{S}}) = -1/n$, and therefore $w(K^m) \ge -1/n$. □

Now we will give a series of results leading to the proof of Proposition 3.2.

Lemma 3.3 *If S is a solid torus with convex boundary, $\#\varGamma_{\partial S} = 2$, and $\operatorname{slope}(\varGamma_{\partial S}) \in \mathbb{Z}$ with its unique tight contact structure ξ, then any Legendrian (p,q) knot $L \subset S$ has* $\operatorname{tw}(L; \partial S) \le 0$.

Proof Notice that this follows immediately from Theorem 3.1, since S is a standard neighborhood, and any Legendrian curve L on a convex torus T must have $\operatorname{tw}(L; T) = \operatorname{tw}(L; \partial S) \le 0$. Alternatively, we can reason in the following way. Recall that Kanda [11] showed that any solid torus with integer slope and two dividing curves has a unique tight contact structure. Suppose that S is a solid torus with convex boundary, $\#\varGamma_{\partial S} = 2$, and $\operatorname{slope}(\varGamma_{\partial S}) = k \in \mathbb{Z}$ with its unique tight contact structure ξ, and that $L \subset S$ is a Legendrian (p,q) knot. Then S is a standard neighborhood of a Legendrian core curve K. Any two standard neighborhoods are contactomorphic, so we can find a neighborhood $N \subset (S^3, \xi_{\mathrm{std}})$ of a Legendrian unknot $U \subset S^3$ with $\operatorname{tb}(U) = -1$, and a contactomorphism $\varphi\colon S \to N$ which sends $\varphi(K) = U$. This contactomorphism sends torus knots to torus knots, so our (p,q) knot L is mapped to a $(p, q - p(k+1))$ knot $\varphi(L)$, as one can easily check. But now $\varphi(L)$ is a torus knot in $(S^3, \xi_{\mathrm{std}})$, and Etnyre and Honda [3] have shown that $\operatorname{tb}(\varphi(L)) \le p(q - p(k+1))$. But

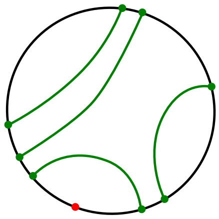


Figure 8: An arbitrary disk with arcs.

we understand how to switch between the Seifert framing and the framing coming from the torus ∂N, that is, $\mathrm{tw}(\varphi(L); \partial N) = \mathrm{tb}(\varphi(L)) - p(q - (k + 1)) \leq 0$. This implies that $\mathrm{tw}(L; \partial S) \leq 0$, since N and S are contactomorphic. □

We can strengthen Lemma 3.3 slightly by dropping the assumption that $\#\Gamma = 2$.

Lemma 3.4 *If S is a solid torus with convex boundary, and* $\mathrm{slope}(\Gamma_{\partial S}) \in \mathbb{Z}$ *with any tight contact structure ξ, then any Legendrian (p, q) knot $L \subset S$ has* $\mathrm{tw}(L; \partial S) \leq 0$.

Proof We will show that (S, ξ) will embed in a tight contact structure $(\widetilde{S}, \tilde{\xi})$ that satisfies the hypothesis of Lemma 3.3, and therefore show that $\mathrm{tw}(L; \partial S) \leq 0$. To this end, we note that we can assume $\mathrm{slope}(\Gamma_{\partial S}) = 0$ by applying a diffeomorphism to S. Recall from [10], that ξ is completely determined by the dividing set Γ_D on a meridional disk D of S. We will build a model situation for S in which we can construct $(\widetilde{S}, \tilde{\xi})$. Since $\#\Gamma_{\partial S} > 2$ we see that $\#\Gamma_D > 1$. Suppose that we have a convex disk D with an arbitrary collection of dividing curves Γ, as in Figure 8.

Let v be a vector field on D that guides the characteristic foliation. We can label the regions in $D \setminus \Gamma$ as either Σ^+ or Σ^- so that no adjacent pair share the same label. There exists an area form ω on D which satisfies that $\pm\mathrm{div}_\omega\, v > 0$ on $\Sigma^\pm$. Assign a 1–form $\lambda = \iota_v \omega$; then we know from Giroux [7] that there exists a function $u\colon D \to \mathbb{R}$

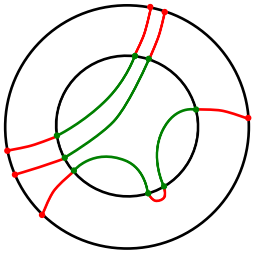


Figure 9: An annulus has been attached, and the number of curves has been reduced by one.

such that $u\,dt + \lambda$ gives rise to a contact structure ξ on $D \times \mathbb{R}$ that is invariant in the $\mathbb{R}$ direction. Moreover, we know from a theorem of Giroux that ξ is tight, since there are no homotopically trivial dividing curves. This invariance means that we can mod out by $\mathbb{Z}$ to obtain a tight contact structure on a solid torus $(D \times \mathbb{R})/\mathbb{Z} = D \times S^1$. The solid torus and contact structure we obtain in this way are contactomorphic to our original (S, ξ), that is, there exist v, ω and $u\colon D \to \mathbb{R}$ for which this construction exactly reproduces (S, ξ).

Now suppose that the number of properly embedded arcs is greater than 1. We would now like to reduce the number of dividing curves by taking a larger disk containing our original D. So we attach an annulus to D to obtain $D_{\mathrm{ext}} = D \cup_\varphi (S^1 \times [0,1])$, where $\varphi\colon S^1 \times \{0\} \to \partial D$ is the gluing map. Denote the endpoints of the properly embedded arcs by $\{x_1, \dots, x_{2k}\}$. Notice that if we fix a point on $p \in \partial D$ and move counterclockwise from p along ∂D, then it must happen that we encounter an x_i followed by an x_{i+1} which are not endpoints of the same curve. If this were not so, then there could only be one curve, which we have supposed not to be the case. Without loss of generality, assume that these two points are x_1 and x_2. Now connect these points by an arc in $S^1 \times [0,1]$. Form arcs from the remaining points $\{x_3, \dots, x_{2k}\}$ to $\partial D_{\mathrm{ext}}$ by using $\{x_i\} \times [0,1]$, as in Figure 9. Notice that D_{ext} has one fewer embedded arc than D. So we can iterate this procedure to obtain a disk $\widetilde{D} \supset D$ which has only 1 properly embedded arc. Call this arc $\widetilde{\Gamma}$. Notice that we can arrange the gluing map φ to be smooth and such that the extension of Γ to $\widetilde{\Gamma}$ is smooth. We can also smoothly extend ω and v to $\widetilde{D}$ so that the singular foliation on $\widetilde{D}$ guided by v has $\widetilde{\Gamma}$ as a dividing curve. We can now build, just as we did above, a contact structure $\tilde{\xi}$ on $\widetilde{D} \times S^1 = \widetilde{S}$ having $\widetilde{D}$ as a convex meridional disk, with convex boundary. Since $\#\Gamma_{\widetilde{D}} = 1$, we have $\mathrm{tb}(\partial \widetilde{D}) = -1$, which in turn implies that $\#\Gamma_{\partial \widetilde{S}} = 2$. Notice that $\tilde{\xi}|_S = \xi$. Also notice that, by construction, the method of reducing the number of dividing curves on ∂S yields $\mathrm{slope}(\Gamma) = \mathrm{slope}(\widetilde{\Gamma})$. Now, by Lemma 3.3, any Legendrian (p,q) knot $L \subset S$ has $\mathrm{tw}(L; \partial S) \le 0$. □

Lemma 3.5 *If ξ is a universally tight contact structure on a solid torus S with convex boundary and $\#\Gamma_{\partial S} = 2$, then any Legendrian (p,q) knot $L \subset S$ has $\mathrm{tw}(L; \partial S) \le 0$.*

Proof By a diffeomorphism of S, we can assume that $\mathrm{slope}(\Gamma_{\partial S}) = -r/s$ where $-\infty \le -r/s \le -1$, and that the meridional slope is $-\infty$. Let $n = \lceil r/s \rceil$. Then since ξ is universally tight, we know that any path in the Farey tessellation describing our contact structure has the property that each jump must be decorated with the same

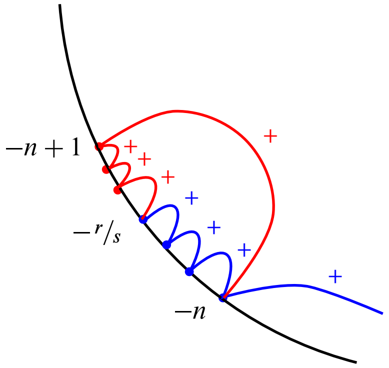


Figure 10: Farey tessellation picture describing the contact structure on our solid torus. The original solid torus, S, is shown in blue, while the red indicates the $T^2 \times I$ which is glued on to obtain the larger solid torus $\widetilde{S}$.

sign by Theorem 2.1. A portion of the Farey tessellation shows this in Figure 10. We can obtain a larger solid torus $\widetilde{S} \supset S$, which is convex, has two dividing curves, and with $\operatorname{slope}(\Gamma_{\partial \widetilde{S}}) = -n+1$ in the following way. Take a shortest path in the Farey tessellation from $-r/s$ to $-n+1$, and decorate each jump with the sign which appears in the description of the contact structure on S. This describes a contact structure on $T^2 \times I$ which extends S to $\widetilde{S}$, and since the signs are all the same we know that $\widetilde{S}$ is tight by Theorem 2.2. Moreover, we see that $\widetilde{S}$ has integer slope giving it a unique tight contact structure. Now we have that $\operatorname{tw}(K; \partial S) \leq 0$ by Lemma 3.3. □

Remark 3.6 In the above proof, we are able to thicken S to a larger solid torus $\widetilde{S} \supseteq S$ with $\operatorname{slope}(\Gamma_{\partial \widetilde{S}}) = -n+1$ because we are thinking of $S = S^1 \times D^2$ abstractly as a contact 3–manifold with convex boundary, and not embedded in any particular contact manifold. There is a shortest path in the standard Farey tessellation picture from any negative rational $-r/s$ to $-n+1$ which describes our contact structure. We

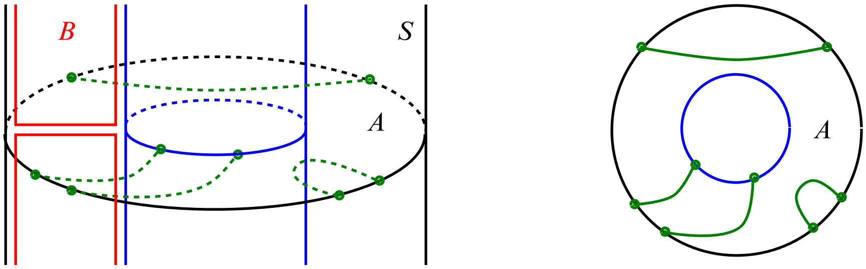


Figure 11: Left: $X = T^2 \times I$. Right: the annulus A and its dividing curves.

are not claiming that if S is a solid torus representing a knot $K \subseteq S^3$ it must always be thickenable in S^3, for example, Etnyre, LaFountain, and Tosun have given examples of nonthickenable tori in [5].

Proposition 3.2 strengthens Lemma 3.5 slightly by dropping the assumption that $\#\Gamma = 2$.

Proof of Proposition 3.2 Suppose we are given a solid torus S with convex boundary, a universally tight contact structure ξ, and we have a Legendrian (p,q) knot L in S. Again, by a diffeomorphism of S, we can assume that $\mathrm{slope}(\Gamma_{\partial S}) = -r/s$, where $-\infty \le -r/s \le -1$, and that the meridional slope is $-\infty$. Let $n = \lceil r/s \rceil$. If

$$\#\Gamma_{\partial S} = 2k > 2,$$

then we can attach a bypass to ∂S along a Legendrian ruling curve to obtain a smaller solid torus $S' \subset S$ which has $\mathrm{slope}(\Gamma_{\partial S'}) = -r/s$ and $\#\Gamma_{\partial S'} = 2k-2$. We can repeat this procedure until we have a solid torus $\widetilde{S} \subset S$ which has $\mathrm{slope}(\Gamma_{\partial \widetilde{S}}) = -r/s$ and $\#\Gamma_{\partial \widetilde{S}} = 2$. Notice that the contact structure on $\widetilde{S}$ is just $\xi|_{\widetilde{S}}$. If we look at a meridional disk $D \subset S$, we know that along ∂D there are $2sk$ intersection points with $\Gamma_{\partial S}$; however, there exists a slope γ for which curves on ∂S of slope γ have exactly $2k$ intersection points with $\Gamma_{\partial S}$. For convenience, change coordinates on S so that $\mathrm{slope}(\gamma) \mapsto -\infty$ and $\mathrm{slope}(\Gamma_{\partial S}) \mapsto 0$. Notice that we have a $T^2 \times I$ layer $X = S \setminus \widetilde{S}$, and we can find a convex annulus A in X with Legendrian boundary of slope γ. We would like to show that the contact structure on X is completely determined by the dividing curves on A. Since $\#\Gamma_{\partial \widetilde{S}} = 2$, $\#\Gamma_{\partial S} = 2k$, and $\mathrm{slope}(\Gamma_{\partial \widetilde{S}}) = \mathrm{slope}(\Gamma_{\partial S}) = 0$, we know that the dividing curves on A must have the form shown in Figure 11, right, by the green arcs.

We know from Giroux [7] that the contact structure on a neighborhood of A is determined by its dividing curves. If we cut X along A, and round corners, we obtain a solid torus Y with convex boundary. Using the edge rounding lemma [10], it is easy to see that $\#\Gamma_{\partial Y} = 2$ and $\mathrm{slope}(\Gamma_{\partial Y}) = -1$. Notice in Figure 11, left, that we have a meridional disk B of Y which we have just seen has $\mathrm{tw}(\partial B) = -1$, and which we can perturb to be convex. There is a unique choice of dividing curves on such a disk. Finally, if we cut Y along B and round corners, we obtain a B^3 with convex boundary, which has a unique tight contact structure from work of Eliashberg [2]. So we have seen that the contact structure of X is determined solely by the dividing curves on A.

Let v be a vector field on A that guides the characteristic foliation. We can label the regions in $A \setminus \Gamma$ as either Σ^+ or Σ^- so that no adjacent pair share the same label. There exists an area form ω on A which satisfies that $\pm\mathrm{div}_\omega\, v > 0$ on $\Sigma^\pm$. Assign a

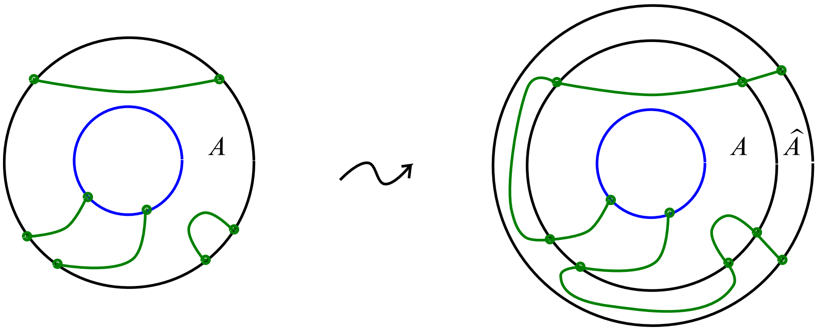


Figure 12: Reducing the number of dividing curves on A by extending with an annulus $\widehat{A}$.

1–form $\lambda = \iota_v\omega$; then we know from Giroux [7] that there exists a function $u\colon A \to \mathbb{R}$ such that $u\,dt + \lambda$ gives rise to a contact structure ξ on $A \times \mathbb{R}$ that is invariant in the $\mathbb{R}$ direction. Moreover, we know from a theorem of Giroux that ξ is tight since there are no homotopically trivial dividing curves. This invariance means that we can mod out by $\mathbb{Z}$ to obtain a tight contact structure on $(A \times \mathbb{R})/\mathbb{Z} = T^2 \times I$. The $T^2 \times I$ layer and contact structure we obtain in this way are contactomorphic to our original (X, ξ); that is, there exist v, ω and $u\colon A \to \mathbb{R}$ for which this construction exactly reproduces (X, ξ).

Now observe that we can smoothly extend A, abstractly, by an annulus $\widehat{A}$ causing the number of dividing curves to be reduced to 2, just as we did with the disk in the proof of Lemma 3.4; see Figure 12.

We can arrange that the extension of Γ_A to $\Gamma_{A\cup\widehat{A}}$ is smooth, and we can also smoothly extend ω and v to a neighborhood of $\widehat{A}$ so that the singular foliation on $\widehat{A}$ guided by v has $\Gamma_{A\cup\widehat{A}}$ as a set of dividing curves. We can now build, just as we did above, a contact structure $\hat{\xi}$ on $(A \cup \widehat{A}) \times S^1 = \widehat{X}$ with convex boundary. Since $\#\Gamma_{A\cup\widehat{A}} = 2$, we see that $\operatorname{tb}(\partial\widehat{A} \cap \partial\widehat{X}) = -1$, which implies that $\#\Gamma_{\partial\widehat{X}} = 2$. Notice that $\hat{\xi}|_X = \xi$. Also notice that, by construction, the method of reducing the number of dividing curves on ∂X yields $\operatorname{slope}(\Gamma_{\partial X}) = \operatorname{slope}(\Gamma_{\partial\widehat{X}})$. But now we have a minimally twisting $T^2 \times I$ layer $\widehat{X}$ whose boundary tori each have two dividing curves with slope 0. Honda [10] showed that there are an integers worth of tight contact structures satisfying these boundary conditions, and that each one is I–invariant. Adding the I–invariant thickened torus $X \cup \widehat{X}$ to $\widetilde{S}$, we get a new solid torus with contact structure contactomorphic to $\xi|_{\widetilde{S}}$, thus universally tight. Clearly S is contained in this solid torus. Now, by Lemma 3.5, L has $\operatorname{tw}(L; \partial S) = \operatorname{tw}(L; \partial S') \leq 0$. □

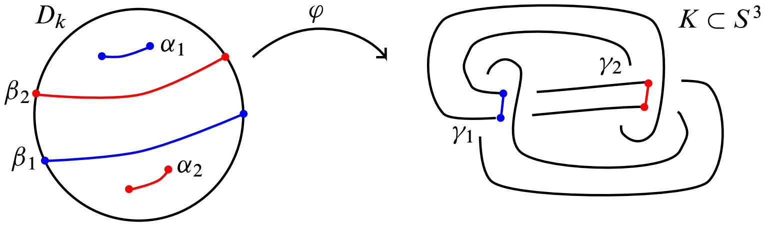


Figure 13: The immersion of a ribbon disk D_K.

4 Building ribbon knots from canceling handles

We are concerned here with ribbon knots, which we take to be the following: A knot $K \subset S^3$ is a *ribbon knot* if there is an immersed disk $\varphi\colon D_K \to S^3$ such that:

(1) $\partial\varphi(D_K) = K$.

(2) All of the double points of $\varphi(D_K) = \widetilde{D}_K$ (we will use the symbol $\sim$ to denote image under φ) occur transversely along arcs $\gamma_i \subset S^3$ whose preimage $\varphi^{-1}(\gamma_i) \subset D_K$ consists of exactly two arcs. One of these, α_i, must be contained entirely in the interior, $\alpha_i \subset \operatorname{int}(D_K)$, and the other, β_i (meant to suggest boundary), must be a properly embedded arc in D_K (ie $\partial\beta_i \subset \partial D_K$ and $\operatorname{int}(\beta_i) \cap \partial D_K = \varnothing$).

An example ribbon disk and its image under φ are shown in Figure 13.

Note that by transversality, the preimages of the γ_i's are 1–dimensional submanifolds of the compact manifold D_K, so there are only finitely many ribbon singularities γ_i.

We want to give a construction of an arbitrary ribbon knot using 1–2–handle-canceling pairs. Given any ribbon knot $K \subset S^3$, it has a ribbon disk D_K by definition. Notice, every ribbon singularity γ_i must appear exactly twice on the ribbon disk: once as a properly embedded arc, and once as an arc contained entirely in the interior of D_K. We will use a common color when picturing these pairs. So a general ribbon disk might look something like the one seen in Figure 14.

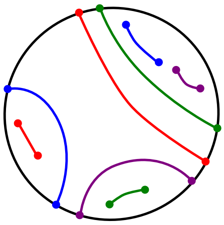


Figure 14: A general ribbon disk example.

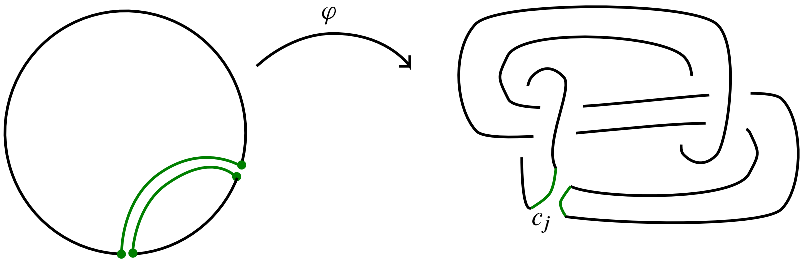


Figure 15: Cutting a ribbon disk along an arc c_j.

We will want to make cuts c_j, by pushing off two parallel copies of an arc in D_K and removing a small ϵ–strip. The result of this cut is shown in Figure 15.

We also need to set up a tool for manipulating ribbon disks and their images. Suppose we have an arc $b \subset \partial D_K$ whose endpoints are the endpoints of one of our β's. Further suppose that the subdisk D they bound contains no other singular points, as in Figure 16.

Let $N = I \times [0, \epsilon]$ be a collar neighborhood of β in D_K such that $(t, 0) = \beta$. We can form a new disk $D_\epsilon = D \cup N$ with boundary

$$\partial D_\epsilon = b \cup (0, s) \cup (1, s) \cup (t, \epsilon)$$

and notice that $\mathrm{int}(\beta) \subset \mathrm{int}(D_\epsilon)$. By choosing $\epsilon > 0$ sufficiently small, we can assume that $\widetilde{D}_\epsilon$ is embedded. Then we can see that D_ϵ guides an isotopy, supported in a small neighborhood of D_K, taking b to (t, ϵ) so that the disk $\overline{D_K - D_\epsilon} = D'_K$ does not contain β. We will refer to such a move as a *disk slide*; Figure 17 shows a typical one.

Theorem 4.1 *Given an arbitrary ribbon knot $K \subset S^3$ with $n \in \mathbb{N}$ ribbon singularities γ_i, we can make $n-1$ or fewer cuts c_j, so that what remains of K is an unlink, and what remains of $\widetilde{D}_K$ is, after n or fewer disk slides, embedded. That is, it is a collection of disjoint disks.*

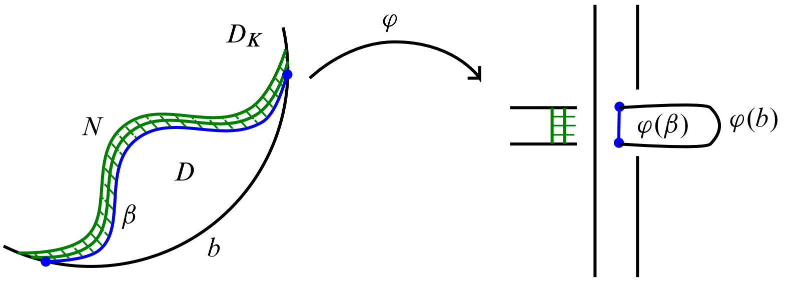


Figure 16: A subdisk and collar neighborhood, and its immersed image under φ.

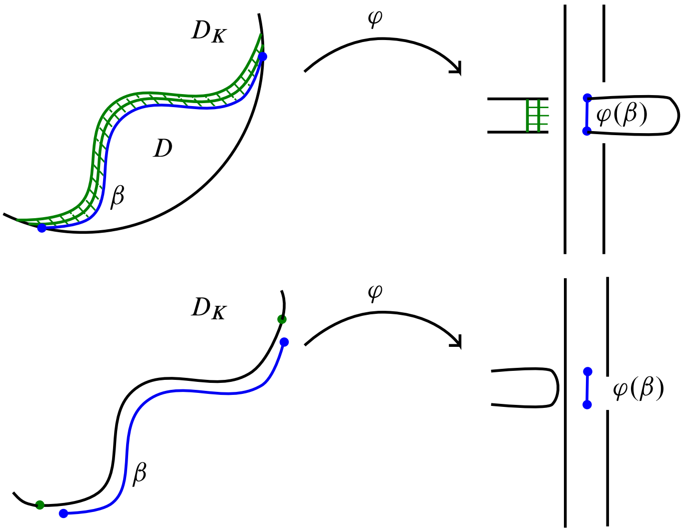


Figure 17: An illustration of a disk slide.

To prove this we will need the following.

Lemma 4.2 *Given a ribbon knot K with n ribbon singularities, if we can find a subdisk $D \subset D_K$ such that*

$$\partial D = (\text{an arc in } \partial D_K) \cup \beta_i$$

for one of our properly embedded arcs β_i, and $\operatorname{int}(D)$ is disjoint from all α's and β's, then a disk slide gives an isotopy of K supported in a small neighborhood of D so that the new slicing disk $\widetilde{D}'_K$ has $n-1$ ribbon singularities.

Proof For reference, let $b_i = \partial D \cap \partial D_K$ so that $\partial D = b_i \cup \beta_i$. Also, let $N = I \times [0, \epsilon]$ be a collar neighborhood of β_i in D_K such that $(t, 0) = \beta_i$, similar to the one shown in Figure 16.

Then we can define a new subdisk $D_\epsilon = D \cup N$ with boundary

$$\partial D_\epsilon = b_i \cup (0, s) \cup (1, s) \cup (t, \epsilon)$$

and notice that $\operatorname{int}(\beta_i) \subset \operatorname{int}(D_\epsilon)$. By choosing $\epsilon > 0$ sufficiently small, we can assume that $\widetilde{D}_\epsilon$ is embedded. Then there is a disk slide taking b_i to (t, ϵ) so that the disk $\overline{D_K - D_\epsilon} = D'_K$ does not contain β_i. But then it also cannot contain α_i, since the preimages of singularities occur in pairs, and hence the singularity γ_i has been eliminated. We also have that the resulting knot $\partial \widetilde{D}'_K$ is isotopic to K. □

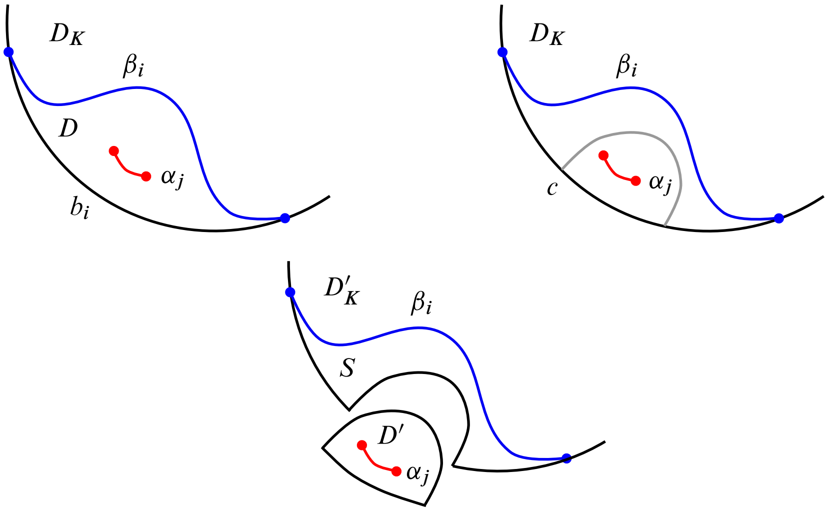


Figure 18: Cutting a ribbon disk.

Notice that Lemma 4.2 says that if we see a boundary parallel arc in D_K with no other singular points between that arc and some portion of ∂D_K, then we can eliminate that arc and its interior partner from the picture by an isotopy of K. Now back to our general picture and the proof of Theorem 4.1.

Proof of Theorem 4.1 We will assume that our ribbon disk is reduced in the sense that, if it were possible to simplify with a disk slide, then we have done so already. We will consider Figure 14 as our prototypical ribbon disk, and recall the convention that for each singularity γ_i, $\varphi^{-1}(\gamma_i)$ consists of $\alpha_i \cup \beta_i$ with β_i properly embedded. Given an arbitrary ribbon knot $K \subset S^3$ with $n \in \mathbb{N}$ ribbon singularities γ_i, and ribbon disk $\varphi\colon D_K \to S^3$, there will always be an "outermost" properly embedded arc β_i. This means that in some subdisk D, whose boundary is β_i together with an arc $b_i \subset \partial D_K$, there are only interior singular points α_j, and no other properly embedded arcs. Figure 18, top left, shows one such case.

Let c be a properly embedded arc in $D \subset D_K$ such that c cuts D into $D' \cup S$ with $\beta_i \subset S$ and D' containing all arcs $\alpha_j \subset D$. We may cut D_K along c so that φ is defined on $D'_K = \overline{D_K - D'}$ and D', and after a small isotopy of $\varphi\,|_{D'}$ we have that $\varphi(D')$ and $\varphi(D'_K)$ are disjoint, as pictured in Figure 18, bottom. Then a disk slide eliminates β_i by Lemma 4.2. Notice that when we eliminate a particular β_i using a disk slide, that automatically eliminates the corresponding α_i since they occur in pairs. Also notice, each cut eliminates at least one β_i, but could allow for the removal of more than one.

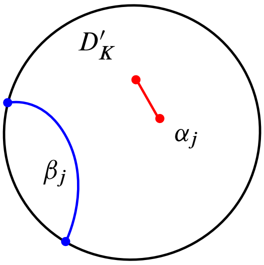


Figure 19: Final iteration.

But after at most $n-1$ cuts we have at most one β_j and its corresponding α_j. Since β_j cuts the disk it sits on into two components, one of them contains no α curves — see Figure 19 — and so β_j can be removed with no further cuts. Thus we never need to make the n^{th} cut since this last β curve may be eliminated by a disk slide without making a cut. Then the image under φ is now n embedded disks whose boundary is an unlink. □

We remark that this gives an upper bound on the number of cuts needed, but there are certainly cases where this number is not optimal as the following example shows.

Example 4.3 Consider the ribbon knot in Figure 20. This knot has $n+2$ ribbon singularities for any $n \in \mathbb{N}$, and yet only one cut (shown in green) will reduce the picture to two disjoint disks.

Now we will introduce handles and obtain a Kirby picture in which our knot K takes a particularly simple form. We assume that the reader is familiar with basic handlebody

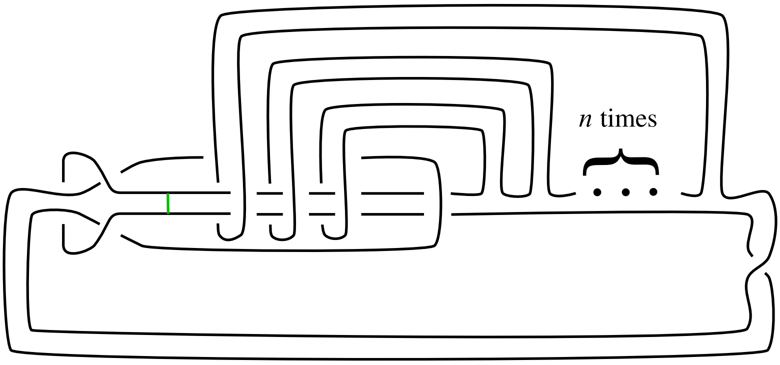


Figure 20: An example ribbon knot with $n+2$ singularities for which a single cut suffices.

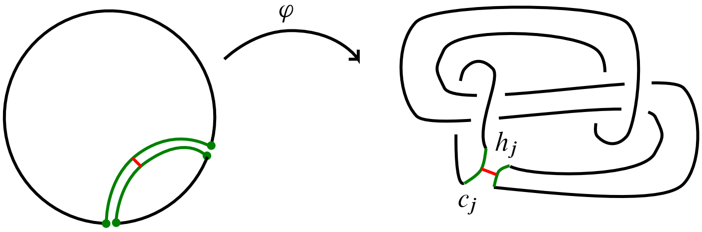


Figure 21: A 2–handle h_j associated to a cut c_j.

theory; an excellent reference for this material is [8]. For every cut c_j, we will attach an arc h_j seen in Figure 21. We will think of h_j as a thin ribbon, which would recover K if glued along. For this reason we will give the arc h_j a framing, by which we mean a parallel arc, and keep track of this framing through any isotopies of K. By a 1–*subhandlebody*, we will mean the subhandlebody consisting of the 0–handles and the 1–handles.

Proof of Theorem 1.12 Using Theorem 4.1, we can make $k < n$ cuts to the ribbon disk to obtain the unlink. So we have a diagram in which there are k disjoint disks, and $k-1$ framed arcs h_j. We know that by taking a band sum along these arcs (paying attention to framings) we can recover our diagram for K. Let K_{cut} be the union of the boundaries of these disks. Now in a small neighborhood of the end points of each h_j we insert the attaching spheres of a 1–handle, letting h_j be the attaching circle of a 2–handle as seen in Figure 22.

This pair cancels by construction, and also has the effect of doing the band sum that recovers K for the cut c_j as seen in Figure 23. Notice that we make two handle slides that free K_{cut} from the 1–handle, and then cancel the pair. Also notice that this has exactly the same effect that a band sum of K_{cut} along h_j would have had.

There is no obstruction to this handle slide and cancellation caused by the possible presence of other handle pairs since the double band sum shown on the left can be

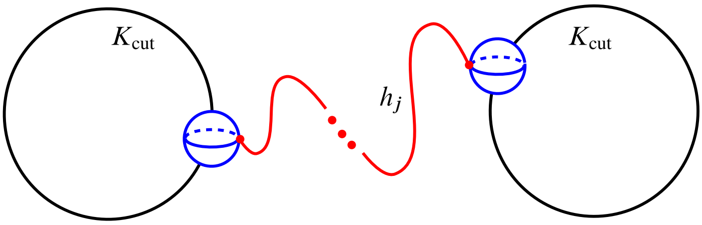


Figure 22: A 1–2–handle-canceling pair.

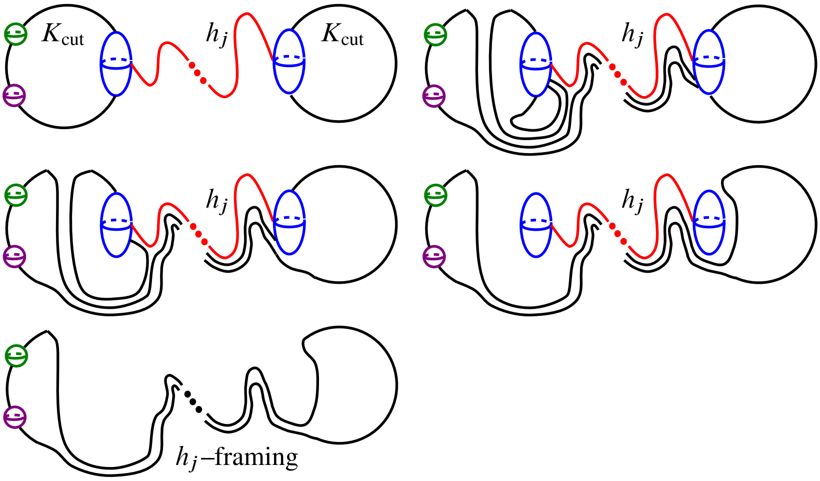


Figure 23: An example of handle cancellation to recover K.

carried out in a small neighborhood of the attaching sphere on the left. So, after $n-1$ or fewer iterations, we have recovered our diagram for K. It is worth noting that framings on 2–handles denote an even number of half twists; therefore the framings on the h_j must be even. If our diagram for K requires an odd number of half twists then we can accommodate this by inserting any number of half twists in one of the disks spanning K_{cut}, shown in Figure 24 for the case of a single half twist.

We would like to think of our diagram in which there are k disjoint disks connected by $k-1$ arcs h_j abstractly as a graph in order to show that K_{cut} can be pulled free of the 1–handles. To do this, we first work in the boundary of the 1–subhandlebody. We think of each of our disjoint disks as a vertex, and put an edge between vertices

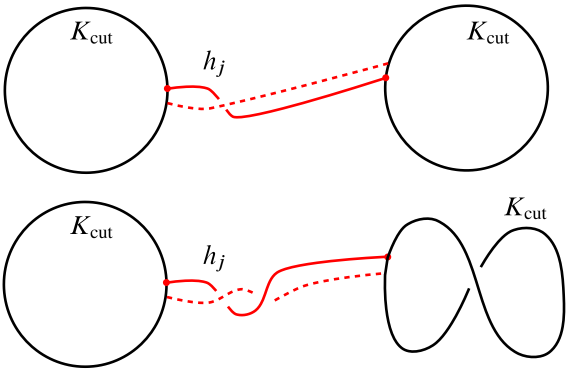


Figure 24: Framing adjustment.

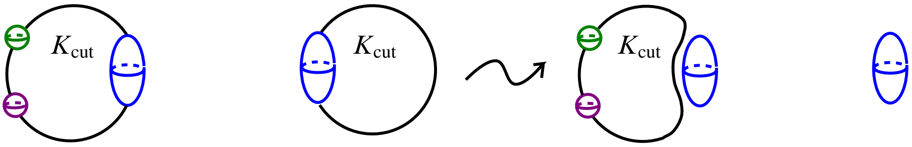


Figure 25: Handle picture corresponding to a univalent vertex of G.

if the corresponding disks are joined by a 1–handle. Notice G embeds in D_K as the "dual" graph to D_K cut along $\varphi^{-1}(c_j)$; that is, there is a vertex in the center of each component of $D_K - \bigcup_{j=1}^{k-1} \varphi^{-1}(c_j)$ and an edge for each $\varphi^{-1}(c_j)$. Then G is homotopy equivalent to D_K, and so we see that $\chi(G) = \chi(D_K) = 1$. It is well known that the Euler characteristic of a connected graph is one if and only if that graph is a tree, so G is a tree. Each univalent vertex of G is now associated to a portion of our picture consisting of two disks connected by a 1–handle, where one disk might have many 1–handle attaching spheres, but the other must have exactly one 1–handle attaching sphere as shown in Figure 25. In the 1–subhandlebody it is clear that K_{cut} may be isotoped off this 1–handle. Notice that the effect of this isotopy on G is to remove the corresponding edge and univalent vertex from the graph. Since G is a tree, we can iterate this procedure revealing that K_{cut} can be pulled completely free of the 1–handles. This may be seen in Figure 26 by simply ignoring the attaching circles of the 2–handles h_j.

The above iteration gives an isotopy of K_{cut} which extends to an ambient isotopy of the boundary of the 1–subhandlebody. This, in turn, induces an isotopy on the attaching circles of the 2–handles h_j, resulting in a 2–handlebody as claimed in Theorem 1.12. See Figure 26. By construction, handle slides and cancellations give us a knot isotopic to $K \subset S^3$. □

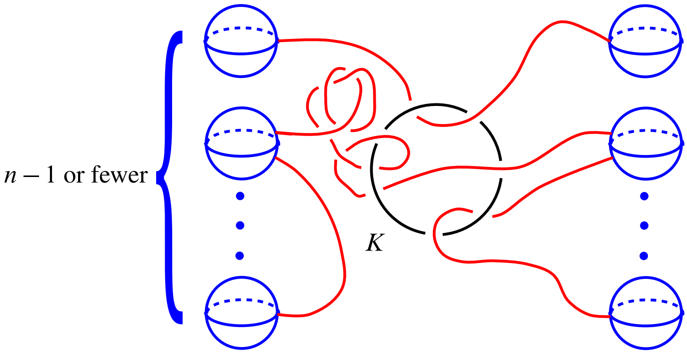


Figure 26: A 2–handlebody picture where K appears as the unknot in the boundary of the 1–subhandlebody.

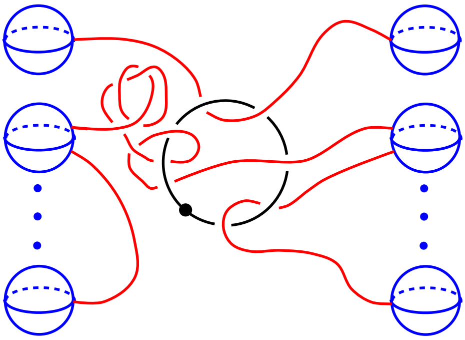


Figure 27: A 2–handlebody picture of the complement of the slice disk for K.

So we have shown that any ribbon knot with n ribbon singularities may be constructed by starting with the unknot in $\#_k S^1 \times S^2$, where $k \le n-1$, and attaching 2–handles to cancel each of the 1–handles in an appropriate manner.

Example 4.4 It is an exercise in Kirby calculus to show that images in Figure 2 are two pictures of the same ribbon knot in S^3.

Corollary 4.5 *In Figure 26, if we replace the unknot in the* 1*–subhandlebody with a dotted circle, then we obtain a picture of the* 4*–manifold which is the complement of the slicing disk in D^4, shown in Figure 27.*

Proof The slicing disk can be seen in the picture as the disk filling the unknot that we have in the 1–subhandlebody. This is because canceling the 1–2–handle pairs not only recovers K, but also recovers the ribbon disk $\widetilde{D}_K$. The definition of the dotted circle notation is that we remove a small neighborhood of the dotted unknot along with a small neighborhood of the disk after pushing it into D^4. And so this is exactly the complement of the slicing disk, $D^4 - \widetilde{D}_K$. □

One nice fact is that, since disk slides, isotopies and handle cancellations can be done locally, and since ribbon knots always bound an immersed ribbon disk, this construction actually works in any 3–manifold. We did not rely on any special properties of S^3 during the process. One can create examples by combining a 2–handlebody picture for a ribbon knot $K \subset S^3$ as in the above construction with a Kirby picture of a 4–manifold W whose boundary is the intended 3–manifold $M^3 = \partial W$. When combining the two pictures, K may be allowed to run across noncanceling 1–handles to form nontrivial examples as shown in Figure 28, where we have a Kirby picture of a 4–manifold whose

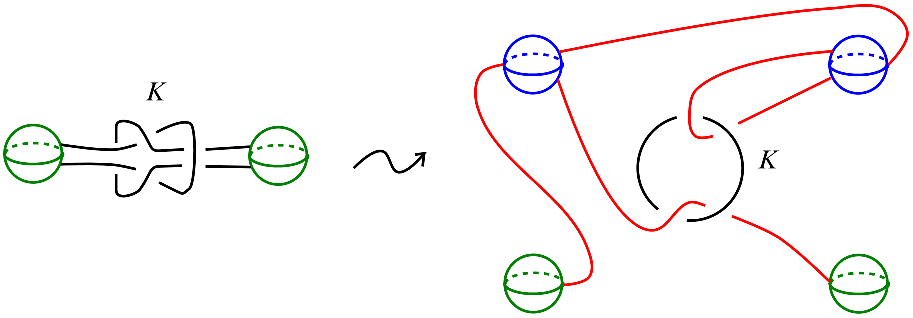


Figure 28: An example ribbon knot in $S^1 \times S^2$ and its decomposition.

boundary is $S^1 \times S^2$. We can see the ribbon disk for K in the image on the left. The image on the right shows the result using the technique developed above.

References

[1] **J Conway**, **J B Etnyre**, **B Tosun**, *Symplectic fillings, contact surgeries, and Lagrangian disks*, Int. Math. Res. Not. 2021 (2021) 6020–6050 MR Zbl

[2] **Y Eliashberg**, *Classification of overtwisted contact structures on 3–manifolds*, Invent. Math. 98 (1989) 623–637 MR Zbl

[3] **J B Etnyre**, **K Honda**, *Knots and contact geometry, I: Torus knots and the figure eight knot*, J. Symplectic Geom. 1 (2001) 63–120 MR Zbl

[4] **J B Etnyre**, **K Honda**, *Cabling and transverse simplicity*, Ann. of Math. 162 (2005) 1305–1333 MR Zbl

[5] **J B Etnyre**, **D J LaFountain**, **B Tosun**, *Legendrian and transverse cables of positive torus knots*, Geom. Topol. 16 (2012) 1639–1689 MR Zbl

[6] **J Etnyre**, **V Vértesi**, *Legendrian satellites*, Int. Math. Res. Not. 2018 (2018) 7241–7304 MR Zbl

[7] **E Giroux**, *Convexité en topologie de contact*, Comment. Math. Helv. 66 (1991) 637–677 MR Zbl

[8] **R E Gompf**, **A I Stipsicz**, *4–Manifolds and Kirby calculus*, Graduate Studies in Math. 20, Amer. Math. Soc., Providence, RI (1999) MR Zbl

[9] **M Hedden**, *Notions of positivity and the Ozsváth–Szabó concordance invariant*, J. Knot Theory Ramifications 19 (2010) 617–629 MR Zbl

[10] **K Honda**, *On the classification of tight contact structures, I*, Geom. Topol. 4 (2000) 309–368 MR Zbl

[11] **Y Kanda**, *The classification of tight contact structures on the* 3*–torus*, Comm. Anal. Geom. 5 (1997) 413–438 MR Zbl

[12] **T Lidman**, **S Sivek**, *Contact structures and reducible surgeries*, Compos. Math. 152 (2016) 152–186 MR Zbl

[13] **H Min**, *A note on uniform thickness property*, in preparation

[14] **L Moser**, *Elementary surgery along a torus knot*, Pacific J. Math. 38 (1971) 737–745 MR Zbl

[15] **K Yasui**, *Maximal Thurston–Bennequin number and reducible Legendrian surgery*, Compos. Math. 152 (2016) 1899–1914 MR Zbl

Powder Springs, GA, United States

andrew.mccullough@gtri.gatech.edu

Received: 7 October 2020 Revised: 30 August 2021

Algebraic & Geometric Topology **23**:6 (2023) 2593–2626
DOI: 10.2140/agt.2023.23.2593
Published: 7 September 2023

Homology of configuration spaces of hard squares in a rectangle

HANNAH ALPERT
ULRICH BAUER
MATTHEW KAHLE
ROBERT MACPHERSON
KELLY SPENDLOVE

We study ordered configuration spaces $C(n; p, q)$ of n hard squares in a $p \times q$ rectangle, a generalization of the well-known "15 puzzle". Our main interest is in the topology of these spaces. Our first result describes a cubical cell complex and proves that it is homotopy equivalent to the configuration space. We then focus on determining for which n, j, p, and q the homology group $H_j[C(n; p, q)]$ is nontrivial. We prove three homology-vanishing theorems, based on discrete Morse theory on the cell complex. Then we describe several explicit families of nontrivial cycles, and a method for interpolating between parameters to fill in most of the picture for "large-scale" nontrivial homology.

55R80; 57Q70, 82B26

1 Introduction

We study the ordered configuration space of n squares in a $p \times q$ rectangle, which we denote by $C(n; p, q)$. The case $n = 15$ and $p = q = 4$ corresponds to the famous "15 puzzle". This puzzle was apparently invented by Noyes Palmer Chapman, a postmaster in Canastota, New York, in 1874; see Sonneveld and Slocum [12]. Already by 1879, the puzzle had been analyzed mathematically by Johnson and Story [10]. They showed that it is not possible, for example, for any sequence of moves to transpose the pieces labeled 14 and 15. Their observation is really a topological one, namely that the configuration space has two connected components.

A natural discrete model for the 15 puzzle is the graph G_{15}, which we describe as follows. The vertices are the aligned positions of the puzzle, corresponding to the 16!

permutations of the 15 pieces and the one hole, and we have an edge between every pair of positions that differ by sliding a piece into the hole.

If we allow arbitrary positions for nonoverlapping squares, then the configuration space for the 15 puzzle is more than 1–dimensional; for instance, there is a three-parameter family of ways to slide horizontally the three pieces in the bottom row. Nevertheless, as a special case of our results here, the configuration space of the 15 puzzle deformation retracts to a one-dimensional subspace homeomorphic to G_{15}.

Having a cell complex structure allows for computing many topological invariants directly. For example, the Betti number β_1 can be computed by counting the number of 0–cells $f_0 = 16!$ and 1–cells $f_1 = 24 \cdot 15!$ of G_{15}, using the fact that $\beta_0 = 2$, and applying the 1–dimensional Euler formula $f_0 - f_1 = \beta_0 - \beta_1$.

In the more general setting, we describe a cubical complex $X(n; p, q)$ and show it is always a deformation retract of the configuration space $C(n; p, q)$. Applying discrete Morse theory on this complex allows us to establish some necessary conditions on where nontrivial homology can appear.

In the following, we always assume that $p, q \geq 1$, $0 \leq n \leq pq$, and $j \geq 0$. We also sometimes use a "large-scale" parametrization, by defining $x = n/pq$ and $y = j/pq$. The quantity x has a physical interpretation as "density", describing the area ratio in the rectangular region that is occupied by squares.

Theorem 1.1 (homology vanishing theorem) *We have*:

(1) *If $j > pq - n$, then $H_j[C(n; p, q)] = 0$.*

(2) *If $j > n$, then $H_j[C(n; p, q)] = 0$.*

(3) *If $j > pq/3$, then $H_j[C(n; p, q)] = 0$.*

Equivalently, *on the large scale*, *if $H_j[C(n; p, q)] \neq 0$ then $y \leq \min\bigl\{1 - x, x, \frac{1}{3}\bigr\}$.*

The cubical cell complex model allows us to do exact computations for small examples. We include a table of Betti numbers in Section 7. Based in part on our computations, we conjecture the following.

Conjecture 1.2 If $H_j[C(n; p, q)] \neq 0$, then

$$j \leq \min\biggl\{pq - n, n - \frac{8n^2}{9pq}, \frac{pq}{4}\biggr\}.$$

Equivalently, we conjecture that if $H_j[C(n; p, q)] \neq 0$, then

$$y \leq \min\biggl\{1 - x, x - \frac{8}{9}x^2, \frac{1}{4}\biggr\}.$$

In Section 6, we describe several families of explicit nontrivial cycles, and a method for interpolating between parameters. We *almost* show that if $y \le \min\{1-x, x-\frac{8}{9}x^2, \frac{1}{4}\}$ there exist n, j, p, and q such that $H_j[C(n;p,q)] \ne 0$. Instead we prove an analogous statement with a piecewise linear approximation of the parabola $y = x - \frac{8}{9}x^2$. Let S be the set of points on the parabola defined by

$$S = \left\{ \left(x, x - \frac{8}{9}x^2 \right) \;\middle|\; x = \frac{3}{4k}, k \ge 1 \right\}.$$

Note that $\left(\frac{3}{4}, \frac{1}{4}\right) \in S$ and $\left(\frac{3}{8}, \frac{1}{4}\right) \in S$. Let I be the closed interval

$$I = \{(x,y) \mid 0 \le x \le 1 \text{ and } y = 0\}.$$

Theorem 1.3 (large-scale homology nonvanishing theorem) *If (x,y) is any rational point in the convex hull of $S \cup I$, then there exist n, p, q, and j such that $x = n/pq$, $y = j/pq$, and $H_j[C(n;p,q)] \ne 0$.*

Theorem 1.3 might suggest the right necessary conditions for nontrivial homology, rather than Conjecture 1.2. We do not currently know of any instance of n, j, p, and q where $H_j[C(n;p,q)] \ne 0$ and (x,y) lies outside of the convex hull of $S \cup I$.

A summary of our main results is illustrated in Figure 1. Although we have made some headway, completely resolving the following question is left as future work.

Question 1.4 What are necessary and sufficient conditions on $(n;j;p,q)$ for

$$H_j[C(n;p,q)] \ne 0?$$

We note that Conjecture 1.2 is only about necessary conditions for nontrivial homology, but at the moment we do not have a good conjecture for necessary and sufficient conditions. The conditions in Conjecture 1.2 by themselves are not sufficient. For example, $\left(\frac{1}{4}, \frac{3}{16}\right)$ is a point in the blue region of Figure 1, corresponding to $n=p=q=4$ and $j=3$. However, it is not true that we have homology whenever $n/pq = \frac{1}{4}$ and $j/pq = \frac{3}{16}$, even when n is arbitrarily large. Suppose that $p=2$ and $q=8k$ for some $k \ge 1$, and $n = 4k$; then we cannot get nontrivial homology with $j = 3k$. The largest j where we will see nontrivial homology is $j = 2k$; by the homotopy equivalence mentioned below, this follows from Theorem 1.2(3) in Alpert, Kahle, and MacPherson [2].

In recent years, there has been increased interest in similar kinds of configuration spaces; see Alpert [1], Baryshnikov, Bubenik, and Kahle [3], and Carlsson, Gorham, Kahle, and Mason [6] for some earlier work on configuration spaces of disks. Plachta

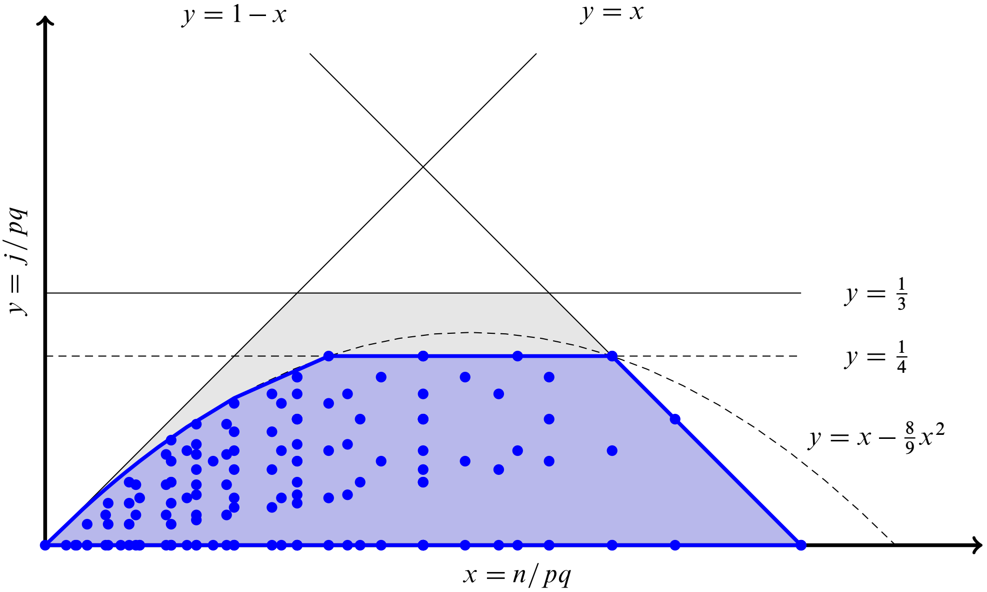


Figure 1: A summary of our main results. The axes are $x = n/pq$ and $y = j/pq$. We show that if (x, y) is outside the shaded region bounded by $y = 1 - x$, $y = x$, and $y = \frac{1}{3}$, then $H_j[C(n; p, q)] = 0$. We show conversely that for every rational point (x, y) in the blue part of the shaded region, there exist n, j, p, and q such that $x = n/pq$, $y = j/pq$, and $H_j[C(n; p, q)] \neq 0$. Each of the blue dots represents a point (x, y) where we computed that $H_j[C(n; p, q)] \neq 0$, with $n \leq 6$.

recently studied configuration spaces of squares in a rectangle [11], using affine Morse–Bott theory, smooth flows, and graphs associated with such configurations. As one application, he showed that under certain conditions the configuration space which we denote by $C(n; p, q)$ is connected. We note that our dimensions of the rectangle, p and q, are always positive integers, but he studies the more general framework where they may be positive real numbers.

What we study here is closely related to the recent paper [2] on configuration spaces of hard disks in an infinite strip, which we now briefly discuss. Let config(n, w) denote the configuration space of n disks of unit diameter in an infinite strip of width w. While we do not prove it here, it is not hard to check that $C(n; p, q)$ is homotopy equivalent to config(n, w) if $q \geq n$ and $p = w$. So the configuration spaces of hard squares in a rectangle we study here are a generalization of the configuration spaces of hard disks in an infinite strip.

Motivated by the notion of phase transitions for hard-spheres systems, definitions are suggested in [2] for homological solid, liquid, and gas regimes. The definitions apply here as well.

Let $\mathrm{Conf}(n;\mathbb{R}^2)$ denote the (ordered) configuration space of points in the plane. We say that $(n; j; p, q)$ is

- in the *homological solid regime* if
$$H_j[C(n;p,q)] = 0,$$
- in the *homological gas regime* if the inclusion map $i\colon C(n;p,q) \to \mathrm{Conf}(n;\mathbb{R}^2)$ induces an isomorphism on homology
$$i_*\colon H_j[C(n;p,q)] \to H_j[\mathrm{Conf}(n;\mathbb{R}^2)], \text{ and}$$
- in the *homological liquid regime* otherwise.

We are mainly concerned with the boundary between trivial and nontrivial homology, ie separating the solid regime from liquid and gas. It will also be interesting to better understand the boundary between the homological liquid and gas regimes, as summarized in the following question.

Question 1.5 What are necessary and sufficient conditions on $(n; j; p, q)$ for the inclusion map $i\colon C(n;p,q) \to \mathrm{Conf}(n;\mathbb{R}^2)$ to induce an isomorphism on homology
$$i_*\colon H_j[C(n;p,q)] \to H_j[\mathrm{Conf}(n;\mathbb{R}^2)]?$$

Acknowledgments

Alpert was supported by NSF-DMS 1802914 during work on this project. Bauer gratefully acknowledges support from the German Research Foundation (DFG) through the Collaborative Research Center SFB/TRR 109 *Discretization in geometry and dynamics*. Kahle gratefully acknowledges the support of NSF-DMS 2005630, NSF-DMS 1839358, and NSF-CCF 1839358. Spendlove was partially supported by the NSF Graduate Research Fellowship Program under grant DGE-1842213 and by EPSRC grant EP/R018472/1.

2 Definitions and notation

The configuration space $C(n;p,q)$ of n unit squares in a $p \times q$ rectangle can be written as a subspace of $\mathbb{R}^{2n}$ by keeping track of the coordinates of the centers of the squares. We select our $p \times q$ rectangle to be the set $\left[\frac{1}{2}, p+\frac{1}{2}\right] \times \left[\frac{1}{2}, q+\frac{1}{2}\right]$ in $\mathbb{R}^2$. Accordingly, we define $C(n; p, q)$ to be the set of all points $(x_1, y_1, \dots, x_n, y_n) \in \mathbb{R}^{2n}$ such that

- $1 \le x_k \le p$ and $1 \le y_k \le q$ for all $1 \le k \le n$, and
- $|x_k - x_\ell| \ge 1$ or $|y_k - y_\ell| \ge 1$ for all $1 \le k < \ell \le n$.

Note that the boundaries of the unit squares can intersect each other or the edges of the rectangle.

We will be working with two ways to draw a grid on the rectangle; these two grids can be thought of as dual to each other, or as offset by $\left(\frac{1}{2}, \frac{1}{2}\right)$. One grid is the integer coordinate grid. The set of possible positions of the center of one square is $[1, p] \times [1, q]$, which we can think of as having vertices at the points where both coordinates are integers, edges between vertices at distance 1, and $(p-1)(q-1)$ square 2–cells. We refer to these integer points as *coordinate grid vertices*, to the edges as *coordinate grid edges*, and to the squares as *coordinate grid squares*. Together we refer to the coordinate grid vertices, edges, and squares as *coordinate grid cells*.

The other grid is the $p \times q$ grid on the rectangle itself. Thinking of the rectangle as a $p \times q$ chessboard, we refer to the unit square centered at each coordinate grid vertex as a *board square*. For each coordinate grid cell, there is a corresponding rectangle of board squares given by taking the union of all unit squares for which the center lies on that coordinate grid cell, as shown in Figure 2. The rectangle corresponding to a coordinate grid vertex is a single board square, the rectangle corresponding to a coordinate grid edge is a pair of adjacent board squares, and the rectangle corresponding to a coordinate grid square is a 2×2 rectangle of board squares.

Let $G(n; p, q)$ be the space $([1, p] \times [1, q])^n$ with its standard cubical complex structure. Here the letter G stands for *grid*. We can think of this space as the set of configurations of labeled squares in the rectangle where the squares are allowed to overlap. As a

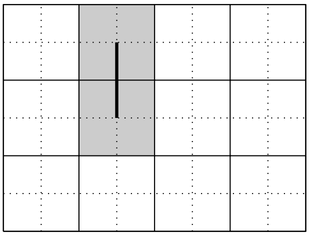


Figure 2: The coordinate grid vertices, at points with integer coordinates, are the centers of the board squares. Here a coordinate grid edge is shown with its corresponding rectangular piece.

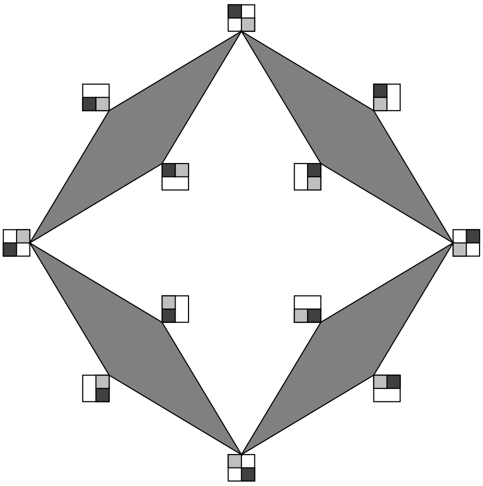


Figure 3: An illustration of the cell complex $X(2;2,2)$. The vertices of the complex are labeled by their corresponding configurations with integer coordinates. Note that in this simple case, the cell complex $X(2;2,2)$ equals the configuration space $C(2;2,2)$, while in general the cell complex $X(n;p,q)$ is only a subspace of the configuration space $C(n;p,q)$.

cubical complex, each cell of $G(n;p,q)$ corresponds to an n–tuple in which each entry is a coordinate grid cell. We can draw the cell of $G(n;p,q)$ by drawing the n corresponding rectangles of board squares. We refer to such a picture as a *rectangle arrangement*, and we refer to the n rectangles as *pieces* in the arrangement. Any list of n rectangles of board squares of sizes 1×1, 1×2, 2×1, and 2×2 is the rectangle arrangement of some cell of $G(n;p,q)$.

We define $X(n;p,q)$ to be the subcomplex of $G(n;p,q)$ consisting of all cells of $G(n;p,q)$ that are fully contained in $C(n;p,q)$. Here the letter X stands for *complex*,

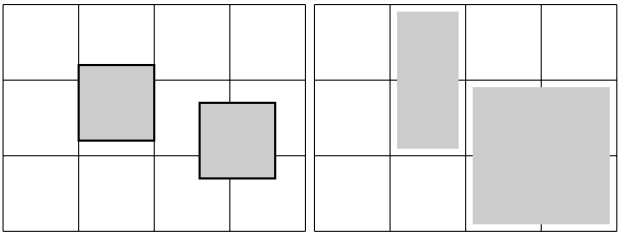


Figure 4: Any configuration where no two squares touch the same board square is in the cell of $X(n;p,q)$ corresponding to the rectangle arrangement that shows which board squares each square touches.

because $X(n;\, p,q)$ is the main cell complex that we work with throughout the paper. It is quick to check that $X(n;\, p,q)$ is equal to the set of cells in which the corresponding rectangle arrangement has none of its pieces overlapping. Given a configuration in $C(n;\, p,q)$, we can check whether it is in $X(n;\, p,q)$ by looking at each unit square in the configuration and drawing the rectangle of board squares that it intersects, as shown in Figure 4. If these rectangles are disjoint, then the configuration is in $X(n;\, p,q)$, and it is in the cell corresponding to the rectangular arrangement that we have just drawn.

3 Homotopy equivalence of the configuration space and complex

The ambient cubical complex $G(n;\, p,q)$ has three kinds of cells: some cells are fully contained in $C(n;\, p,q)$ and together form $X(n;\, p,q)$, some cells are partially in $C(n;\, p,q)$, and some cells are disjoint from $C(n;\, p,q)$. We will define a deformation retraction from $C(n;\, p,q)$ to $X(n;\, p,q)$ by considering the cells of $G(n;\, p,q)$ that are partially in $C(n;\, p,q)$ one at a time. To do this, we define local coordinates on each of these cells and give a criterion in those local coordinates for which points are in $C(n;\, p,q)$ and which points are not.

We define a function $\operatorname{snap}\colon \mathbb{R} \to \mathbb{R}$ by $\operatorname{snap}(x) = \frac{1}{2}(\lfloor x \rfloor + \lceil x \rceil)$. In other words, we have $\operatorname{snap}(k) = k$ for all $k \in \mathbb{Z}$, and if $x \in (k, k+1)$, then $\operatorname{snap}(x) = k + \frac{1}{2}$. We can also define $\operatorname{snap}\colon \mathbb{R}^d \to \mathbb{R}^d$ for any dimension d, by applying snap to each coordinate separately.

If $z = (x_1, y_1, \ldots, x_n, y_n) \in \mathbb{R}^{2n}$ is a point in the complex $G(n;\, p,q)$, then $\operatorname{snap}(z)$ is the barycenter of the unique cubical cell of $G(n;\, p,q)$ whose interior contains z. Geometrically, if (x_i, y_i) is the center of a unit square, then $\operatorname{snap}(x_i, y_i)$ is the center of the corresponding rectangle of the board squares that it touches. Note that snap is idempotent, $\operatorname{snap}(\operatorname{snap}(z)) = \operatorname{snap}(z)$, and z is a barycenter of some grid cell in $G(n;\, p,q)$ if and only if $z = \operatorname{snap}(z)$.

3.1 Containment of cells in the configuration space

We can check whether a given cell of $G(n;\, p,q)$ has empty intersection with $C(n;\, p,q)$ by looking at pairs of pieces, case by case, in its corresponding rectangle arrangement. Figure 5 shows which pairs of pieces prevent a cell from having any configurations in $C(n;\, p,q)$; in each case, the barycenter of the corresponding cell is not a configuration in $C(n;\, p,q)$. For each pair of pieces, there is no way to fit a unit square in the interior of each piece while keeping the two unit squares disjoint. (Two unit squares can fit if

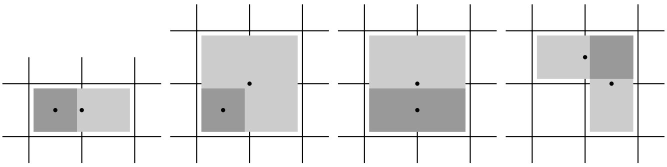


Figure 5: If the unit squares at the centers of the rectangular pieces overlap, then the corresponding cell in $G(n;p,q)$ does not contain any configurations in $C(n;p,q)$. The darker gray indicates where the two pieces overlap, and the black dots give the centers of the pieces.

they touch the boundaries of the rectangles, but the resulting configuration is in the boundary of the specified open cell, not inside it.)

Figure 6 shows the four remaining ways for two pieces in a rectangle arrangement to overlap; for these, the corresponding cell is partially in $C(n;p,q)$, and the barycenter is a configuration in $C(n;p,q)$. The following lemma summarizes how to check whether a given cell of $G(n;p,q)$ is partially in $C(n;p,q)$.

Lemma 3.1 *Let $z = \mathrm{snap}(z) = (x_1, y_1, \dots, x_n, y_n) \in G(n;p,q)$ be the barycenter of an open cell σ of $G(n;p,q)$. Then*:

(1) *σ has a nonempty intersection with $C(n;p,q)$ if and only if its barycenter z lies in the configuration space $C(n;p,q)$, or equivalently, the ℓ^∞ distance between (x_ℓ, y_ℓ) and (x_k, y_k) is at least 1 for all $1 \le k < \ell \le n$*:

$$\max(|x_\ell - x_k|, |y_\ell - y_k|) \ge 1.$$

(2) *σ is fully contained in $C(n;p,q)$, and hence a cell of $X(n;p,q)$, if and only if, for all $1 \le k < \ell \le n$, the corresponding pieces do not overlap, or equivalently,*

$$\lfloor \max(x_k, x_\ell) \rfloor > \lceil \min(x_k, x_\ell) \rceil \quad \textit{or} \quad \lfloor \max(y_k, y_\ell) \rfloor > \lceil \min(y_k, y_\ell) \rceil.$$

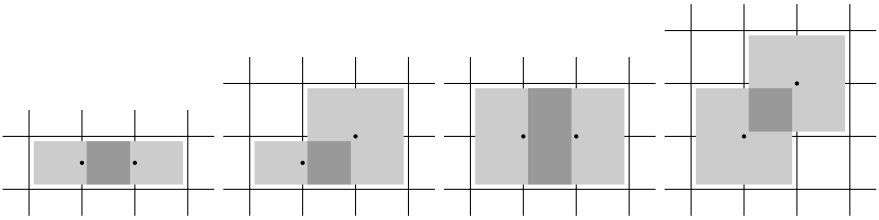


Figure 6: If a given cell of $G(n;p,q)$ is partially contained in $C(n;p,q)$, then some pair of overlapping pieces in the rectangle arrangement must overlap in one of the four ways shown. The darker gray indicates where the two pieces overlap, and the black dots give the centers of the pieces.

Proof To check the first claim, we observe that if any point $z \in \sigma$ is in $C(n; p, q)$, then $\mathrm{snap}(z) \in C(n; p, q)$ as well. This is because for any $x_1, x_2 \in \mathbb{R}$, if $x_2 - x_1 \geq 1$, then $\mathrm{snap}(x_2) - \mathrm{snap}(x_1) \geq 1$ as well.

For the second statement, note that piece k covers the board squares with centers in $[\lfloor x_k \rfloor, \lceil x_k \rceil] \times [\lfloor y_k \rfloor, \lceil y_k \rceil]$, and piece ℓ covers the board squares with centers in $[\lfloor x_\ell \rfloor, \lceil x_\ell \rceil] \times [\lfloor y_\ell \rfloor, \lceil y_\ell \rceil]$. The two pieces overlap if and only if the intervals $[\lfloor x_k \rfloor, \lceil x_k \rceil]$ and $[\lfloor x_\ell \rfloor, \lceil x_\ell \rceil]$ overlap and the intervals $[\lfloor y_k \rfloor, \lceil y_k \rceil]$ and $[\lfloor y_\ell \rfloor, \lceil y_\ell \rceil]$ also overlap. □

We say that a subcomplex of a regular CW complex is a *full subcomplex* if it is maximal with respect to its vertex set.

Corollary 3.2 *The complex $X(n; p, q)$ is a full subcomplex of the ambient cubical complex $G(n; p, q)$.*

An equivalent description for when an open cell σ is partially in $C(n; p, q)$ can be obtained from examining the cases in Figure 6. Let $b = (i_1, j_1, \dots, i_n, j_n) \in G(n; p, q)$ be the barycenter of σ, and note that the coordinates i_k and j_k are half-integers. Then σ is partially in $C(n; p, q)$ if and only if

(1) for all k and ℓ we have $\max(|i_\ell - i_k|, |j_\ell - j_k|) \geq 1$, and

(2) there is a pair k, ℓ such that

  (a) $|i_\ell - i_k| = 1$ and $|j_\ell - j_k| < 1$, and i_k and i_ℓ are not integers, or

  (b) $|j_\ell - j_k| = 1$ and $|i_\ell - i_k| < 1$, and j_k and j_ℓ are not integers, or

  (c) $|i_\ell - i_k| = |j_\ell - j_k| = 1$ and none of i_k, i_ℓ, j_k, and j_ℓ are integers.

3.2 Membership in the configuration space using local coordinates

The next lemma specifies how to use local coordinates to check, for an open cell partially in $C(n; p, q)$, whether a given point in the cell is in $C(n; p, q)$. Given an open cell σ of $G(n; p, q)$, we can specify the points $z \in \sigma$ in terms of the local coordinates $z - \mathrm{snap}(z) \in \left(-\frac{1}{2}, \frac{1}{2}\right)^{2n}$. Not every point in $\left(-\frac{1}{2}, \frac{1}{2}\right)^{2n}$ corresponds to a point in the cell, because for each coordinate of the barycenter $\mathrm{snap}(z)$ that is an integer, the corresponding coordinate in $z - \mathrm{snap}(z)$ is zero.

Let b be the barycenter of cell σ, and let $I(\sigma)$ be the set of indices of noninteger coordinates of b. The number of elements of $I(\sigma)$ is the dimension of σ. Let $\left(-\frac{1}{2}, \frac{1}{2}\right)^{I(\sigma)}$ denote the coordinate subspace of $\left(-\frac{1}{2}, \frac{1}{2}\right)^{2n}$ given by letting the $I(\sigma)$ coordinates vary and requiring the remaining coordinates (corresponding to the integer coordinates in b) to be zero. We have $z \in \sigma$ if and only if $z - b \in \left(-\frac{1}{2}, \frac{1}{2}\right)^{I(\sigma)}$, in which case $b = \mathrm{snap}(z)$.

A point of $G(n;p,q)$ is in $C(n;p,q)$ if and only if no two of the n specified squares intersect. Thus, we check the local coordinates for two of the n squares at a time to see whether those two squares overlap.

Lemma 3.3 *Let σ be an open cell of $G(2;p,q)$ that is partially in $C(2;p,q)$, and let $z=(x_1,y_1,x_2,y_2)$ be a point in the interior of σ. Then $(x_1,y_1,x_2,y_2)\in C(2;p,q)$ if and only if one of the following conditions holds*:

- $|\mathrm{snap}(x_2)-\mathrm{snap}(x_1)|=1$ *and* $(x_2-\mathrm{snap}(x_2))-(x_1-\mathrm{snap}(x_1))$ *is zero or has the same sign as* $\mathrm{snap}(x_2)-\mathrm{snap}(x_1)$, *or*
- $|\mathrm{snap}(y_2)-\mathrm{snap}(y_1)|=1$ *and* $(y_2-\mathrm{snap}(y_2))-(y_1-\mathrm{snap}(y_1))$ *is zero or has the same sign as* $\mathrm{snap}(y_2)-\mathrm{snap}(y_1)$.

In the fourth case in Figure 6, where the two pieces are 2 × 2 rectangles intersecting at one board square, either condition in the lemma may hold, so the intersection of $C(n;p,q)$ with the cell of $G(2;p,q)$ is the union of solutions to two linear inequalities. In the other three cases, the centers of the two pieces have only one coordinate that differs by 1, so the intersection of $C(n;p,q)$ with the cell is the set of solutions to one linear inequality.

Proof A point (x_1,y_1,x_2,y_2) is in $C(2;p,q)$ if and only if either $|x_2-x_1|\geq 1$ or $|y_2-y_1|\geq 1$. Note that the function snap is weakly order-preserving, meaning that $x_2-x_1\geq 0$ implies $\mathrm{snap}(x_2)-\mathrm{snap}(x_1)\geq 0$. Thus, by symmetry of x_1 and x_2 as well as (x_1,x_2) and (y_1,y_2), it suffices to show that $x_2-x_1\geq 1$ if and only if both $\mathrm{snap}(x_2)-\mathrm{snap}(x_1)=1$ and $(x_2-\mathrm{snap}(x_2))-(x_1-\mathrm{snap}(x_1))\geq 0$. The latter condition straightforwardly implies the former.

Conversely, $x_2-x_1\geq 1$ clearly implies $\mathrm{snap}(x_2)-\mathrm{snap}(x_1)\geq 1$. Further, the assumption that σ is only partially in $C(2;p,q)$ rules out the case $\mathrm{snap}(x_2)-\mathrm{snap}(x_1)>1$, as in this case we would necessarily have $\lfloor\mathrm{snap}(x_2)\rfloor>\lceil\mathrm{snap}(x_1)\rceil$, and Lemma 3.1 would imply that σ is fully contained in $C(2;p,q)$. Thus we get $\mathrm{snap}(x_2)-\mathrm{snap}(x_1)=1$ and $(x_2-\mathrm{snap}(x_2))-(x_1-\mathrm{snap}(x_1))\geq 0$ as desired. □

3.3 Construction of the deformation retraction

The next lemma gives the main step in constructing the deformation retraction from $C(n;p,q)$ to $X(n;p,q)$.

Lemma 3.4 *Let σ be an open cell of $G(n;p,q)$ that is partially in $C(n;p,q)$. Then we have that $\partial\sigma\cap C(n;p,q)$ is a deformation retract of $\overline{\sigma}\cap C(n;p,q)$.*

Proof Let $z = (x_1, y_1, \dots, x_n, y_n)$ be a point in the open cell σ. If we want to check whether z is in $C(n; p, q)$, then Lemma 3.3 gives a set of inequalities on the local coordinates $z - \operatorname{snap}(z) = (u_1, v_1, \dots, u_n, v_n)$ within the open cell σ that we can evaluate. For each pair of pieces k, ℓ in the rectangle arrangement for σ, the lemma specifies zero, one, or two inequalities of the form

- $u_k \ge u_\ell$ or $u_k \le u_\ell$,
- $v_k \ge v_\ell$ or $v_k \le v_\ell$.

The case of zero inequalities comes from the case where the two rectangular pieces do not intersect. The case of one inequality comes from the case where they intersect in one of the first three ways shown in Figure 6. And, the case of two inequalities comes from the fourth case in Figure 6, where the pieces are both 2×2 squares and they have one board square in common; in this case, the local coordinate $z - \operatorname{snap}(z)$ needs to satisfy either or both of the two inequalities. Together, we refer to the inequalities from the lemma as the inequalities associated to σ. Note that the above inequalities are stated for local coordinates, but at the same time, the coordinates of the barycenter

$$b = (i_1, j_1, \dots, i_n, j_n)$$

satisfy the same set of inequalities, even strictly. This property will be crucial for our argument.

We define the deformation retraction as follows. Let λ be large enough that $(1/\lambda)b$ is in $\left(-\frac{1}{2}, \frac{1}{2}\right)^{2n}$; we can take $\lambda = 2(p+q)$. Let m be the coordinate projection of $(1/\lambda)b$ onto $\left(-\frac{1}{2}, \frac{1}{2}\right)^{I(\sigma)}$, that is, for each integer coordinate of b, we set the corresponding coordinate of m to be zero. Since m is a positively scaled version of b, it inherits the magical quality of satisfying all of the inequalities associated to σ, and since b satisfies all those inequalities strictly, $-m$ has the magical quality of violating all of the inequalities associated to σ. (Note that every coordinate appearing in the inequalities associated to σ is in $I(\sigma)$.) Thus the point $b + m$ in σ is in the configuration space $C(n; p, q)$, while the point $b - m$ is not.

The deformation retraction now pushes every point $z \in \sigma$ outward along a ray from $b - m$ until it hits $\partial\sigma$. In other words, the vector from $b - m$ to z is given by $z - b + m$, so as time t increases from 0, we set

$$z_t = z + t(z - b + m),$$

until we reach the maximum t for which $z + t(z - b + m)$ is in $\overline{\sigma}$, and then the point no longer moves. Formally, we can define T_z to be the positive value such that

$z + T_z(z - b + m) \in \partial\sigma$. (Because σ is a cube and hence star-shaped around any interior point, for any point in σ and any nonzero vector within σ starting at that point, there is a unique nonnegative multiple of that vector that reaches $\partial\sigma$.) If d is the distance within σ from $b - m$ to $C(n; p, q)$, then the vector $z - b + m$ from $b - m$ to z has length at least d, and any vector from z to $\partial\sigma$ has length at most $\operatorname{diam}(\sigma)$, so $T_z \le (1/d)\operatorname{diam}(\sigma) \le 2n/d$.

Using this notation, the deformation retraction is defined as

$$\begin{aligned} F\colon \bar{\sigma} \cap C(n; p, q) \times \left[0, \frac{2n}{d}\right] &\to \bar{\sigma} \cap C(n; p, q), \\ (z, t) &\mapsto z + \min(t, T_z)(z - b + m). \end{aligned}$$

We still need to check that if $z \in C(n; p, q)$, then $z_t \in C(n; p, q)$ for all t, in order to ensure that the homotopy remains in $\bar{\sigma} \cap C(n; p, q)$. This is equivalent to checking that

$$z_t - b = (z - b) + t(z - b + m) = (1 + t)(z - b) + tm$$

satisfies a sufficient collection of the inequalities associated to σ. We claim that $z_t - b$ satisfies every one of the inequalities that $z - b$ satisfies; since this collection of inequalities is sufficient for z to be in $C(n; p, q)$, it is also sufficient for z_t to be in $C(n; p, q)$. Indeed, the inequalities are linear with no constant term, so given two points satisfying the inequalities, any linear combination of them with positive coefficients also satisfies the inequalities. Because $z - b$ satisfies a sufficient set of inequalities and m satisfies all of the inequalities associated to σ, this implies that $(1 + t)(z - b) + tm$ also satisfies the same set of inequalities as $z - b$. Thus, z_t is in $C(n; p, q)$ for every $t \le T_z$, and the map F that we have defined is indeed a deformation retraction from $\bar{\sigma} \cap C(n; p, q)$ to $\partial\sigma \cap C(n; p, q)$. □

Putting all the cells together, we obtain a deformation retraction from $C(n; p, q)$ to $X(n; p, q)$.

Theorem 3.5 *The subcomplex $X(n; p, q)$ is a deformation retract of the configuration space $C(n; p, q)$.*

Proof Order the cells σ of $G(n; p, q)$ that are partially in $C(n; p, q)$ so that their dimensions are nonincreasing. Then, cell by cell in order, we use Lemma 3.4 to obtain a deformation retraction from $\bar{\sigma} \cap C(n; p, q)$ to $\partial\sigma \cap C(n; p, q)$. Concatenating these deformation retractions gives a deformation retraction from $C(n; p, q)$ to the set $X(n; p, q)$ of cells completely contained in $C(n; p, q)$. □

4 Discrete Morse theory

In this section, we describe a discrete gradient vector field on $X(n; p, q)$, in the sense of Forman's discrete Morse theory [7], and characterize its critical cells. The analysis of which cells are critical is based on what we call the *apex* of a cell. The apex of a cell, as shown in Figure 7, is the 0–dimensional face that is obtained by replacing each piece by its upper-right corner square — in particular, the apex of any 0–dimensional cell is that 0–cell itself. We use discrete Morse theory to collapse our cell complex so that among the cells remaining, at most one cell has any given apex.

Theorem 4.1 *There is a discrete gradient vector field on $X(n; p, q)$ with the properties*:

(1) *Every matched pair consists of two cells with the same apex.*

(2) *Among the cells with a given apex*, *at most one cell is critical* (*unmatched*).

(3) *The matching is S_n–equivariant*: *if cells e_1 and e_2 are a matched pair*, *and we apply the same permutation to the labels in the rectangle arrangements of e_1 and e_2*, *then the two resulting cells are also a matched pair.*

The proof relies on constructing what we call the apex graph, which facilitates the enumeration of all the cells with a given apex. In Lemmas 4.2 and 4.3 we prove the basic properties of the apex graph, and in Lemma 4.4 we define the matching for Theorem 4.1 in the language of the apex graph. After that, it is straightforward to finish the proof of Theorem 4.1.

Given any rectangle arrangement, we describe the locations of the pieces according to the coordinates of their upper-right corner squares, so we say that a piece is at (i, j) if

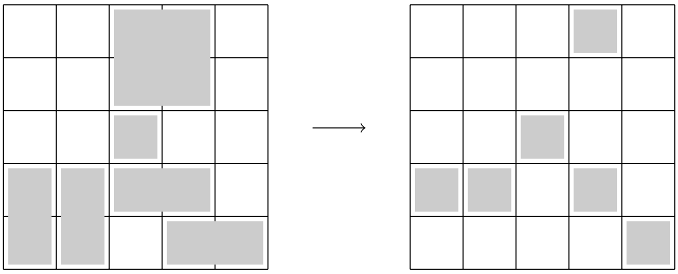


Figure 7: The apex of a rectangle arrangement replaces each piece by its upper-right corner. The correspondence does not depend on the labels of the pieces, so the labels are not shown.

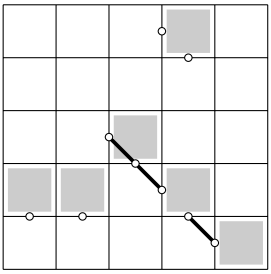


Figure 8: To find the apex graph, we place one vertex for each direction that a piece in the apex can extend, and draw edges between directions where the pieces cannot extend simultaneously.

its upper-right corner is in column i (from left to right) and row j (from bottom to top) of our $p \times q$ rectangle. (Alternatively, the center of the upper-right board square has coordinates (i, j) in the plane.) Giving the coordinates of each piece is the same as specifying the apex of our cell. To distinguish cells with the same apex, we need to specify, for each piece, whether it has height 1 or 2 and whether it has width 1 or 2. Not all these possibilities give rise to valid rectangle arrangements, because some pieces may overlap or hang off the board. For each possible apex, we construct the *apex graph* to record these possible conflicts, as shown in Figure 8.

The apex graph has at most two vertices per piece. If our apex has a piece at (i, j), then we let $\bigl(i - \frac{1}{2}, j\bigr)$ — the center of the left edge of the (i, j) board square — be a vertex of the apex graph if and only if the piece at (i, j) can have width 2 in some cell with that apex, that is, if $i > 1$ and there is no piece at $(i - 1, j)$. Similarly, we let $\bigl(i, j - \frac{1}{2}\bigr)$ — the center of the lower edge of the (i, j) board square — be a vertex if and only if there is a piece at (i, j), we have $j > 1$, and there is no piece at $(i, j - 1)$.

The edges of the apex graph record which of the width 2 or height 2 options would conflict with each other. A piece at (i, j) can have width 2 or height 2 but not both when there are no pieces at $(i-1, j)$ and $(i, j-1)$, but there is a piece at $(i-1, j-1)$. In this case we draw an edge between the vertices $\bigl(i - \frac{1}{2}, j\bigr)$ and $\bigl(i, j - \frac{1}{2}\bigr)$. The other possible conflict is between pieces at (i, j) and $(i - 1, j + 1)$. If there is no piece at $(i-1, j)$, then the (i, j) piece may have width 2, and the $(i-1, j+1)$ piece may have height 2, but not both simultaneously. In this case we draw an edge between $\bigl(i - \frac{1}{2}, j\bigr)$ and $\bigl(i - 1, j + \frac{1}{2}\bigr)$. These two types of edges give all the edges in the apex graph.

Lemma 4.2 *Each apex graph is a disjoint union of path graphs.*

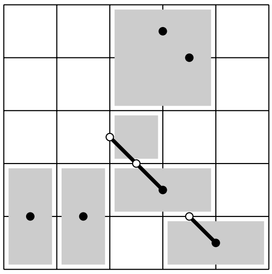


Figure 9: Cells with a given apex correspond to independent sets in the apex graph; we select the vertices corresponding to the directions where the apex pieces extend. Here, the vertices in the independent set are drawn filled, and the other vertices are drawn empty.

Proof The two types of edges have the same slope and length when drawn on the coordinate lattice. Any graph that can be drawn in this way is a disjoint union of paths. Note that some of the paths may be single vertices. □

Lemma 4.3 *The set of cells with a given apex is in bijection with the set of independent sets in its apex graph. One cell is a face of another if and only if the independent set corresponding to the first cell under this bijection is a subset of the independent set corresponding to the second cell.*

Proof For a cell with a given apex, we find the corresponding subset of vertices in the apex graph by considering each piece in the associated rectangle arrangement, say at (i,j), selecting vertex $\bigl(i-\frac{1}{2},j\bigr)$ if the piece has width 2, and selecting vertex $\bigl(i,j-\frac{1}{2}\bigr)$ if it has height 2, as in Figure 9. The construction guarantees that these are in fact vertices of the apex graph and that no two of them share an edge.

For the converse, suppose that we have an independent set in the apex graph. We select our pieces to have width 2 and/or height 2 according to which vertices are in the independent set, and we want to check whether the pieces overlap or hang off the board. Consider the (i,j) piece. It cannot hang off the board or overlap with a piece at $(i-1,j)$ or $(i,j-1)$, because the vertices corresponding to those possibilities are not in the apex graph. It cannot overlap with a piece at $(i-1,j-1)$ because that would mean choosing both vertices $\bigl(i-\frac{1}{2},j\bigr)$ and $\bigl(i,j-\frac{1}{2}\bigr)$, which would be adjacent. And, it cannot overlap with a piece at $(i-1,j+1)$, because that would mean choosing both vertices $\bigl(i-\frac{1}{2},j\bigr)$ and $\bigl(i-1,j+\frac{1}{2}\bigr)$, which would be adjacent. Symmetrically, by swapping the roles of the two pieces, we see that the piece at (i,j) also cannot overlap with the pieces at $(i+1, j)$, $(i, j+1)$, $(i+1, j+1)$, or $(i+1, j-1)$. This exhausts all the possibilities for how two pieces of width and height at most 2 might overlap, and shows that we have a bijection.

For the second property, suppose that cells e and f have the same apex. Then f is a face of e if and only if every piece of width 2 in f also has width 2 in e, and every piece of height 2 in f also has height 2 in e. This is equivalent to the condition that the independent set corresponding to f is a subset of the independent set corresponding to e. □

Thinking of the cells as independent sets in the apex graph suggests how to think about pairing them up. The dimension of a cell is equal to the number of vertices in the independent set corresponding to that cell. So, if cells e and f have the same apex, then f is a face of e with $\dim f = \dim e - 1$ if and only if the independent set of f is a subset of the independent set of e and the two sets differ by one vertex. When two independent sets differ by one vertex, we say that they are *adjacent*.

Lemma 4.4 *Given a disjoint union of paths, there is a matching on the set of independent sets such that every matched pair of independent sets are adjacent and at most one independent set is unmatched.*

Proof We start by proving the statement for one connected path of k vertices. We express the independent sets as binary strings of length k with no consecutive 1's, so that 0 indicates that the vertex in that position is not part of the independent set, and 1 indicates that the vertex is part of the independent set. The matching is defined recursively. For $k = 1$ the strings are 0 and 1, which we match as a pair. For $k = 2$ the strings are 00, 01, and 10; we match 00 with 10 and leave 01 unmatched. For $k > 2$, each string begins with 00, 10, or 010. We match the strings beginning with 00 to the strings beginning with 10 such that each matched pair differs only in the first bit. Then, for the strings beginning with 010 we ignore the first three bits and use the matching for the $k-3$ case.

The result is that for $k \equiv 1 \bmod 3$ all strings are matched, for $k \equiv 0 \bmod 3$ the only unmatched string consists of repeating copies of 010, and for $k \equiv 2 \bmod 3$ the only unmatched string consists of repeating copies of 010 followed by 01 at the end. This proves the lemma for the case of one path.

For several disjoint paths, we select some ordering on them. Given an independent set, if its restriction to each path agrees with the unmatched independent set from

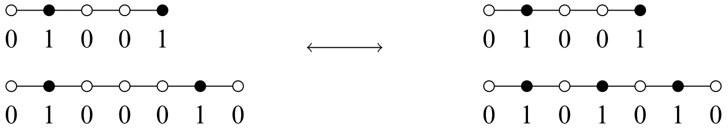


Figure 10: Given an independent set on a disjoint union of paths, to find its match we select the first component that is not critical, ignore any 010 prefixes, and flip the first bit of the remainder.

the one-path case, we leave it unmatched. Otherwise, we find the first path P where this is not true. To find the matching independent set, we keep all the other paths as they are and alter the set on P to be the matching set from the one-path case, as in Figure 10. There is an unmatched independent set if and only if none of the paths has 1 mod 3 vertices, and in this case the unmatched set corresponds to repeating 010 on each path. □

To finish the proof of Theorem 4.1, we need to check that the matching we have just defined determines a discrete gradient vector field with the properties we are looking for.

Proof of Theorem 4.1 The discrete vector field is defined as follows. Given a cell, we find its apex and the apex graph. Encoding the original rectangle arrangement as an independent set in the apex graph (Lemma 4.3), we find the matching independent set (Lemma 4.4) if there is one, and decode to get another cell with the same apex. Properties (1) and (3) are automatic from the construction, and Property (2) is a consequence of Lemma 4.4.

We still need to check that the discrete vector field is gradient. We want to show that there does not exist a cycle of cells $e_1, f_1, e_2, f_2, \dots, e_r, f_r, e_{r+1} = e_1$ such that every e_i and f_i are a matched pair (where, in particular, e_i is a face of f_i), and every f_i is a face of e_{i+1} with $\dim f_i = \dim e_{i+1} - 1$. We observe that because f_i is a face of e_{i+1}, if the apex of e_{i+1} is not equal to the apex of f_i, then it differs by moving some piece one square left or down. Every pair e_i and f_i have the same apex, so as the sequence continues, the apex keeps moving leftward and downward, making it impossible to have a cycle unless all cells in it have the same apex.

Thus we may assume that the cells $e_1, f_1, \dots, e_r, f_r$ all have the same apex. We can encode these cells as independent sets in the apex graph. To go from e_1 to f_1, we delete one vertex v from the independent set of e_1, and to go from f_1 to e_2, we add one vertex w to the independent set of f_1. Remembering the ordering of the paths and

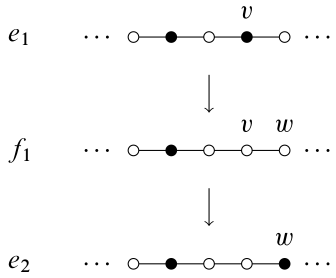


Figure 11: In a sequence of cells with the same apex, alternating between two consecutive dimensions, with consecutive pairs alternating between matched and incident, the corresponding independent sets look more and more like the unmatched set, and thus cannot cycle.

vertices in the apex graph, we observe that up until v, the independent set for f_1 agrees with the unmatched independent set, so any added vertices there would destroy the property of being an independent set. Thus w cannot be at or before v. If the vertex immediately after v is on the same path, then w can be that vertex. But w cannot be anywhere else after v, because if so, then the matching independent set to e_2, which we have supposed is f_2, would have v added rather than a vertex subtracted — it would have the wrong dimension — giving a contradiction unless w is immediately after v.

Thus we cannot have a cycle $e_1, f_1, \dots, e_r, f_r, e_{r+1} = e_1$, because each successive item agrees more and more with the unmatched set, as in Figure 11: the independent set of e_1 agrees before v, the independent set of f_1 agrees through v, the independent set of e_2 agrees through w, and so on. So, our discrete vector field is gradient and has all three desired properties. □

We prove one last theorem in this section, which helps with assessing the dimension of critical cells in the following section. For the following, we divide each unit square in the $p \times q$ grid into two *half-squares* by drawing a diagonal line from the upper-right to the lower-left corner.

Theorem 4.5 *There is a function r that assigns a set of half-squares to each vertex of the apex graph, with the properties:*

(1) *For any vertex v, the set $r(v)$ has four half-squares if v is the only vertex of a path, three half-squares if v is the first or last vertex of a path, and two half-squares otherwise.*

(2) *The sets $r(v)$ are disjoint for all v.*

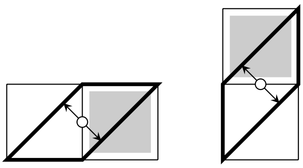


Figure 12: Each vertex is assigned the two half-squares it touches. If vertex v has neighbors in both directions, then $r(v)$ contains only these two half-squares.

Proof Recalling that we can draw each vertex of the apex graph as the midpoint of an edge between a square occupied by an apex piece and an unoccupied square, we set $r(v)$ to contain both of the half-squares that v touches, as in Figure 12.

We think of the vertices as ordered first by the sum of coordinates and then by the column coordinate, so that the ordering starts in the lower-left corner and goes right and down along diagonals. If v is the first or last vertex of a path, we need to find another half-square to add to $r(v)$. There are several cases, shown in Figure 13:

(1) If v is the first vertex of a path and is on a vertical edge, we add in the remainder of the square to the left of v.

(2) If v is the last vertex of a path and is on a horizontal edge, we add in the remainder of the square below v.

(3) If v is the first vertex of a path and is on a horizontal edge and there is no vertex on the preceding (above-left) edge, we add in the remainder of the square above v.

(4) If v is the last vertex of a path and is on a vertical edge and there is no vertex on the following (below-right) edge, we add in the remainder of the square to the right of v.

(5) If v is the first vertex of a path and is on a horizontal edge and there is a (nonadjacent) vertex on the preceding edge, we add the half-square to the left of the square below v.

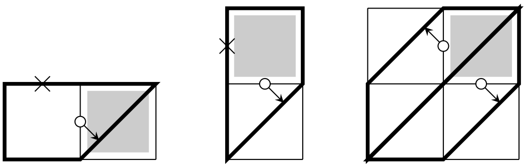


Figure 13: For the first vertex of a path, rules (1), (3), and (5) specify how to add a third half-square; the × symbol indicates an absence of vertex. The third picture also shows an instance of rule (6). For the last vertex of a path, rules (2), (4), and (6) are analogous.

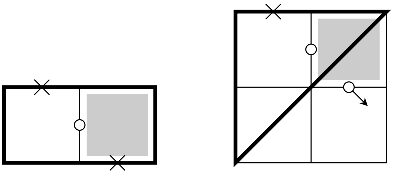


Figure 14: If the only vertex of a path is on a vertical edge, rules (1) and (4) or rules (1) and (6) assign four half-squares to that vertex. If the vertex is on a horizontal edge, rules (2) and (3) or rules (2) and (5) are analogous.

(6) If v is the last vertex of a path and is on a vertical edge and there is a (nonadjacent) vertex on the following edge, we add the half-square below the square to the left of v.

In the case where v is the only vertex of the path, if v is on a vertical edge, then either rules (1) and (4) or rules (1) and (6) apply, as shown in Figure 14, and if v is on a horizontal edge, then either rules (2) and (3) or rules (2) and (5) apply, so that $r(v)$ has four half-squares in total. This completes the definition of $r(v)$.

We need to check that no half-square has been assigned twice. To do this, we consider the assignment from the point of view of each square. Consider a square that is occupied by a piece in the apex arrangement. It may have vertices on its left or lower edges. If it has both vertices, then half of the square is assigned to each vertex. If it has one vertex, then all of the square is assigned to that vertex, by rule (3) or rule (4). If it has no vertex, then none of the rules assign that square to any vertex.

Similarly, consider a square that is unoccupied in the apex arrangement. It may have vertices on its right or upper edges. If it has both vertices, then half of the square is assigned to each vertex. If it has one vertex, then all of the square is assigned to that vertex, by rule (1) or rule (2). (Note that rules (1) and (2) cannot apply to an unoccupied square with both vertices, because in this case the two vertices would be adjacent.) If our unoccupied square has no vertices, then we divide the square in half. The lower-right half gets assigned by rule (5) to the same vertex (if any) as the half-square to its right, and the upper-left half gets assigned by rule (6) to the same vertex (if any) as the half-square above it.

In each case, only one rule can apply to each half-square, so each half-square can be assigned to only one vertex. □

5 Homology-vanishing theorems

The existence of the cell complex $X(n;\, p,q)$ and the discrete gradient on it allow us to establish a number of homology-vanishing results.

Theorem 5.1 *If $j > pq-n$, then $H_j[C(n;\, p,q)] = 0$.*

Proof This is almost immediate from the homotopy equivalence $C(n;p,q) \sim X(n;p,q)$. Consider the dimensions of the cells in $X(n;\, p,q)$. A cell is indexed by a collection of n nonoverlapping rectangular pieces in a $p \times q$ grid. A 1×1 piece contributes 0 to the dimension of the cell, a 1×2 or 2×1 piece contributes 1, and a 2×2 piece contributes 2. The total area of the pieces is at most pq. So the largest dimension of a cell is at most $pq-n$. By the definition of cellular homology, there is no homology above the dimension of the cell complex itself, so $H_j[C(n;\, p,q)] = 0$ for $j > pq-n$. □

Theorem 5.2 *If $j > n$, then $H_j[C(n;\, p,q)] = 0$.*

Proof This follows from the properties of the discrete gradient described in Section 4. Every cell is indexed by a collection of nonoverlapping rectangular pieces, and each piece is one of 1×1, 1×2, 2×1, or 2×2. The analysis of the gradient shows that there are no critical cells indexed by a collection of pieces including a 2×2 piece. (When such a cell is encoded as an independent set in the apex graph as in Lemma 4.3, the 2×2 piece corresponds to two vertices of the independent set that are consecutive but not adjacent. However, the proof of Lemma 4.4 implies that in a critical cell, the first vertex of each path is never part of the corresponding independent set.) Hence the dimension of a critical cell is at most n. □

Theorem 5.3 *If $j > \frac{1}{3}pq$, then $H_j[C(n;\, p,q)] = 0$.*

Proof This follows from Theorem 4.5. A critical cell corresponds to an independent set that on each path looks like $010\ldots010$ or $010\ldots01$ depending on whether the number of vertices in the path is 0 or 2 mod 3. The dimension of the critical cell is the number of vertices in the independent set. If the path has k vertices, then the independent set has $\frac{1}{3}k$ vertices in the first case, and has $\frac{1}{3}(k+1)$ vertices in the second case.

For a path of k vertices, the theorem allocates half-squares with a total area of $k+1$ to the vertices of the path. The independent set for that path contributes at most $\frac{1}{3}(k+1)$ to the dimension of the critical cell. Thus, in total, the dimension of the critical cell is at most one third of the total area allocated to all the paths in the apex graph, and thus is at most $\frac{1}{3}pq$. □

Putting together Theorems 5.1, 5.2, and 5.3, we have proved Theorem 1.1.

Theorem 1.1 (homology vanishing theorem) *We have*:

(1) *If $j > pq - n$, then $H_j[C(n; p, q)] = 0$.*

(2) *If $j > n$, then $H_j[C(n; p, q)] = 0$.*

(3) *If $j > pq/3$, then $H_j[C(n; p, q)] = 0$.*

Equivalently, on the large scale, if $H_j[C(n; p, q)] \neq 0$ then $y \leq \min\bigl\{1 - x, x, \frac{1}{3}\bigr\}$.

6 Nontrivial homology

In this section, our main aim is to prove Theorem 1.3. We give several explicit constructions of nontrivial cycles, and then a method for interpolating between parameters.

Lemma 6.1 *The points $\bigl(\frac{1}{2}, \frac{1}{4}\bigr)$ and $\bigl(\frac{3}{4}, \frac{1}{4}\bigr)$ are attainable.*

Proof Figure 15 shows a cycle in $H_1[C(2; 2, 2)]$. More precisely, the figure illustrates a piecewise-linear map $i : S^1 \to C(2; 2, 2)$, where we linearly interpolate at constant speed between the positions shown. Then if $[\sigma]$ is a generator of $H_1(S^1)$, the cycle we are describing is the image $i_*([\sigma])$.

To show that this cycle is nontrivial, consider the map $f : C(2; 2, 2) \to S^1$, where one takes the angle the line from the center of square 1 to the center of square 2 makes

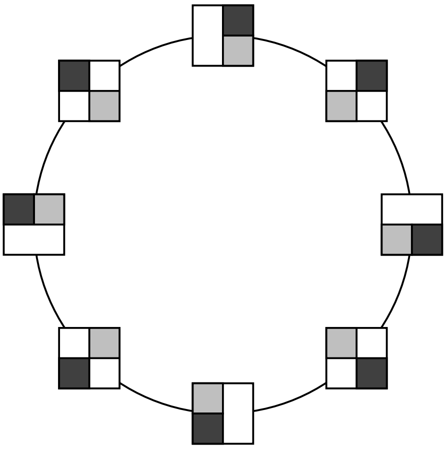


Figure 15: A nontrivial cycle in $H_1[C(2; 2, 2)]$. This realizes the point $(x, y) = \bigl(\frac{1}{2}, \frac{1}{4}\bigr)$.

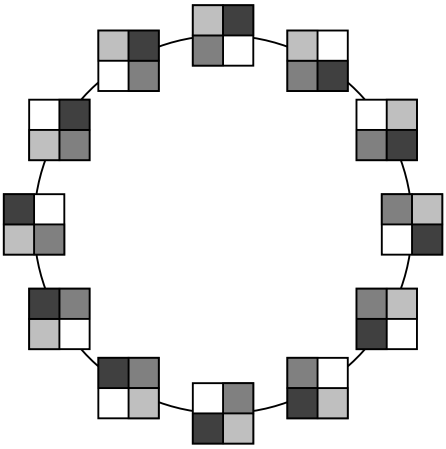


Figure 16: A nontrivial cycle in $H_1[C(3;2,2)]$. This realizes the point $(x,y)=\left(\frac{3}{4},\frac{1}{4}\right)$.

with the x–axis. In other words, define

$$f(x_1,y_1,x_2,y_2)=\frac{1}{\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}}(x_2-x_1,y_2-y_1).$$

The composition $f\circ i$ is a degree-one map $S^1\to S^1$, and in particular the induced map $(f\circ i)_*$ is an isomorphism on H_1. So then i must be injective on H_1.

Similarly, Figure 16 shows a cycle in $H_1[C(3;2,2)]$. The figure illustrates a piecewise-linear map $i:S^1\to C(3;2,2)$, and the cycle we are interested in is the image. This also represents a nontrivial cycle in $H_1[C(3;2,2)]$. Indeed, we have a natural projection map to $C(2;2,2)$ where one forgets the coordinates of the third square, and then the argument above shows that the image of the cycle is still nontrivial in this projection. □

The following lemma will be superseded later in this section by a stronger result, but we present the lemma and proof as a warmup, and we will also reuse the main construction in its proof later.

Lemma 6.2 *The point*

$$(x,y)=\left(\frac{k}{k^2},\frac{k-1}{k^2}\right)=\left(\frac{1}{k},\frac{1}{k}-\frac{1}{k^2}\right)$$

is attainable for every $k\geq 1$.

Proof Consider first Figure 17. We illustrate a 2×2 square and a 1×1 square orbiting each other in a 3×3 grid, as in Figure 15. We can then put two 1×1 squares inside the

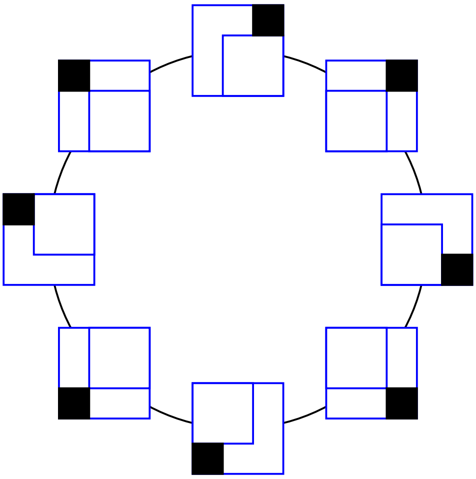


Figure 17: A 2×2 square and a 1×1 square orbiting each other in a 3×3 grid.

2×2 square, and these can orbit each other independently. So, together these motions describe a map $i: T^2 \to C(3; 3, 3)$. This map is illustrated in Figures 18 and 19. By induction, we can embed a $(k-1)$–dimensional torus realizing a nontrivial cycle in $C(k; k, k)$ for every $k \geq 2$.

The same argument as before gives that this represents a nontrivial class in $H_2[C(3; 3, 3)]$. Indeed, compose with a map $f: C(3; 3, 3) \to T^2 = S^1 \times S^1$ which assigns to the first coordinate the angle between the line segment from the center of square 1 to the center of square 2 and the x–axis. Similarly, the map assigns to the second coordinate the angle between the line segment from the center of square 1 to the center of square 3 and the x–axis. This is a degree-one map $T^2 \to T^2$. The induced map $(f \circ i)_*$ is an isomorphism on homology, and so i_* is injective. □

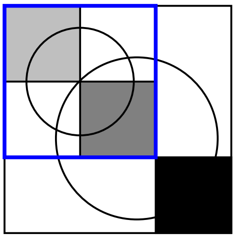


Figure 18: A map $T^2 \to C(3; 3, 3)$. The light gray and dark gray squares orbit each other inside the blue 2×2 square as in Figure 15, while the black and blue squares orbit each other independently as in Figure 17.

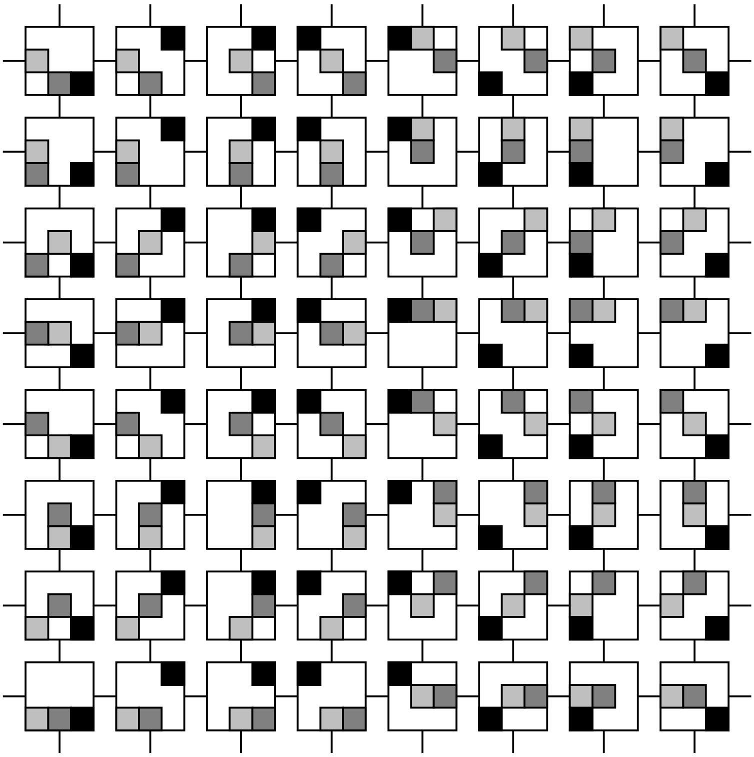


Figure 19: Another view of the map $i\colon T^2 \to C(3;3,3)$ visualized in Figure 18. The image of the fundamental class of the torus is a nontrivial cycle in $H_2[C(3;3,3)]$.

By Lemma 6.2, there are infinitely many points realized on the parabola $y = x - x^2$. The following lemma improves on that result, showing that there are infinitely many points realized on the parabola $y = x - \frac{8}{9}x^2$.

Lemma 6.3 *The point*

$$(x,y) = \left(\frac{3}{4k}, \frac{3}{4k} - \frac{1}{2k^2}\right)$$

is attainable for every $k \geq 1$.

Proof The case $k = 1$ is already covered by Lemma 6.1. For any $k \geq 2$, we can embed a $(k-1)$–dimensional torus in $C(k;k,k)$. Now consider the configuration space $C(3k;2k,2k)$. We can divide the $2k \times 2k$ grid into four $k \times k$ grids. Inside each, we use k squares to embed a $(k-1)$–torus as in the proof of Lemma 6.2. This describes a $(3k-3)$–torus, and the three $k \times k$ squares can orbit each other in the $2k \times 2k$ grid, giving one more dimension. So putting it all together, we have a $(3k-2)$–torus. This

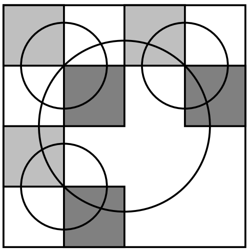


Figure 20: A map $T^4 \to C(6;4,4)$. The three pairs of squares orbit each other in three 2×2 squares, and these three 2×2 squares orbit each other in the 4×4 square, as in Figure 16. The image of this map gives a nontrivial cycle in $H_4[C(6;4,4)]$, realizing the point $(x,y) = \bigl(\frac{3}{8},\frac{1}{4}\bigr)$.

realizes the point

$$(x,y) = \left(\frac{3k}{4k^2}, \frac{3k-2}{4k^2}\right) = \left(\frac{3}{4k}, \frac{3}{4k} - \frac{1}{2k^2}\right).$$

The case $k=2$ is illustrated in Figure 20, and the case $k=3$ in Figure 21. □

Lemma 6.4 *The point* $(x,0)$ *is attainable for every rational* x *with* $0 \le x \le 1$.

Proof Indeed, suppose x is a rational point in $[0,1]$, and write $x = a/b$, where a is a nonnegative integer, b is a positive integer, and $a \le b$. Set $n = ab$ and $p = q = b$. By assumption, we have $a \le b$, so $n \le pq$ and the configuration space $C(n;p,q)$ is nonempty, so $H_0[C(n;p,q)] \ne 0$. □

Finally, we show that we can rationally interpolate between all the points we have described. Let S be the set of points

$$S = \left\{ \left(\frac{3}{4k}, \frac{3}{4k} - \frac{1}{2k^2}\right) \,\middle|\, k \ge 1 \right\}.$$

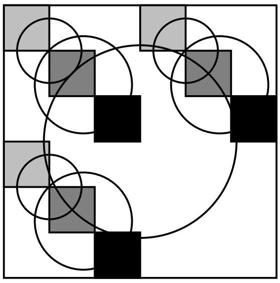


Figure 21: A map $T^7 \to C(9;6,6)$, realizing the point $(x,y) = \bigl(\frac{1}{4},\frac{7}{36}\bigr)$.

Let I be the closed interval

$$I = \{(x,y) \mid 0 \le x \le 1 \text{ and } y \ge 0\}.$$

Theorem 1.3 (large-scale homology nonvanishing theorem) *If (x,y) is any rational point in the convex hull of $S \cup I$, then there exist n, p, q, and j such that $x = n/pq$, $y = j/pq$, and $H_j[C(n;p,q)] \ne 0$.*

Proof By Cartheodory's theorem, if (r_1,r_2) is in the convex hull of $S \cup I$, then (r_1,r_2) is in the convex hull of three points of $S \cup I$. Write (r_1,r_2) as a rational convex combination of these three points, ie

$$(r_1,r_2) = \lambda_1(u_1,v_1) + \lambda_2(u_2,v_2) + \lambda_3(u_3,v_3)$$

with

(1) $(u_1,v_1),(u_2,v_2),(u_3,v_3) \in S \cup I$,

(2) $0 \le \lambda_1,\lambda_2,\lambda_3 \le 1$ with $\lambda_1+\lambda_2+\lambda_3 = 1$, and

(3) $\lambda_1,\lambda_2,\lambda_3$ all rational.

By the previous lemmas, (u_i,v_i) is realizable as a nontrivial homology class for hard squares in a square for $i = 1,2,3$. Let n_i, p_i, and j_i be such that $u_i = n_i/p_i^2$ and $v_i = j_i/p_i^2$ for $i = 1,2,3$. Let $\lambda_i = a_i/b_i$ for $i = 1,2,3$. Set

$$P = p_1p_2p_3, \quad B = b_1b_2b_3, \quad R = PB,$$

then let

$$N = r_1R^2 \quad \text{and} \quad J = r_2R^2.$$

If we can find a nontrivial class in

$$H_J[C(N;R,R)],$$

we are done.

Partition the $R \times R$ square into B^2 smaller squares, each of dimension $P \times P$. In a λ_1 fraction of these smaller squares (ie in $\lambda_1B^2 = a_1b_1b_2^2b_3^2$ of them), we realize (u_1,v_1) as follows. Further partition each $P \times P$ square into $p_2^2p_3^2$ squares, of dimension $p_1 \times p_1$. In each of these squares, we can place n_1 squares and can then describe a map from a torus giving a nontrivial class in $H_{j_1}[C(n_1;p_1,p_1)]$. So in total, we place

$$(a_1b_1b_2^2b_3^2)(p_2^2p_3^2)n_1 = (\lambda_1B^2)\left(\frac{P^2}{p_1^2}\right)n_1 = \lambda_1\left(\frac{n_1}{p_1^2}\right)(P^2B^2) = \lambda_1u_1R^2$$

squares, and get a map from the torus of dimension

$$(a_1b_1b_2^2b_3^2)(p_2^2p_3^2)j_1 = (\lambda_1B^2)\left(\frac{P^2}{p_1^2}\right)j_1 = \lambda_1\left(\frac{j_1}{p_1^2}\right)(P^2B^2) = \lambda_1v_1R^2.$$

Similarly, in a λ_2 fraction of these $P \times P$ squares we can realize (u_2, v_2) by dividing up into $p_1^2 p_3^2$ smaller squares of dimension $p_2 \times p_2$, and in a λ_3 fraction of the $P \times P$ squares we realize (u_3, v_3).

Altogether, we have used

$$\lambda_1 u_1 R^2 + \lambda_2 u_2 R^2 + \lambda_3 u_3 R^2 = r_1 R^2 = N$$

squares, and defined an embedded torus of dimension

$$\lambda_1 v_1 R^2 + \lambda_2 v_2 R^2 + \lambda_3 v_3 R^2 = r_2 R^2 = J.$$

This describes a cycle in

$$H_J[C(N; P, P)],$$

as desired. The cycle is nontrivial as before — we can compose with a map to T^j such that the composed map $T^j \to T^j$ has degree one. □

7 Betti number computations for small n, p, q

We compute the Betti numbers $\beta_j[C(n; p, q)]$ for $n \le 6$ and $p \le q \le n$. These are provided in Table 1. Another view of the Betti numbers for $n = 6$ and $j = 2$ with p and q varying is illustrated in Figure 22. Finally, in Table 2 we record information about the size of the complex $X(n; p, q)$ in the form of its *f–vector* $(f_0, f_1, f_2, \dots)$, where f_i is the number of i–dimensional cells in $X(n; p, q)$. All of our computations are using coefficients in the prime field $\mathbb{Z}/2\mathbb{Z}$.

For our computations we employ three different software packages, and we dedicate a small section to each one. The first is a Python/Sage Jupyter notebook which uses the discrete Morse vector field of Section 4. The second is a branch of the pyCHomP package, available at [9] specifically for computing the Betti numbers for these configuration spaces. The third is the DIPHA package with a custom script to build the configuration cell complex. Finally, note that in the case when $n = q$ the configuration space $C(n; p, q)$ is homotopy equivalent to the configuration space of disks in a strip addressed in [2]; in this case, one can use the Salvetti complex to compute the Betti numbers as done in [2].

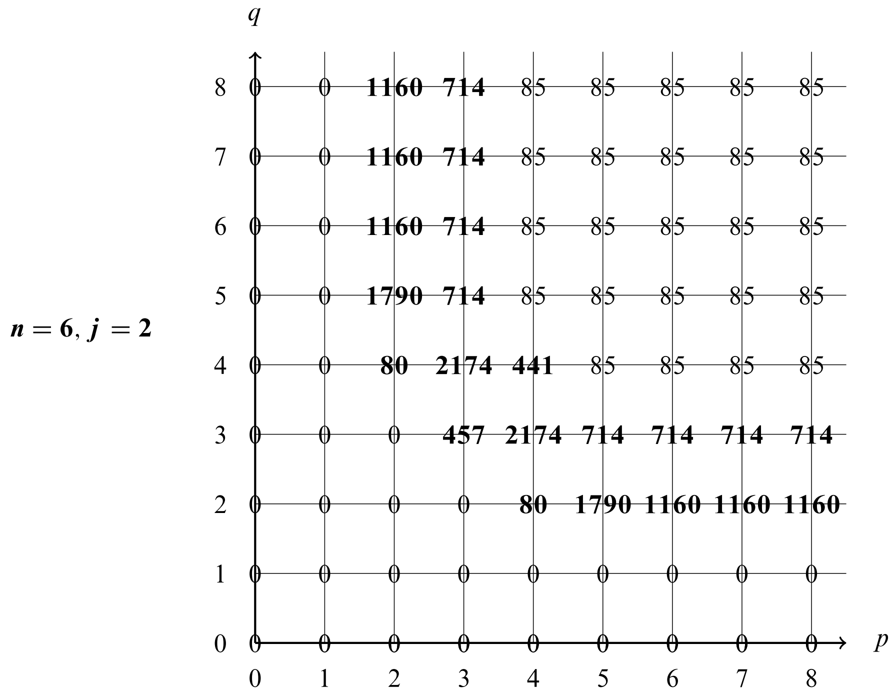
7.1 Discrete Morse theory Sage notebook

Using the discrete gradient vector field from Section 4, we compute the collapsed Morse chain complex for $X(n; p, q)$ as follows. The idea is first to find the critical

n	p	q	β_0	β_1	β_2	β_3	β_4	β_5
2	2	2	1	1	0	0	0	0
3	2	2	**2**	**2**	0	0	0	0
3	2	3	1	**7**	0	0	0	0
3	3	3	1	3	2	0	0	0
4	2	2	**24**	0	0	0	0	0
4	2	3	1	**49**	0	0	0	0
4	2	4	1	**31**	**6**	0	0	0
4	3	3	1	**12**	**11**	0	0	0
4	3	4	1	6	**29**	0	0	0
4	4	4	1	6	11	6	0	0
5	2	3	**2**	**122**	0	0	0	0
5	2	4	1	**161**	**40**	0	0	0
5	2	5	1	**111**	**110**	0	0	0
5	3	3	1	**68**	**67**	0	0	0
5	3	4	1	10	**249**	0	0	0
5	3	5	1	10	**169**	**40**	0	0
5	4	4	1	10	**71**	**62**	0	0
5	4	5	1	10	35	**146**	0	0
5	5	5	1	10	35	50	24	0
6	2	3	**720**	0	0	0	0	0
6	2	4	1	**2241**	**80**	0	0	0
6	2	5	1	**351**	**1790**	0	0	0
6	2	6	1	**351**	**1160**	**90**	0	0
6	3	3	1	**458**	**457**	0	0	0
6	3	4	1	15	**2174**	0	0	0
6	3	5	1	15	**714**	**1429**	0	0
6	3	6	1	15	**714**	**780**	**80**	0
6	4	4	1	15	**441**	**457**	**30**	0
6	4	5	1	15	85	**1541**	**30**	0
6	4	6	1	15	85	**1066**	**275**	0
6	5	5	1	15	85	**465**	**394**	0
6	5	6	1	15	85	225	**875**	0
6	6	6	1	15	85	225	274	120

Table 1: The Betti numbers of $C(n; p, q)$ for $2 \leq n \leq 6$. The homological liquid regime is indicated in bold.

cells and then to compute their boundaries in the Morse complex. However, it turns out that most of this process depends very little on p and q. Thus, in order to compute for various p and q without duplicating effort, we first compute the Morse complex for $X(n; n, n)$. The Morse complex of each $X(n; p, q)$ for $1 \leq p$ and $q \leq n$ turns out to be a subcomplex of the Morse complex for $X(n; n, n)$, obtained by selecting only the critical cells for which the apex is in the $p \times q$ rectangle. This is because of the properties of



q \ p	0	1	2	3	4	5	6	7	8
8	0	0	**1160**	**714**	85	85	85	85	85
7	0	0	**1160**	**714**	85	85	85	85	85
6	0	0	**1160**	**714**	85	85	85	85	85
5	0	0	**1790**	**714**	85	85	85	85	85
4	0	0	**80**	**2174**	**441**	85	85	85	85
3	0	0	0	**457**	**2174**	**714**	**714**	**714**	**714**
2	0	0	0	0	**80**	**1790**	**1160**	**1160**	**1160**
1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Figure 22: Another view of the Betti numbers. Let $n = 6$ and $j = 2$, and let p and q be the horizontal and vertical axes. Then the solid regime is in the lower-left, the gas regime is in the upper-right, and the liquid regime (in bold) is in between. If $p \geq n$, then the inclusion map $C(n; p, q) \hookrightarrow C(n; p+1, q)$ induces an isomorphism on homology. Similarly, if $q \geq n$ then the inclusion map $C(n; p, q) \hookrightarrow C(n; p, q+1)$ induces an isomorphism on homology.

our discrete gradient vector field. Namely, we know that if a cell's apex fits into a $p \times q$ rectangle, so does every boundary cell of that cell (the apex takes upper-right corners, and the $p \times q$ rectangle grows from the lower-left). Together with the fact that every two paired cells have the same apex, this implies that the $X(n; p, q)$ Morse complex is a subcomplex of the $X(n; p', q')$ Morse complex whenever $p \leq p'$ and $q \leq q'$. The construction of the discrete gradient vector field guarantees that no apex that skips a row or column can be the apex of a critical cell — this is because every apex graph with an isolated vertex has an even number of independent sets — so the $X(n; n, n)$ Morse complex is sufficiently large to contain the Morse complexes for all $X(n; p, q)$.

Thus, the code computes as follows. First, we list all ways of placing n squares in an $n \times n$ grid. Then, we check which of these arrangements are the apex of a critical cell. For each critical cell, we compute its boundary in the Morse complex by

n	p	q	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
2	2	2	12	16	4						
3	2	2	24	24							
3	2	3	120	252	144	18					
3	3	3	504	1512	1560	624	72				
4	2	3	360	672	264						
4	2	4	1680	4800	4464	1488	120				
4	3	3	3024	10 080	11 520	5184	720				
4	3	4	11 880	48 960	76 608	56 448	19 536	2688	96		
4	4	4	43 680	209 664	402 336	393 120	206 232	56 640	7728	576	24
5	2	3	720	840							
5	2	4	6720	18 000	14 280	3120					
5	2	5	30 240	109 200	141 600	79 200	17 520	960			
5	3	3	15 120	50 400	55 200	22 080	2160				
5	3	4	95 040	42 8400	735 840	600 600	234 720	38 040	1680		
5	3	5	360 360	1 887 600	3 979 800	4 322 880	2 561 160	800 400	114 960	5280	
5	4	4	524 160	2 882 880	6 448 200	7 538 400	4 928 640	1 793 280	345 240	33 120	1440
6	3	3	60 480	181 440	161 280	40 320					

Table 2: The f–vectors for $X(n;\, p,q)$ for small n, p, and q.

applying discrete gradient flow to its original boundary in $X(n;n,n)$. Doing this for every critical cell gives all boundary coefficients for the Morse complex of $X(n;n,n)$, computed as integers with signs. Then we restrict to smaller $p\times q$ rectangles, producing subcomplexes of the Morse complex. For each p and q, we compute the Betti numbers from the ranks of the boundary matrices and the dimensions of the chain groups; because the matrices have integer entries, to specify the coefficient field for homology, we only need to specify the field for the rank computation, which can be done over $\mathbb{Q}$ or modulo any choice of prime. The Sage notebook is available online.[1]

We found that the code runs quickly for $n \le 5$ and agrees with our other computation methods; for $n \le 6$ it becomes slow and would require more speed optimization.

7.2 PyCHomP

We briefly review the computations involved in pyCHomP, which may be used to compute the homology of $X(n;\, p,q)$ with $\mathbb{Z}/2\mathbb{Z}$ coefficients.

Let $(P,\le)$ be the total order with $P=\{0,1\}$ and $0\le 1$. As $X(n;\, p,q)$ is a subcomplex of $G(n;\, p,q)$, there is an order-preserving map ν from the face poset $(G(n;\, p,q),\le)$

[1] https://gist.github.com/ubauer/87e7ee1462966127e9837c4747829a4a

to $(P, \leq)$ given by

$$\nu(\sigma) = \begin{cases} 0 & \text{if } \sigma \in X(n; p, q), \\ 1 & \text{if } \sigma \notin X(n; p, q). \end{cases}$$

In order to construct the map ν, we use Lemma 3.1 to determine whether a cell belongs to the cubical complex $X(n; p, q)$. The complex $G(n; p, q)$ together with the map ν defines a P–graded cell complex; see [8]. PyCHomP uses iterated algebraic–discrete Morse theory to reduce $G(n; p, q)$ to a (chain-homotopy equivalent) P–graded cell complex $(A(n; p, q), \mu)$ characterized by the condition that $\partial^A|_{\mu^{-1}(p)} = 0$ for $p \in P$. This condition implies that the j–dimensional Betti number of $X(n; p, q)$ is precisely the number of j–dimensional cells in $\mu^{-1}(0)$; see [8, Example 4.30] for more detail.

Theorems 5.1–5.3 suggest that speedups can be obtained for any code which computes homology starting from the complex $X(n; p, q)$ by not considering cells above a certain dimension. The branch of pyCHomP available at [9] incorporates these speedups; pyCHomP is able to compute the Betti numbers for all the examples in Table 1. A Jupyter notebook which sets up the computation of Betti numbers for $X(n; p, q)$ is available online.[2]

7.3 DIPHA

Finally, we describe the homology computation of $X(n; p, q)$ using DIPHA, a software package for computing persistent homology in a distributed setting [4; 5]. DIPHA supports the computation of persistent homology for *lower star filtrations* of cubical grids such as $G(n; p, q)$. The data determining a lower star filtration is a real-valued function f on the vertices of $G(n; p, q)$, ie the integer points in $([1, p] \times [1, q])^n$. The filtration then consists of the full subcomplexes of the ambient cube complex $G(n; p, q)$ induced by sublevel sets $f^{-1}(-\infty, t]$ of the function f.

Our computations make use of the fact that $X(n; p, q)$ is a full subcomplex of the ambient cube complex $G(n; p, q)$ (see Corollary 3.2). In other words, the complex $X(n; p, q)$ is determined by the set of all configurations in $C(n; p, q)$ with integer coordinates. Thus, it suffices to enumerate all permutations of all n–element subsets of the $p \times q$ possible integer coordinates for the cubes. The input to DIPHA consists of the characteristic function of this set as a subset of all vertices of $G(n; p, q)$. A Mathematica file for generating the input to DIPHA is available online.[3]

[2] https://github.com/kellyspendlove/pyCHomP/blob/config/doc/config/ConfigSpacePaper.ipynb

[3] https://gist.github.com/ubauer/01934ad494eeb6e9ef66ca14e0301fe9

References

[1] **H Alpert**, *Restricting cohomology classes to disk and segment configuration spaces*, Topology Appl. 230 (2017) 51–76 MR Zbl

[2] **H Alpert**, **M Kahle**, **R MacPherson**, *Configuration spaces of disks in an infinite strip*, J. Appl. Comput. Topol. 5 (2021) 357–390 MR Zbl

[3] **Y Baryshnikov**, **P Bubenik**, **M Kahle**, *Min-type Morse theory for configuration spaces of hard spheres*, Int. Math. Res. Not. 2014 (2014) 2577–2592 MR Zbl

[4] **U Bauer**, **M Kerber**, **J Reininghaus**, *DIPHA, a distributed persistent homology algorithm* (2014) Available at `https://github.com/DIPHA/dipha`

[5] **U Bauer**, **M Kerber**, **J Reininghaus**, *Distributed computation of persistent homology*, from "Proceedings of the 16th workshop on algorithm engineering and experiments (ALENEX '14)", SIAM, Philadelphia, PA (2014) 31–38 Zbl

[6] **G Carlsson**, **J Gorham**, **M Kahle**, **J Mason**, *Computational topology for configuration spaces of hard disks*, Phys. Rev. E 85 (2012) art. id. 011303

[7] **R Forman**, *Morse theory for cell complexes*, Adv. Math. 134 (1998) 90–145 MR Zbl

[8] **S Harker**, **K Mischaikow**, **K Spendlove**, *A computational framework for connection matrix theory*, J. Appl. Comput. Topol. 5 (2021) 459–529 MR Zbl

[9] **S Harker**, **K Spendlove**, *pyCHomP* (*computational homology project with python bindings*) (2020) `https://github.com/kellyspendlove/pyCHomP/tree/config`

[10] **W W Johnson**, **W E Story**, *Notes on the "15" puzzle*, Amer. J. Math. 2 (1879) 397–404 MR Zbl

[11] **L Plachta**, *Configuration spaces of squares in a rectangle*, Algebr. Geom. Topol. 21 (2021) 1445–1478 MR Zbl

[12] **D Sonneveld**, **J Slocum**, *The* 15 *puzzle*: *how it drove the world crazy*, Solcum Puzzle Foundation, Beverly Hills, CA (2006)

Department of Mathematics and Statistics, Auburn University
Auburn, AL, United States

Department of Mathematics, Technical University of Munich
Munich, Germany

Department of Mathematics, Ohio State University
Columbus, OH, United States

School of Mathematics, Institute for Advanced Study
Princeton, NJ, United States

Mathematical Institute, University of Oxford
Oxford, United Kingdom

`hcalpert@auburn.edu, mail@ulrich-bauer.org, mkahle@math.osu.edu,`
`rdm@ias.edu, kelly.spendlove@gmail.com`

Received: 20 November 2020 Revised: 4 October 2021

Algebraic & Geometric Topology **23**:6 (2023) 2627–2672
DOI: 10.2140/agt.2023.23.2627
Published: 7 September 2023

Nonorientable link cobordisms and torsion order in Floer homologies

SHERRY GONG
MARCO MARENGON

We use unoriented versions of instanton and knot Floer homology to prove inequalities involving the Euler characteristic and the number of local maxima appearing in nonorientable cobordisms, which mirror results of a recent paper by Juhász, Miller and Zemke concerning orientable cobordisms. Most of the subtlety in our argument lies in the fact that maps for nonorientable cobordisms require more complicated decorations than their orientable counterparts. We introduce unoriented versions of the band unknotting number and the refined cobordism distance and apply our results to give bounds on these based on the torsion orders of the Floer homologies. Finally, we show that the difference between the unoriented refined cobordism distance of a knot K from the unknot and the nonorientable slice genus of K can be arbitrarily large.

57K18; 57K16

1 Introduction

A classical problem in low-dimensional topology is the study of embedded orientable surfaces in 4–manifolds. The special case of surfaces with boundary has been a particularly popular topic for a very long time, and it includes for example questions pertaining to the slice genus of a knot or the complexity of a knot or link cobordism.

On the other hand, the study of *nonorientable* surfaces and knot cobordisms in $I \times S^3$ has received increasing attention in the last decade — see Batson [3], Ozsváth, Stipsicz and Szabó [25], Golla and Marengon [8] and Fan [7] — and there are now several bounds to the nonorientable slice genus of a knot. However, if a knot bounds a nonorientable surface of a given "genus", it is not clear how complicated the embedding must be. We tackle this question by proving a nonorientable analogue of a recent result of Juhász, Miller and Zemke. In a recent paper [11], they proved an inequality involving the number of local maxima and the genus appearing in an oriented knot cobordism using a version of knot Floer homology. Here we prove similar inequalities for nonorientable knot cobordisms using the torsion orders of unoriented versions of knot Floer homology and instanton Floer homology.

As for knot Floer homology, we use Ozsváth, Stipsicz and Szabó's *unoriented knot Floer homology* HFK' [26], which is a module over $\mathbb{F}[U]$. For a knot $K \subset S^3$ we define its unoriented knot Floer torsion order as

$$\mathrm{Ord}_U(K) = \min\{n \geq 0 \mid U^n \cdot \mathrm{Tors} = 0\},$$

where $\mathrm{Tors} \subset \mathrm{HFK}'(K)$ denotes the $\mathbb{F}[U]$–torsion subgroup.

In the instanton setting, we use Kronheimer and Mrowka's *instanton Floer homology with local coefficients*, denoted by $I^\sharp(K)$, which is a module over a Noetherian domain $\mathcal{S}$ which has a special element P [17]. We will restrict our attention to certain domains $\mathcal{S}$, for which $I^\sharp(K)$ is functorial for nonorientable knot cobordisms with singular bundles represented by surfaces ω with $\partial\omega$ on the cobordism. In this case, it can be shown that, for a knot K and for the torsion part Tors of $I^\sharp(K)$, there is a positive integer n such that $P^n \cdot \mathrm{Tors} = 0$. Thus, we define

$$\mathrm{Ord}_I(K) = \min\{n \geq 0 \mid P^n \cdot \mathrm{Tors} = 0\}.$$

For a nonorientable surface Σ with n boundary components, recall that its *nonorientable genus* is

$$\gamma(\Sigma) = 2 - \chi(\Sigma) - n.$$

For example, $\mathbb{RP}^2$ (with an arbitrary number of punctures) has nonorientable genus 1. Note that, for nonorientable knot cobordisms, $\gamma(\Sigma) = -\chi(\Sigma)$. With this notation in mind, we state our main theorem:

Theorem 1.1 *Let K_1 and K_2 be knots in S^3. Suppose that there is a nonorientable knot cobordism Σ in $I \times S^3$ from K_1 to K_2 with M local maxima. Then*

$$\mathrm{Ord}_I(K_1) \leq \max\{\mathrm{Ord}_I(K_2), M\} + \gamma(\Sigma) \tag{1}$$

and

$$\mathrm{Ord}_U(K_1) \leq \max\{\mathrm{Ord}_U(K_2), M\} + \gamma(\Sigma). \tag{2}$$

From a formal viewpoint, Theorem 1.1 is analogous to [11, Theorem 1.1]. The proof of Theorem 1.1 uses the functorial properties of HFK' and $I^\sharp$ (see [7; 17]), in a similar way as [11, Theorem 1.1] relies on knot Floer homology cobordism maps; see Juhász [10] and Zemke [35]. Despite being inspired by [11, Theorem 1.1], the proof of Theorem 1.1 must necessarily deviate from it. Recall that, in order to define a cobordism map in knot Floer homology, one needs to choose a properly embedded 1–manifold on the surface, often called the set of *decorations*. In [11], the chosen decorations were a pair of parallel arcs, which make the computations of the cobordism maps more tractable. This choice does not work for nonorientable cobordisms in HFK', so we are forced to choose different decorations, which make the cobordism map more complicated. To circumvent this problem, we relate the resulting nonorientable cobordism to an orientable one, then use a stabilisation lemma proved by Ian Zemke (see Lemma 5.4). In the case of $I^\sharp$, for a cobordism Σ to define a map, one needs a surface ω with boundary $\partial\omega \subset \Sigma$. The natural choice for orientable cobordisms would be $\omega = \varnothing$, in which case [11] applies verbatim to the case of $I^\sharp$. The map can be defined for the nonorientable surfaces we are interested in, but it will usually vanish. To overcome this problem, we choose a particular ω that allows us to control the induced map.

Remark 1.2 While Theorem 1.1 is stated for nonorientable cobordisms, both inequalities also hold for *orientable* cobordisms. The proof follows verbatim from [11], by replacing knot Floer homology with the desired Floer theory.

We prove several applications of Theorem 1.1, which mirror those of [11, Theorem 1.1].

1.1 Nonorientable ribbon cobordism

A knot cobordism in $I \times S^3$ is called *ribbon* if it has no local maxima. For example, a ribbon concordance (ie a cobordism of genus 0) from the unknot to a knot K is equivalent to a ribbon disc for K. Theorem 1.1 has a straightforward application to nonorientable ribbon cobordisms:

Corollary 1.3 *Let K_1 and K_2 be knots in S^3. Suppose that there is a nonorientable ribbon cobordism Σ in $I \times S^3$ from K_1 to K_2. Then*

$$\mathrm{Ord}_I(K_1) \le \mathrm{Ord}_I(K_2) + \gamma(\Sigma) \quad \textit{and} \quad \mathrm{Ord}_U(K_1) \le \mathrm{Ord}_U(K_2) + \gamma(\Sigma).$$

1.2 The nonorientable refined cobordism distance

The standard cobordism distance between two knots K_1 and K_2 is $d_o(K_1, K_2) = 2g_4(K_1 \# \overline{K}_2)$, where g_4 denotes the standard slice genus. This is not a distance on the set of knots, because concordant knots have distance 0, but it descends to a well-defined distance on the concordance group; see Baader [2]. In [11], Juhász, Miller and Zemke define a refined cobordism distance on the set of knots, and give lower bounds to it in terms of the torsion order in knot Floer homology.

There are analogous nonorientable notions too. For an (orientable or nonorientable) cobordism Σ in $I \times S^3$ from K_1 to K_2 with m local minima and M local maxima, let

$$|\Sigma| = \max\{m, M\} - \chi(\Sigma).$$

Definition 1.4 Given knots $K_1, K_2, \subset S^3$, we define the *standard nonorientable cobordism distance d_u* and the *refined nonorientable cobordism distance d_u^r* between them as

$$d_u(K_1, K_2) = \min\{-\chi(\Sigma)\} \quad \text{and} \quad d_u^r(K_1, K_2) = \min\{|\Sigma|\},$$

where in both cases Σ varies over all nonorientable connected cobordisms and all genus-0 orientable cobordisms (ie concordances).

Remark 1.5 The orientable counterparts — the *standard orientable cobordism distance d_o* from [2] and the *refined orientable cobordism distance d_o^r* from [11] — are defined in the same way as in Definition 1.4, but the surface Σ now varies over all *orientable* connected cobordisms. One can also define analogous notions d_a and d_a^r, which we can call *all-surface cobordism distances*, where Σ varies over all (orientable or nonorientable) connected cobordisms.

It is immediate to see that d_o, d_u and d_a are distances on the concordance group and d_o^r, d_u^r and d_a^r are distances on the set of knots.

Theorem 1.1 implies the following lower bounds:

Corollary 1.6 *If K_1 and K_2 are knots in S^3, then*

$$|\mathrm{Ord}_I(K_1) - \mathrm{Ord}_I(K_2)| \leq d_u^r(K_1, K_2)$$

and

$$|\mathrm{Ord}_U(K_1) - \mathrm{Ord}_U(K_2)| \leq d_u^r(K_1, K_2).$$

In view of Remark 1.2, one can in fact replace d_u^r with d_a^r. However, for orientable cobordisms, one can also use the standard versions of instanton and knot Floer homology, which should give better bounds.

We use Corollary 1.6 to show that the difference between $d_u^r(K_1, K_2)$ and $d_u(K_1, K_2)$ can be arbitrarily large.

Corollary 1.7 *For all $\gamma \geq 1$ and $m \geq 1$, there exists a knot $K_{\gamma,m}$ with $d_u(K_{\gamma,m}, U_1) = \gamma_4(K_{\gamma,m}) = \gamma$ and such that $d_u^r(K_{\gamma,m}, U_1) \geq \gamma + m$.*

Thus, each nonorientable surface $\Sigma \subset B^4$ with $\partial\Sigma = K_{\gamma,m}$ and $\gamma(\Sigma) = \gamma$ has at least m local minima (with respect to the radial function).

The knots $K_{\gamma,m}$ that we consider in the proof of Corollary 1.7 are a subfamily of torus knots for which Ord_U can be computed explicitly.

1.3 The unoriented band-unlinking number

For a knot K in S^3, the *oriented band-unknotting number* $u_b(K)$ is defined as the minimum number of oriented band surgeries that turn K into the unknot. This was called the SH(2)–unknotting number by Hoste, Nakanishi and Taniyama [9]. Its unoriented counterpart $u_b^u(K)$, called the $H(2)$–unknotting number in [9], seems to predate $u_b(K)$ in the literature, since Lickorish proved that there exist knots with $u_b^u(K) > 1$ in [22]. Note that, in the definition of $u_b^u(K)$, we allow both orientable and nonorientable band surgeries.

Juhász, Miller and Zemke [11] introduced a variation, called the *oriented band-unlinking number* $\mathrm{ul}_b(K)$, which is defined as the minimum number of oriented band surgeries that turn K into an unlink. Of course, $\mathrm{ul}_b(K) \leq u_b(K)$, and they proved that $\mathrm{Ord}_v(K) \leq \mathrm{ul}_b(K)$ for all knots K in S^3. Using Theorem 1.1, we can derive a similar result for the corresponding unoriented notion.

Definition 1.8 The *unoriented band-unlinking number* $\mathrm{ul}_b^u(K)$ of a knot K in S^3 is defined as the minimum number of (orientable or nonorientable) band surgeries that turn K into an unlink.

Clearly, we have

$$\begin{array}{ccc} \mathrm{ul}_b^u(K) & \leq & u_b^u(K) \\ \rotatebox{-90}{$\leq$} & & \rotatebox{-90}{$\leq$} \\ \mathrm{ul}_b(K) & \leq & u_b(K) \end{array}$$

Corollary 1.9 *For a knot K in S^3,*

$$\mathrm{Ord}_I(K) \leq \mathrm{ul}_b^u(K) \quad \textit{and} \quad \mathrm{Ord}_U(K) \leq \mathrm{ul}_b^u(K).$$

Remark 1.10 Wong (personal communication, 2020) has informed us of a proof, using methods analogous to Alishahi and Eftekhary [1], that, if there is a cobordism

$\Sigma \subset I \times S^3$ from K_1 and K_2 (no matter whether orientable or nonorientable) with m minima, b saddles and M maxima, then

$$|\mathrm{Ord}_U(K_1) - \mathrm{Ord}_U(K_2)| \leq m + b + M.$$

Since the unlink has vanishing torsion order, this would recover the inequality of Corollary 1.9 involving Ord_U.

Ours is one of several recent papers related to ribbon cobordisms. Zemke [33] showed that knot Floer homology obstructs ribbon concordance, a result that prompted a flurry of interesting results in this area, including Levine and Zemke [21], Miller and Zemke [23], Daemi, Lidman, Vela-Vick and Wong [5], Kang [13] and Caprau, González, Lee, Lowrance, Sazdanović and Zhang [4]. Other papers in the area are Sarkar's paper on the ribbon distance [31] and the already-cited paper of Juhász, Miller and Zemke [11], which is the closest paper to ours.

Organisation

The first two sections of the paper are on instanton Floer homology: we review the necessary background in Section 2, and we prove the main instanton technical result (Proposition 3.3) in Section 3. In the following two sections we do the same for knot Floer homology: after a review in Section 4, we prove the main knot Floer technical result (Proposition 5.5) in Section 5. In Section 6 we prove Theorem 1.1 and the applications discussed in the introduction (Corollaries 1.3, 1.6, and 1.9). Finally, in Section 7 we compute the torsion order Ord_U for a subfamily of torus knots and prove Corollary 1.7.

Acknowledgements

We warmly thank Haofei Fan for his support and help. Special acknowledgements go to Ciprian Manolescu, who suggested we investigate nonorientable cobordisms, and to C-M Michael Wong, who shared his work with us and who gave us helpful comments on an early draft of the paper. We are also extremely grateful and indebted to Ian Zemke, who proved Lemma 5.4. Finally, we would like to thank the referee for their careful reading and helpful comments. Marengon was partially supported by NSF FRG grant DMS-1563615 and the Max Planck Institute for Mathematics. Gong was partially supported by NSF grant DMS-2055736.

2 Background on instanton homology with local systems

2.1 Instanton homology groups

Kronheimer and Mrowka introduced singular instanton homology with local systems in [15], and introduced several more involved variants of it in [17]. We will be working with a variant from the latter. Let us now review the relevant definitions and properties, following [17; 18].

Let Y be a closed, oriented 3–manifold, let L be a link in Y, let y_0 be a basepoint in Y, and let B_{y_0} be a ball around y_0 that does not intersect L. Let $\theta_0 \subset Y$ be a standard θ–web in B_{y_0}. Let ω be a 1–dimensional submanifold of Y which consists of components that are circles disjoint from L and B_{y_0} and arcs which have endpoints on L and are otherwise disjoint from L.

Then there is an associated space $\mathcal{B}^\sharp(Y, L)_\omega$ of SO(3)–connections on Y which are singular at $L \cup \theta_0$, lift to SU(2) away from the $L \cup \omega \cup \theta_0$, and are such that the SU(2)–holonomy around ω is -1 and the SU(2)-holonomies around components of L and arcs of θ_0 are conjugate to $I \in \mathrm{SU}(2)$, when we regard SU(2) as the unit quaternions and 1, I, J and K are the fundamental quaternion units.

The local system Γ is defined using three maps $h_i \colon \mathcal{B}^\sharp(Y, L)_\omega \to \mathbb{R}/\mathbb{Z}$ for $i = 1, 2, 3$, given by taking holonomy along the three arcs of the θ–web, which gives three maps to SU(2), and then composing with a character $\mathrm{SU}(2) \to U(1) = \mathbb{R}/\mathbb{Z}$ to get maps to $\mathbb{R}/\mathbb{Z}$. Let $\mathcal{R} = \mathbb{F}_2[\mathbb{Z}^3]$ be the group ring, which we can also write as the ring of Laurent polynomials in three variables,

$$\mathcal{R} = \mathbb{F}_2[T_1^{\pm 1}, T_2^{\pm 1}, T_3^{\pm 1}].$$

Then Γ is defined as the pullback via (h_1, h_2, h_3) of a particular local system over $(\mathbb{R}/\mathbb{Z})^3$ with fibre the free rank 1 module over $\mathcal{R}$. For a commutative ring $\mathcal{S}$ and a homomorphism $\sigma \colon \mathcal{R} \to \mathcal{S}$, let Γ_σ denote the induced local system of $\mathcal{S}$–modules.

The instanton homology group $I^\sharp(Y, L; \Gamma_\sigma)_\omega$ is defined as the Floer homology of $\mathcal{B}^\sharp(Y, L)_\omega$ with a perturbed Chern Simons functional and with the local system Γ_σ. (In [17], there is an additional map $h_0 \colon \mathcal{B}^\sharp(Y, L)_\omega \to \mathbb{R}/\mathbb{Z}$ coming from taking holonomy along the link itself, and $\mathcal{R}$ is defined to be $\mathbb{F}_2[\mathbb{Z}^4]$, but, for our purposes, we will only be using the local system coming from h_1, h_2 and h_3.)

2.2 Maps induced by cobordisms with dots

We now review the functoriality of $I^\sharp(Y, L; \Gamma_\sigma)_\omega$. Keeping previous notation, let $\sigma\colon \mathcal{R} \to \mathcal{S}$ be a map of commutative rings.

For $i = 1, 2$, let Y_i denote a closed, oriented 3–manifolds, with a link L_i and a 1–manifold ω_i embedded in Y_i with boundary on L_i and otherwise not intersecting L_i.

For a cobordism of pairs (X, S) from (Y_1, L_1) to (Y_2, L_2), and ω a 2–manifold with corners whose boundary pieces are ω_1 and ω_2 in Y_1 and Y_2, respectively, together with arcs and circles in S, there is an induced map

$$I^\sharp(X, S; \Gamma_\sigma)_\omega\colon I^\sharp(Y_1, L_1; \Gamma_\sigma)_{\omega_1} \to I^\sharp(Y_2, L_2; \Gamma_\sigma)_{\omega_2}$$

of $\mathcal{S}$–modules.

This functoriality can be extended to morphisms given by cobordisms of pairs with dots on the surfaces. That is, for a cobordism of pairs (X, S), define a *dot* on S to be an interior point $p \in S$ along with an orientation of $T_p S$. Then, for dots $p_1, p_2, \dots, p_d$ on S, there is an induced map of $\mathcal{S}$–modules

$$I^\sharp(X, S, p_1, p_2, \dots, p_d; \Gamma_\sigma)_\omega\colon I^\sharp(Y_1, L_1; \Gamma_\sigma)_{\omega_1} \to I^\sharp(Y_2, L_2; \Gamma_\sigma)_{\omega_2}.$$

In our computations, we will always have $Y_1, Y_2 = S^3$ and $X = S^3 \times [0, 1]$. Moreover, we will be using the same Γ_σ. Thus, we will denote our cobordisms by

$$I^\sharp(S, p_1, p_2, \dots, p_d)_\omega = I^\sharp(X, S, p_1, p_2, \dots, p_d; \Gamma_\sigma)_\omega.$$

2.3 Properties of the induced maps

Before going over some of the properties of the maps of $\mathcal{S}$–modules induced by cobordisms, let us recall two particular elements of the rings $\mathcal{R}$ and $\mathcal{S}$. Writing $\mathcal{R} = \mathbb{F}_2[T_1^{\pm 1}, T_2^{\pm 1}, T_3^{\pm 1}]$, the elements P and Q are given by

$$P = T_1 T_2 T_3 + T_1 T_2^{-1} T_3^{-1} + T_1^{-1} T_2 T_3^{-1} + T_1^{-1} T_2^{-1} T_3$$

and

$$Q = T_1^2 + T_1^{-2} + T_2^2 + T_2^{-2} + T_3^2 + T_3^{-2}.$$

For $\sigma\colon \mathcal{R} \to \mathcal{S}$, the elements $\sigma(P), \sigma(Q) \in \mathcal{S}$ will also be denoted P and Q, respectively.

(a) [18, Lemma 3.2] Let S be an oriented cobordism. Suppose S' is obtained from S by adding an internal 1–handle connecting points $p, q \in S$, where p and q both have the same orientation as S. Then

$$I^\sharp(S') = I^\sharp(S,p) + I^\sharp(S,q) + PI^\sharp(S).$$

Here and throughout we assume that $\omega = \varnothing$ when it is not denoted.

(b) [17, Lemma 4.2] Let (S,ω) be a cobordism between (L_1,ω_1) and (L_2,ω_2). Let R_+ and R_- be the two standard embedded copies of $\mathbb{R}P^2$ in S^4 with self-intersection $+2$ and -2, respectively. Let π be a disk whose boundary is the generator of $H_1(\mathbb{R}P^2)$. Then

$$I^\sharp(S \mathbin{\#} R_+)_{\omega+\pi} = I^\sharp(S)_\omega \quad \text{and} \quad I^\sharp(S \mathbin{\#} R_-)_{\omega+\pi} = PI^\sharp(S)_\omega.$$

(c) **Künneth formula for split links** [14, Section 5.5; 16, Section 2.2; 17, Section 5.3] Let L be a split link, so that $L = L_0 \amalg L_1$, and L_0 and L_1 are contained in disjoint balls in S^3. Then

$$I^\sharp(L) \simeq I^\sharp(L_0) \otimes I^\sharp(L_1),$$

and this is natural with respect to cobordisms with dots.

This is shown in [14, Section 5.5] using a version of excision without local coefficients, Hopf link instead of a θ–web, and without dots. There is an argument in [16, Section 2.2] for why it does not matter whether one uses a θ–web or a Hopf link, and it is explained in [17, Section 5.3] why it still works with local coefficients. The proof of functoriality in [14] carries over with no problems to the situation of cobordisms with dots.

(d) [17, Section 5.2] Let U_l be the l–component unlink. Then $I^\sharp(U_0)$ is a free module of rank 1 over $\mathcal{S}$, which we write as $I^\sharp(U_0) = \mathcal{S}u_0$, and $I^\sharp(U_1)$ is the free module over $\mathcal{S}$ of rank 2, which we write as $I^\sharp(U_1) = \mathcal{S}u_+ \oplus \mathcal{S}u_-$. For D the standard disk viewed as a cobordism from the empty link to the unknot, and q a point with orientation compatible with the choice of orientation of the knot,

$$I^\sharp(D)(u_0) = u_+ \quad \text{and} \quad I^\sharp(D,q)(u_0) = u_-.$$

Moreover, if D_o is the standard disk viewed as a cobordism from the unknot to the empty link, and q a point with orientation compatible with the choice of orientation of the knot,

$$I^\sharp(D_o)(u_-) = 1, \quad I^\sharp(D_o)(u_+) = 0, \quad I^\sharp(D_o,q)(u_+) = 1, \quad I^\sharp(D_o,q)(u_-) = P.$$

For U_l, by the previous point, we have

$$I^\sharp(U_l) = (\mathcal{S}u_+ \oplus \mathcal{S}u_-)^{\otimes l}.$$

(e) [17, Section 5.4] Let m and Δ denote the standard "pair of pants" cobordisms between the two-component unlink U_2 and the unknot U_1, the merge

$$m\colon U_2 \to U_1$$

and the split

$$\Delta\colon U_1 \to U_2.$$

The map on $I^\sharp$ induced by m (with no dots) is given by

$$u_+ \otimes u_+ \mapsto u_+, \quad u_\pm \otimes u_\mp \mapsto u_-, \quad u_- \otimes u_- \mapsto Pu_- + Qu_+, \tag{3}$$

and the map induced by Δ (with no dots) is given by

$$u_+ \mapsto u_+ \otimes u_- + u_- \otimes u_+ + Pu_+ \otimes u_+, \quad u_- \mapsto u_- \otimes u_- + Qu_+ \otimes u_+. \tag{4}$$

3 A technical result for instantons

In this section we prove the main technical result for instanton Floer homology $I^\sharp$ which we will use to prove Theorem 1.1. To do so, we will use a classical result in Morse theory, Lemma 3.2 below. We state it in the most convenient form for us, and give a quick sketch of its proof.

Definition 3.1 Given a knot K in S^3 and a band B for K, ie an embedded rectangle B in S^3 which intersects K in two opposite sides, we say that B is *orientable with respect to K* if the knot K and the result of band surgery on K along B can be given coherent orientations (equivalently, if surgering K along B gives a two-component link).

Lemma 3.2 *Let $\Sigma \subset I \times S^3$ be a nonorientable cobordism between knots K and K' with m local minima, b saddles and M local maxima. Then, after an isotopy rel boundary, we can break it into a sequence of cobordisms as follows:*

(a) *m births (from $K_1 = K$ to L_1);*

(b) *m band surgeries that join the various components of the link (from L_1 to K'_1);*

(c) *$b - (m + M + 1)$ band surgeries orientable with respect to K'_1 (this cobordism ends with a knot or a two-component link L');*

(d) *1 band surgery nonorientable with respect to K'_1 (this cobordism goes from L' to a knot K'_2);*

(e) *M band surgeries that split the knot K'_2 into $M + 1$ components;*

(f) *M deaths.*

Moreover, in this decomposition, the attaching arcs of the b bands on K_1' can be assumed to be all disjoint, and we can assume that both attaching arcs of the nonorientable band are already contained in K_1.

Sketch of the proof We can arrange all births to appear first and all deaths to appear last (steps (a) and (f)). We can also find bands that connect the various components (steps (b) and (e)). Thus, we can restrict to the part of the cobordism between K_1' and K_2', which consists of saddles (ie band surgeries). Note that both K_1' and K_2' are knots.

If all bands were orientable with respect to K_1', then all Σ would be orientable, so there is at least one band nonorientable with respect to K_1'.

Arrange for all bands from K_1' to K_2' to appear at the same time.

If there is more than one band nonorientable with respect to K_1', pick one of them (call it B) and slide it following the surgery of K_1' along all the other bands. When B slides over an orientable band, it stays nonorientable. When B slides over a nonorientable band, it becomes orientable. Note that eventually it must slide over a nonorientable band because K_2' is connected, so $K_2' \setminus B$ consists of just two arcs.

Repeat until you have only one nonorientable band left.

If B is the unique nonorientable band, then you can slide its endpoints along L' so that they are disjoint from all the other (oriented) bands, so we can think of it as in K_1. □

The main technical result of this section, needed to prove Theorem 1.1, is the following proposition:

Proposition 3.3 *Let S be a cobordism from K to K' with m local minima, b saddles, and M local maxima. Then there is a surface ω that meets S cleanly and only at $\partial\omega \subset S$, whose boundary is a circle in S such that, for $\overline{\omega}$ its mirror, we have*

$$P^M I^\sharp(\overline{S} \circ S)_{\omega\cup\overline{\omega}} = P^{b-m}\,\mathrm{Id}. \tag{5}$$

Towards this goal, let us start by doing some computations of maps induced by cobordisms with ω.

First let us understand the dependence of $I^\sharp(\Sigma)_\omega$ on ω when ω is a surface with boundary on Σ which intersects Σ cleanly and only at $\partial\omega \subset S$. Note that, for a link L in S^3, up to isomorphism, $I^\sharp(S^3, L)_\omega$ depends only on the homology of $[\partial\omega] \in H_0(L; \mathbb{Z}/2)$, because it counts flat connections and instantons on spaces determined by the homology class.

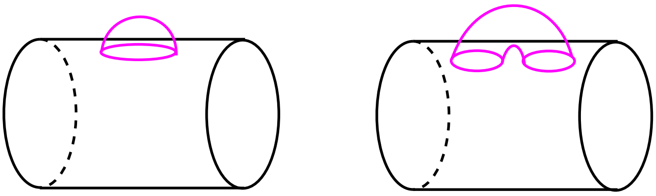


Figure 1: Cylinders with the magenta surfaces depicting ω.

Similarly, $I^\sharp(\Sigma)_\omega$ depends only on the homology class $[\partial\omega] \in H_1(\Sigma, \mathbb{Z}/2)$. This is because the map counts instantons on a moduli space built from Σ, $[\partial\omega] \in H^1(\Sigma, \mathbb{Z}/2)$ and $[\omega, \partial\omega] \in H_2(X, \Sigma, \mathbb{Z}/2)$, and $H_1(\Sigma) \simeq H_2(X, \Sigma)$ for $X = S^3 \times \mathbb{R}$.

From here, we can see that, for a cylinder Σ and ω given by either a small disk or a small tube with boundary on Σ, as in Figure 1, $I^\sharp(\Sigma)_\omega$ induces the identity: here $\partial\omega$ is trivial in $H_1(\Sigma)$, and $[\omega, \partial\omega]$ is also trivial in $H_2(X, \Sigma, \mathbb{Z}/2)$.

When the cobordism in Figure 2, left, is composed with its inverse, the map induced is the identity. Moreover, up to isomorphism, $I^\sharp(U, \omega)$ depends only on $[\partial\omega] \in H_0(U; \mathbb{Z}/2)$, so the two ends of the cobordism have the same instanton Floer homology. Thus, the cobordism in Figure 2, left, induces an isomorphism.

We will call the two generators of the instanton Floer homology of the unknot with an arc ω on the right, which is depicted in Figure 2, centre, x_+ and x_-, so that, in the $u_\pm$ and $x_\pm$ bases, the cobordism depicted in Figure 2, left, is the identity matrix.

The cobordism from the two-component unlink to itself induced by two standard cylinders with ω as a tube between them, as depicted in Figure 2, right, induces the identity map, because, in this situation, $(\omega, \partial\omega)$ is trivial in homology in $(S^3 \times I, \Sigma)$.

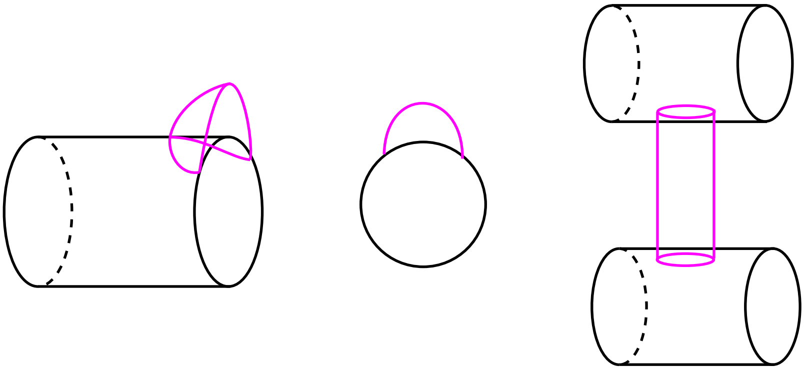


Figure 2

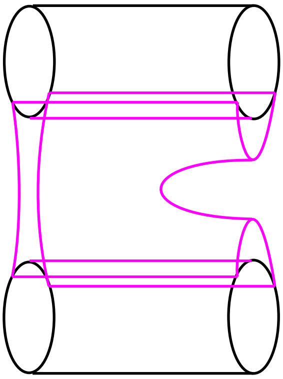


Figure 3

The same is true for the map depicted in Figure 3 precomposed with its mirror. Thus, the map induced by the cobordism depicted in Figure 3 is an isomorphism whose inverse is its mirror image. Here, we are identifying the link with ω on the right end of Figure 3 with the unlink with empty ω via the isomorphism induced by Figure 2, left, and the link with ω on the left has isomorphic instanton Floer homology.

For the link on the left in Figure 3, its homology is then a free module of rank 4 over $\mathcal{S}$. Let $\{x_{++}, x_{+-}, x_{-+}, x_{--}\}$ be a basis of this homology, so that, if we choose the basis $\{x_+ \otimes x_+, x_+ \otimes x_-, x_- \otimes x_+, x_- \otimes x_-\}$ for the two-component unlink on the right, the matrix the cobordism induces is the identity. (Recall that $x_\pm$ are the basis elements of the instanton homology of the unknot with an arc, so the cobordism of Figure 2, left, induces the identity matrix.)

A central step in our proof will be dealing with a cobordism that flips an unknot but does not change ω. To describe this, consider a link L with decoration ω which has an unknot component U that is split from the rest of L; we may isotope U so that it is a geometric circle. Suppose that ω has two endpoints on U, p and q, which we may isotope to be the endpoints of a diameter of U. Then the flip cobordism is a cobordism in $I \times S^3$ that is traced by the isotopy obtained by rotating U by π about the diameter pq. So this is an isotopy that does not change ω and reverses the orientation of one of the two components.

Claim 1 *The map on the instanton homology of U_2 with ω consisting of two arcs, each going between the two components, that results from flipping one of the unknots (as described above) in a way that does not change ω, is the identity map.*

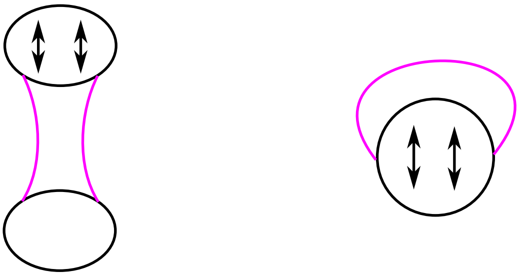


Figure 4

Proof By composing with the isomorphisms induced by the cobordism depicted in Figure 3, if Φ is the matrix associated to the flip in the basis $\{x_{++}, x_{+-}, x_{-+}, x_{--}\}$, then

$$\Phi = \begin{bmatrix} a & 0 & b & 0 \\ 0 & a & 0 & b \\ c & 0 & d & 0 \\ 0 & c & 0 & d \end{bmatrix},$$

where $\Phi_1 = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is the flip on the unknot with an arc in the basis $\{x_+, x_-\}$, depicted in Figure 4, right. This is because Φ is the matrix for $\Phi_1 \otimes \mathrm{Id}$ for Id the identity map, in the basis $\{x_+ \otimes x_+, x_+ \otimes x_-, x_- \otimes x_+, x_- \otimes x_-\}$, and we are using the basis of the instanton homology of Figure 4, left, corresponding to this basis.

Now let us compute some of the entries of Φ_1. Note that if we pre- or postcompose Φ_1 with caps like those in Figure 5, we get back the cap itself. These caps induce the maps $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $[0\ 1]$, so, from these compositions, we can see that $a = d = 1$ and $c = 0$.

Note that, if we did not have ω, then we could do the same argument with a cap with a dot, and, using the fact that doing a flip and then a cap with a dot is the same as doing a negative dot, we would be able to get the remaining entry, b, and recover [17, Proposition 5.8], in which the flip map does *not* induce the identity. However, because we have ω here, this does not work: the flip changes which side of $\partial\omega$ the dot is on.

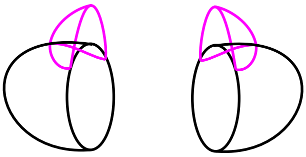


Figure 5

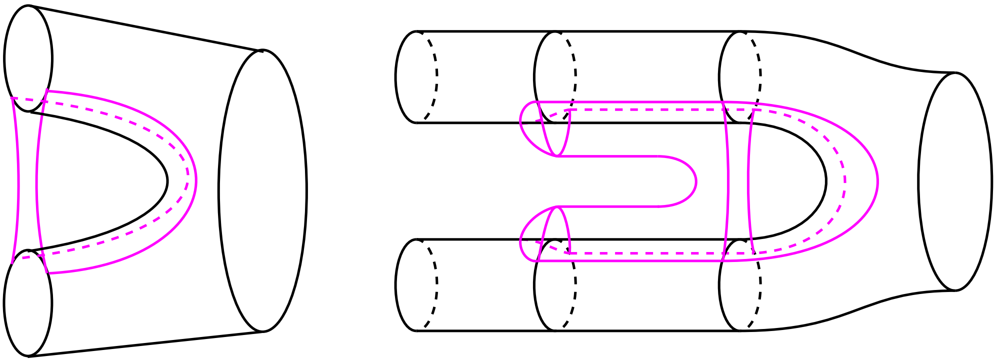


Figure 6

Going back to our computation, we now have that

$$\Phi = \begin{bmatrix} 1 & 0 & b & 0 \\ 0 & 1 & 0 & b \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We would now like to show that b is 0. Consider the pair of pants cobordism with ω as two half-disks from the unlink with two arcs going between components to the unknot, as depicted in Figure 6, left. Because we can precompose with isomorphisms to make a regular merge with a null-homotopic disk on top, as in Figure 6, right, we see that Figure 6, left, induces the same as the merge map, if we use the basis $\{x_{++}, x_{+-}, x_{-+}, x_{--}\}$. Here we are using that the reverse of the map in Figure 3 is also the identity matrix with our choice of basis.

Thus, in this basis, it induces the map

$$m = \begin{bmatrix} 1 & 0 & 0 & Q \\ 0 & 1 & 1 & P \end{bmatrix}.$$

Similarly, the reverse of this cobordism induces the same map as Δ, so it induces

$$\Delta = \begin{bmatrix} P & Q \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, composing $m \circ \Phi \circ \Delta$, we get the map

$$m \circ \Phi \circ \Delta = \begin{bmatrix} b + P & 0 \\ 0 & b + P \end{bmatrix}.$$

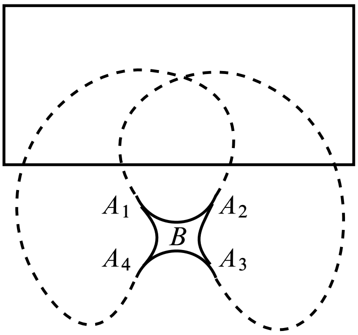


Figure 7: The cobordism S is cylindrical on the dotted part.

However, if we compose these cobordisms, we get a Klein bottle, which is a connected sum of $\mathbb{R}P^2_+$ and $\mathbb{R}P^2_-$, with ω given by two disks, one on each $\mathbb{R}P^2$, such that the boundary circle of each disk is the generator of that $H_1(\mathbb{R}P^2)$. It is shown in [17] that this Klein bottle with these ω induces the map $P \cdot \mathrm{Id}$, so $b = 0$, as desired. □

Claim 2 *Let $S \subset S^3 \times [0,2]$ be a cobordism from K_1 to K_2 such that in $S^3 \times [0,1]$ it is the cylinder on K_1 and in $S^3 \times [1,2]$ it consists of adding a single nonorientable band. More precisely, we may consider a band $B \subset S^3$ with vertices $A_1, A_2, A_3, A_4 \in S^3$ with A_1A_2 and A_3A_4 on K_1, as in Figure 7. In $S^3 \times [1,2]$, S then looks like $(K_1 \setminus (A_1A_2 \cup A_3A_4)) \times [1,2]$ away from the band $B \times [1,2]$, and within the band it goes from $A_1A_2 \cup A_3A_4$ at time 1 to $A_2A_3 \cup A_4A_1$ at time 2.*

The cobordism is depicted in frames in Figure 8.

Then there is a surface ω with boundary in the interior of S such that ω meets S only at the boundary, where they meet cleanly, and such that, for $\overline{S}$ the reverse of S with corresponding $\overline{\omega}$,

$$I^\sharp(\overline{S})_{\overline{\omega}} \circ I^\sharp(S)_\omega = P \cdot \mathrm{Id}\colon I^\sharp(K_1) \to I^\sharp(K_1).$$

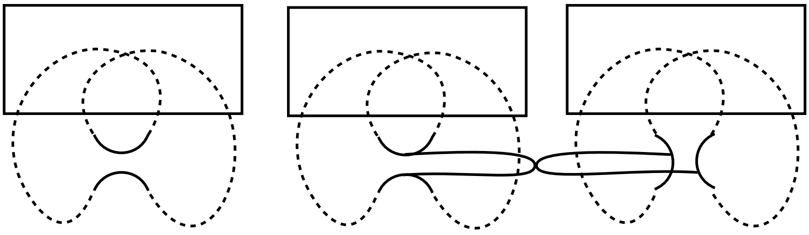


Figure 8

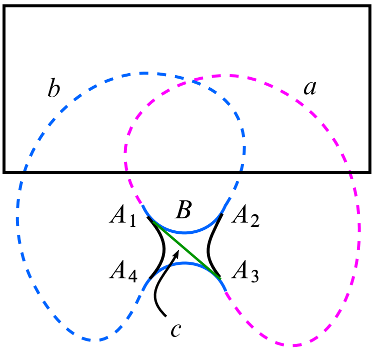


Figure 9: The cobordism S is cylindrical on the dotted part.

Proof Observe that A_1 and A_3 split K_1 into two parts, which we call a and b (these are coloured magenta and blue, respectively, in Figure 9). Let c be the diagonal on the band that goes from A_1 to A_3.

Consider $a \cup c$ as a knot in S^3 and let F_0 be a Seifert surface of it. Then F_0 is a surface with corners, with boundary $a \cup c$, and which meets b at the ends, A_1 and A_3. We may isotope a, b and c so that F_0 meets b cleanly at the ends and transversely in the interior, as in Figure 10.

If we choose an orientation of F_0 and b, then the intersection points may have positive or negative sign. We can increase the number of positive or negative intersection points without changing the isotopy type of the embedding of $K_1 \cup c$ into S^3 by twisting b around A_1 or A_3, as in Figure 11. Let us do this, adding either positive or negative intersections as needed until there are the same number of positive as negative interior intersection points between b and F_0.

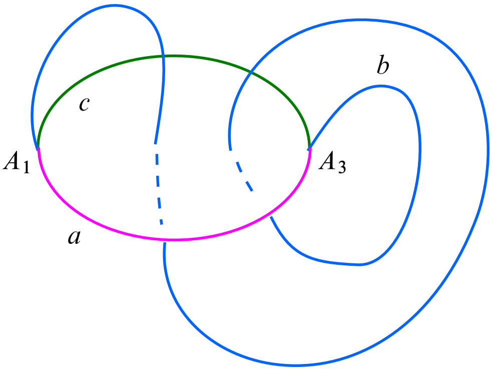


Figure 10: Here F_0 is depicted as a disk though it could have higher genus.

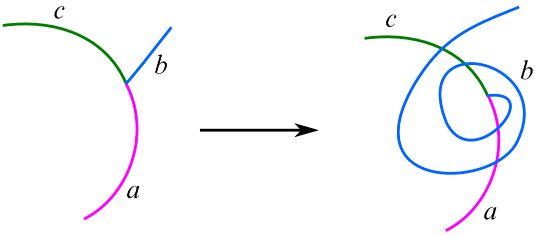


Figure 11

Now say that the intersection points are $A_1, p_1, p_2, \dots p_{2k}, A_3$, in order along b. Then, if p_i and p_{i+1} are intersection points with opposite sign, we may remove a small disk around each of p_i and p_{i+1} and replace it with a small tube around the part of b that goes from p_i to p_{i+1}, thus reducing the number of intersection points. We may continue in this manner, removing adjacent opposite-sign intersection points until none remain.

We now have a surface, which we call F_1 with boundary $a \cup c$, which intersects b only at A_1 and A_3, where the intersection is clean.

We now consider a surface $F_2 \subset S^3 \times [0,2]$ with boundary on S which is given by the union of $F_1 \subset S^3 \times \{1\}$ with a disk sitting between $c \times 1 \subset S^3 \times [1,2]$ and S, as in Figure 12.

Then this F_2 can have its corners smoothed out to a surface with boundary ω.

Let us now show that, for this ω, we have

$$I^\sharp(\overline{S})_{\overline{\omega}} \circ I^\sharp(S)_\omega = P \cdot \mathrm{Id}\colon I^\sharp(K_1) \to I^\sharp(K_1).$$

Let Σ denote the composition of S with $\overline{S}$, and let $\omega_\Sigma = \omega \cup \overline{\omega}$ be the decoration on this cobordism. See Figure 13.

Let γ denote the circle composed of the cocore of the band and its mirror, depicted in blue in Figure 13. A regular neighbourhood of γ in Σ is a tube, represented in

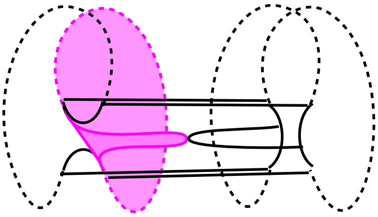


Figure 12

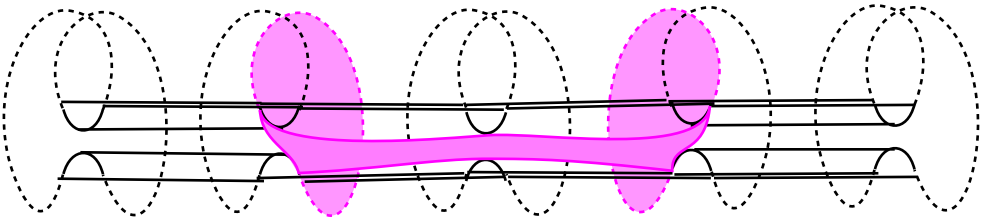


Figure 13: This is S with ω composed with the reverses $\overline{S}$ with $\overline{\omega}$.

Figure 14. If we cut the surface along γ, we get the twice-punctured cylinder as a cobordism from K_1 to itself.

Figure 14 shows $\partial\omega_\Sigma$ as well. The mod 2 homology class $[\partial\omega_\Sigma]$ on the surface Σ is the same as $[\gamma]$. One way to see this is to perform surgery on $\partial\omega_\Sigma$ along the green arc in Figure 14: this operation does not change the homology class and it yields a curve which is easily checked to be isotopic to γ in Σ.

Let Σ' be the cobordism obtained from Σ by inserting a flip in the tube in the centre of Figure 14, with the same decoration ω_Σ. Using Claim 1, we will see below that $I^\sharp(\Sigma)_{\omega_\Sigma} = I^\sharp(\Sigma')_{\omega_\Sigma}$. However, the curve $\partial\omega_\Sigma$ is homologically trivial in Σ'. One can check this again by doing surgery on the green arc, but this time the extra flip ensures that the obtained curve is not γ, but a homotopically trivial one. Thus, $I^\sharp(\Sigma')_{\omega_\Sigma} = I^\sharp(\Sigma')_{\varnothing}$,

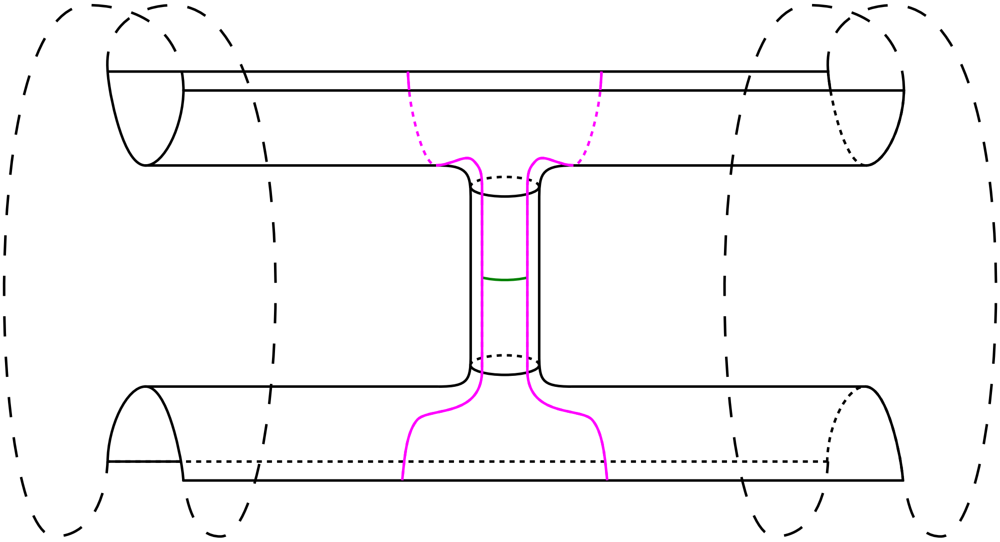


Figure 14

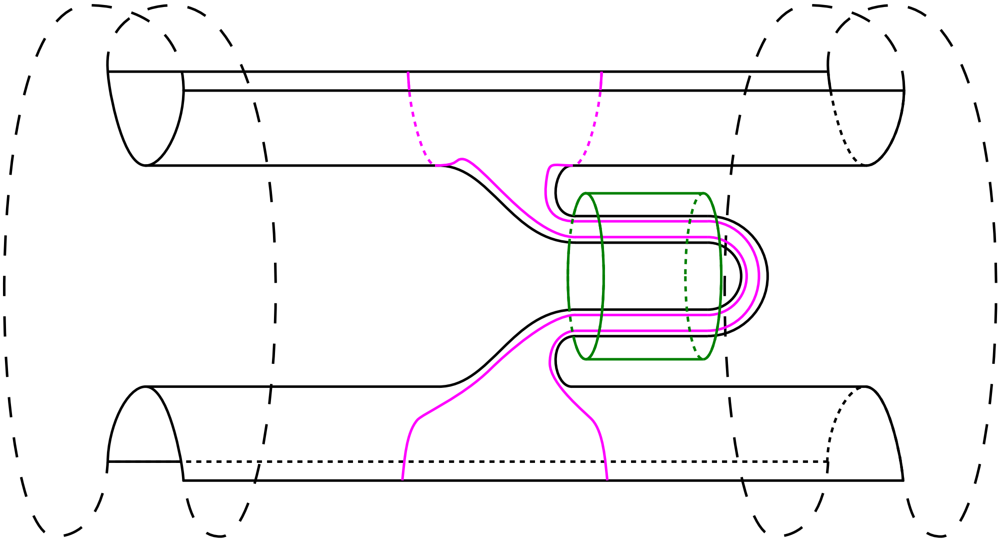


Figure 15

since the map depends only on $[\partial\omega_\Sigma]$. If $\omega = \varnothing$, one can apply the neck-cutting relation (property (a) in Section 2.3) to obtain that

$$I^\sharp(\Sigma') = P \cdot I^\sharp(I \times K_1) = P \cdot \mathrm{Id}_{I^\sharp(K_1)}.$$

We still have to show that $I^\sharp(\Sigma)_{\omega_\Sigma} = I^\sharp(\Sigma')_{\omega_\Sigma}$. To see this, isotope the tube in the middle as shown in Figure 15.

Let's restrict our attention to the piece contained in the cylinder in green, which is the identity cobordism on a two-component unlink. By Claim 1, the map induced by this cobordism is the same that we get if we introduce a flip on one of the two components. Since instanton Floer maps respect composition of cobordisms and disjoint unions, the map induced by the whole cobordism is not affected by the insertion of the flip, ie $I^\sharp(\Sigma)_{\omega_\Sigma} = I^\sharp(\Sigma')_{\omega_\Sigma}$. □

Claim 3 *Let S be a cobordism from a knot K_1 to a knot K_2 such that S consists of only b bands. That is, there are no births nor deaths. Then there is a surface ω with boundary on S such that*

$$I^\sharp(\bar{S})_{\bar{\omega}} \circ I^\sharp(S)_\omega = P^b \cdot \mathrm{Id}.$$

Proof We proceed by induction on b. The base case $b = 0$ is obvious.

For the inductive step, we divide into two cases. If S is orientable, then the statement holds for ω empty, because the cobordism $\overline{S} \circ S$ is the same as the cylinder on K_1 with b orientable tubes, and the result follows from the tube-cutting formula.

In the case that S is not orientable, at least one of the bands of S must be nonorientable with respect to K. In this case, let us write $S = S_r \circ S_u$, where S_u is a cobordism consisting of the nonorientable band and S_r is the rest of the cobordism, which may or may not be orientable.

Then, by the induction hypothesis, there is some ω_r such that

$$I^\sharp(\overline{S}_r)_{\overline{\omega}_r} \circ I^\sharp(S_r)_{\omega_r} = P^{b-1} \cdot \mathrm{Id}.$$

Applying Claim 2, there is a surface ω_u with boundary on S_u such that

$$I^\sharp(\overline{S}_u)_{\overline{\omega}_u} \circ I^\sharp(S_u)_{\omega_u} = P \cdot \mathrm{Id}.$$

The statement

$$I^\sharp(\overline{S})_{\overline{\omega}} \circ I^\sharp(S)_{\omega} = P^{b} \cdot \mathrm{Id}$$

now follows. □

Now we can proceed with the proof of Proposition 3.3.

Proof of Proposition 3.3 Applying Lemma 3.2, we may break S into

(a) m births (from K_1 to L_1);

(b) m band surgeries that join the various components of the link (from L_1 to K_1');

(c) $b-(m+M)$ band surgeries which may or may not be orientable, ending in a knot K_2';

(d) M band surgeries that split the knot K_2' into $M+1$ components;

(e) M deaths.

Let us call the cobordisms corresponding to the five steps $S_1, S_2, \dots S_5$. We may isotope the cobordism in $S^3 \times \mathbb{R}$ so that S_i is in $S^3 \times [i, i+1]$.

We will choose ω to be in $S^3 \times [2,3]$, so that its boundary is in S_3 as in Claim 3, so that

$$I^\sharp(\overline{S}_3)_{\overline{\omega}} \circ I^\sharp(S_3)_{\omega} = P^{b-m-M} \cdot \mathrm{Id}.$$

The proof now proceeds the same way as the proof of [11, Proposition 4.1]. The main argument is by considering a cobordism Σ that comes from adding M tubes connecting points on the death caps to their mirrors.

By [18, Lemma 3.2], for a connected, oriented cobordism Σ, if Σ' is obtained from Σ by adding a tube between points p and q, then

$$I^\sharp(\Sigma') = I^\sharp(\Sigma, p) + I^\sharp(\Sigma, q) + PI^\sharp(\Sigma) = PI^\sharp(\Sigma), \tag{6}$$

where the second equality is because Σ is connected, so $I^\sharp(\Sigma, p)$ and $I^\sharp(\Sigma, q)$ induce the same map, and, since we are working over characteristic two, they cancel.

Let Σ_1 denote the cobordism that takes $\overline{S} \circ S$ and adds M tubes, one for each death, connecting a point in the death to its reverse, so that $\Sigma_1 = S_1 S_2 S_3 S_4 \overline{S}_4 \overline{S}_3 \overline{S}_2 \overline{S}_1$. Applying (6) for each death, to the part of the cobordism from K_3 to itself coming from doing S_4, S_5 and their reverses, we see that

$$I^\sharp(\Sigma_1) = P^M I^\sharp(\overline{S} \circ S).$$

Here, we are allowed to use the above result because $S_4 S_5 \overline{S}_5 \overline{S}_4$ and $S_4 \overline{S}_4$ are both orientable and connected.

In $\Sigma_1 = (S_1 S_2 S_3 S_4)\overline{(S_1 S_2 S_3 S_4)}$, M splitting bands of S_4 and their reverses, cap off the ends, and call the resulting cobordism Σ_2. Then, for the same reason as above, we have

$$I^\sharp(\Sigma_1) = P^M I^\sharp(\Sigma_2),$$

because again $S_3 S_4 \overline{S}_4 \overline{S}_3$ and $S_3 \overline{S}_3$ are both orientable and connected.

Now we have $\Sigma_2 = (S_1 S_2 S_3)\overline{(S_1 S_2 S_3)}$.

Because of our construction of ω, $S_3 \overline{S}_3$ with $\omega \cup \overline{\omega}$ falls under the setting of Claim 3, so the map it induces is $P^{b-m-M} \cdot \mathrm{Id}_{I^\sharp(K_2)}$. Thus, if we let $\Sigma_3 = (S_1 S_2)\overline{(S_1 S_2)}$, then

$$I^\sharp(\Sigma_2) = P^{b-m-M} I^\sharp(\Sigma_3).$$

Now Σ_3 is given by a cylinder on K_1 and m S^2's, with m tubes, with the tubes connecting the S^2's and the cylinder in a tree-like fashion. Applying the tube-cutting formula, Lemma 3.2 of [18], and observing that a sphere without any dots induces the zero map while a sphere with one dot induces the identity, we see that $I^\sharp(\Sigma_3)$ induces the same map as the cylinder, which is to say the identity.

Putting all of this together, we get

$$P^M I^\sharp(\overline{S} \circ S) = I^\sharp(\Sigma_1) = P^{b-m} \cdot \mathrm{Id},$$

as desired. □

4 Background on unoriented knot Floer homology

Unoriented knot Floer homology was introduced by Ozsváth, Stipsicz and Szabó [25; 26]. Fan [7] showed that a nonorientable cobordism (with some extra data) induces maps on the unoriented knot Floer homology. We now review the relevant definitions, following mostly [35; 7].

4.1 Zemke's oriented TQFT

Cobordism maps in link Floer homology were first defined by Juhász [10]. Here we use Zemke's setup [35], specified to unoriented link Floer homology in the case $Y = S^3$.

Definition 4.1 An *oriented multibased link* in S^3 is a triple $\mathbb{L} = (L, \boldsymbol{w}, \boldsymbol{z})$ consisting of an oriented, embedded link $L \subset S^3$, with two disjoint collections of basepoints $\boldsymbol{w}$ and $\boldsymbol{z}$ on L, such that each component of L has at least two basepoints and the basepoints alternate between those in $\boldsymbol{w}$ and those in $\boldsymbol{z}$ as one traverses a component of L.

To an oriented multibased link $\mathbb{L}$, Zemke's most general construction gives a curved $\mathbb{F}[U_{\boldsymbol{w}}, V_{\boldsymbol{z}}]$–complex $\mathcal{CFL}^-(\mathbb{L})$ up to $\mathbb{F}[U_{\boldsymbol{w}}, V_{\boldsymbol{z}}]$–equivariant chain homotopy. Here $\mathbb{F}[U_{\boldsymbol{w}}, V_{\boldsymbol{z}}]$ denotes the polynomial ring generated by a U variable for each $\boldsymbol{w}$ basepoint and a V variable for each $\boldsymbol{z}$ basepoint. The curved complex is also endowed with gradings and a filtration.

In our case, we only need a simpler version of Zemke's complex, namely unoriented link Floer homology. This is defined as

$$\mathrm{CFL}'(\mathbb{L}) := \mathcal{CFL}^-(\mathbb{L}) \otimes_{\mathbb{F}[U_{\boldsymbol{w}}, V_{\boldsymbol{z}}]} \mathbb{F}[U],$$

where all variables act on $\mathbb{F}[U]$ as multiplication by U. For the reader familiar with Heegaard Floer homology, this is the free $\mathbb{F}[U]$–module generated by the intersection points $\mathbb{T}_{\boldsymbol{\alpha}} \cap \mathbb{T}_{\boldsymbol{\beta}}$ in the symmetric product, with differential given by

$$\partial \boldsymbol{x} = \sum_{\boldsymbol{y} \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}} \sum_{\substack{\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y}) \\ \mu(\phi) = 1}} \#\widehat{\mathcal{M}}(\phi) \cdot U^{n_{\boldsymbol{o}}(\phi)} \cdot \boldsymbol{y}, \tag{7}$$

where $n_{\boldsymbol{o}}(\phi) = \sum_{w \in \boldsymbol{w}} n_w(\phi) + \sum_{z \in \boldsymbol{z}} n_z(\phi)$.

Definition 4.2 For a doubly based knot $\mathbb{K} = (K, w, z)$, we also use the notation $\mathrm{CFK}'(\mathbb{K})$ and $\mathrm{HFK}'(\mathbb{K})$ for $\mathrm{CFL}'(\mathbb{K})$ and $\mathrm{HFL}'(\mathbb{K})$, respectively.

If $\mathbb{K}_1 = (K, w_1, z_1)$ and $\mathbb{K}_2 = (K, w_2, z_2)$ are two doubly based knots with the same underlying knot K, then $\mathrm{HFK}'(\mathbb{K}_1)$ and $\mathrm{HFK}(\mathbb{K}_2)$ are noncanonically isomorphic as $\mathbb{F}[U]$–modules. Thus, the following number is well defined:

Definition 4.3 If K is a knot, we define its *unoriented torsion order* as

$$\mathrm{Ord}_U(K) = \min\{n \geq 0 \mid U^n \cdot \mathrm{Tors} = 0\},$$

where Tors is the torsion submodule of $\mathrm{HFK}'(\mathbb{K})$, considered as a module over $\mathbb{F}[U]$. Here $\mathbb{K}$ is any doubly based knot with underlying knot K.

Remark 4.4 CFL' enjoys the following properties:

(a) $\mathrm{CFL}'(\mathbb{L})$ is a genuine chain complex (ie its curvature vanishes), so one can compute its homology $\mathrm{HFL}'(\mathbb{L})$, known as the *unoriented link Floer homology* of $\mathbb{L}$. This is still an $\mathbb{F}[U]$–module.

(b) For a doubly based knot $\mathbb{K} = (K, w, z)$, $\mathrm{HFK}'(\mathbb{K}) \cong \mathbb{F}[U] \oplus \mathrm{Tors}$, where Tors is the torsion as an $\mathbb{F}[U]$–module.

(c) For a doubly based unknot $\mathbb{U}_1 = (U_1, w, z)$, $\mathrm{HFK}'(\mathbb{U}_1) \cong \mathbb{F}[U]$.

(d) Given doubly based knots $\mathbb{K}_1$ and $\mathbb{K}_2$,

$$\mathrm{CFK}'(\mathbb{K}_1 \# \mathbb{K}_2) = \mathrm{CFK}'(\mathbb{K}_1) \otimes_{\mathbb{F}[U]} \mathrm{CFK}'(\mathbb{K}_2).$$

As a consequence, for knots K_1 and K_2 in S^3,

$$\mathrm{Ord}_U(K_1 \# K_2) = \max\{\mathrm{Ord}_U(K_1), \mathrm{Ord}_U(K_2)\}.$$

(e) If $\overline{\mathbb{L}}$ is the mirror of $\mathbb{L}$ (with the same basepoints), then, by [26, Proposition 2.17],

$$\mathrm{CFL}'(\overline{\mathbb{L}}) = \hom_{\mathbb{F}[U]}(\mathrm{CFL}'(\mathbb{L}), \mathbb{F}[U]).$$

As a consequence, for a knot K in S^3,

$$\mathrm{Ord}_U(\overline{K}) = \mathrm{Ord}_U(K).$$

Definition 4.5 If $\mathbb{L}_1 = (L_1, \boldsymbol{w}_1, \boldsymbol{z}_1)$ and $\mathbb{L}_2 = (L_2, \boldsymbol{w}_2, \boldsymbol{z}_2)$ are two oriented multi-based links, an (*oriented*) *decorated link cobordism* from $\mathbb{L}_1$ to $\mathbb{L}_2$ is a pair $\mathbb{S} = (\Sigma, \mathcal{A})$ such that:

(a) $\Sigma \subset I \times S^3$ is a properly embedded, compact, oriented surface with $\Sigma \cap \{0\} \times S^3 = \{0\} \times (-L_1)$ and $\Sigma \cap \{1\} \times S^3 = \{1\} \times L_2$.

(b) $\mathcal{A} \subset \Sigma$ is a properly embedded 1–manifold, which we refer to as the *decorations*.

(c) The components of $\Sigma \setminus \mathcal{A}$ are partitioned into two subsurfaces, $\Sigma_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{z}}$, which meet along $\mathcal{A}$.

(d) Each component of $L_i \setminus \mathcal{A}$ contains exactly one basepoint of $\boldsymbol{w}_i \sqcup \boldsymbol{z}_i$.

(e) $\boldsymbol{w}_1 \sqcup \boldsymbol{w}_2 \subset \Sigma_{\boldsymbol{w}}$ and $\boldsymbol{z}_1 \sqcup \boldsymbol{z}_2 \subset \Sigma_{\boldsymbol{z}}$.

Definition 4.6 The identity (decorated link) cobordism $\mathrm{id}_{\mathbb{L}}$ from $\mathbb{L} = (L, \boldsymbol{w}, \boldsymbol{z})$ to itself is given by the surface $\Sigma = I \times L$ with decorations $\mathcal{A} = I \times Q$, where $Q \subset L \setminus (\boldsymbol{w} \cup \boldsymbol{z})$ is a finite set such that the inclusion induces an isomorphism in π_0.

By the work of Zemke [35], an oriented decorated link cobordism $\mathbb{S}$ from $\mathbb{L}_1$ to $\mathbb{L}_2$ induces an $\mathbb{F}[U]$–equivariant map

$$F^Z_{\mathbb{S}} : \mathrm{HFL}'(\mathbb{L}_1) \to \mathrm{HFL}'(\mathbb{L}_2).$$

Remark 4.7 The map $F^Z_{\mathbb{S}}$ enjoys the following properties:

(a) $F^Z_{\mathbb{S}}$ is invariant under isotopy of Σ in $I \times S^3$ while fixing the boundary, and under isotopy of $\mathcal{A}$ in Σ while keeping

$$\partial\mathcal{A} \subset (L_1 \setminus (\boldsymbol{w}_1 \cup \boldsymbol{z}_1)) \cup (L_2 \setminus (\boldsymbol{w}_2 \cup \boldsymbol{z}_2)).$$

(b) If $\mathrm{id}_{\mathbb{L}}$ is the identity cobordism from $\mathbb{L}$ to itself, then

$$F^Z_{\mathrm{id}_{\mathbb{L}}} = \mathrm{id}_{\mathrm{HFL}'(\mathbb{L})}.$$

(c) If $\mathbb{S}_1$ and $\mathbb{S}_2$ are oriented decorated link cobordisms from $\mathbb{L}_1$ to $\mathbb{L}_2$ and from $\mathbb{L}_2$ to $\mathbb{L}_3$, respectively, then one can stack $\mathbb{S}_2$ on top of $\mathbb{S}_1$ (after isotoping the decorations so that they match on the $\mathbb{L}_2$ level), and obtain a new oriented decorated link cobordism $\mathbb{S}_2 \circ \mathbb{S}_1$ from $\mathbb{L}_1$ to $\mathbb{L}_3$. In such a case,

$$F^Z_{\mathbb{S}_2 \circ \mathbb{S}_1} = F^Z_{\mathbb{S}_2} \circ F^Z_{\mathbb{S}_1}.$$

(d) If $\mathbb{S}' = (\Sigma', \mathcal{A}')$ is obtained from $\mathbb{S} = (\Sigma, \mathcal{A})$ by attaching a tube with both feet in $\Sigma_{\boldsymbol{z}}$ (or both feet in $\Sigma_{\boldsymbol{w}}$), then $F^Z_{\mathbb{S}'} = U \cdot F^Z_{\mathbb{S}}$.

4.2 Fan's unoriented TQFT

By the work of Fan [7], the link Floer TQFT can be extended to the nonorientable case. We review the relevant definitions.

Definition 4.8 A *disoriented link* in S^3 is a tuple $\mathcal{L} = (L, \boldsymbol{p}, \boldsymbol{q})$ consisting of an unoriented, embedded link $L \subset S^3$, with two disjoint collections of points $\boldsymbol{p}$ and $\boldsymbol{q}$ on L, called the *dividing set*, such that each component of L has at least two points in the dividing set and the points in the dividing set alternate between those in $\boldsymbol{p}$ and those in $\boldsymbol{q}$ as one traverses a component of L.

Each component of $L \setminus (\boldsymbol{p} \cup \boldsymbol{q})$ is given a canonical orientation from $\boldsymbol{q}$ to $\boldsymbol{p}$. We denote the oriented manifold $L \setminus (\boldsymbol{p} \cup \boldsymbol{q})$ by $\boldsymbol{l}$. Note that these orientations do not glue to an orientation of L.

As it is customary, we consider isotopic disoriented knots as different disoriented knots. It is well known that isotopies can induce nontrivial maps in knot Floer homology, such as the moving basepoint maps [30; 32].

Definition 4.8 looks exactly the same as Definition 4.1, except that the link is now unoriented. However, we emphasise that the basepoints $\boldsymbol{w} \cup \boldsymbol{z}$ from Definition 4.1 are ontologically different from the dividing set from Definition 4.8. From a Morse-theoretical viewpoint, the former arise as the intersection between L and the middle level surface of a Morse function, whereas the latter are the index-0 and index-3 critical points of the function.

However, we can define a notion of compatibility between oriented decorated links and disoriented links.

Definition 4.9 We say that an oriented decorated link $\mathbb{L} = (L, \boldsymbol{w}, \boldsymbol{z})$ and a disoriented link $\mathcal{L} = (L, \boldsymbol{p}, \boldsymbol{q})$ are *compatible* if

- the underlying unoriented link L is the same (but note that in $\mathbb{L}$ it also comes with an orientation);
- $\boldsymbol{p} \cup \boldsymbol{q}$ is disjoint from $\boldsymbol{w} \cup \boldsymbol{z}$;
- each component of $L \setminus (\boldsymbol{p} \cup \boldsymbol{q})$ contains exactly one basepoint in $\boldsymbol{w} \cup \boldsymbol{z}$;
- the components of $L \setminus (\boldsymbol{w} \cup \boldsymbol{z})$ containing the $\boldsymbol{p}$ point are oriented from z to w (with orientation induced from $\mathbb{L}$).

Remark 4.10 Every disoriented link admits a (noncanonical) compatible oriented decorated link. Likewise, every oriented decorated link admits a (noncanonical) compatible disoriented link.

If two oriented decorated links $\mathbb{L}_1$ and $\mathbb{L}_2$ are compatible with the disoriented link $\mathcal{L}$, then $\mathrm{HFL}'(\mathbb{L}_1)$ and $\mathrm{HFL}'(\mathbb{L}_2)$ are canonically isomorphic. Thus, we can define $\mathrm{HFL}'(\mathcal{L})$ as $\mathrm{HFL}'(\mathbb{L})$ for any $\mathbb{L}$ compatible with $\mathcal{L}$. (More precisely, $\mathrm{HFL}'(\mathcal{L})$ is the transitive system over all compatible oriented decorated links.)

Note that $\mathrm{HFL}'(\mathcal{L})$ does not depend on the orientation chosen on L. If $(L, \boldsymbol{w}, \boldsymbol{z})$ is a compatible oriented decorated link, then the orientation reversal L^r is also part of a compatible oriented decorated link, namely $(L^r, \boldsymbol{z}, \boldsymbol{w})$. The swap of the $\boldsymbol{w}$ and $\boldsymbol{z}$ basepoints does not affect the homology, since the differential was defined to be symmetric in $\boldsymbol{w}$ and $\boldsymbol{z}$ (see (7)). This justifies the name *unoriented knot Floer homology* used in [26].

Remark 4.11 Fan [7] defines other categories of unoriented links, which he calls *bipartite links* and *bipartite disoriented links*. These are essential to define a TQFT framework for disoriented links, but we do not recall them here.

We now revise the cobordism maps defined by Fan [7].

Definition 4.12 A *disoriented link cobordism* from $\mathcal{L}_1 = (L_1, \boldsymbol{p}_1, \boldsymbol{q}_1)$ to $\mathcal{L}_2 = (L_2, \boldsymbol{p}_2, \boldsymbol{q}_2)$ is a pair $\mathcal{S} = (\Sigma, \mathcal{M})$ such that

(a) $\Sigma \subset I \times S^3$ is a properly embedded, compact surface with $\Sigma \cap \{0\} \times S^3 = \{0\} \times (-L_1)$ and $\Sigma \cap \{1\} \times S^3 = \{1\} \times L_2$;

(b) $\mathcal{M} \subset \Sigma$ is a properly embedded, compact, oriented 1–manifold, which we refer to as the *motion* of the critical points;

(c) the components of $\Sigma \backslash \mathcal{M}$ are compact, oriented surfaces with orientation induced by $\mathcal{M}$;

(d) $\partial \mathcal{M} = \boldsymbol{q}_1 - \boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{q}_2$.

Note that, with the orientation given in point (c), $\partial(\Sigma \setminus \mathcal{M}) = \boldsymbol{l}_2 - \boldsymbol{l}_1 + 2\mathcal{M}$. The surface Σ does not need to be oriented.

There is a natural notion of identity cobordism, in the same spirit as Definition 4.6. We do not write this definition explicitly.

By the work of Fan [7], an disoriented link cobordism $\mathcal{S}$ from $\mathcal{L}_1$ to $\mathcal{L}_2$ induces an $\mathbb{F}[U]$–equivariant map

$$F^F_{\mathcal{S}} \colon \mathrm{HFL}'(\mathcal{L}_1) \to \mathrm{HFL}'(\mathcal{L}_2).$$

Remark 4.13 The map $F^F_{\mathcal{S}}$ enjoys the following properties:

(a) $F^F_{\mathcal{S}}$ is invariant under isotopy of Σ in $I \times S^3$ while fixing the boundary, and under isotopy of $\mathcal{M}$ in Σ while fixing the boundary.

(b) If $\mathrm{id}_{\mathcal{L}}$ is the identity cobordism from $\mathcal{L}$ to itself, then

$$F^F_{\mathrm{id}_{\mathcal{L}}} = \mathrm{id}_{\mathrm{HFL}'(\mathcal{L})}.$$

(c) If $\mathcal{S}_1$ and $\mathcal{S}_2$ are oriented disoriented link cobordisms from $\mathcal{L}_1$ to $\mathcal{L}_2$ and from $\mathcal{L}_2$ to $\mathcal{L}_3$, respectively, then one can stack $\mathcal{S}_2$ on top of $\mathcal{S}_1$, and obtain a new oriented disoriented link cobordism $\mathcal{S}_2 \circ \mathcal{S}_1$ from $\mathcal{L}_1$ to $\mathcal{L}_3$. In such a case,

$$F^F_{\mathcal{S}_2 \circ \mathcal{S}_1} = F^F_{\mathcal{S}_2} \circ F^F_{\mathcal{S}_1}.$$

4.3 Relation between Zemke's TQFT and Fan's TQFT

Definition 4.14 For $i = 1, 2$, suppose that $\mathbb{L}_i = (L_i, \boldsymbol{w}_i, \boldsymbol{z}_i)$ and $\mathcal{L}_i$ are compatible. We say that a decorated link cobordism $\mathbb{S} = (\Sigma, \mathcal{A})$ from $\mathbb{L}_1$ to $\mathbb{L}_2$ and a disoriented link cobordism $\mathcal{S} = (\Sigma, \mathcal{M})$ from $\mathcal{L}_1$ to $\mathcal{L}_2$ are *compatible* if

- the underlying unoriented surface Σ is the same (but note that in $\mathbb{S}$ it also comes with an orientation);
- after isotoping $\mathcal{A}$ without crossing $\boldsymbol{w}_1 \sqcup \boldsymbol{z}_1 \sqcup \boldsymbol{w}_2 \sqcup \boldsymbol{z}_2$, $\mathcal{A} = \mathcal{M}$.

Remark 4.15 For $i = 1, 2$, suppose that $\mathbb{L}_i$ and $\mathcal{L}_i$ are compatible. Moreover, suppose that $\mathbb{S}$ is a decorated link cobordism from $\mathbb{L}_1$ to $\mathbb{L}_2$ and $\mathcal{S}$ is a compatible disoriented link cobordism from $\mathcal{L}_1$ to $\mathcal{L}_2$. Then

$$F^Z_{\mathbb{S}} = F^F_{\mathcal{S}}.$$

5 A technical result for HFK′

5.1 The flip cobordism in HFK′

Definition 5.1 The standard disoriented unknot is $\mathcal{U}_1 = (U_1, p, q)$, where $U_1 = \{x^2 + y^2 = 1\} \cap \{z = 0\}$, $p = (1, 0, 0)$ and $q = (-1, 0, 0)$.

Definition 5.2 The *flip cobordism* $\mathcal{F} = (\Sigma_{\mathcal{F}}, \mathcal{M}_{\mathcal{F}})$ from the standard disoriented unknot $\mathcal{U}_1 = (U_1, p, q)$ to itself is the disoriented cobordism traced by the isotopy obtained by rotating U_1 by π along the x–axis. The points p and q stay fixed throughout the isotopy, so we can set $\mathcal{M}_{\mathcal{F}} = I \times \{p, q\}$.

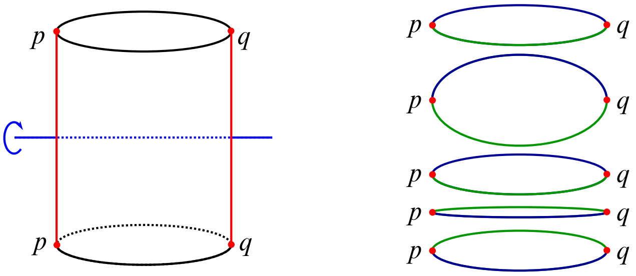


Figure 16: Our notation for the flip cobordism (left) and a few sections of the cobordism (right). We use different colours for the two components of $U_1 \setminus \{p, q\}$ to help the visualisation.

Note that the surface underlying a flip cobordism is orientable, although no orientation of the surface restricts to the same orientation on the two standard disoriented unknots on the boundary.

Lemma 5.3 *The map $F_{\mathcal{F}}^{F}$ induced by the flip cobordism is the identity map on* $\mathrm{HFK}'(\mathcal{U}_1) \cong \mathbb{F}[U]$.

Proof The fourth iteration $\mathcal{F}^4$ is the disoriented cobordism traced by a 4π rotation about the x–axis. Since $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}/2\mathbb{Z}$, the rotation by 4π is isotopic to the identity. Thus, $\mathcal{F}^4$ is isotopic to the identity cobordism, and

$$(F_{\mathcal{F}}^{F})^4 = F_{\mathcal{F}^4}^{F} = \mathrm{id}_{\mathbb{F}[U]}\,. \tag{8}$$

The map

$$F_{\mathcal{F}}^{F} : \mathbb{F}[U] \to \mathbb{F}[U]$$

is U–equivariant, so it is completely determined by the image of 1. If we set $p(U) := F_{\mathcal{F}}^{F}(1) \in \mathbb{F}[U]$, equation (8) implies that $(p(U))^4 = 1$. Since every invertible element of $\mathbb{F}[U]$ must be in $\mathbb{F}$, we deduce $p(U) = 1$. □

5.2 A stabilisation lemma

In this subsection only, we will need to work in a more general setting than the one outlined in Section 4.

First, we will consider decorated links $\mathbb{L}$ in a 3–manifold Y, and decorated link cobordisms $(\Sigma, \mathcal{A})$ in a 4–manifold W. In Section 4, we have stated the definitions of decorated link and decorated link cobordism only when $Y = S^3$ and $W = I \times S^3$. The general definitions are only needed in this subsection, and they can be found in

[35, Definitions 2.1 and 2.4]. Again in this subsection only, we will consider the full chain complex $\mathcal{CFL}^-(Y, \mathbb{K})$ associated to a decorated knot, which is a complex over $\mathbb{F}[U, V]$, up to chain homotopy equivalence.

We will also use the homological action on link Floer homology. (See [27, Section 4.2.5] for the original definition of Heegaard Floer homology, Theorem 3.1 of [29] for the cobordism action, and [34, Section 12.2] for its extension to link Floer homology.) Given a decorated link $\mathbb{L}$ in a 3–manifold Y, there is a homological action

$$A\colon \Lambda^*(H_1(Y;\mathbb{Z})/\mathrm{Tors}) \otimes \mathcal{CFL}^-(Y, \mathbb{L}) \to \mathcal{CFL}^-(Y, \mathbb{L}),$$

and, given a decorated link cobordism $(\Sigma, \mathcal{A})$ in a 4–manifold W, there is a version of the cobordism map incorporating the homological action,

$$F^H_{W,\Sigma,\mathcal{A}}\colon \Lambda^*(H_1(W;\mathbb{Z})/\mathrm{Tors}) \otimes \mathcal{CFL}^-(Y_1, \mathbb{L}_1) \to \mathcal{CFL}^-(Y_2, \mathbb{L}_2).$$

(We use the notation F^H to distinguish it from the cobordism map F^Z, which does not incorporate the homological action.) If the 4–manifold W is obtained by adding 1–handles to B^4, then the map F^H can be recovered from F^Z by postcomposing with the homological action on Y_1. The following lemma, which is needed to establish Proposition 5.5, was proved by Ian Zemke. A related argument appeared in [12, Section 5] (see in particular [12, Lemma 5.3]):

Lemma 5.4 *Let $\Sigma = I \times K \subset I \times S^3$ be the identity cobordism from the knot K to itself, and let Σ' denote a surface obtained by adding a compressible 1–handle to Σ. If $\gamma \subset \Sigma$ denotes an embedded arc joining the feet of the 1–handle, define decorations $\mathcal{A}'$ on Σ' as two parallel embedded arcs from $\{0\} \times K$ to $\{1\} \times K$ such that:*

- *$\mathcal{A}'$ does not intersect γ.*
- *Each arc of $\mathcal{A}'$ crosses the cocore of the 1–handle exactly once.*
- *The arcs of $\mathcal{A}'$ join the points $(0, p)$ and $(0, q)$ to the points $(1, p)$ and $(1, q)$ in $I \times K$, respectively.*
- *The decorations $\mathcal{A}$, obtained by restricting $\mathcal{A}'$ to Σ and by reconnecting each pair of arcs with an arc parallel to γ, are isotopic rel boundary to a product decoration $I \times \{p, q\}$.*

Then, if $\mathbb{K} = (K, w, z)$ for some points w and z alternated to p and q, the cobordism map

$$F^Z_{\Sigma',\mathcal{A}'}\colon \mathrm{HFK}'(\mathbb{K}) \to \mathrm{HFK}'(\mathbb{K})$$

coincides with the map $U \cdot \mathrm{id}_{\mathrm{HFK}'(\mathbb{K})}$.

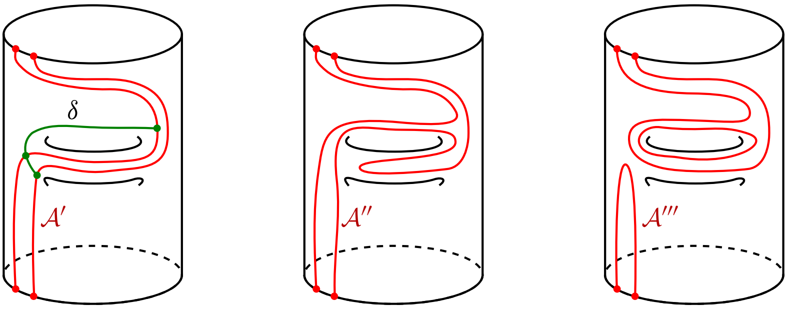


Figure 17: The three decorations $\mathcal{A}'$, $\mathcal{A}''$ and $\mathcal{A}'''$ on the surface Σ' appearing in the bypass relation arising from the arc δ.

Proof If a_1 and a_2 denote the two components of $\mathcal{A}'$, let δ be an arc on Σ' which starts from a_1 near a foot of the 1–handle, then traverses a_2, follows γ to the other foot of the 1–handle, and ends on a_2. See Figure 17, left, for an illustration.

We apply Zemke's bypass relation on a disc $\Delta \subset \Sigma'$ obtained as a regular neighbourhood of the arc δ in Σ'. If $\mathcal{A}''$ and $\mathcal{A}'''$ denote the other decorations appearing in the bypass relation as in Figure 17, we have that

$$F^{Z}_{\Sigma',\mathcal{A}'} = F^{Z}_{\Sigma',\mathcal{A}''} + F^{Z}_{\Sigma',\mathcal{A}'''}.$$

The decorations $\mathcal{A}''$ can be isotoped away from the cocore of the 1–handle. After compressing the 1–handle, the surface becomes isotopic to Σ and the decorations become isotopic to the product decorations $\mathcal{A}$. Thus, by Remark 4.7(d),

$$F^{Z}_{\Sigma',\mathcal{A}''} = U \cdot F^{Z}_{\Sigma,\mathcal{A}} = U \cdot \mathrm{id}_{\mathrm{HFK}'(\mathbb{K})}.$$

Thus, we only need to show that $F^{Z}_{\Sigma',\mathcal{A}'''} = 0$. From this point until the end of the proof we will work on the chain level $\mathcal{CFL}^-(\mathbb{K})$, considered as an $\mathbb{F}[U,V]$–complex, up to chain homotopy equivalence. (The variable U is associated to the basepoint w.)

We split the cobordism $(I \times S^3, \Sigma', \mathcal{A}''')$ as the composition of two cobordisms. The first one, which we call $\mathcal{W}_1 = (W_1, \Sigma_1, \mathcal{A}_1)$, is obtained by taking as W_1 the (disjoint) union of a regular neighbourhood of $\{0\} \times S^3$ and a neighbourhood of $\gamma \cup c$ (where c denotes the core of the 1–handle) containing the 1–handle entirely. Note that the latter component of W_1 is diffeomorphic to $S^1 \times D^3$. The decorated surface $(\Sigma_1, \mathcal{A}_1)$ is obtained by intersecting W_1 with $(\Sigma', \mathcal{A}''')$. The second cobordism $\mathcal{W}_2$ is obtained by taking the closure of the complement of W_1 in $I \times S^3$. Thus, we have

$$F^{Z}_{\Sigma,\mathcal{A}'''} = F^{H}_{(I\times S^3,\Sigma',\mathcal{A}''')} = F^{H}_{\mathcal{W}_2} \circ F^{H}_{\mathcal{W}_1}.$$

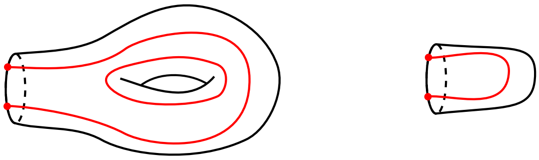


Figure 18: Left: the decorated surface $(\widetilde{\Sigma}_1, \widetilde{\mathcal{A}}_1)$. The punctured torus $\widetilde{\Sigma}_1$ sits in $S^1 \times D^3$ in such a way that its longitude generates $H_1(S^1 \times D^3)$, while its meridian is null-homotopic. Right: the decorated surface $(\overline{\Sigma}_1, \overline{\mathcal{A}}_1)$.

(In the first equality we used the fact that $H_1(I \times S^3) = 0$.)

We focus on the map $F^{\boldsymbol{H}}_{\mathcal{W}_1}$. Since W_1 has two connected components (one of which is an identity cobordism over $\mathbb{K}$), the map splits as a tensor product

$$F^{\boldsymbol{H}}_{\mathcal{W}_1} = \mathrm{id}_{\mathcal{CFL}^-(\mathbb{K})} \otimes F^{\boldsymbol{H}}_{\widetilde{\mathcal{W}}_1}, \tag{9}$$

where $\widetilde{\mathcal{W}}_1 = (S^1 \times D^3, \widetilde{\Sigma}_1, \widetilde{\mathcal{A}}_1)$ is a cobordism from the empty link in the empty 3–manifold to a doubly pointed unknot $\mathbb{U}$ in $S^1 \times S^2$, illustrated in Figure 18, left.

The knot Floer complex $\mathcal{CFL}^-(S^1 \times S^2, \mathbb{U})$ is generated over $\mathbb{F}[U, V]$ by two homogeneous elements $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$. Their $\mathbf{gr}_{\boldsymbol{w}}$ and $\mathbf{gr}_{\boldsymbol{z}}$ gradings (as defined in [34]) are given by

$$(\mathbf{gr}_{\boldsymbol{w}}, \mathbf{gr}_{\boldsymbol{z}})(\boldsymbol{x}_\pm) = \left(\pm\tfrac{1}{2}, \pm\tfrac{1}{2}\right).$$

For grading reasons [34], we have

$$F^{\boldsymbol{H}}_{\widetilde{\mathcal{W}}_1}(1) = k \cdot \boldsymbol{x}_- \tag{10}$$

for some $k \in \mathbb{Z}/2\mathbb{Z}$. An explicit computation of the action of $H_1(S^1 \times S^2) = \mathbb{Z}\langle\zeta\rangle$ shows that $A(\zeta \otimes \boldsymbol{x}_+) = \boldsymbol{x}_-$. From this fact, a direct computation shows that

$$F^{\boldsymbol{H}}_{\overline{\mathcal{W}}_1}(\zeta \otimes 1) = \boldsymbol{x}_-, \tag{11}$$

where $\overline{\mathcal{W}}_1 = (S^1 \times S^2, \overline{\Sigma}_1, \overline{\mathcal{A}}_1)$ is the cobordism shown in Figure 18, right.

Recall that the cobordism $\mathcal{W}_1$ is the disjoint union of an identity cobordism over $\mathbb{K}$ and the cobordism $\widetilde{\mathcal{W}}_1$. If $\widehat{\mathcal{W}}_1$ denotes the cobordism obtained by replacing the $\widetilde{\mathcal{W}}_1$ component with $\overline{\mathcal{W}}_1$, then, by combining (9), (10) and (11), we have

$$F^{\boldsymbol{H}}_{\mathcal{W}_1}(\boldsymbol{x}) = \boldsymbol{x} \otimes F^{\boldsymbol{H}}_{\widetilde{\mathcal{W}}_1}(1) = k \cdot \boldsymbol{x} \otimes F^{\boldsymbol{H}}_{\overline{\mathcal{W}}_1}(\zeta \otimes 1) = k \cdot F^{\boldsymbol{H}}_{\widehat{\mathcal{W}}_1}(\zeta \otimes \boldsymbol{x}). \tag{12}$$

Finally, let $\widehat{\mathcal{W}}$ denote the composition of $\widehat{\mathcal{W}}_1$ and $\mathcal{W}_2$. Note that the 4–manifold underlying $\widehat{\mathcal{W}}$ is still $I \times S^3$ (the same as $\mathcal{W}$), since the replacement of $\mathcal{W}_1$ with $\widehat{\mathcal{W}}_1$

did not affect the underlying 4–manifold. Then, by (12),

$$F^{Z}_{\Sigma,\mathcal{A}'''}(\boldsymbol{x}) = F^{H}_{\mathcal{W}_2} \circ F^{H}_{\mathcal{W}_1}(\boldsymbol{x}) = k \cdot F^{H}_{\mathcal{W}_2} \circ F^{H}_{\widehat{\mathcal{W}}_1}(\zeta \otimes \boldsymbol{x}) = k \cdot F^{H}_{\widehat{\mathcal{W}}}(\iota_*(\zeta) \otimes \boldsymbol{x}) = 0.$$

The last term vanishes because the map $\iota_* \colon H_1(S^1 \times D^3) \to H_1(I \times S^3) = 0$ induced by the inclusion of $\widehat{\mathcal{W}}_1$ into $\widehat{\mathcal{W}}$ must map ζ to 0. □

5.3 The main theorem in HFK′

Proposition 5.5 *Suppose that Σ is a connected nonorientable knot cobordism from K_1 to K_2 in $I \times S^3$ with m local minima, b saddles and M local maxima, and let $\overline{\Sigma}$ denote the mirrored upside down cobordism from K_2 to K_1.*

Then there are choices of motions of the critical points such that the disoriented knot cobordisms $\mathcal{S} = (\Sigma, \mathcal{M}_1)$ and $\bar{\mathcal{S}} = (\overline{\Sigma}, \mathcal{M}_2)$ can be composed to $\bar{\mathcal{S}} \circ \mathcal{S}$, and

$$U^M \cdot F^F_{\bar{\mathcal{S}}} \circ F^F_{\mathcal{S}} = U^{b-m} \cdot \mathrm{id}_{\mathrm{HFK}'(K_1)}. \tag{13}$$

Proof Using Lemma 3.2, we can break the cobordism Σ into the composition of cobordisms labelled (a)–(f). Let K_1' and K_2' be the knots after steps (b) and (d), respectively, as in the statement of Lemma 3.2, and let L' be the link after step (c). Note that L' differs from K_2' by a single band surgery.

By removing the two attaching arcs of the nonorientable band B from L', we are left with two arcs γ and δ. If L' is a knot, let p_a, q_a, p_b, q_b be points on γ, appearing in this order, such that p_a and q_a are close to one end of γ and p_b and q_b are close to the other end of γ, so that all the intersections of $L' \cap \gamma$ with the oriented bands are between p_b and q_a. See Figure 19, left. If instead L' is a two-component link, let p_a and q_a be on γ and p_b and q_b be on δ such that they are near opposite corners of the band and p_a (resp. q_b) is closer to the band than q_a (resp. p_b). See Figure 20.

Let $\mathcal{S}$ be the disoriented cobordism from (K_1, p_a, q_a) to (K_2, p_b, q_a) obtained by endowing Σ with the following motion of basepoints:

- On steps (a)–(c), the motion consists of straight arcs $I \times \{p_a, q_a\}$.
- On step (d), the motion consists of a straight arc $I \times \{q_a\}$ and of an arc that starts from p_a, goes through the nonorientable saddle, and ends at p_b (see Figure 19).
- On steps (e)–(f), the motion consists of straight arcs $I \times \{p_b, q_a\}$.

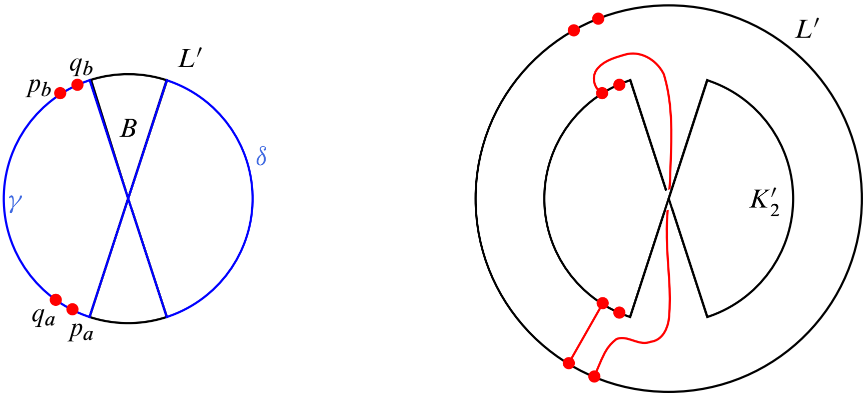


Figure 19: Left: the circle represents L' after step (c) in the case it is a knot. Right: step (d) of the cobordism Σ, from L' to K_2', together with the motion chosen to define $\mathcal{S}$.

A crucial condition in Definition 4.12 is that each component of $\Sigma \setminus \mathcal{M}_1$ must be orientable. In fact, we show that $\Sigma \setminus \mathcal{M}_1$ consists of a single and orientable component. If L' is a knot, one can check from Figure 19 that $\Sigma \setminus \mathcal{M}_1$ restricted to step (d) has a single component; in steps (a)–(c) the surface Σ is orientable and the motion is given by two parallel arcs, so there are two components of $\Sigma \setminus \mathcal{M}_1$, which are then glued to the unique component in step (d); steps (e)–(f) define a concordance of disoriented knots, which does not change the abstract topology of the disoriented cobordism. The compatibility of the orientation of $\Sigma \setminus \mathcal{M}_1$ with the orientation of $\mathcal{M}_1$ is dealt with in an analogous way.

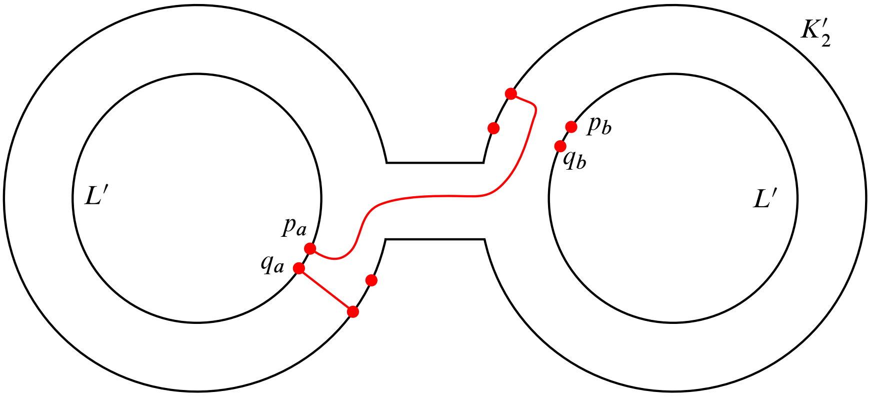


Figure 20: Step (d) of $\mathcal{S}$ in the case L' is a two-component link, represented above by the two inner circles.

If L' is a two-component link, then one should consider Figure 20 instead. Let ζ be the closed component of L' containing p_b and q_b (which appears in Figure 20, right), and let ε be the component containing p_a and q_a, minus the short arc connecting p_a and q_a (which appears in Figure 20, right). From Figure 20, it is immediate to check that $\Sigma \setminus \mathcal{M}_1$ has two components in step (d), which deformation retract on ζ and on ε. Since K_1 is a knot, $\Sigma \setminus \mathcal{M}_1$ in steps (a)–(c) also has two components: a "small" rectangular one, S, spanned by the short arc connecting p_a and q_a, and a large one, L, the complement of it, which contains all the genus. When you glue steps (a)–(c) to step (d), the rectangular component S is glued to the component containing ζ, without affecting the topology, whereas the component L glues to both components of step (d). Thus, we see that there is only one component of $\Sigma \setminus \mathcal{M}_1$. Its orientability and the compatibility with the orientation of $\mathcal{M}_1$ is left to the reader (it basically follows from the fact that cutting along $\mathcal{M}_1$ effectively cuts the nonorientable saddle, leaving an orientable cobordism). As before, we do not worry about steps (e)–(f), since they define a concordance, which does not change the abstract topology.

We next introduce a disoriented cobordism $\bar{\mathcal{S}}$ from (K_2, p_b, q_a) to (K_1, p_a, q_a) with underlying surface $\overline{\Sigma}$. To define it, we play the steps of the cobordism $\mathcal{S}$ in reverse order, but we use a different motion of basepoints:

- On the reversed steps (f)–(e), the motion consists of straight arcs $I \times \{p_b, q_a\}$.
- On the reversed step (d), the motion consists of a straight arc $I \times \{p_b\}$ and of an arc that starts from q_a, goes through the (dual) nonorientable saddle, and ends at q_b.
- On the reversed steps (c)–(a), the motion consists of straight arcs $I \times \{p_b, q_b\}$.
- Finally, in a collar of the K_1 boundary component, the motion of the basepoints brings p_b and q_b back to p_a and q_a.

Note that $\bar{\mathcal{S}}$ is not $\mathcal{S}$ turned upside down as disoriented cobordisms (even if the disoriented knots at the boundary are not the same).

We also define a disoriented cobordism $\mathcal{T}^u$ from (K_1, p_a, q_a) to (K_1, p_a, q_a), obtained in three steps:

- The first step is the same disoriented cobordism as in Figure 19, except that the knot L' is replaced with K_1; more explicitly, the surface Σ in the first step consists of the cylinder $I \times K_1$, with the nonorientable band B attached on the upper end (recall that by Lemma 3.2 all bands have disjoint attaching arcs, so

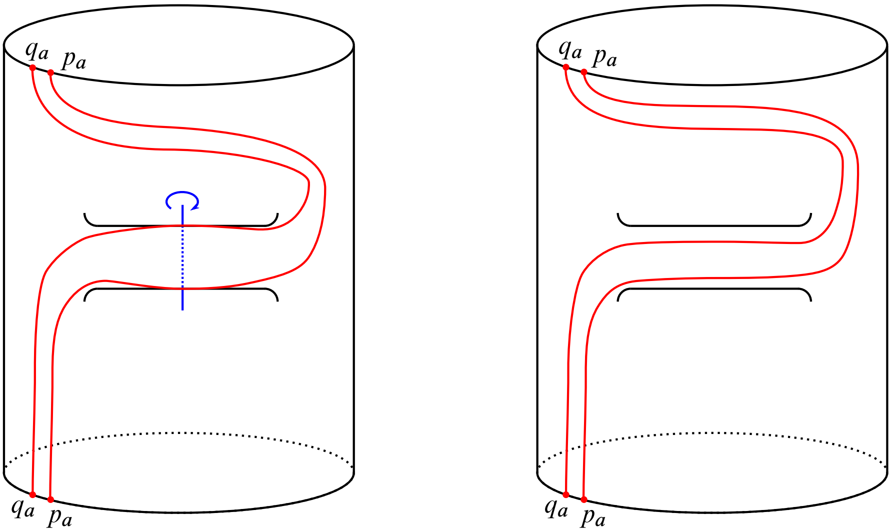


Figure 21: The disoriented cobordism $\mathcal{T}^u$ (left) and the disoriented cobordism $\mathcal{T}^o$ (right).

we can attach B to K_1), and the motion of the basepoints consists of $I \times \{q_a\}$ and of an arc that starts from p_a, goes through the band, and ends at p_b.

- The surface in the second step is simply the surface from the first step turned upside down, and the motion consists of a straight arc $I \times \{p_b\}$ and of an arc that starts from q_a, goes through the (dual) band, and ends at q_b.
- Finally, in a collar of the end boundary component, the motion of the basepoints brings p_b and q_b back to p_a and q_a.

Note that the surface Σ of the disoriented cobordism $\mathcal{T}^u$ is a genus-1 surface, consisting of a cylinder $I \times K_1$ with a flipped tube attached to it. The flipped tube is made up of the two nonorientable bands. See Figure 21, left.

Lastly, we define a variant of $\mathcal{T}^u$: the disoriented cobordism $\mathcal{T}^o$ from (K_1, p_a, q_a) to (K_1, p_a, q_a) is obtained by replacing the flipped tube in $\mathcal{T}^u$ with an orientable tube, so that the underlying surface Σ is orientable (in other words, the nonorientable bands are replaced with orientable bands); the motion of the basepoints divides Σ into a disc and a punctured torus; see Figure 21, right. Note that Lemma 5.3 implies that $F^F_{\mathcal{T}^u} = F^F_{\mathcal{T}^o}$, since it is possible to isolate a flip cobordism.

In order to prove Proposition 5.5 we argue in a similar way as in [11, Proposition 4.1]: we define a cobordism $\mathcal{G}^u$, and we compute the map $F^F_{\mathcal{G}^u}$ in two different ways, which will be the two sides of equation (13).

The disoriented cobordism $\mathcal{G}^u$, from (K_1, p_a, q_a) to itself, is obtained by playing all the steps of $\mathcal{S}$ except (f) followed by all the reversed steps of $\bar{\mathcal{S}}$ except (f). (In $\bar{\mathcal{S}}$ we also play the basepoint moving step in a collar of K_1.)

Since the attaching arc of the unoriented band can be isotoped in K_1, we can move step (d) of $\mathcal{S}$ and the corresponding reversed step of $\bar{\mathcal{S}}$ up past all the other steps of $\mathcal{S}$ and $\bar{\mathcal{S}}$. These two steps together make up the cobordism $\mathcal{T}^u$, which we can replace with the orientable cobordism $\mathcal{T}^o$. The replacement yields a new disoriented cobordism $\mathcal{G}^o$ from $\mathcal{G}^u$, with $F^F_{\mathcal{G}^o} = F^F_{\mathcal{G}^u}$. The advantage of $\mathcal{G}^o$ over $\mathcal{G}^u$ is that the underlying surface of the former is orientable, so $F^F_{\mathcal{G}^o} = F^Z_{\mathbb{G}}$ for a compatible decorated link cobordism $\mathbb{G}$, and we can use the properties of Zemke's TQFT on $F^Z_{\mathbb{G}}$, in particular the one about compressing discs.

Note that in the definition of $\mathcal{G}^u$ (or $\mathcal{G}^o$) we do not play the M deaths of $\mathcal{S}$ (step (f)) and the M births of $\bar{\mathcal{S}}$, obtained by mirroring the deaths of $\mathcal{S}$. Thus, $\bar{\mathcal{S}} \circ \mathcal{S}$ is obtained from $\mathcal{G}^u$ by compressing M discs with boundary in the complement of the motion of the basepoints. By transiting through their orientable replacements, and by Remark 4.7(d), we get

$$F^F_{\mathcal{G}^u} = U^M \cdot \circ F^F_{\bar{\mathcal{S}}} \circ F^F_{\mathcal{S}}. \tag{14}$$

On the other hand, we saw earlier that the cobordism $\mathcal{G}^o$ can be rearranged so that we have $\mathcal{T}^o$ at the top. The first part consists of a cylindrical cobordism from (K_1, p_a, q_a) to itself with $b-1-m$ tubes added, as in [11]. (The -1 here comes from the fact that we have moved the nonorientable band to the top of the cobordism.) Thus, we can compress the cobordism $\mathcal{G}^o$ $b-1-m$ times to get $\mathcal{T}^o$, so

$$F^F_{\mathcal{G}^o} = U^{b-1-m} \cdot F^F_{\mathcal{T}^o}.$$

But the cobordism $\mathcal{T}^o$ is of the form studied in Lemma 5.4, so the map that it induces is multiplication by U. Thus,

$$F^F_{\mathcal{G}^u} = F^F_{\mathcal{G}^o} = U^{b-m} \cdot \mathrm{id}_{\mathrm{HFL}'(K_1)}. \tag{15}$$

By combining (14) and (15), we finish the proof. □

Remark 5.6 The careful reader will note that the motions of the basepoints play an important role in the proof of Proposition 5.5. This is by contrast with Proposition 4.1 of [11], where the decorations of the cobordism were the simplest possible, ie two parallel arcs from the bottom to the top. In the unoriented setting it is impossible to choose two parallel arcs as the motion of basepoints, otherwise the cobordism would not fall in the correct category.

6 Applications

In this section we prove Theorem 1.1, which we restate below, together with its corollaries.

Theorem 1.1 *Let K_1 and K_2 be knots in S^3. Suppose that there is a nonorientable knot cobordism Σ in $I \times S^3$ from K_1 to K_2 with M local maxima. Then*

$$\operatorname{Ord}_I(K_1) \leq \max\{\operatorname{Ord}_I(K_2), M\} + \gamma(\Sigma) \tag{1}$$

and

$$\operatorname{Ord}_U(K_1) \leq \max\{\operatorname{Ord}_U(K_2), M\} + \gamma(\Sigma). \tag{2}$$

Proof The proof closely follows that of [11, Theorem 1.1].

Add decorations on Σ and $\overline{\Sigma}$ (in the instanton or unoriented knot Floer sense) to obtain cobordisms with decorations $\mathcal{S}$ and $\bar{\mathcal{S}}$ that satisfy the relation in Proposition 3.3 or Proposition 5.5,

$$v^M \cdot F_{\bar{\mathcal{S}}} \circ F_{\mathcal{S}} = v^{b-m} \cdot \mathrm{id}_{H(K_1)}. \tag{16}$$

Here m is the number of local minima and b is the number of saddles on Σ, H is either $I^\sharp$ or HFK', F denotes the corresponding map induced by an unoriented cobordism with decorations, and v denotes the relevant variable P or U.

Suppose that $x \in H(K_1)$ is a torsion element. Since $F_{\mathcal{S}}(x)$ must be torsion in $H(K_2)$, $v^l \cdot F_{\bar{\mathcal{S}}} \circ F_{\mathcal{S}}(x) = F_{\bar{\mathcal{S}}}(v^l \cdot F_{\mathcal{S}}(x)) = 0$ whenever $l \geq \operatorname{Ord}(K_2)$. Thus, in view of (16), $v^{l+b-m-M} \cdot x = 0$ whenever $l \geq \max\{\operatorname{Ord}(K_2), M\}$. Since this holds for every torsion element $x \in H(K_1)$, we obtain

$$\operatorname{Ord}(K_1) \leq \max\{\operatorname{Ord}(K_2), M\} + b - m - M,$$

and we conclude by noticing that $\gamma(\Sigma) = -\chi(\Sigma) = b - m - M$. □

We now focus on the proofs of the corollaries from the introduction. Corollary 1.3 follows immediately from Theorem 1.1 by setting $M = 0$, so we move directly to the following corollary, about the refined unoriented cobordism distance.

Recall that, for a cobordism Σ in $I \times S^3$ from K_1 to K_2, we define

$$|\Sigma| = \max\{m, M\} - \chi(\Sigma),$$

and that the *refined nonorientable cobordism distance* between two knots K_1 and K_2 is given by

$$d_u^r(K_1, K_2) = \min\{|\Sigma|\},$$

where Σ varies over all connected nonorientable cobordisms and oriented concordances from K_1 to K_2.

Corollary 1.6 *If K_1 and K_2 are knots in S^3, then*

$$|\mathrm{Ord}_I(K_1) - \mathrm{Ord}_I(K_2)| \le d_u^r(K_1, K_2)$$

and

$$|\mathrm{Ord}_U(K_1) - \mathrm{Ord}_U(K_2)| \le d_u^r(K_1, K_2).$$

Proof The proof follows closely that of [11, Corollary 1.5]. Given a cobordism Σ from K_1 to K_2 with M maxima and m minima of the kind considered in the definition of d_u^r, by Theorem 1.1 (if Σ is nonorientable) and Remark 1.2 (if Σ is an orientable concordance), we have

$$\mathrm{Ord}(K_1) \le \max\{\mathrm{Ord}(K_2), M\} - \chi(\Sigma) \le \mathrm{Ord}(K_2) + M - \chi(\Sigma),$$

where Ord is either Ord_I or Ord_U. From here we get

$$\mathrm{Ord}(K_1) - \mathrm{Ord}(K_2) \le M - \chi(\Sigma) \le \max\{m, M\} - \chi(\Sigma),$$

and we conclude by exchanging the roles of K_1 and K_2, and taking the minimum on the right-hand side. □

Recall that the *unoriented band-unlinking number* $\mathrm{ul}_b^u(K)$ of a knot K in S^3 is defined as the minimum number of (orientable or nonorientable) band surgeries that turn K into an unlink.

Corollary 1.9 *For a knot K in S^3,*

$$\mathrm{Ord}_I(K) \le \mathrm{ul}_b^u(K) \quad \textit{and} \quad \mathrm{Ord}_U(K) \le \mathrm{ul}_b^u(K).$$

Proof The proof is similar to that of [11, Corollary 1.6]. If $b = \mathrm{ul}_b^u(K)$, one can build a cobordism Σ from K to the unknot U with b saddles and M local maxima, by attaching b bands to K to get an $(M+1)$–component unlink and then capping off M components of the latter. By applying Theorem 1.1, and Remark 1.2 if necessary (ie if Σ is orientable), we get (for $I^\sharp$ or HFK')

$$\mathrm{Ord}(K) \le \max\{\mathrm{Ord}(U), M\} - \chi(\Sigma) = M - \chi(\Sigma),$$

since the unknot has vanishing torsion order in both $I^\sharp$ and HFK'. We conclude by noticing that $\chi(\Sigma) = M - b$. □

7 Examples

Lemma 7.1 *For the torus knot $T_{n,n+1}$, $\mathrm{Ord}_U(T_{n,n+1}) = \lfloor \frac{1}{2}n \rfloor$.*

Proof Any torus knot is an L–space knot, so its Alexander polynomial determines the full knot Floer complex CFK^∞ up to chain homotopy equivalence, and in turn the unoriented knot Floer homology. See [28; 25; 11].

If K is an L–space knot, its Alexander polynomial takes the form

$$\Delta_K(t) = \sum_{k=0}^{2l} (-1)^k t^{\alpha_k} \tag{17}$$

for a decreasing sequence of integers $\alpha_0, \dots, \alpha_{2l}$. Let $d_1, \dots, d_{2l}$ denote the gaps, ie $d_k = \alpha_k - \alpha_{k-1}$. Then the full knot Floer complex is (up to chain homotopy equivalence) a staircase $\mathbb{F}[U, U^{-1}]$–module, generated by $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{2l}$, with

$$\partial \boldsymbol{x}_{2k} = 0, \quad \partial \boldsymbol{x}_{2k+1} = \boldsymbol{x}_{2k} + \boldsymbol{x}_{2k+2}.$$

Moreover, the filtration over $\mathbb{Z} \oplus \mathbb{Z}$ is determined up to an overall shift by the following properties:

- The element $\boldsymbol{x}_{2k+1}$ has the same j–filtration as $\boldsymbol{x}_{2k}$, but the i–filtration differs by d_{2k+1}.
- The element $\boldsymbol{x}_{2k+1}$ has the same i–filtration as $\boldsymbol{x}_{2k+2}$, but the j–filtration differs by d_{2k+2}.

Then the unoriented knot Floer complex $\mathrm{CFK}'(K)$ (up to chain homotopy equivalence) is generated over $\mathbb{F}[U]$ by $\boldsymbol{y}_0, \dots, \boldsymbol{y}_{2l}$, with differential

$$\partial \boldsymbol{y}_{2k} = 0, \quad \partial \boldsymbol{y}_{2k+1} = U^{d_{2k+1}} \cdot \boldsymbol{y}_{2k} + U^{d_{2k+2}} \cdot \boldsymbol{y}_{2k+2}.$$

In the language of [6], this is a *standard complex associated to a graded root*. Graded roots were introduced by Némethi [24] to study HF^+ of plumbed 3–manifolds. We instead consider the "upside-down" graded roots as in [6], which are used to describe HF^-. Note that our generators $\boldsymbol{y}_0, \dots, \boldsymbol{y}_{2l+1}$ were called $v_1, \alpha_1, v_2, \alpha_2, \dots, \alpha_{n-1}, v_n$ in [6]. The numbers d_i determine the graded root up to an overall shift: the (relative) grading is given by

$$\chi(\boldsymbol{y}_{2k}) - \chi(\boldsymbol{y}_{2k+1}) = d_{2k+1}, \quad \chi(\boldsymbol{y}_{2k+2}) - \chi(\boldsymbol{y}_{2k+1}) = d_{2k+2}.$$

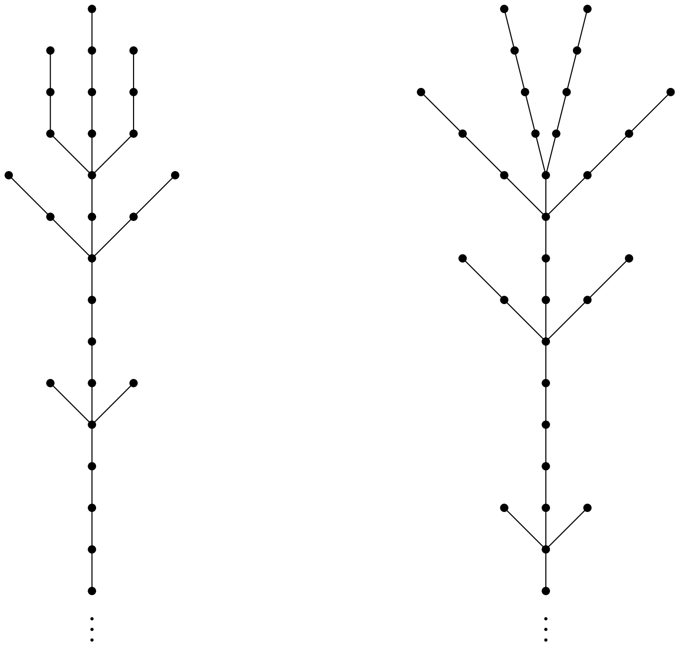


Figure 22: The graded roots homotopy equivalent to $\mathrm{CFK}'(T_{7,8})$ and $\mathrm{CFK}'(T_{8,9})$, respectively. Each dot denotes a generator of the complex over $\mathbb{F}$, and the edges encode the U–action: for a dot $\boldsymbol{x}$, $U \cdot \boldsymbol{x}$ is the dot you get by following the edge exiting from the bottom of the dot $\boldsymbol{x}$. The height of the dot denotes its (relative) Maslov grading, and the U–action decreases the Maslov grading by 2. Note that when n is odd (eg $T_{7,8}$), there is one branch of length $\lceil \frac{1}{2}n \rceil$ and two branches of length $\lfloor \frac{1}{2}n \rfloor$, whereas when n is even (eg $T_{8,9}$) there are two branches of length $\frac{1}{2}n$.

We now determine the numbers d_i in the case of the torus knot $T_{n,n+1}$. Recall that the Alexander polynomial of $T_{p,q}$ is

$$\Delta_{p,q}(t) = \frac{(t^{pq}-1)\cdot(t-1)}{(t^p-1)\cdot(t^q-1)}.$$

The coefficients of $\Delta_{p,q}(t)$ have been computed in the general case (see for example [20, (1.6) and (2.16)] or [19]). In our case, $p = n$ and $q = n+1$, and $\Delta_{p,q}$ is simple enough to be computed explicitly. After simplifying

$$\Delta_{n,n+1} = \frac{(x^n)^n + (x^n)^{n-1} + \cdots x^n + 1}{x^n + x^{n-1} + \cdots + x + 1},$$

one can carry out the long division explicitly and find that $\Delta_{n,n+1}$ is in the form of (17), with

$$(d_1, d_2, d_3, d_4, \ldots, d_{2l-1}, d_{2l}) = (1, n-1, 2, n-2, \ldots, n-1, 1).$$

From this one can check that the graded root has a picture with n branches, of lengths $1, 2, \dots, 2, 1$. See Figure 22. The longest branch is in the middle, of length $\lceil \frac{1}{2}n \rceil$. This is also the top graded branch, so it generates the infinite tower. Thus, the order of HFK' is given by the next longest branch, which has length $\lfloor \frac{1}{2}n \rfloor$. Thus, $\mathrm{Ord}_U(T_{n,n+1}) = \lfloor \frac{1}{2}n \rfloor$. □

We now restrict the attention to the torus knots of the form $T_{2r-1,2r}$.

Remark 7.2 Batson [3] first proved that $\gamma_4(T_{2r-1,2r}) = r-1$. This can be proved with any of the bounds from [3; 26; 8] (for $T_{2r-1,2r}$ or $\overline{T}_{2r-1,2r}$), which all give the same sharp obstruction. We choose to use υ from [26] because it is an additive quantity, like the knot signature. In [26, Theorem 1.2], Ozsváth, Stipsicz and Szabó proved that, for a knot K in S^3,

$$\gamma_4(K) \geq \upsilon(K) - \tfrac{1}{2}\sigma(K). \tag{18}$$

Batson [3] computed that $\sigma(T_{2r-1,2r}) = -2r^2 + 2$, and, by [26, Theorem 1.3], one can compute $\upsilon(T_{2r-1,2r}) = -r^2 + r$. Thus,

$$\upsilon(T_{2r-1,2r}) - \tfrac{1}{2}\sigma(T_{2r-1,2r}) = r-1. \tag{19}$$

We now restate and prove Corollary 1.7 from the introduction:

Corollary 1.7 *For all $\gamma \geq 1$ and $m \geq 1$, there exists a knot $K_{\gamma,m}$ with $d_u(K_{\gamma,m}, U_1) = \gamma_4(K_{\gamma,m}) = \gamma$ and such that $d_u^r(K_{\gamma,m}, U_1) \geq \gamma + m$.*

Thus, each nonorientable surface $\Sigma \subset B^4$ with $\partial\Sigma = K_{\gamma,m}$ and $\gamma(\Sigma) = \gamma$ has at least m local minima (with respect to the radial function).

Proof Let $K_{\gamma,m} = T_{2r-1,2r} \# \overline{T}_{2s-1,2s}$ for $r = \gamma + m$ and $s = m$. By (18), the additivity of σ and υ, and (19), we have

$$\begin{aligned}\gamma_4(K_{\gamma,m}) &\geq \big(\upsilon(T_{2r-1,2r}) - \tfrac{1}{2}\sigma(T_{2r-1,2r})\big) - \big(\upsilon(T_{2s-1,2s}) - \tfrac{1}{2}\sigma(T_{2s-1,2s})\big) \\ &= (r-1) - (s-1) = r - s = \gamma.\end{aligned}$$

On the other hand, Batson showed that there is a sequence of $r-s$ unoriented band surgeries from $T_{2r-1,2r}$ to $T_{2s-1,2s}$ [3, Figure 7]. Thus, we get a sequence of $r-s$ unoriented band surgeries from $K_{\gamma,m}$ to $T_{2s-1,2s} \# \overline{T}_{2s-1,2s}$, which is slice, so $\gamma_4(K_{\gamma,m}) = r - s$.

Now let Σ be a (possibly nonorientable) surface $\Sigma \subset B^4$ with $\partial\Sigma = K_{\gamma,m}$ and $b_1(\Sigma) = \gamma$. Theorem 1.1 gives a lower bound on the number of local minima. More precisely, if Σ has n local minima, by removing a small ball from B^4 we get a cobordism from $K_{\gamma,m}$ to the unknot U_1 with $M = n-1$ maxima (note that the cobordism is upside down, so the minima are turned into maxima, and one of them disappears when we remove the ball). Thus, Theorem 1.1 implies that

$$\mathrm{Ord}_U(K_{\gamma,m}) \le (n-1) + (r-s) = n - s + r - 1.$$

We also know that

$$\mathrm{Ord}_U(K_{\gamma,m}) = \max\{\mathrm{Ord}_U(T_{2r-1,2r}), \mathrm{Ord}_U(T_{2s-1,2s})\} = r - 1$$

by Remark 4.4(d)–(e) and Lemma 7.1, so we get

$$n \ge s = m.$$

The statement about d_u and d_u^r follows from the computation of $\gamma_4(K_{\gamma,m})$ above and from Corollary 1.6. □

Remark 7.3 We do not know if the bound on $d_u^r(K_{\gamma,m}, U_1)$ and on the number of minima of Σ in Corollary 1.7 is sharp on the knots used in the proof of the corollary. Recall that we set

$$K_{\gamma,m} := T_{2r-1,2r} \# \overline{T}_{2s-1,2s}$$

for $r = \gamma + m$ and $s = m$. Batson showed that with γ bands we can get to $K_{0,m} = T_{2s-1,2s} \# \overline{T}_{2s-1,2s}$, and Juhász, Miller and Zemke showed that $K_{0,m}$ bounds a ribbon disc with $2m-1$ local minima. Thus,

$$d_u^r(K_{\gamma,m}, U_1) \le \gamma + 2m - 2.$$

We conjecture that this inequality is actually an equality.

References

[1] **A Alishahi**, **E Eftekhary**, *Knot Floer homology and the unknotting number*, Geom. Topol. 24 (2020) 2435–2469 MR Zbl

[2] **S Baader**, *Scissor equivalence for torus links*, Bull. Lond. Math. Soc. 44 (2012) 1068–1078 MR Zbl

[3] **J Batson**, *Nonorientable slice genus can be arbitrarily large*, Math. Res. Lett. 21 (2014) 423–436 MR Zbl

[4] **C Caprau**, **N González**, **C R S Lee**, **A M Lowrance**, **R Sazdanović**, **M Zhang**, *On Khovanov homology and related invariants*, from "Research directions in symplectic and contact geometry and topology" (B Acu, C Cannizzo, D McDuff, Z Myer, Y Pan, L Traynor, editors), Assoc. Women Math. Ser. 27, Springer (2021) 273–292 MR Zbl

[5] **A Daemi**, **T Lidman**, **D S Vela-Vick**, **C-M M Wong**, *Ribbon homology cobordisms*, Adv. Math. 408 (2022) art. id. 108580 MR Zbl

[6] **I Dai**, **C Manolescu**, *Involutive Heegaard Floer homology and plumbed three-manifolds*, J. Inst. Math. Jussieu 18 (2019) 1115–1155 MR Zbl

[7] **H Fan**, *Unoriented cobordism maps on link Floer homology*, PhD thesis, University of California, Los Angeles (2019) Available at `https://escholarship.org/uc/item/9823t959`

[8] **M Golla**, **M Marengon**, *Correction terms and the nonorientable slice genus*, Michigan Math. J. 67 (2018) 59–82 MR Zbl

[9] **J Hoste**, **Y Nakanishi**, **K Taniyama**, *Unknotting operations involving trivial tangles*, Osaka J. Math. 27 (1990) 555–566 MR Zbl

[10] **A Juhász**, *Cobordisms of sutured manifolds and the functoriality of link Floer homology*, Adv. Math. 299 (2016) 940–1038 MR Zbl

[11] **A Juhász**, **M Miller**, **I Zemke**, *Knot cobordisms, bridge index, and torsion in Floer homology*, J. Topol. 13 (2020) 1701–1724 MR Zbl

[12] **A Juhász**, **I Zemke**, *New Heegaard Floer slice genus and clasp number bounds*, preprint (2020) arXiv 2007.07106

[13] **S Kang**, *Link homology theories and ribbon concordances*, Quantum Topol. 13 (2022) 183–205 MR Zbl

[14] **P B Kronheimer**, **T S Mrowka**, *Khovanov homology is an unknot-detector*, Publ. Math. Inst. Hautes Études Sci. 113 (2011) 97–208 MR Zbl

[15] **P B Kronheimer**, **T S Mrowka**, *Knot homology groups from instantons*, J. Topol. 4 (2011) 835–918 MR Zbl

[16] **P B Kronheimer**, **T S Mrowka**, *A deformation of instanton homology for webs*, Geom. Topol. 23 (2019) 1491–1547 MR Zbl

[17] **P B Kronheimer**, **T S Mrowka**, *Instantons and Bar-Natan homology*, Compos. Math. 157 (2021) 484–528 MR Zbl

[18] **P B Kronheimer**, **T S Mrowka**, *Instantons and some concordance invariants of knots*, J. Lond. Math. Soc. 104 (2021) 541–571 MR Zbl

[19] **T Y Lam**, **K H Leung**, *On the cyclotomic polynomial $\Phi_{pq}(X)$*, Amer. Math. Monthly 103 (1996) 562–564 MR Zbl

[20] **H W Lenstra, Jr**, *Vanishing sums of roots of unity*, from "Proc. Bicentennial Congress Wiskundig Genootschap, II" (P C Baayen, D van Dulst, J Oosterhoff, editors), Math. Centre Tracts 101, Math. Centrum, Amsterdam (1979) 249–268 MR Zbl

[21] **A S Levine**, **I Zemke**, *Khovanov homology and ribbon concordances*, Bull. Lond. Math. Soc. 51 (2019) 1099–1103 MR Zbl

[22] **W B R Lickorish**, *Unknotting by adding a twisted band*, Bull. London Math. Soc. 18 (1986) 613–615 MR Zbl

[23] **M Miller**, **I Zemke**, *Knot Floer homology and strongly homotopy-ribbon concordances*, Math. Res. Lett. 28 (2021) 849–861 MR Zbl

[24] **A Némethi**, *On the Ozsváth–Szabó invariant of negative definite plumbed* 3*–manifolds*, Geom. Topol. 9 (2005) 991–1042 MR Zbl

[25] **P S Ozsváth**, **A I Stipsicz**, **Z Szabó**, *Concordance homomorphisms from knot Floer homology*, Adv. Math. 315 (2017) 366–426 MR Zbl

[26] **P S Ozsváth**, **A I Stipsicz**, **Z Szabó**, *Unoriented knot Floer homology and the unoriented four-ball genus*, Int. Math. Res. Not. 2017 (2017) 5137–5181 MR Zbl

[27] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158 MR Zbl

[28] **P Ozsváth**, **Z Szabó**, *On knot Floer homology and lens space surgeries*, Topology 44 (2005) 1281–1300 MR Zbl

[29] **P Ozsváth**, **Z Szabó**, *Holomorphic triangles and invariants for smooth four-manifolds*, Adv. Math. 202 (2006) 326–400 MR Zbl

[30] **S Sarkar**, *Moving basepoints and the induced automorphisms of link Floer homology*, Algebr. Geom. Topol. 15 (2015) 2479–2515 MR Zbl

[31] **S Sarkar**, *Ribbon distance and Khovanov homology*, Algebr. Geom. Topol. 20 (2020) 1041–1058 MR Zbl

[32] **I Zemke**, *Quasistabilization and basepoint moving maps in link Floer homology*, Algebr. Geom. Topol. 17 (2017) 3461–3518 MR Zbl

[33] **I Zemke**, *Knot Floer homology obstructs ribbon concordance*, Ann. of Math. 190 (2019) 931–947 MR Zbl

[34] **I Zemke**, *Link cobordisms and absolute gradings on link Floer homology*, Quantum Topol. 10 (2019) 207–323 MR Zbl

[35] **I Zemke**, *Link cobordisms and functoriality in link Floer homology*, J. Topol. 12 (2019) 94–220 MR Zbl

Department of Mathematics, Texas A&M University
College Station, TX, United States

Alfréd Rényi Institute for Mathematics
Budapest, Hungary

sgongli@math.tamu.edu, marengon@renyi.hu

http://www.people.tamu.edu/~sgongli, https://users.renyi.hu/~marengon/

Received: 1 December 2020 Revised: 12 October 2021

Algebraic & Geometric Topology **23**:6 (2023) 2673–2713
DOI: 10.2140/agt.2023.23.2673
Published: 7 September 2023

A uniqueness theorem for transitive Anosov flows obtained by gluing hyperbolic plugs

FRANÇOIS BÉGUIN
BIN YU

In work with C Bonatti, we defined a general procedure to build new examples of Anosov flows in dimension 3. The procedure consists in gluing together some building blocks, called *hyperbolic plugs*, along their boundary in order to obtain a closed three-manifold endowed with a complete flow. The main theorem of that work states that (under some mild hypotheses) it is possible to choose the gluing maps so the resulting flow is Anosov. Here we show a uniqueness result for Anosov flows obtained by such a procedure. Roughly speaking, we show that the orbital equivalence class of these Anosov flows is insensitive to the precise choice of the gluing maps used in the construction. The proof relies on a coding procedure, which we find interesting for its own sake, and follows a strategy that was introduced by T Barbot in a particular case.

37D20; 57M99

1 Introduction

In a previous paper, written with C Bonatti [5], we have proved a result allowing one to construct transitive Anosov flows in dimension 3 by "gluing hyperbolic plugs along their boundaries". The purpose here is to study Anosov flows obtained by such a construction. We focus our attention on the diffeomorphisms that are used to glue together the boundaries of the hyperbolic plugs. We aim to understand what is the impact of the choice of these diffeomorphisms on the dynamics of the resulting Anosov flows. We will see that two gluing diffeomorphisms that are "strongly isotopic" yield some Anosov flows that are orbitally equivalent. In other words, in [5], we have proved the *existence* of Anosov flows constructed by a certain gluing procedure, and the goal here is to prove a *uniqueness result* for these Anosov flows.

In order to state some precise questions and results, we need to introduce some terminology. A *hyperbolic plug* is a pair (U, X), where U is a (not necessarily connected) compact three-dimensional manifold with boundary and X is a vector field on U, transverse to ∂U and such that the maximal invariant set $\Lambda_X := \bigcap_{t\in\mathbb{R}} X^t(U)$ is a saddle hyperbolic set for the flow (X^t). Given such a hyperbolic plug (U, X), we decompose ∂U as the disjoint union of an *entrance boundary* $\partial^{\text{in}}U$ (the union of the connected components of ∂U where the vector field X is pointing into U) and an *exit boundary* $\partial^{\text{out}}U$ (the union of the connected components of ∂U where the vector field X is pointing out of U). The stable lamination $W^s(\Lambda_X)$ of the maximal invariant set Λ_X intersects transversally the entrance boundary $\partial^{\text{in}}U$ and is disjoint from the exit boundary $\partial^{\text{out}}U$. Hence, $L^s_X := W^s(\Lambda_X) \cap \partial U$ is a one-dimensional lamination embedded in the surface $\partial^{\text{in}}U$. Similarly, $L^u_X := W^u(\Lambda_X) \cap \partial U$ is a one-dimensional lamination embedded in the surface $\partial^{\text{out}}U$. We call L^s_X and L^u_X the *entrance lamination* and the *exit lamination* of the hyperbolic plug (U, X). It can be proved that these laminations are quite simple:

(i) They contain only finitely many compact leaves.

(ii) Every half noncompact leaf is asymptotic to a compact leaf.

(iii) Each compact leaf may be oriented such that its holonomy is a contraction.

Hyperbolic plugs should be thought as the basic blocks of a building game, our goal being to build some Anosov flows by gluing a collection of such basic blocks together. From a formal viewpoint, a finite collection of hyperbolic plugs can always be viewed as a single nonconnected hyperbolic plug. For this reason, it is enough to consider a single hyperbolic plug (U, X) and a gluing diffeomorphism $\psi\colon \partial^{\text{out}}U \to \partial^{\text{in}}U$. For such (U, X) and ψ, the quotient space $M := U/\psi$ is a closed three-manifold, and the incomplete flow (X^t) on U induces a complete flow (Y^t) on M. The purpose of [5] was to describe some sufficient conditions on U, X and ψ for (Y^t) to be an Anosov flow. We will now explain these conditions.

We say that a one-dimensional lamination L is *filling a surface S* if every connected component C of $S \setminus L$ is "a strip whose width tends to 0 at both ends"; more precisely, C is simply connected, the accessible boundary of C consists of two distinct noncompact leaves ℓ^- and ℓ^+ of L, and these two leaves ℓ^- and ℓ^+ are asymptotic to each other at both ends. We say that two laminations L_1 and L_2 embedded in the same surface S are *strongly transverse* if they are transverse to each other and, moreover, every connected component C of $S \setminus (L_1 \cup L_2)$ is a topological disc whose boundary ∂C consists of exactly four arcs α_1, α_2, α'_1 and α'_2, where α_1 and α'_1 are arcs of leaves of the lamination L_1 and α_2 and α'_2 are arcs of leaves of the lamination L_2. We say that a hyperbolic plug (U, X) has *filling laminations* if the entrance lamination L^s_X is filling the surface $\partial^{\text{in}}U$ and the exit lamination L^u_X is filling the surface $\partial^{\text{out}}U$. Given a hyperbolic plug (U, X), we say that a gluing diffeomorphism $\psi\colon \partial^{\text{out}}U \to \partial^{\text{in}}U$ is *strongly transverse* if the laminations L^s_X and $\psi_* L^u_X$ (both embedded in the surface $\partial^{\text{in}}U$) are strongly transverse. If (U, X_1) and (U, X_2) are two hyperbolic plugs with the same underlying manifold U and $\psi_1, \psi_2\colon \partial^{\text{out}}U \to \partial^{\text{in}}U$ are two gluing diffeomorphisms, we say that the triples (U, X_1, ψ_1) and (U, X_2, ψ_2) are *strongly isotopic* if one can find a continuous one-parameter family $\{(U, X_t, \psi_t)\}_{t\in[1,2]}$ such that (U, X_t) is a hyperbolic plug and $\psi_t\colon \partial^{\text{out}}U \to \partial^{\text{in}}U$ is a strongly transverse gluing map for every t. The main technical result of [5] can be stated as follows:

Theorem 1.1 *Let (U, X_0) be a hyperbolic plug with filling laminations such that the maximal invariant set of (U, X_0) contains neither attractors nor repellers, and let $\psi_0\colon \partial^{\text{out}}U \to \partial^{\text{in}}U$ be a strongly transverse gluing diffeomorphism. Then there exist a hyperbolic plug (U, X) with filling laminations and a strongly transverse gluing diffeomorphism $\psi\colon \partial^{\text{out}}U \to \partial^{\text{in}}U$ such that (U, X_0, ψ_0) and (U, X, ψ) are strongly isotopic, and such that the vector field Y induced by X on the closed manifold $M := U/\psi$ is Anosov.*

The idea of building transitive Anosov flows by gluing hyperbolic plugs goes back to [7], where Bonatti and R Langevin consider a very simple hyperbolic plug (U, X) whose maximal invariant set is a single isolated periodic orbit and are able to find an explicit gluing diffeomorphism $\psi\colon \partial^{\text{out}}U \to \partial^{\text{in}}U$ such that the vector field Y induced by X on the closed manifold $M := U/\psi$ generates a transitive Anosov flow. This example was later generalized by T Barbot, who defined a infinite family of transitive Anosov flows which he calls *BL flows*. These examples are obtained by considering the same very simple hyperbolic plug (U, X) as Bonatti and Langevin, but more general gluing diffeomorphisms.

Theorem 1.1 naturally raises the following question (see [5, Question 1.7]): *In the statement of Theorem 1.1, is the Anosov vector field Y well defined up to orbitally equivalence*? (Recall that two Anosov flows are said to be *orbitally equivalent* if there exists a homeomorphism between their phase space mapping the oriented orbits of the first flow to the oriented orbits of the second one.) One of the main purposes of the present paper is to provide a positive answer to this question. More precisely, we will prove the following:

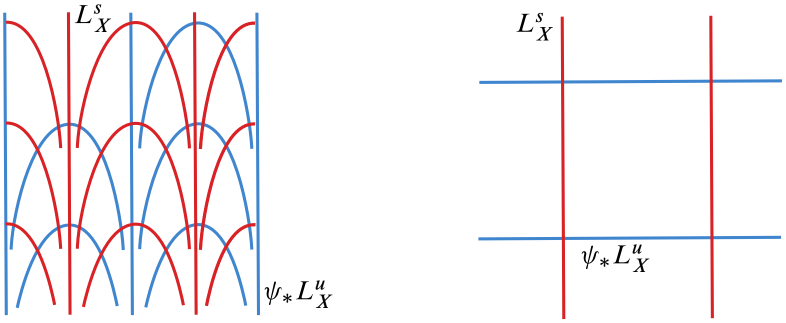


Figure 1: Two examples of strongly transverse gluing diffeomorphisms. On the left-hand side, the laminations are filling. The right-hand side corresponds to Bonatti and Langevin's example.

Theorem 1.2 *Let (U, X_1, ψ_1) and (U, X_2, ψ_2) be two hyperbolic plugs endowed with strongly transverse gluing diffeomorphisms. Let Y_1 and Y_2 be the vector fields induced by X_1 and X_2 on the closed manifolds $M_1 := U/\psi_1$ and $M_2 := U/\psi_2$. Suppose that*:

(0) *The manifolds U, M_1 and M_2 are orientable.*

(1) *Both Y_1 and Y_2 are transitive Anosov vector fields.*

(2) *The triples (U, X_1, ψ_1) and (U, X_2, ψ_2) are strongly isotopic.*

Then the flows (Y_1^t) and (Y_2^t) are orbitally equivalent.

Remark 1.3 In the statement of Theorem 1.2, we do not require that the hyperbolic plugs (U, X_1) and (U, X_2) have filling laminations. So Theorem 1.2 concerns a class of Anosov flows which is larger than the class of Anosov flows provided by Theorem 1.1. For example, Bonatti and Langevin's classical example and its generalizations by Barbot (BL flows) satisfy the hypotheses of Theorem 1.2.

Remark 1.4 On the other hand, we require the Anosov vector fields Y_1 and Y_2 to be transitive. The result is probably still true without this assumption. Nevertheless, at some point of our proof, we will need some leaves of the weak (un)stable foliations of Y_1 and Y_2 to be dense. This denseness is not true in general for nontransitive Anosov vector fields. Note that [5, Proposition 1.6] provides a sufficient condition for an Anosov vector field constructed by gluing some hyperbolic plugs to be transitive.

Remark 1.5 A possible application of Theorem 1.2 is to get some finiteness results. Suppose we are given a hyperbolic plug (U, X) and a diffeomorphism $\psi_0 : \partial^{\text{out}} U \to \partial^{\text{in}} U$. Consider the partition of the isotopy class of ψ_0 into strong isotopy classes. Although we did not write down a complete proof, it seems to us that this partition is finite.

Roughly speaking:

- The stable lamination $L^s_X = W^s(\Lambda_X) \cap \partial^{\mathrm{in}} U$ have finitely many compact leaves which cut $\partial^{\mathrm{in}} U$ in finitely many annuli $A^s_1, \dots, A^s_k$.
- The unstable lamination $L^u_X = W^u(\Lambda_X) \cap \partial^{\mathrm{out}} U$ have finitely many compact leaves which cut $\partial^{\mathrm{out}} U$ in finitely many annuli $A^u_1, \dots, A^u_\ell$.
- It seems that (except in a finite number of some very specific situations) the strong isotopy class of a gluing map ψ (isotopic to ψ_0) only depends on whether the annulus $\psi(A^u_i)$ intersects the annulus A^s_j for each (i, j) (which would of course imply that there are only finitely many possible strong isotopy classes for ψ_0.

Assume that the partition in strong isotopy classes is indeed finite. By Theorem 1.2, this means the following: up to orbital equivalence, there are only finitely many transitive Anosov flows that are built using the hyperbolic plug (U, X) and a gluing map $\psi \colon \partial^{\mathrm{out}} U \to \partial^{\mathrm{in}} U$ isotopic to ψ_0. A further consequence should be that, if we consider some given hyperbolic plugs $(U_1, X_1), \dots, (U_n, X_n)$ such that $U_1, \dots, U_n$ are hyperbolic manifolds, and if we consider a manifold M, then, up to orbital equivalence, there should only finitely many transitive Anosov flows on M that are obtained by gluing $(U_1, X_1), \dots, (U_n, X_n)$.

An analog of Theorem 1.2 was proved by Barbot in the much more restrictive context of BL flows (see [2, Theorem B(2)]). Barbot's result can actually be considered as a particular case of Theorem 1.2: it corresponds to the case where the maximal invariant set of the hyperbolic plug (U_i, X_i) is a single isolated periodic orbit for $i = 1, 2$. Our proof of Theorem 1.2 roughly follows the same strategy as that of Barbot's result, but is far more intricate and requires some important new ingredients since we manipulate general hyperbolic plugs.

The proof is based on a coding procedure that we will describe now. Consider a hyperbolic plug (U, X) and a strongly transverse gluing diffeomorphism $\psi \colon \partial^{\mathrm{out}} U \to \partial^{\mathrm{in}} U$. Let Y be the vector field induced by X on the closed manifold $M := U/\psi$, and assume that the flow (Y^t) is a transitive Anosov flow. The projection in M of ∂U is a closed surface transverse to the orbits of the Anosov flow (Y^t); we denote this surface by S. The projection in M of the entrance lamination of the plug (U, X) is a lamination in the surface S; we denote it by L^s. Consider the universal cover $\widetilde{M}$ of the manifold M and the lifts $(\widetilde{Y}^t)$, $\widetilde{S}$ and $\widetilde{L}^s$ of (Y^t), S and L^s. We will consider the (countable) alphabet $\mathcal{A}$ whose letters are the connected components of $\widetilde{S} \setminus \widetilde{L}^s$, and the symbolic space Σ whose elements are bi-infinite words on the alphabet $\mathcal{A}$. We will construct a coding map χ from (a dense subset of) the surface $\widetilde{S}$ to the symbolic space Σ, commuting with the natural actions of the fundamental group of M, and conjugating the Poincaré first return map of the flow $(\widetilde{Y}^t)$ on the surface $\widetilde{S}$ to the shift map on the symbolic space Σ. If Λ denotes the projection in M of the maximal invariant set of the plug (U, X), and $\widetilde{\Lambda}$ denotes the lift of Λ in $\widetilde{M}$, then the map χ is defined at every point of $\widetilde{S}$ which is neither in the stable nor in the unstable lamination of $\widetilde{\Lambda}$. This means that the dynamics of the flow (Y^t) can be decomposed into two parts: on the one hand, the orbits that converge towards to the maximal invariant set Λ in the past or in the future; on the other hand, the dynamics that is well described by the coding map χ.

Remark 1.6 Besides being the cornerstone of the proof of Theorem 1.2, this coding procedure is interesting for its own sake. Indeed, it allows one to understand the behaviour of the recurrent orbits of the Anosov flow (Y^t) that intersect the surface S (ie which do not correspond to recurrent orbits of the incomplete flow (X^t)). In a forthcoming paper [6], we will use this coding procedure to describe the free homotopy classes of theses orbits, and build new examples of transitive Anosov flows.

Let us now explain how this coding procedure yields a proof of Theorem 1.2. For $i = 1, 2$, we get a symbolic space Σ_i and a coding map χ_i with values in Σ_i. The strong isotopy between (U, X_1, ψ_1) and (U, X_2, ψ_2) implies that there is a natural map between the symbolic spaces Σ_1 and Σ_2. Together with the coding maps, this yields a conjugacy between the Poincaré return maps of the flows $(\widetilde{Y}_1^t)$ and $(\widetilde{Y}_2^t)$ on the surfaces $\widetilde{S}_1$ and $\widetilde{S}_2$. Unfortunately, this conjugacy is not well defined on the whole surfaces $\widetilde{S}_1$ and $\widetilde{S}_2$. So we need to extend it. In order to do that, we introduce some (partial) preorders on the leaf spaces of the lifts of the stable/unstable foliations of the Anosov flows (Y_1^t) and (Y_2^t), and prove that the conjugacy preserves these preorders. This is quite delicate since the coding maps χ_1 and χ_2 do not behave very well with respect to these preorders. Once the extension has been achieved, we obtain a homeomorphism between the orbits spaces of the flows $(\widetilde{Y}_1^t)$ and $(\widetilde{Y}_2^t)$ that is equivariant with respect to the actions of the fundamental groups of the manifolds M_1 and M_2. Using a classical result, this implies that the Anosov flows (Y_1^t) and (Y_2^t) are orbitally equivalent.

2 Coding procedure

In this section, we will consider a transitive Anosov flow obtained by gluing hyperbolic plugs. Our goal is to define a coding procedure for the orbits of this Anosov flow.

Actually, this coding procedure will only describe the behaviour of the orbits which do not remain in $\operatorname{int}(U)$ forever.

2.1 Setting

We consider a hyperbolic plug (U, X). Recall that this means that U is a (not necessarily connected)[1] compact three-dimensional manifold with boundary, and X is a vector field on U, transverse to ∂U, such that the maximal invariant set

$$\Lambda_X := \bigcap_{t\in\mathbb{R}} X^t(U)$$

is a saddle hyperbolic set for the flow of X. We decompose the boundary of U as

$$\partial U := \partial^{\mathrm{in}} U \sqcup \partial^{\mathrm{out}} U,$$

where $\partial^{\mathrm{in}} U$ (resp. $\partial^{\mathrm{out}} U$) is the union of the connected component of ∂U where X is pointing into (resp. out of) U. The stable manifold theorem implies that $W^s_X(\Lambda_X)$ and $W^u_X(\Lambda_X)$ are two-dimensional laminations transverse to ∂U. Moreover, $W^s_X(\Lambda_X)$ is obviously disjoint from $\partial^{\mathrm{out}} U$ and $W^u_X(\Lambda_X)$ is obviously disjoint from $\partial^{\mathrm{in}} U$. As a consequence,

$$\begin{aligned} L^s_X &:= W^s_X(\Lambda_X) \cap \partial U = W^s_X(\Lambda_X) \cap \partial^{\mathrm{in}} U, \\ L^u_X &:= W^u_X(\Lambda_X) \cap \partial U = W^u_X(\Lambda_X) \cap \partial^{\mathrm{out}} U \end{aligned}$$

are one-dimensional laminations embedded in the surfaces $\partial^{\mathrm{in}} U$ and $\partial^{\mathrm{out}} U$, respectively. Note that L^s_X can be described as the set of points in $\partial^{\mathrm{in}} U$ whose forward (X^t)–orbit remains in U forever, ie does not intersect $\partial^{\mathrm{out}} U$. Similarly, L^u_X is the set of points in $\partial^{\mathrm{out}} U$ whose backward (X^t)–orbit remains in U forever, ie does not intersect $\partial^{\mathrm{in}} U$. These characterizations of L^s_X and L^u_X allow us to define a map

$$\theta_X \colon \partial^{\mathrm{in}} U \setminus L^s_X \to \partial^{\mathrm{out}} U \setminus L^u_X,$$

where $\theta_X(x)$ is the (unique) point of intersection the (X^t)–orbit of x with the surface $\partial^{\mathrm{out}} U$. Clearly, θ_X is a homeomorphism between $\partial^{\mathrm{in}} U \setminus L^s_X$ and $\partial^{\mathrm{out}} U \setminus L^u_X$. We call θ_X the *crossing map* of the plug (U, X).

In order to create a closed manifold equipped with a transitive Anosov flow, we consider a diffeomorphism

$$\psi \colon \partial^{\mathrm{out}} U \to \partial^{\mathrm{in}} U.$$

[1] Hence, a finite collection of hyperbolic plugs can always be considered as a single, nonconnected, hyperbolic plug.

The quotient space

$$M := U/\psi$$

is a closed three-dimensional topological manifold. We denote by $\pi\colon U \to M$ the natural projection map. The topological manifold M can equipped with a differential structure (compatible with the differential structure of U) so that the vector field

$$Y := \pi_* X$$

is well defined (and as smooth as X). We adopt the following hypotheses:

(0) The manifolds U and M are orientable.

(1) The flow (Y^t) is a transitive Anosov flow on the manifold M.

(2) The diffeomorphism ψ is a strongly transverse gluing diffeomorphism.

Recall that (2) means that the laminations L^s_X and $\psi_*(L^u_X)$ are transverse in the surface $\partial^{\mathrm{in}} U$ and moreover that every connected component C of $\partial^{\mathrm{in}} U \setminus (L^s_X \cup \psi_*(L^u_X))$ is a topological disc whose boundary ∂C consists of exactly four arcs α^s, $\alpha^{s\prime}$, α^u and $\alpha^{u\prime}$, where α^s and $\alpha^{s\prime}$ are arcs of leaves of L^s_X and α^u and $\alpha^{u\prime}$ are arcs of leaves $\psi_*(L^u_X))$.

Remark 2.1 We insist on the fact that (2) implies that every connected components of $\partial^{\mathrm{in}} U \setminus (L^s_X \cup \psi_*(L^u_X))$ is a topological disc, even if some of the connected components of $\partial^{\mathrm{in}} U \setminus L^s_X$ and $\partial^{\mathrm{in}} U \setminus \psi_*(L^u_X)$ might be annuli (eg in Bonatti and Langevin's construction). Further properties which follow from (0)–(2) will be stated and proven in Section 2.2. Anyhow, recall that the second part of [5] as well as [7] or [2] provide many examples of hyperbolic plugs (U, X) and gluing maps ψ for which (0)–(2) are satisfied.

We define

$$S := \pi(\partial^{\mathrm{in}} U) = \pi(\partial^{\mathrm{out}} U), \quad \Lambda := \pi(\Lambda_X), \quad L^s := \pi_*(L^s_X), \quad L^u := \pi_*(L^u_X).$$

By construction, S is a closed surface, embedded in the manifold M, transverse to the vector field Y. The set Λ is the union of the orbits of (Y^t) that do not intersect the surface S. It is an invariant saddle hyperbolic set for the Anosov flow (Y^t). Our assumptions imply that L^s and L^u are two strongly transverse one-dimensional laminations in the surface S. The lamination L^s (resp. L^u) can be described as the set of points in S whose forward (resp. backward) (Y^t)–orbit does not intersect S.

Similarly, L^u is a strict subset of $W^u(\Lambda) \cap S$. The homeomorphism θ_X induces a homeomorphism

$$\theta = (\pi|_{\partial^{\mathrm{out}}U}) \circ \theta_X \circ (\pi|_{\partial^{\mathrm{in}}U})^{-1} \colon S \setminus L^s \to S \setminus L^u.$$

Note that θ is nothing but the Poincaré first return map of the orbits of the Anosov flow (Y^t) on the surface S.

Since (Y^t) is an Anosov flow, it comes with a stable foliation $\mathcal{F}^s$ and an unstable foliation $\mathcal{F}^u$. These are two-dimensional foliations, transverse to each other, and transverse to the surface S. Hence, they induce two transverse one-dimensional foliations

$$F^s := \mathcal{F}^s \cap S \quad \text{and} \quad F^u := \mathcal{F}^u \cap S$$

on the surface S. Clearly, L^s and L^u are sublaminations (ie union of leaves) of the foliations F^s and F^u, respectively.

In order to code the orbits of the Anosov flow (Y^t), we cannot work directly in the manifold M; we need to unfold the leaves of the laminations L^s and L^u by lifting them to the universal cover of M. We denote this universal cover by $p \colon \widetilde{M} \to M$, and we denote by

$$\widetilde{S}, \quad \widetilde{\Lambda}, \quad \widetilde{W}^s(\Lambda), \quad \widetilde{W}^u(\Lambda), \quad \widetilde{L}^s, \quad \widetilde{L}^u, \quad \widetilde{\mathcal{F}}^s, \quad \widetilde{\mathcal{F}}^u, \quad \widetilde{F}^s, \quad \widetilde{F}^u$$

the complete lifts of the surface S, the hyperbolic set Λ, the laminations $W^s(\Lambda)$, $W^u(\Lambda)$, L^s and L^u, and the foliations $\mathcal{F}^s$, $\mathcal{F}^u$, F^s and F^u. We insist that $\widetilde{S}$ is the *complete* lift of S; that is, $\widetilde{S} := p^{-1}(S)$. In particular, $\widetilde{S}$ has infinitely many connected components. By construction, $\widetilde{F}^s$ and $\widetilde{F}^u$ are two transverse one-dimensional foliations on the surface $\widetilde{S}$, and $\widetilde{L}^s$ and $\widetilde{L}^u$ are sublaminations of $\widetilde{F}^s$ and $\widetilde{F}^u$, respectively. We also lift the vector field Y to a vector field $\widetilde{Y}$ on M. Of course, $\widetilde{Y}$ is transverse to the surface $\widetilde{S}$, so we can consider the Poincaré return map

$$\widetilde{\theta} \colon \widetilde{S} \setminus \widetilde{L}^s \to \widetilde{S} \setminus \widetilde{L}^u$$

of the orbits of $(\widetilde{Y}^t)$ on the surface $\widetilde{S}$. Obviously, $\widetilde{\theta}$ is a lift of the map θ.

2.2 Connected components of $\widetilde{S} \setminus \widetilde{L}^s$

We next collect some information about the connected components of $\widetilde{S} \setminus \widetilde{L}^s$ and the action of the Poincaré map $\widetilde{\theta}$ on these connected components. This information will be used in Section 2.3. Let us start by the topology of the surface $\widetilde{S}$.

Proposition 2.2 *Every connected component of $\widetilde{S}$ is a properly embedded topological plane.*

Proof The surface S is transverse to the Anosov flow (Y^t). Hence, S is a collection of incompressible tori in M (see eg [8, Corollary 2.2]). □

This allows us to describe the topology of the leaves of the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$:

Proposition 2.3 *Every leaf of the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$ is a properly embedded topological line. A leaf of $\widetilde{F}^s$ and a leaf of $\widetilde{F}^u$ intersect in no more than one point.*

Proof The first assertion follows immediately from Proposition 2.2: it is a classical consequence of the Poincaré–Hopf theorem that the leaves of a foliation of a plane are properly embedded topological lines.

The second assertion is again a consequence Proposition 2.2, together with the transversality of the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$. To prove it, we argue by contradiction: Consider a leaf ℓ^s of $\widetilde{F}^s$ and a leaf ℓ^u of $\widetilde{F}^u$, and assume that ℓ^s and ℓ^u intersect at more than one point. Then one can find two arcs $\alpha^s \subset \ell^s$ and $\alpha^u \subset \ell^u$ which share the same endpoints and have disjoint interiors. The union $\alpha^s \cup \alpha^s$ is a simple closed curve in $\widetilde{S}$. Since every connected component of $\widetilde{S}$ is a topological plane, $\alpha^s \cup \alpha^s$ bounds a topological disc $C \subset \widetilde{S}$. Consider two copies of C, and glue them along α^s in order to obtain a new topological disc D. The boundary of D is the union of two copies of α^u, and hence is piecewise smooth. The foliation $\widetilde{F}^s$ provides a one-dimensional foliation on D, which is topologically transverse to boundary ∂D. This contradicts the Poincaré–Hopf theorem. □

The next three propositions below concern the action of the Poincaré map $\widetilde{\theta}$ on the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$ and the laminations $\widetilde{L}^s$ and $\widetilde{L}^u$. We recall that $\widetilde{L}^s$ and $\widetilde{L}^u$ are sublaminations (ie union of leaves) of the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$, respectively.

Proposition 2.4 *The Poincaré map $\widetilde{\theta}\colon \widetilde{S} - \widetilde{L}^s \to \widetilde{S} - \widetilde{L}^u$ preserves the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$.*

Remark 2.5 Proposition 2.4 states that the foliation $(\widetilde{F}^s)|_{\widetilde{S}-\widetilde{L}^s}$ is mapped by $\widetilde{\theta}$ to the foliation $(\widetilde{F}^s)|_{\widetilde{S}-\widetilde{L}^u}$. The leaves of $(\widetilde{F}^s)|_{\widetilde{S}-\widetilde{L}^s}$ are full leaves of the foliation $\widetilde{F}^s$. On the contrary, a leaf of the foliation $(\widetilde{F}^s)|_{\widetilde{S}-\widetilde{L}^u}$ is never a full leaf of $\widetilde{F}^s$ (because every leaf of $\widetilde{F}^s$ is "cut into infinitely many pieces" by the transverse lamination $\widetilde{L}^u$). As a consequence, $\widetilde{\theta}$ maps leaves of $\widetilde{F}^s$ to pieces of leaves of $\widetilde{F}^s$. Similarly, $\widetilde{\theta}$ maps pieces of leaves of $\widetilde{F}^u$ to full leaves of $\widetilde{F}^u$.

Proof of Proposition 2.4 Recall that $\widetilde{F}^s$ is defined as the intersection of the foliation $\widetilde{\mathcal{F}}^s$ with the transverse surface $\widetilde{S}$. The foliation $\widetilde{\mathcal{F}}^s$ is leafwise invariant under the flow $(\widetilde{Y}^t)$. As a consequence, $\widetilde{F}^s = \widetilde{\mathcal{F}}^s \cap \widetilde{S}$ is invariant under the Poincaré return map of $(\widetilde{Y}^t)$ on $\widetilde{S}$. □

Proposition 2.6 *For every $n \geq 0$, $\bigcup_{p=0}^{n} \widetilde{\theta}^{-p}(\widetilde{L}^s)$ is a closed sublamination of the foliation $\widetilde{F}^s$.*

Proof The foliation $\widetilde{F}^s$ is invariant under the Poincaré map $\widetilde{\theta}\colon \widetilde{S} - \widetilde{L}^s \to \widetilde{S} - \widetilde{L}^u$. Since $\widetilde{L}^s$ is a union of leaves of $\widetilde{F}^s$, it follows that $\widetilde{\theta}^{-1}(\widetilde{L}^s)$ is a union of leaves of $\widetilde{F}^s$. Moreover, since $\widetilde{L}^s$ is a closed subset of $\widetilde{S}$, its preimage $\widetilde{\theta}^{-1}(\widetilde{L}^s)$ must be a closed subset of $\widetilde{S} - \widetilde{L}^s$ (remember that $\widetilde{\theta}$ is well defined on $\widetilde{S} - \widetilde{L}^s$). Therefore $\bigcup_{p=0}^{1} \widetilde{\theta}^{-p}(\widetilde{L}^s)$ is a closed subset of $\widetilde{S}$. So $\bigcup_{p=0}^{1} \widetilde{\theta}^{-p}(\widetilde{L}^s)$ is a closed union of leaves of $\widetilde{F}^s$, ie a closed sublamination of $\widetilde{F}^s$. Repeating the same arguments, one proves by induction that $\bigcup_{p=0}^{n} \widetilde{\theta}^{-p}(\widetilde{L}^s)$ is a closed sublamination of $\widetilde{F}^s$ for every $n \geq 0$. □

Proposition 2.7

$$\bigcup_{p=0}^{\infty} \widetilde{\theta}^{-p}(\widetilde{L}^s) = \widetilde{W}^s(\Lambda) \cap \widetilde{S}.$$

Proof By definition, $W^s(\Lambda) \cap S$ is the set of all points $x \in S$ such that the forward orbit of x converges towards the set Λ, which is disjoint from S. As a consequence, for every point $x \in W^s(\Lambda) \cap S$, the forward orbit of x intersects the surface S only finitely many times, say $p(x)$ times. We have observed that L^s is the set of all points $y \in S$ such that the forward orbit of y does not intersect S and converges towards the set Λ (see Section 2.1). It follows that, for every $x \in W^s(\Lambda) \cap S$, the last intersection point $\theta^{p(x)}$ of the forward orbit of x with S is in L^s. This proves the inclusion $W^s(\Lambda) \cap S \subset \bigcup_{p=0}^{\infty} \theta^{-p}(L^s)$. The converse inclusion is straightforward. Hence, $\bigcup_{p=0}^{\infty} \theta^{-p}(L^s) = W^s(\Lambda) \cap S$. The equality $\bigcup_{p=0}^{\infty} \widetilde{\theta}^{-p}(\widetilde{L}^s) = \widetilde{W}^s(\Lambda) \cap \widetilde{S}$ follows by lifting to the universal cover. □

Of course, $\widetilde{W}^s(\Lambda) \cap \widetilde{S}$ and $\widetilde{W}^u(\Lambda) \cap \widetilde{S}$ are unions of leaves of the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$, respectively. But these sets are not closed. More precisely:

Proposition 2.8 *Both $\widetilde{W}^s(\Lambda) \cap \widetilde{S}$ and $\widetilde{S} - \widetilde{W}^s(\Lambda)$ are dense in $\widetilde{S}$.*

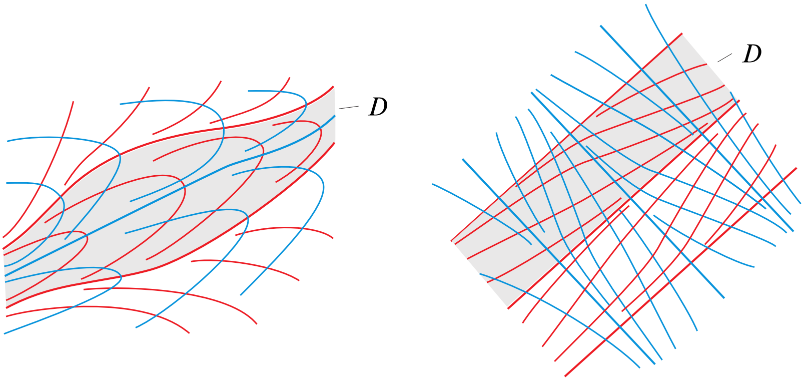


Figure 2: Left: a proper stable strip. Right: a trivially bifoliated proper stable strip.

Proof Recall that (Y^t) is a transitive Anosov flow on M. Hence, every leaf of the weak stable foliation $\mathcal{F}^s$ is dense in M. Since both $W^s(\Lambda)$ and $M \setminus W^s(\Lambda)$ are nonempty unions of leaves of the foliation $\mathcal{F}^s$, and since the leaves of $\mathcal{F}^s$ are transversal to the surface S, it follows that both $W^s(\Lambda) \cap S$ and $S \setminus W^s(\Lambda)$ are dense in S. Lifting to the universal cover, we obtain that $\widetilde{W}^s(\Lambda) \cap \widetilde{S}$ and $\widetilde{S} - \widetilde{W}^s(\Lambda)$ are dense in $\widetilde{S}$. □

Of course, the analogs of Propositions 2.6, 2.7 and 2.8 for $\widetilde{L}^u$ and $W^u(\widetilde{\Lambda})$ hold ($\widetilde{\theta}^{-p}$ should be replaced by $\widetilde{\theta}^p$ in Propositions 2.6 and 2.7). We will now describe the topology of the connected components of $\widetilde{S} \setminus \widetilde{L}^s$. We first introduce some vocabulary.

Definition 2.9 We call a *proper stable strip* every topological open disc D of $\widetilde{S}$ whose boundary is the union of two leaves of the foliation $\widetilde{F}^s$.

If D is a proper stable strip, one can easily construct a homeomorphism h from the closure of D to $\mathbb{R} \times [-1, 1]$. We will need the following stronger notion:

Definition 2.10 We say that a proper stable strip D is *trivially bifoliated* if there exists a homeomorphism h from the closure of D to $\mathbb{R} \times [-1, 1]$ mapping the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$ to the horizontal and vertical foliations on $\mathbb{R} \times [-1, 1]$.

Of course, *proper unstable strips* and *trivially bifoliated proper unstable strips* can be defined similarly. The proposition below gives a fairly precise description of the positions of the connected components of $\widetilde{S} - \widetilde{L}^s$ with respect to the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$:

Proposition 2.11 *Every connected component of $\widetilde{S} - \widetilde{L}^s$ is a trivially bifoliated proper stable strip bounded by two leaves of the lamination $\widetilde{L}^s$.*

Proof Let D be a connected component of $\widetilde{S} - \widetilde{L}^s$. Denote by P the connected component of $\widetilde{S}$ containing D. Since P is a topological plane (Proposition 2.2), and since each leaf of $\widetilde{L}^s$ is a properly embedded topological line (Proposition 2.3) which separates P into two connected components, it follows that D is a topological disc. The boundary of D is a union of leaves of $\widetilde{L}^s$ (which we call *the boundary leaves of* D). We denote by $\overline{D}$ the closure of D.

Claim 1 *Let ℓ^u be a leaf of the foliation $\widetilde{F}^u$ intersecting $\overline{D}$, and α^u be a connected component of $\ell^u \cap \overline{D}$. Then α^u is an arc joining two different boundary leaves of D.*

Let R be a connected component of $D \setminus \widetilde{L}^u$ such that α^u is included in the closure $\overline{R}$ of R (actually R is unique, but we will not use this fact). Observe that R is a connected component of $\widetilde{S} - (\widetilde{L}^s \cup \widetilde{L}^u)$. Our assumptions (specifically the strong transversality of the gluing map ψ) imply that R is a relatively compact topological disc whose boundary ∂R is made of four arcs α^s_-, α^s_+, α^u_- and α^u_+, where α^s_- and α^s_- are disjoint and lie in some leaves of $\widetilde{L}^s$, and where α^u_- and α^u_+ are disjoint and lie in some leaves of $\widetilde{L}^u$. Loosely speaking, R is a rectangle with two sides α^s_- and α^s_+ in $\widetilde{L}^s$ and two sides α^u_- and α^u_+ in $\widetilde{L}^u$. Proposition 2.3 implies that ℓ^u intersects α^s_- and α^s_+ at no more than one point. Since ℓ^u is a proper line and $\overline{R}$ is a compact set, it follows that α^u must be an arc going from α^s_- to α^s_+. Using again Proposition 2.3, it also follows that α^s_- to α^s_+ cannot be in the same leaf of $\widetilde{F}^s$. The claim is proved.

Claim 2 *D has exactly two boundary leaves.*

In order to prove this claim, we endow the foliation $\widetilde{F}^u$ with an orientation (this is possible since $\widetilde{F}^u$ is a foliation on a collection of topological planes). For every $x \in \overline{D}$, we denote by $\ell^u(x)$ the leaf of the foliation $\widetilde{F}^u$ passing through x, and denote by $\alpha^u(x)$ the connected component of $\ell^u_x \cap \overline{D}$ containing x. Note that $\ell^u(x)$ and $\alpha^u(x)$ are oriented by the orientation of $\widetilde{F}^u$. By Claim 1, $\alpha^u(x)$ is an arc whose endpoints lie on two boundary leaves $\ell^s_-(x)$ and $\ell^s_+(x)$ of D. By transversality of the foliations $\widetilde{F}^u$ and $\widetilde{F}^s$, the maps $x \mapsto \ell^s_-(x)$ and $x \mapsto \ell^s_+(x)$ are locally constant. Since $\overline{D}$ is connected, these maps are constant. In other words, one can find two boundary leaves ℓ^s_- and ℓ^s_+ of D such that $\alpha^u(x)$ is an arc from ℓ^s_- to ℓ^s_+ for every $x \in \overline{D}$. It follows that ℓ^s_- and ℓ^s_+ are the only accessible boundary leaves of D: otherwise, one can consider another boundary leaf ℓ^s, take a point $x \in \ell^s$, and get a contradiction since one end of α^u_x is on ℓ^s. As a further consequence, the accessible boundary of D is closed (recall that ℓ^s_- and ℓ^s_+ are properly embedded lines), and therefore coincides with the boundary of D. We finally conclude that ℓ^s_- and ℓ^s_+ are the only boundary leaves of D, and Claim 2 is proved.

Claims 1 and 2 already imply that D is a proper stable strip bounded by two leaves ℓ^s_- and ℓ^s_+ of $\widetilde{L}^s$. We are left to prove that D is trivially bifoliated. Recall that $\widetilde{S}$ is a topological plane (Proposition 2.2), and that ℓ^s_- and ℓ^s_+ are properly embedded topological lines (Proposition 2.3). By easy planar topology, it follows that there exists a homeomorphism h from $\overline{D}$ to $\mathbb{R} \times [-1, 1]$ mapping ℓ^s_- and ℓ^s_+ to $\mathbb{R} \times \{-1\}$ and $\mathbb{R} \times \{1\}$, respectively. Claim 1 implies that $h_*(\widetilde{F}^u_{\overline{D}})$ is a foliation of $\mathbb{R} \times [-1, 1]$ by arcs going from $\mathbb{R} \times \{-1\}$ and $\mathbb{R} \times \{1\}$. One can easily construct a self-homeomorphism h' of $\mathbb{R} \times [-1, 1]$ mapping this foliation on the vertical foliation of $\mathbb{R} \times [-1, 1]$. Up to replacing h by $h' \circ h$, we will assume that h maps $\widetilde{F}^u_{\overline{D}}$ on the vertical foliation of $\mathbb{R} \times [-1, 1]$. Now we consider a leaf ℓ^s of the foliation $\widetilde{F}^s$ included in $\overline{D}$. According to Proposition 2.3, ℓ^s intersects each leaf of $\widetilde{F}^u$ at no more than one point. Hence, $h(\ell^s)$ intersects each vertical segment in $\mathbb{R} \times [-1, 1]$ at no more than one point. Let E be the set of $t \in \mathbb{R}$ such that $h(\ell^s)$ intersects the vertical segment $\{t\} \times [-1, 1]$. Since ℓ^s is a proper topological line transversal to $\widetilde{F}^u$, the set E_t must be open and closed in $\mathbb{R}$. Therefore, $h(\ell^s)$ intersects every vertical segment in $\mathbb{R} \times [-1, 1]$ at exactly one point. In other words, the leaves of $h_*(\widetilde{F}^s_{\overline{D}})$ are graphs over the first coordinate in $\mathbb{R} \times [-1, 1]$. One can easily modify the homeomorphism h so that $h_*(\widetilde{F}^s_{\overline{D}})$ is the horizontal foliation of $\mathbb{R} \times [-1, 1]$. Hence, D is a trivially bifoliated proper stable strip. □

Of course, the unstable analog of Proposition 2.11 holds true: every connected component of $\widetilde{S} - \widetilde{L}^u$ is a trivially bifoliated proper unstable strip bounded by two leaves of the lamination $\widetilde{L}^u$. On the other hand, $\widetilde{\theta}$ maps connected components of $\widetilde{S} - \widetilde{L}^s$ to connected component of $\widetilde{S} - \widetilde{L}^u$. So, we obtain:

Corollary 2.12 *If D is a connected component of $\widetilde{S} - \widetilde{L}^s$, then $\widetilde{\theta}(D)$ is a trivially bifoliated proper unstable strip, disjoint from $\widetilde{L}^u$, bounded by two leaves of the lamination $\widetilde{L}^u$.*

The following proposition describes the action of $\widetilde{\theta}$ on the connected components of $\widetilde{S} - \widetilde{L}^s$:

Proposition 2.13 *Let D be a connected component of $\widetilde{S} - \widetilde{L}$, and D' be any trivially bifoliated proper stable strip. Assume that $D \cap \widetilde{\theta}^{-1}(D')$ is nonempty. Then $D \cap \widetilde{\theta}^{-1}(D')$ is a trivially bifoliated proper stable substrip of D.*

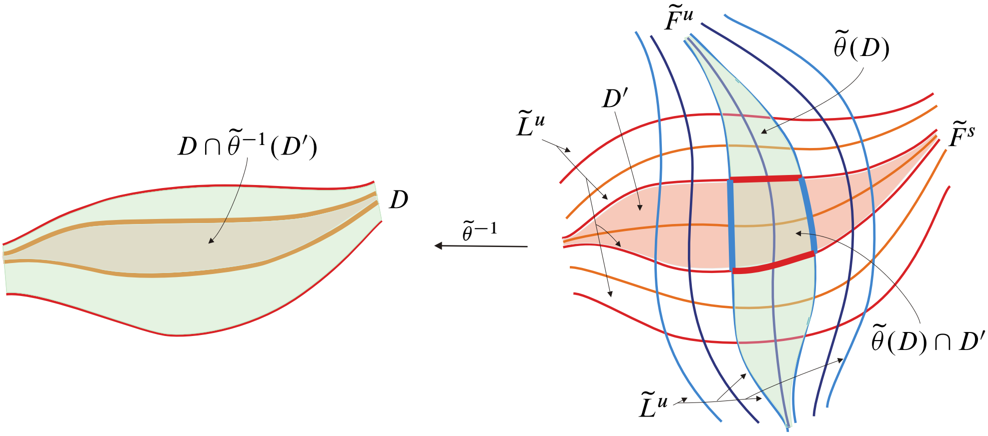


Figure 3: The proof of Proposition 2.13.

Proof We call a *trivially bifoliated rectangle* every topological open disc $R\subset\widetilde{S}$ such that there exists a homeomorphism from the closure of R to $[-1,1]^2$ mapping the restrictions of $\widetilde{F}^s$ and $\widetilde{F}^u$ to the horizontal and vertical foliations of $[-1,1]^2$. In particular, the boundary of such a trivially bifoliated rectangle is made of two stable sides and two unstable sides.

According to Corollary 2.12, $\widetilde{\theta}(D)$ is a trivially bifoliated proper unstable strip, disjoint from $\widetilde{L}^u$, bounded by two leaves of $\widetilde{L}^u$. By assumption, D' is a trivially bifoliated proper stable strip. It easily follows that $\widetilde{\theta}(D)\cap D'$ is a trivially bifoliated rectangle, disjoint from $\widetilde{L}^u$, whose unstable sides are in $\widetilde{L}^u$ (see Figure 3). Observe that the interiors of two stable sides of $\widetilde{\theta}(D)\cap D'$ are full leaves of $\widetilde{F}^s|_{\widetilde{S}-\widetilde{L}^u}$. Hence:

($\star$) $\theta(D)\cap D'$ is a connected subset of $\theta(D)$ and the boundary of $\theta(D)\cap D'$ in $\theta(D)$ is made of two disjoint leaves of $\widetilde{F}^s|_{\widetilde{S}-\widetilde{L}^u}$.

Now recall that $\widetilde{\theta}^{-1}$ is a homeomorphism from $\widetilde{S}-\widetilde{L}^u$ to $\widetilde{S}-\widetilde{L}^s$, mapping leaves of $\widetilde{F}^s|_{\widetilde{S}-\widetilde{L}^u}$ to full leaves of $\widetilde{F}^s$ (see Proposition 2.4 and Remark 2.5). Also observe that $D\cap\widetilde{\theta}^{-1}(D')$ is a subset of D. As a consequence, property ($\star$) implies:

($\star'$) $D\cap\widetilde{\theta}^{-1}(D')$ is a connected subset of D, and the boundary of $D\cap\widetilde{\theta}^{-1}(D')$ is made of two disjoint leaves of $\widetilde{F}^s$.

Since D is a trivially foliated proper stable strip D, Property ($\star'$) clearly implies that $D\cap\widetilde{\theta}^{-1}(D')$ is a trivially bifoliated proper stable substrip of D. See Figure 3. □

2.3 The coding procedure

In this section, we will use the connected components of $\widetilde{S}\setminus\widetilde{L}^s$ to describe the itinerary of the orbits the flow $(\widetilde{Y}^t)$ that do not belong to $\widetilde{W}^s(\Lambda)\cup\widetilde{W}^u(\Lambda)$. We consider the alphabet

$$\mathcal{A} := \{\text{connected components of } \widetilde{S}\setminus\widetilde{L}^s\},$$

and the symbolic spaces

$$\begin{aligned}\Sigma^s &= \{\overline{D}^s = (D_p)_{p\geq 0} \mid D_p \in \mathcal{A} \text{ and } \widetilde{\theta}(D_p)\cap D_{p+1}\neq\varnothing \text{ for every } p\},\\ \Sigma^u &= \{\overline{D}^u = (D_p)_{p<0} \mid D_p \in \mathcal{A} \text{ and } \widetilde{\theta}(D_p)\cap D_{p+1}\neq\varnothing \text{ for every } p\},\\ \Sigma &= \{\overline{D} = (D_p)_{p\in\mathbb{Z}} \mid D_p \in \mathcal{A} \text{ and } \widetilde{\theta}(D_p)\cap D_{p+1}\neq\varnothing \text{ for every } p\}.\end{aligned}$$

In order to define the coding maps, we need to introduce some leaf spaces. We will denote by f^s the leaf space of the foliation $\widetilde{F}^s$ (equipped with the quotient topology). We will denote by $f^{s,\infty}$ the subset of f^s made of the leaves that are not in $\widetilde{W}^s(\Lambda)$. Similarly, we denote by f^u the leaf space of $\widetilde{F}^u$, and by $f^{u,\infty}$ the subset fo f^u made of the leaves that are not in $\widetilde{W}^u(\Lambda)$. Finally, we denote by $\widetilde{S}^\infty$ the set of points in $\widetilde{S}$ that are neither in $\widetilde{W}^s(\Lambda)$ nor in $\widetilde{W}^u(\Lambda)$. That is,

$$\begin{aligned}f^{s,\infty} &= \{\text{leaves of } \widetilde{F}^s \text{ that are not in } \widetilde{W}^s(\Lambda)\},\\ f^{u,\infty} &= \{\text{leaves of } \widetilde{F}^u \text{ that are not in } \widetilde{W}^u(\Lambda)\},\\ \widetilde{S}^\infty &= \widetilde{S} - (\widetilde{W}^s(\Lambda)\cup\widetilde{W}^u(\Lambda)).\end{aligned}$$

By Proposition 2.7, if $\ell^s \in f^{s,\infty}$, then $\widetilde{\theta}^p(\ell^s)$ is included in a connected component of $\widetilde{S}-\widetilde{L}^s$ for every $p\geq 0$. Similarly, if $\ell^u\in f^{u,\infty}$, then $\widetilde{\theta}^p(\ell^u)$ is included in a connected component of $\widetilde{S}-\widetilde{L}^u$ for every $p\leq 0$. Since $\widetilde{\theta}^{-1}$ maps homeomorphically $\widetilde{S}-\widetilde{L}^u$ to $\widetilde{S}-\widetilde{L}^s$, we deduce that, if $\ell^u\in f^{u,\infty}$, then $\widetilde{\theta}^p(\ell^u)$ is included in a connected component of $\widetilde{S}-\widetilde{L}^s$ for every $p<0$. As a further consequence, if x is a point of $\widetilde{S}^\infty$, then $\widetilde{\theta}^p(x)$ is in a connected component of $\widetilde{S}-\widetilde{L}^s$ for every $p\in\mathbb{Z}$. This shows that the following coding maps are well defined:

$$\begin{aligned}\chi^s\colon f^{s,\infty}\to\Sigma^s, &\quad \ell^s\mapsto\overline{D}^s=(D_p)_{p\geq 0}, &&\text{where } \widetilde{\theta}^p(\ell^s)\subset D_p \text{ for every } p\geq 0;\\ \chi^u\colon f^{u,\infty}\to\Sigma^u, &\quad \ell^u\mapsto\overline{D}^u=(D_p)_{p<0}, &&\text{where } \widetilde{\theta}^p(\ell^u)\subset D_p \text{ for every } p<0;\\ \chi\colon \widetilde{S}^\infty\to\Sigma, &\quad x\mapsto\overline{D}=(D_p)_{p\in\mathbb{Z}}, &&\text{where } \widetilde{\theta}^p(x)\in D_p \text{ for every } p\in\mathbb{Z}.\end{aligned}$$

The following proposition is an important ingredient of the proof of Theorem 1.2:

Proposition 2.14 *The maps χ^s, χ^u and χ are bijective.*

Lemma 2.15 (1) *For every $\overline{D}^s = (D_p)_{p\geq 0} \in \Sigma^s$, the set $\bigcap_{p\geq 0} \widetilde{\theta}^{-p}(D_p)$ is a stable leaf $\ell^s \in f^{s,\infty}$.*

(2) *For every $\overline{D}^u = (D_p)_{p<0} \in \Sigma^u$, the set $\bigcap_{p<0} \widetilde{\theta}^{-p}(D_p)$ is an unstable leaf $\ell^u \in f^{u,\infty}$.*

(3) *For every $\overline{D} = (D_p)_{p\in\mathbb{Z}} \in \Sigma$, the set $\bigcap_{p\in\mathbb{Z}} \widetilde{\theta}^{-p}(D_p)$ is a single point $x \in \widetilde{S}^{\infty}$.*

Remark 2.16 Lemma 2.15 is completely false if we replace the connected components of $\widetilde{S}\setminus\widetilde{L}^s$ by the connected components of $S\setminus L^s$ (and $\widetilde{\theta}$ by θ). For example, if $(D_p)_{p\geq 0}$ is a sequence of connected components of $S \setminus L^s$, then $\bigcap_{p\geq 0} \theta^{-p}(D_p)$, if not empty, will be the union of uncountably many leaves of the foliation F^s. This is the reason why we need to work in the universal cover of M.

Proof of Lemma 2.15 Let us prove the first item. Consider a sequence $\overline{D}^s = (D_p)_{p\geq 0} \in \Sigma^s$. By Proposition 2.11, D_0 is a trivially bifoliated proper stable strip. Proposition 2.13 and a straightforward induction imply that, for every $n \in \mathbb{N}$, the set $\bigcap_{p=0}^{n} \widetilde{\theta}^{-p}(D_p)$ is a substrip of D_0. So $\left(\bigcap_{p=0}^{n} \widetilde{\theta}^{-p}(D_p)\right)_{n\geq 0}$ is a decreasing sequence of substrips of the trivially bifoliated proper stable strip D_0. It easily follows that $\bigcap_{p\geq 0} \widetilde{\theta}^{-p}(D_p)$ is a substrip of D_0. In particular, $\bigcap_{p\geq 0} \widetilde{\theta}^{-p}(D_p)$ is a connected union of leaves of $\widetilde{F}^s$. On the other hand, since $D_0, D_1, \dots$ are connected components of $\widetilde{S} - \widetilde{L}^s$, the set $\bigcap_{\geq 0} \widetilde{\theta}^{-p}(D_p)$ is disjoint from $\bigcup_{p\geq 0} \widetilde{\theta}^{-p}(\widetilde{L}^s) = \widetilde{W}^s(\Lambda) \cap \widetilde{S}$ (see Proposition 2.7). But $\widetilde{W}^s(\Lambda) \cap \widetilde{S}$ is dense in $\widetilde{S}$ (Proposition 2.8). It follows that $\bigcap_{p\geq 0} \widetilde{\theta}^{-p}(D_p)$ must be a single leaf of $\widetilde{F}^s$. This completes the proof of (1).

Item (2) follows from exactly the same arguments as (1). In order to prove the last item, we consider a sequence $\overline{D} = (D_p)_{p\in\mathbb{Z}}$ in Σ. According to (1)–(2), $\bigcap_{p\geq 0} \widetilde{\theta}^{-p}(D_p)$ is a leaf ℓ^s of the foliation $\widetilde{F}^s$ and $\bigcap_{p<0} \widetilde{\theta}^{-p}(D_p)$ is a leaf ℓ^u of the foliation $\widetilde{F}^u$. Since $\overline{D} = (D_p)_{p\in\mathbb{Z}}$ is in Σ, the intersection $D_0 \cap \widetilde{\theta}^{(}D_{-1})$ is not empty. Since D_0 is a trivially bifoliated proper stable strip (Proposition 2.11) and $\widetilde{\theta}^{(}D_{-1})$ is a trivially bifoliated proper unstable strip (Corollary 2.12), every leaf of $\widetilde{F}^s$ in D_0 intersects every leaf of $\widetilde{F}^u$ in $\widetilde{\theta}(D_{-1})$ at exactly one point. In particular, $\bigcap_{p\in\mathbb{Z}} \widetilde{\theta}^{-p}(D_p) = \ell^s \cap \ell^u$ is made of exactly one point x. Since the leaves ℓ^s and ℓ^u are disjoint from $\widetilde{W}^s(\Lambda)$ and $\widetilde{W}^u(\Lambda)$, respectively, the point x must be in $\widetilde{S}^{\infty}$. □

Proof of Proposition 2.14 Lemma 2.15 allows us to define some inverse maps for χ^s, χ^u and χ. Therefore, χ^s, χ^u and χ are bijective. □

Deck transformation preserve the surface $\widetilde{S}$, the foliations $\widetilde{\mathcal{F}}^s$ and $\widetilde{\mathcal{F}}^u$, and the laminations $W^s(\widetilde{\Lambda})$ and $W^u(\widetilde{\Lambda})$. This induces some natural actions of $\pi_1(M)$ on the set $\widetilde{S}^\infty$, on the leaf spaces $f^{s,\infty}$ and $f^{u,\infty}$, on the alphabet $\mathcal{A}$, and therefore on the symbolic spaces Σ, Σ^s and Σ^u. From the definition of the coding maps, one easily checks that:

Proposition 2.17 *The coding maps χ, χ^s and χ^u commute with the actions of the fundamental group of M on $\widetilde{S}^\infty$ f^s, f^u, Σ Σ^s and Σ^u.*

The definition of the coding maps also implies that:

Proposition 2.18 *The coding map χ (resp. χ^s and χ^u) conjugates the action of the Poincaré first return map $\widetilde{\theta}$ on $\widetilde{S}^\infty$ (resp. f^s and f^u) to the left shift on the symbolic space Σ (resp. Σ^s and Σ^u).*

Given an integer $n \geq 0$ and some connected components $D_0^0, \dots, D_n^0$ of $\widetilde{S} - \widetilde{L}^s$, we define the cylinder

$$[D_0^0 \dots D_n^0]^s := \{(D_p)_{p \geq 0} \in \Sigma^s \mid D_p = D_p^0 \text{ for } 0 \leq p \leq n\}.$$

Similarly, given $n < 0$ and some connected components $D_n^0, \dots, D_{-1}^0$ of $\widetilde{S} - \widetilde{L}^s$, we define the cylinder

$$[D_n^0 \dots D_{-1}^0]^u := \{(D_p)_{p < 0} \in \Sigma^u \mid D_p = D_p^0 \text{ for } n \leq p \leq -1\}.$$

The following proposition will be used in the next subsection:

Proposition 2.19 (1) *For $n \geq 0$ and $D_0, \dots, D_n \in \mathcal{A}$, the set*

$$(\chi^s)^{-1}([D_0 D_1 \dots D_n]^s) = \bigcap_{0 \leq p \leq n} \widetilde{\theta}^{-p}(D_p)$$

is either empty or a substrip of the trivially foliated proper stable strip D_0 bounded by two leaves of $\widetilde{\theta}^{-n}(\widetilde{L}^s)$.

(2) *For $n < 0$ and $D_n, \dots, D_{-1} \in \mathcal{A}$, the set*

$$(\chi^u)^{-1}([D_n D_{n+1} \dots D_{-1}]) = \bigcap_{-n \leq p \leq -1} \widetilde{\theta}^{-p+1}(D_p)$$

is a substrip of the trivially foliated proper unstable strip $\widetilde{\theta}(D_{-1})$ bounded by two leaves of $\widetilde{\theta}^{K-1}(\widetilde{L^u})$.

Proof This follows from the arguments of the proof of Lemma 2.15. □

2.4 Partial orders on the leaf spaces and the symbolic spaces

We will now describe a partial preorder on the leaf space f^s. The preservation of this partial preorder will be a fundamental ingredient of our proof of Theorem 1.2 in Section 3.

Let us start by choosing some orientations. First of all, we choose an orientation of the hyperbolic plug U. The orientation of U, together with the vector field X, provides an orientation of ∂U: if ω is a 3–form defining the orientation on U, then the 2–form $i_X U$ defines the orientation on ∂U. The orientation of U induces an orientation of the manifold $M = U/\psi$ (we have assumed that the manifold M is orientable, which is equivalent to assuming that the gluing map ψ preserves the orientation of ∂U), and the orientation of ∂U induces an orientation of the surface $S = \pi(\partial^{\mathrm{in}} U) = \pi(\partial^{\mathrm{out}} U)$. The orientations of M and S induce some orientations on $\widetilde{M}$ and $\widetilde{S}$. Now, since every connected component of $\widetilde{S}$ is a topological plane, the foliation $\widetilde{F}^s$ is orientable. We fix an orientation of $\widetilde{F}^s$. This automatically induces an orientation of the foliation $\widetilde{F}^u$ as follows: the orientation of $\widetilde{F}^u$ is chosen so that, if Z^s and Z^u are vector fields tangent to $\widetilde{F}^s$ and $\widetilde{F}^u$, respectively, and pointing in the direction of the orientation of the leaves, then the frame field (Z^s, Z^u) is positively oriented with respect to the orientation of $\widetilde{S}$.

Remarks 2.20 (1) By construction, the orientations of the manifold $\widetilde{M}$ and the surface $\widetilde{S}$ are related as follows: if ω is a 3–form defining the orientation on $\widetilde{M}$, then the 2–form $i_{\widetilde{Y}} \widetilde{M}$ defines the orientation on $\widetilde{S}$. As a consequence, the Poincaré return map $\widetilde{\theta}$ of the orbits of $\widetilde{Y}$ on $\widetilde{S}$ preserves the orientation of $\widetilde{S}$.

(2) Consequently, for any connected component D of $\widetilde{S} - \widetilde{L}^s$, if the Poincaré map $\widetilde{\theta}|_D$ preserves (resp. reverses) the orientation of the foliation $\widetilde{F}^s$, then it also preserves (resp. reverses) the orientation of the foliation $\widetilde{F}^u$.

Let ℓ be a leaf of the foliation $\widetilde{F}^s$, contained in a connected component $\widetilde{S}_\ell$ of $\widetilde{S}$. Recall that $\widetilde{S}_\ell$ is a topological plane, and ℓ is a properly embedded line in $\widetilde{S}_\ell$. As a consequence, $\widetilde{S}_\ell \setminus \ell$ has two connected components.

Definition 2.21 We denote by $L(\ell)$ and $R(\ell)$ the two connected components of $\widetilde{S} \setminus \ell$ so that the oriented leaves of $\widetilde{F}^u$ crossing ℓ go from $L(\ell)$ towards $R(\ell)$. The points of $L(\ell)$ are said to be *on the left* of ℓ; the points of $R(\ell)$ are said to be *on the right* of ℓ.

Now we can define a preorder on the leaf space f^s.

Definition 2.22 (preorder on f^s) Given two leaves $\ell \neq \ell'$ of the foliation $\widetilde{F}^s$, we write $\ell \prec \ell'$ if there exists an arc of a leaf of $\widetilde{F}^u$ with endpoints $a \in \ell$ and $a' \in \ell'$ such that the orientation of $\widetilde{F}^u$ goes from a towards a'.

Proposition 2.23 *$\prec$ is a preorder on f^s: the relations $\ell \prec \ell'$ and $\ell' \prec \ell$ are incompatible.*

Proof The relation $\ell \prec \ell'$ implies that the leaf ℓ' is on the right of ℓ; that is, $\ell' \subset R(\ell)$. Similarly, the relation $\ell \prec \ell'$ implies $\ell' \subset L(\ell)$. The proposition follows since $L(\ell) \cap R(\ell) = \varnothing$. □

The proposition below is very easy to prove, but fundamental (it will be used in a crucial way to extend some conjugating maps in the next section, see Corollary 3.12):

Proposition 2.24 *$\prec$ is a local total order on f^s: for every leaf ℓ_0 of $\widetilde{F}^s$, there exists a neighbourhood $\mathcal{V}_0$ of ℓ_0 in f^s such that any two different leaves $\ell, \ell' \in \mathcal{V}_0$ are comparable (ie satisfy either $\ell \prec \ell'$ or $\ell' \prec \ell$).*

Proof Consider a leaf ℓ_0 of $\widetilde{F}^s$ and a leaf ℓ^u of $\widetilde{F}^u$ such that $\ell^u \cap \ell_0 \neq \varnothing$. By transversality of the foliations $\widetilde{F}^s$ and $\widetilde{F}^u$, there exists a neighbourhood $\mathcal{V}_0$ of ℓ_0 in f^s such that ℓ^u crosses every leaf in $\mathcal{V}_0$. As a consequence, any two different leaves $\ell, \ell' \in \mathcal{V}_0$ are comparable for the preorder $\prec$. □

The proposition below shows that the preorder $\prec$ is "compatible" with the connected components decomposition of $\widetilde{S} - \widetilde{L}^s$:

Proposition 2.25 *Given two different elements D and D' of $\mathcal{A}$, the following are equivalent:*

(1) *There exist some leaves $\ell_0, \ell'_0 \in f^s$ such that $\ell_0 \subset D$, $\ell'_0 \subset D'$ and $\ell_0 \prec \ell'_0$.*

(2) *All leaves $\ell, \ell' \in f^s$ such that $\ell \subset D$ and $\ell' \subset D'$ satisfy $\ell \prec \ell'$.*

Proof Assume that (1) is satisfied. Since $\ell_0 \prec \ell'_0$, there must be a leaf ℓ^u of the foliation $\widetilde{F}^u$ intersecting both ℓ_0 and ℓ'_0. Proposition 2.11 implies that $\alpha := \ell^u \cap D$ and $\alpha' := \ell^u \cap D'$ are two disjoint arcs in the leaf ℓ^u. Consider some leaves ℓ and ℓ' of $\widetilde{F}^s$ contained in D and D', respectively. Again Proposition 2.11 implies that ℓ intersects ℓ^u at some point $a_\ell \in \alpha$ and ℓ' intersects ℓ^u at some point $a_{\ell'} \in \alpha'$. Since $\ell_0 \preceq \ell'_0$, the orientation of ℓ^u goes from α towards α', and hence from a_ℓ towards $a_{\ell'}$. This shows that $\ell \prec \ell'$. □

Definition 2.26 (preorder on $\mathcal{A}$) Given two different elements D and D' of $\mathcal{A}$, we write $D \prec D'$ if there exist some leaves $\ell_0, \ell'_0 \in f^s$ such that $\ell_0 \subset D$, $\ell'_0 \subset D'$ and $\ell_0 \prec \ell'_0$.

Definition 2.27 (preorder on Σ^s) The partial preorder $\prec$ on $\mathcal{A}$ induces a lexicographic partial preorder on $\Sigma^s \subset \mathcal{A}^{\mathbb{N}}$, which will also be denoted by $\prec$: for $\overline{D} = (D_p)_{p\geq 0}$ and $\overline{D'} = (D'_p)_{p\geq 0}$ in Σ^s, we write $\overline{D} \prec \overline{D'}$ if and only if there exists $p_0 \geq 0$ such that $D_p = D'_p$ for $p \in \{0, \dots, p_0 - 1\}$ and $D_{p_0} \prec D'_{p_0}$.

We have defined a preorder on the leaf space f^s (Definition 2.22) and a preorder on the symbolic space Σ^s (Definition 2.27). It is natural to wonder whether the coding map $\chi^s \colon f^{s,\infty} \to \Sigma^s$ is compatible with these preorders or not. For pedagogical reasons, we first consider the simple situation where the two-dimensional foliation $\mathcal{F}^u$ is orientable:

Proposition 2.28 *Assume that the unstable foliation $\mathcal{F}^u$ is orientable. Then the coding map $\chi^s \colon f^{s,\infty} \to \Sigma^s$ preserves the preorders, ie for $\ell, \ell' \in f^{s,\infty}$, $\ell \prec \ell'$ if and only if $\chi^s(\ell) \prec \chi^s(\ell')$.*

Proof Since the two-dimensional foliation $\mathcal{F}^u$ is orientable, its lift $\widetilde{\mathcal{F}}^u$ is also orientable. Recall that the vector field $\widetilde{Y}$ is tangent to the leaves of the foliation $\widetilde{\mathcal{F}}^u$. So the orientability of the two-dimensional foliation $\widetilde{\mathcal{F}}^u$ implies that the return map $\widetilde{\theta}$ of the orbits of the vector field $\widetilde{Y}$ on the surface $\widetilde{S}$ preserves the orientation of the one-dimensional foliation $\widetilde{F}^u = \widetilde{\mathcal{F}}^u \cap \widetilde{S}$.

Consider two leaves $\ell, \ell' \in f^{s,\infty}$ such that $\ell \prec \ell'$. Let $\chi^s(\ell) = (D_p)_{p\geq 0}$ and $\chi^s(\ell) = (D'_p)_{p\geq 0}$. Recall that this means that

$$\ell = \bigcap_{p\geq 0} \widetilde{\theta}^{-p}(D_p) \quad \text{and} \quad \ell' = \bigcap_{p\geq 0} \widetilde{\theta}^{-p}(D'_p).$$

Consider the integer $p_0 = \min\{p \geq 0 \mid D_p \neq D'_p\}$ and the set

$$\widehat{D} := \bigcap_{p=0}^{p_0-1} \widetilde{\theta}^{-p}(D_p).$$

Both the leaves ℓ and ℓ' are included in $\widehat{D}$, and, according to Proposition 2.19, $\widehat{D}$ is a trivially bifoliated proper stable strip. So we can consider an arc α^u of a leaf ℓ^u of the foliation $\widetilde{F}^u$ such that α^u is included in the trivially bifoliated proper stable strip $\widehat{D}$ and the ends a and a' of α^u are on ℓ and ℓ', respectively. Since $\ell \prec \ell'$, the orientation of $\widetilde{F}^u$ goes from a towards a'. Now observe that $\widehat{D}$ is a connected component of $\widetilde{S} - \bigcup_{p=0}^{p_0-1} \widetilde{\theta}^p(\widetilde{L}^s)$. As a consequence, the map $\widetilde{\theta}^{p_0}$ is well defined on $\widetilde{D}$. In particular, we can consider $\beta^u := \widetilde{\theta}^{p_0}(\alpha^u)$. Observe that β^u is an arc of a leaf of the foliation $\widetilde{F}^u$. Its ends $b := \widetilde{\theta}^{p_0}(a)$ and $b' := \widetilde{\theta}^{p_0}(a')$ are respectively in $\widetilde{\theta}^{p_0}(\ell) \subset D_{p_0}$ and $\widetilde{\theta}^{p_0}(\ell') \subset D'_{p_0}$. Since the return map $\widetilde{\theta}^{p_0}$ preserves the orientation of the foliation $\widetilde{F}^u$, the orientation of $\widetilde{F}^u$ goes from b towards b'. It follows that $\widetilde{\theta}^{p_0}(\ell) \prec \widetilde{\theta}^{p_0}(\ell')$ and therefore $D_{p_0} \prec D'_{p_0}$. As a further consequence,

$$\chi^s(\ell) = (D_0, D_1, \dots, D_{p_0-1}, D_{p_0}, \dots) \prec (D_0, D_1, \dots, D_{p_0-1}, D'_{p_0}, \dots) = \chi^s(\ell').$$

This completes the proof of the implication $\ell \prec \ell' \Longrightarrow \chi^s(\ell) \prec \chi^s(\ell')$. The converse implication follows from the very same arguments in reversed order. □

In general, the relationship between the order on the leaf space f^s and the symbolic space Σ^s is more complicated:

Proposition 2.29 *Let ℓ and ℓ' be two different elements of $f^{s,\infty}$. Let $(D_p)_{p\geq 0} := \chi^s(\ell)$ and $(D'_p)_{p\geq 0} := \chi^s(\ell')$. Let p_0 be the smallest integer p such that $D_p \neq D'_p$.*

(1) *If the map $\widetilde{\theta}^{p_0}|_{\bigcap_{p=0}^{p_0-1} \widetilde{\theta}^{-p}(D_p)}$ preserves the orientation of the foliation $\widetilde{F}^u$, then*

$$\ell \prec \ell' \iff D_{p_0} \prec D'_{p_0} \iff \chi^s(\ell) \prec \chi^s(\ell').$$

(2) *If the map $\widetilde{\theta}^{p_0}|_{\bigcap_{p=0}^{p_0-1} \widetilde{\theta}^{-p}(D_p)}$ reverses the orientation of the foliation $\widetilde{F}^u$, then*

$$\ell \prec \ell' \iff D'_{p_0} \prec D_{p_0} \iff \chi^s(\ell') \prec \chi^s(\ell).$$

Proof The arguments are exactly the same as in the proof of Proposition 2.28. □

3 Topological equivalence of Anosov flows

We will now prove Theorem 1.2 with the help of the coding procedure implemented in Section 2.

3.1 A simplification

We begin by explaining why it is enough to prove Theorem 1.2 in the particular case where the vector fields X_1 and X_2 coincide.

Let (U, X_1, ψ_1) and (U, X_2, ψ_2) be two triples satisfying the hypotheses of Theorem 1.2. In particular, (U, X_1, ψ_1) and (U, X_2, ψ_2) are strongly isotopic. This means that there exists a continuous one-parameter family $\{(U, X_t, \psi_t)\}_{t\in[1,2]}$ such that (U, X_t) is a hyperbolic plug and $\psi_t\colon \partial^{\mathrm{out}}U \to \partial^{\mathrm{in}}U$ is a strongly transverse gluing map for every t. By standard hyperbolic theory, hyperbolic plugs are structurally stable. Hence, this means that we can find a continuous family $(h_t)_{t\in[1,2]}$ of self-homeomorphisms of U such that $h_1 = \mathrm{Id}$ and h_t induces an orbital equivalence between X_1 and X_t. For $t \in [1, 2]$, define

$$\widehat{\psi}_t := (h_t|_{\partial^{\mathrm{in}}U})^{-1} \circ \psi_t \circ (h_t|_{\partial^{\mathrm{out}}U})$$

and observe that $\widehat{\psi}_1 = \psi_1$. For sake of clarity, let $X := X_1$. Then:

- The triples $(U, X, \widehat{\psi}_1)$ and $(U, X, \widehat{\psi}_2)$ are strongly isotopic; the strong isotopy is given by the continuous path $\{(U, X, \widehat{\psi}_t)\}_{t\in[1,2]}$.
- For $t \in [1, 2]$, the flow induced by the vector field X on the manifold $\widehat{M}_t := U/\widehat{\psi}_t$ is orbitally equivalent to the flow induced by the vector field X_t on the manifold $M_t := U/\psi_t$; the orbital equivalence is induced by the homeomorphism h_t.

This shows that the hypotheses and the conclusion of Theorem 1.2 are satisfied for the triples (U, X_1, ψ_1) and (U, X_2, ψ_2) if and only if they are satisfied for the triples $(U, X, \widehat{\psi}_1)$ and $(U, X, \widehat{\psi}_2)$. This allows us to replace the vector fields X_1 and X_2 by a single vector field X in the proof of Theorem 1.2.

3.2 Setting

From now on, we consider a hyperbolic plug (U, X) endowed with two strongly transverse gluing diffeomorphisms $\psi_1, \psi_2\colon \partial^{\mathrm{out}}U \to \partial^{\mathrm{in}}U$. We denote by $\Lambda := \bigcap_{t\in\mathbb{R}} X^t(U)$ the maximal invariant set of the plug (U, X). For $i = 1, 2$, the quotient space $M_i := U/\psi_i$ is a closed three-dimensional manifold, and X induces a vector field Y_i on M_i. We assume that the hypotheses of Theorem 1.2 are satisfied; that is:

(0) The manifolds U, M_1 and M_2 are orientable.

(1) For $i = 1, 2$, the flow (Y_i^t) of the vector field Y_i is a transitive Anosov flow.

(2) The gluing maps ψ_1 and ψ_2 are strongly isotopic, ie there exists an isotopy $(\psi_s)_{s\in[1,2]}$ such that, for every s, the laminations L^s and $\psi_s(L^u_X)$ are strongly transverse.

In order to prove Theorem 1.2, we have to construct a homeomorphism $H\colon M_1 \to M_2$ mapping the oriented orbits of the Anosov flow (Y_1^t) to the orbits of the Anosov flow (Y_2^t). The construction will be divided into several steps.

3.3 Starting point of the construction: diffeomorphisms $\phi_{\mathrm{in}}, \phi_{\mathrm{out}}\colon S_1 \to S_2$

For $i = 1, 2$, we denote by π_i the projection of U on the closed three-dimensional manifold $M_i = U/\psi_i$. We denote by

$$S_i = \pi_i(\partial^{\mathrm{in}} U) = \pi_i(\partial^{\mathrm{out}} U)$$

the projection of the boundary of U. The surface S_i is endowed with the strongly transverse laminations

$$L_i^s := \pi_i(L_X^s) \quad \text{and} \quad L_i^u := \pi_i(L_X^u).$$

The maps $\pi_i|_{\partial^{\mathrm{in}} U}\colon \partial^{\mathrm{in}} U \to S_i$ and $\pi_i|_{\partial^{\mathrm{out}} U}\colon \partial^{\mathrm{out}} U \to S_i$ are invertible. This provides us with two diffeomorphisms

$$\phi_{\mathrm{in}} := \pi_2|_{\partial^{\mathrm{in}} U} \circ (\pi_1|_{\partial^{\mathrm{in}} U})^{-1}\colon S_1 \to S_2 \quad \text{and} \quad \phi_{\mathrm{out}} := \pi_2|_{\partial^{\mathrm{out}} U} \circ (\pi_1|_{\partial^{\mathrm{out}} U})^{-1}\colon S_1 \to S_2.$$

The diffeomorphisms ϕ_{in} and ϕ_{out} are the starting point of our construction. Observe that, at this step, we are very far from getting an orbital equivalence. Indeed, ϕ_{in} and ϕ_{out} are in no way compatible with the actions of the flows (Y_1^t) and (Y_2^t) (ie they do not conjugate the Poincaré return maps of (Y_1^t) and (Y_2^t) on the surfaces S_1 and S_2).

Nevertheless, the definitions of the diffeomorphisms ϕ_{in} and ϕ_{out} imply that

$$\phi_{\mathrm{in}}(L_1^s) = \pi_2|_{\partial^{\mathrm{in}} U} \circ (\pi_1|_{\partial^{\mathrm{in}} U})^{-1}(L_1^s) = \pi_2(L_X^s) = L_2^s.$$
$$\phi_{\mathrm{out}}(L_1^u) = \pi_2|_{\partial^{\mathrm{out}} U} \circ (\pi_1|_{\partial^{\mathrm{out}} U})^{-1}(L_1^u) = \pi_2(L_X^u) = L_2^u.$$

Remark 3.1 Be careful: in general, $\phi_{\mathrm{in}}(L_1^u) \neq L_2^u$ and $\phi_{\mathrm{out}}(L_1^s) \neq L_2^s$.

On the other hand, the strong isotopy connecting the gluing maps ψ_1 and ψ_2 can be used to construct an isotopy between the diffeomorphisms ϕ_{in} and ϕ_{out}:

Proposition 3.2 *There exists a continuous family $(\phi_t)_{t\in[0,1]}$ of diffeomorphisms from S_1 to S_2 such that $\phi_0 = \phi_{\mathrm{out}}$, such that $\phi_1 = \phi_{\mathrm{in}}$ and such that the laminations $\phi_t(L_1^u)$ and L_2^s are strongly transverse for every t.*

Proof By assumption, the gluing maps ψ_1 and ψ_2 are connected by a continuous path $(\psi_s)_{s\in[1,2]}$ of diffeomorphisms from $\partial^{\mathrm{out}} U$ to $\partial^{\mathrm{in}} U$ such that the laminations $\psi_s(L^u)$ and L^s are strongly transverse for every s. For $t \in [0, 1]$, we set

$$\phi_t := \pi_2|_{\partial^{\mathrm{out}} U} \circ \psi_2^{-1} \circ \psi_{2-t} \circ (\pi_1|_{\partial^{\mathrm{out}} U})^{-1}.$$

From this formula, we immediately get

$$\phi_0 = \pi_2|_{\partial^{\mathrm{out}} U} \circ (\pi_1|_{\partial^{\mathrm{out}} U})^{-1} = \phi_{\mathrm{out}}.$$

Plugging the equality $\pi_i|_{\partial^{\mathrm{in}}U} \circ \psi_i = \pi_i|_{\partial^{\mathrm{out}}U}$ into the definition of ϕ_1, we get

$$\phi_1 = \pi_2|_{\partial^{\mathrm{out}}U} \circ \psi_2^{-1} \circ \psi_1 \circ (\pi_1|_{\partial^{\mathrm{out}}U})^{-1} = \pi_2|_{\partial^{\mathrm{in}}U} \circ (\pi_1|_{\partial^{\mathrm{in}}U})^{-1} = \phi_{\mathrm{in}}.$$

We know that the laminations L^s_X and $\psi_{2-t}(L^u_X)$ are strongly transverse for every t. As a consequence, the laminations

$$\pi_2|_{\partial^{\mathrm{out}}U} \circ \psi_2^{-1}(L^s_X) = \pi_2|_{\partial^{\mathrm{in}}U}(L^s_X) = L^s_2$$

and

$$\pi_2|_{\partial^{\mathrm{out}}U} \circ \psi_2^{-1} \circ \psi_{2-t}(L^u_X) = \phi_t \circ \pi_1|_{\partial^{\mathrm{out}}U}(L^u_X) = \phi_t(L^u_1)$$

are strongly transverse for every t. □

It is important to observe that the diffeomorphism ϕ_{in} can be obtained as the restriction of a diffeomorphism from M_1 to M_2:

Proposition 3.3 *The diffeomorphism* $\phi_{\mathrm{in}}\colon S_1 \to S_2$ *is the restriction of a diffeomorphism* $\Phi_{\mathrm{in}}\colon M_1 \to M_2$.

Proof Once again, we use the existence of a continuous path $(\psi_s)_{s\in[1,2]}$ of diffeomorphisms from $\partial^{\mathrm{out}}U$ to $\partial^{\mathrm{in}}U$ connecting the gluing maps ψ_1 and ψ_2. We consider a collar neighbourhood V of $\partial^{\mathrm{out}}U$ in U, and a diffeomorphism $\xi\colon \partial^{\mathrm{out}}U \times [0,1] \to V$ of V such that $\xi(\partial^{\mathrm{out}}U \times \{0\}) = \partial^{\mathrm{out}}U$. We define a diffeomorphism $\overline{\Phi}_{\mathrm{in}}\colon U \to U$ by setting $\overline{\Phi}_{\mathrm{in}}(\xi(x,t)) := \psi_{2-t}^{-1} \circ \psi_1(x)$ for every $(x,t) \in \partial^{\mathrm{out}}U \times [0,1]$, and $\overline{\Phi}_{\mathrm{in}} = \mathrm{Id}$ on $U \setminus V$. By construction, this diffeomorphism satisfies

$$\overline{\Phi}_{\mathrm{in}} = \begin{cases} \mathrm{Id} & \text{on } \partial^{\mathrm{in}}U, \\ \psi_2^{-1} \circ \psi_1 & \text{on } \partial^{\mathrm{out}}U. \end{cases}$$

As a consequence, the relation $\pi_2 \circ \overline{\Phi}_{\mathrm{in}} = \overline{\Phi}_{\mathrm{in}} \circ \pi_1$ holds, and therefore $\overline{\Phi}_{\mathrm{in}}$ induces a diffeomorphism $\Phi_{\mathrm{in}}\colon M_1 \to M_2$. Since $\overline{\Phi}_{\mathrm{in}} = \mathrm{Id}$ on $\partial^{\mathrm{in}}U$, it follows that $\Phi_{\mathrm{in}}|_{S_1} = \pi_2|_{\partial^{\mathrm{in}}U} \circ (\pi_2|_{\partial^{\mathrm{in}}U})^{-1} = \phi_{\mathrm{in}}$, as desired. □

Now, we introduce the return maps on the surface S_1 and S_2. We first consider the crossing map of the plug (U, X)

$$\theta_X\colon \partial^{\mathrm{in}}U \setminus L^s \to \partial^{\mathrm{out}}U \setminus L^u.$$

By definition, $\theta_X(x)$ is the unique intersection point of the forward (X^t)–orbit of the point x with the surface $\partial^{\mathrm{out}}U$. For $i = 1, 2$, the map θ_X induces a map

$$\theta_i := \pi_i|_{\partial^{\mathrm{out}}U} \circ \theta_X \circ (\pi_i|_{\partial^{\mathrm{in}}U})^{-1}\colon S_i \setminus L^s_i \to S_i \setminus L^u_i.$$

This map θ_i is just the Poincaré return map of the flow (Y^t_i) on the surface S_i.

Proposition 3.4 *The diffeomorphisms θ_1, θ_2, ϕ_{in} and ϕ_{out} are related by*

$$\theta_2 \circ \phi_{\mathrm{in}} = \phi_{\mathrm{out}} \circ \theta_1.$$

Proof This follows immediately from the formulas defining θ_1, θ_2, ϕ_{in} and ϕ_{out}. □

Now we lift all the objects to the universal covers of M_1 and M_2. We pick a point $x_1 \in M_1$ which will serve as the basepoint of the fundamental group of the manifold M_1. The point $x_2 := \Phi_{\mathrm{in}}(x_1)$ will be used as the basepoint of fundamental group of the manifold M_2. The diffeomorphism Φ_{in} provides us with an isomorphism $(\Phi_{\mathrm{in}})_*$ between the fundamental groups $\pi_1(M_1, x_1)$ and $\pi_1(M_2, x_2)$. For $i = 1, 2$, we denote by $p_i \colon \widetilde{M}_i \to M_i$ the universal cover of the manifold M_i. We denote by $\widetilde{Y}_i$ the lift of the vector field Y_i on $\widetilde{M}_i$. Observe that $\widetilde{Y}_i$ is equivariant under the action of $\pi_1(M_i, x_i)$: for $\gamma \in \pi_1(M_i, x_i)$, one has $\widetilde{Y}_i(\gamma\tilde{x}) = D_{\tilde{x}}\gamma.\widetilde{Y}_i(\tilde{x})$. We denote by $\widetilde{S}_i$ the complete lift of the surface S_i (ie $\widetilde{S}_i := p_i^{-1}(S_i)$).

We denote by $\widetilde{L}_i^s$ and $\widetilde{L}_i^u$ the complete lifts of the laminations L_i^s and L_i^u. We denote by

$$\tilde{\theta}_i \colon \widetilde{S}_i \setminus \widetilde{L}_i^s \to \widetilde{S}_i \setminus L_i^u$$

the first return map of the flow of the vector field $\widetilde{Y}_i$ on the surface $\widetilde{S}_i$. Clearly, $\tilde{\theta}_i$ is a lift of the map θ_i. Moreover, $\tilde{\theta}_i$ commutes with the deck transformations:

$$\tilde{\theta}_i \circ \gamma = \gamma \circ \tilde{\theta}_i \quad \text{for every } \gamma \in \pi_1(M_i, x_i). \tag{1}$$

This commutation relation is an immediate consequence of the equivariance of $\widetilde{Y}_i$ (see above). Now we fix a lift $\widetilde{\Phi}_{\mathrm{in}} \colon \widetilde{M}_1 \to \widetilde{M}_2$ of the diffeomorphism Φ_{in} (note that, unlike what happens for θ_1 and θ_2, there is no canonical lift of Φ_{in}). Recall that the diffeomorphism Φ_{in} maps the surface S_1 to the surface S_2, and that the restriction of Φ_{in} to S_1 coincides with ϕ_{in}. As a consequence, the lift $\widetilde{\Phi}_{\mathrm{in}}$ maps the surface $\widetilde{S}_1$ to $\widetilde{S}_2$, and the restriction of $\widetilde{\Phi}_{\mathrm{in}}$ to $\widetilde{S}_1$ is a lift $\tilde{\phi}_{\mathrm{in}}$ of the diffeomorphism ϕ_{in}. By construction, this lift satisfies

$$\tilde{\phi}_{\mathrm{in}} \circ \gamma = (\Phi_{\mathrm{in}})_*(\gamma) \circ \tilde{\phi}_{\mathrm{in}} \quad \text{for every } \gamma \in \pi_1(M_1, x_1). \tag{2}$$

Now recall that, according to Proposition 3.2, there exists a continuous arc $(\phi_t)_{t\in[0,1]}$ of diffeomorphisms from S_1 to S_2 such that $\phi_0 = \phi_{\mathrm{in}}$ and $\phi_1 = \phi_{\mathrm{out}}$, and such that the laminations $\phi_t(L_1^u)$ and L_2^s are strongly transverse for every t. We lift this isotopy, starting at the lift $\tilde{\phi}_{\mathrm{in}}$ of $\phi_{\mathrm{in}} = \phi_0$. This yields a continuous arc $(\tilde{\phi}_t)_{t\in[0,1]}$ of diffeomorphisms from $\widetilde{S}_1$ to $\widetilde{S}_2$ such that $\tilde{\phi}_0 = \tilde{\phi}_{\mathrm{in}}$ and such that the laminations $\tilde{\phi}_t(\widetilde{L}_1^u)$ and $\widetilde{L}_2^s$ are strongly transverse for every t. The diffeomorphism $\tilde{\phi}_{\mathrm{out}} := \tilde{\phi}_1$ is a lift of the diffeomorphism ϕ_{out}. By continuity, the relation (2) remains true if we replace $\widetilde{\phi}_{\mathrm{in}} = \widetilde{\phi}_0$ by $\widetilde{\phi}_t$ for any $t \in [0,1]$. In particular, the diffeomorphism $\widetilde{\phi}_{\mathrm{out}}$ satisfies

$$\widetilde{\phi}_{\mathrm{out}} \circ \gamma = (\Phi_{\mathrm{in}})_*(\gamma) \circ \widetilde{\phi}_{\mathrm{out}} \quad \text{for every } \gamma \in \pi_1(M_1, x_1). \tag{3}$$

Proposition 3.5 *The diffeomorphisms* $\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\phi}_{\mathrm{in}}$ *and* $\widetilde{\phi}_{\mathrm{out}}$ *are related by*

$$\widetilde{\theta}_2 \circ \widetilde{\phi}_{\mathrm{in}} = \widetilde{\phi}_{\mathrm{out}} \circ \widetilde{\theta}_1.$$

Proof According to Proposition 3.4, the diffeomorphisms $\theta_2 \circ \phi_{\mathrm{in}}$ and $\phi_{\mathrm{out}} \circ \theta_1$ coincide. Hence, the diffeomorphisms $\widetilde{\theta}_2 \circ \widetilde{\phi}_{\mathrm{in}}$ and $\widetilde{\phi}_{\mathrm{out}} \circ \widetilde{\theta}_1$ are two lifts of the same diffeomorphism. It follows that there exists a deck transformation $\gamma_0 \in \pi_1(M_2, y_0)$ such that

$$\widetilde{\theta}_2 \circ \widetilde{\phi}_{\mathrm{in}} = \gamma_0 \circ \widetilde{\phi}_{\mathrm{out}} \circ \widetilde{\theta}_1.$$

Now consider a deck transformation $\gamma \in \pi_1(M_1, x_0)$. On the one hand, using (2) and (1), we get

$$\widetilde{\theta}_2 \circ \widetilde{\phi}_{\mathrm{in}} \gamma = \widetilde{\theta}_2 \circ (\Phi_{\mathrm{in}})_*(\gamma) \circ \widetilde{\phi}_{\mathrm{in}} = (\Phi_{\mathrm{in}})_*(\gamma) \circ \widetilde{\theta}_2 \circ \widetilde{\phi}_{\mathrm{in}} = ((\Phi_{\mathrm{in}})_*(\gamma) \cdot \gamma_0) \circ \widetilde{\phi}_{\mathrm{out}} \circ \widetilde{\theta}_1.$$

On the other hand, using (1) and (3), we get

$$\widetilde{\theta}_2 \circ \widetilde{\phi}_{\mathrm{in}} \circ \gamma = \gamma_0 \circ \widetilde{\phi}_{\mathrm{out}} \circ \widetilde{\theta}_1 \circ \gamma = \gamma_0 \circ \widetilde{\phi}_{\mathrm{out}} \circ \gamma \circ \widetilde{\theta}_1 = (\gamma_0 \cdot (\Phi_{\mathrm{in}})_*(\gamma)) \circ \widetilde{\phi}_{\mathrm{out}} \circ \widetilde{\theta}_1.$$

Hence,

$$(\Phi_{\mathrm{in}})_*(\gamma) \cdot \gamma_0 = \gamma_0 \cdot (\Phi_{\mathrm{in}})_*(\gamma).$$

Since $(\Phi_{\mathrm{in}})_*(\gamma)$ ranges over the whole fundamental group $\pi_1(M_2, y_0)$, it follows that γ_0 is in the centre of the fundamental group $\pi_1(M_2, y_0)$. If $\gamma_0 \neq \mathrm{Id}$, this implies that $\pi_1(M_2, y_0)$ has a nontrivial centre. It follows that M_2 is a Seifert manifold (see eg [1, Theorem 2.5.5]). Then an easy generalization of a well-known theorem of É Ghys implies that, up to finite cover, the Anosov flow (X_2^t) must be topologically equivalent to the geodesic flow on the unit tangent bundle of a closed hyperbolic surface (see [9] or [3, théorème 3.1]). This is clearly impossible, since X_2 admits a transverse torus (any connected component of the surface S_2 is such a torus). As a consequence, γ_0 must be the identity, and the desired relation $\widetilde{\theta}_2 \circ \widetilde{\phi}_{\mathrm{in}} = \widetilde{\phi}_{\mathrm{in}} \circ \widetilde{\theta}_1$ is proved. □

3.4 Construction of maps $\Delta^s : f_1^{s,\infty} \to f_2^{s,\infty}$ and $\Delta^u : f_1^{u,\infty} \to f_2^{u,\infty}$

In Section 2, we have defined some symbolic spaces which allow us to code certain orbits of certain Anosov flows. Let us introduce these symbolic space in our particular setting. For $i = 1, 2$, we consider the alphabet

$$\mathcal{A}_i := \{\text{connected components of } \widetilde{S}_i \setminus \widetilde{L}_i^s\},$$

and the symbolic space

$$\Sigma_i := \{(D_p)_{p\in\mathbb{Z}} \mid D_p \in \mathcal{A}_i \text{ and } \widetilde{\theta}_i(D_p) \cap D_{p+1} \neq \varnothing \text{ for every } p\}.$$

In order to code stable and unstable leaves, we consider the subspaces Σ_i^s and Σ_i^u of Σ_i defined by

$$\Sigma_i^s := \{(D_p)_{p\geq 0} \mid D_p \in \mathcal{A}_i \text{ and } \widetilde{\theta}_i(D_p) \cap D_{p+1} \neq \varnothing \text{ for every } p\}$$

and

$$\Sigma_i^u := \{(D_p)_{p< 0} \mid D_p \in \mathcal{A}_i \text{ and } \widetilde{\theta}_i(D_p) \cap D_{p+1} \neq \varnothing \text{ for every } p\}.$$

Proposition 3.6 *Let D_1 and D_1' be two elements of $\mathcal{A}_1$. Let $D_2 := \widetilde{\phi}_{\text{in}}(D_1)$ and $D_2' := \widetilde{\phi}_{\text{in}}(D_1')$. Then $\widetilde{\theta}_1(D_1)$ intersects D_1' if and only if $\widetilde{\theta}_2(D_2)$ intersects D_2'.*

Proof We have the sequence of equivalences

$$\begin{aligned}\widetilde{\theta}_1(D_1) \cap D_1' \neq \varnothing &\iff \widetilde{\phi}_{\text{in}}(\widetilde{\theta}_1(D_1)) \cap \widetilde{\phi}_{\text{in}}(D_1') \neq \varnothing \\ &\iff \widetilde{\phi}_{\text{out}}(\widetilde{\theta}_1(D_1)) \cap \widetilde{\phi}_{\text{in}}(D_1') \neq \varnothing \\ &\iff \widetilde{\theta}_2(\widetilde{\phi}_{\text{in}}(D_1)) \cap \widetilde{\phi}_{\text{in}}(D_1') \neq \varnothing \\ &\iff \widetilde{\theta}_2(D_2) \cap D_2' \neq \varnothing.\end{aligned}$$

The first equivalence is straightforward. The last one is nothing but the definition of the connected components D_2 and D_2'. The third equivalence follows from Proposition 3.5. It remains to prove the second equivalence. For that purpose, observe that $\widetilde{\theta}_1(D_1)$ is a strip bounded by two leaves of $\widetilde{L}_1^u$, and $\widetilde{\phi}_{\text{in}}(D_1')$ is a strip bounded by two leaves of $\widetilde{L}_2^s$. Now recall that there exists an isotopy $(\widetilde{\phi}_t)_{t\in[0,1]}$ joining $\widetilde{\phi}_{\text{in}}$ to $\widetilde{\phi}_{\text{out}}$ such that the lamination $\widetilde{\phi}_t(\widetilde{L}_1^u)$ is strongly transverse to the lamination $\widetilde{L}_2^s$. It follows that $\widetilde{\phi}_{\text{out}}(\widetilde{\theta}_1(D_1))$ intersects $\widetilde{\phi}_{\text{in}}(D_1')$ if and only if $\widetilde{\phi}_{\text{in}}(\widetilde{\theta}_1(D_1))$ intersects $\widetilde{\phi}_{\text{in}}(D_1')$. □

Now we consider the map

$$(\widetilde{\phi}_{\text{in}})^{\otimes\mathbb{Z}} \colon \mathcal{A}_1^{\mathbb{Z}} \to \mathcal{A}_2^{\mathbb{Z}}, \quad (D_p)_{p\in\mathbb{Z}} \mapsto (\widetilde{\phi}_{\text{in}}(D_p))_{p\in\mathbb{Z}}.$$

As an immediate consequence of Proposition 3.6, we get:

Corollary 3.7 $(\widetilde{\phi}_{\text{in}})^{\otimes\mathbb{Z}} \colon \mathcal{A}_1^{\mathbb{Z}} \to \mathcal{A}_2^{\mathbb{Z}}$ *maps Σ_1 to Σ_2.*

Corollary 3.7 entails that $(\widetilde{\phi}_{\text{in}})^{\otimes \mathbb{Z}_{\geq 0}}$ maps Σ_1^s to Σ_2^s, and $(\widetilde{\phi}_{\text{in}})^{\otimes \mathbb{Z}_{<0}}$ maps Σ_1^u to Σ_2^u. Hence, the map $\widetilde{\phi}_{\text{in}}$ builds a bridge between the symbolic spaces associated to the vector field Y_1 and those associated to the vector field Y_2.

Let us recall the definition of the coding maps constructed in Section 2.3. For $i = 1, 2$, we denote by $\mathcal{F}_i^s$ and $\mathcal{F}_i^u$ the weak stable and the weak unstable foliations of the Anosov flow (Y_i^t) on the manifold M_i. These two-dimensional foliations induce two one-dimensional foliations F_i^s and F_i^u on the surface S_i. We denote by $\widetilde{F}_i^s$ and $\widetilde{F}_i^u$ the lifts of F_i^s and F_i^u on $\widetilde{S}_i$. We denote by f_i^s and f_i^u the leaf spaces of the foliations $\widetilde{F}_i^s$ and $\widetilde{F}_i^u$. We denote by $f_i^{s,\infty}$ the subset of f_i^s made of the leaves that are not in $\widetilde{W}^s(\Lambda_i)$ (recall that $\widetilde{W}^s(\Lambda_i)$ is a union of leaves of $\widetilde{\mathcal{F}}_i^s$ and therefore $\widetilde{W}^s(\Lambda_i) \cap \widetilde{S}_i$ is a union of leaves of F_i^s). Similarly, we denote by $f_i^{u,\infty}$ the subset of f_i^u made of the leaves that are not in $\widetilde{W}^u(\Lambda_i)$. The construction of Section 2.3 provides two bijective coding maps

$$\chi_i^s \colon \tilde{f}_i^{s,\infty} \to \Sigma_i^s, \quad \ell \mapsto (D_p)_{p \geq 0}, \quad \text{where } \widetilde{\theta}_i^p(\ell) \subset D_p \text{ for every } p \geq 0,$$

and

$$\chi_i^u \colon \tilde{f}_i^{u,\infty} \to \Sigma_i^u, \quad \ell \mapsto (D_p)_{p<0}, \quad \text{where } \widetilde{\theta}_i^p(\ell) \subset D_p \text{ for every } p < 0.$$

Hence, we obtain two natural bijective maps

$$\Delta^s := (\chi_2^s)^{-1} \circ (\widetilde{\phi}_{\text{in}})^{\otimes \mathbb{Z}_{\geq 0}} \circ \chi_1^s \colon \tilde{f}_1^{s,\infty} \to \tilde{f}_2^{s,\infty}$$

and

$$\Delta^u := (\chi_2^u)^{-1} \circ (\widetilde{\phi}_{\text{in}})^{\otimes \mathbb{Z}_{<0}} \circ \chi_1^u \colon \tilde{f}_1^{u,\infty} \to \tilde{f}_2^{u,\infty}.$$

3.5 Extension of the maps Δ^s and Δ^u

We wish to extend the map Δ^s in order to obtain a bijective map between the leaf spaces $\tilde{f}_1^s$ and $\tilde{f}_2^s$. Observe that Δ^s is already defined from a dense subset of $\tilde{f}_1^s$ onto a dense subset of $\tilde{f}_2^s$. We will prove that Δ^s preserves the orders on $\tilde{f}_1^s$ and $\tilde{f}_2^s$. Of course, these are only partial orders. Nevertheless, according to Proposition 2.24, every leaf of $\widetilde{F}_1^s$ (resp. $\widetilde{F}_2^s$) admits a neighbourhood in $\tilde{f}_1^s$ (resp. $\tilde{f}_2^s$) which is totally ordered. As a consequence, the preservation of the order will be sufficient to extend Δ^s.

Our first task is to write a precise definition of the partial orders on $\tilde{f}_1^s$ and $\tilde{f}_2^s$. First we choose an orientation of the lamination $L_X^u \subset \partial^{\text{out}} U$. Pushing this orientation by the maps π_1 and π_2, this defines some orientations of the laminations $L_1^u = (\pi_1)_*(L_X^u) \subset S_1$ and $L_2^u = (\pi_2)_*(L_X^u) \subset S_2$. Since L_i^u is a sublamination of the foliation F_i^u (and since L_i^u intersects every connected component of S_i), the orientations of the laminations L_1^u and L_2^u define some orientations of the foliations F_1^u and F_2^u. Finally, these orientations can be lifted, providing orientations of the lifted foliations $\widetilde{F}_1^u$ and $\widetilde{F}_2^u$. It is important to notice that our choice of orientations for $\widetilde{F}_1^u$ and $\widetilde{F}_2^u$ are not independent from each other. More precisely, the orientations are chosen so that $\phi_{\mathrm{out}} = \pi_2|_{\partial^{\mathrm{out}}U} \circ (\pi_2|_{\partial^{\mathrm{out}}U})^{-1}$ maps the orientation of the lamination L_1^u to the orientation of the lamination L_2^u, and therefore:

(4) $\widetilde{\phi}_{\mathrm{out}}$ maps the orientated lamination $\widetilde{L}_1^u$ to the orientated lamination $\widetilde{L}_2^u$.

As explained in Section 2.4, the orientation of the foliation $\widetilde{F}_i^u$ induces a partial order $\prec_i$ on the leaf space $\widetilde{f}_i^s$ defined as follows: given two leaves $\ell_i, \ell_i' \in \widetilde{f}_i^s$ satisfy $\ell_i \prec_i \ell_i'$ if there exists an arc segment of an oriented leaf of $\widetilde{F}_i^u$ going from a point of ℓ_i to a point of ℓ_i'. Proposition 2.23 proves that this indeed defines an order on $\widetilde{f}_i^s$. Moreover, this order on $\widetilde{f}_i^s$ induces a partial order on the alphabet $\mathcal{A}_i$: given two elements D_i and D_i' of $\mathcal{A}_i$, we write $D_i \prec_i D_i'$ if there exists a leaf $\widetilde{\alpha}_i$ of $\widetilde{F}_i^s$ included in D_i and a leaf $\widetilde{\alpha}_i'$ of $\widetilde{F}_i^s$ included in D_i' such that $\widetilde{\alpha}_i \prec_i \widetilde{\alpha}_i'$. Proposition 2.25 shows that we can replace "there exists" by "for every" in this definition. It follows that $\prec_i$ is indeed a partial order on $\mathcal{A}_i$. Now comes the technical result which will allow us to extend the map Δ^s:

Proposition 3.8 *The map* $\Delta^s \colon (f_1^{s,\infty}, \prec_1) \to (f_2^{s,\infty}, \prec_2)$ *is order-preserving.*

In order to prove Proposition 3.8, we need several intermediary results.

Lemma 3.9 *The map* $\widetilde{\phi}_{\mathrm{in}} \colon (\mathcal{A}_1, \prec_1) \to (\mathcal{A}_2, \prec_2)$ *is order-preserving.*

Proof Consider two elements D_1 and D_1' of $\mathcal{A}_1$. Assume that $D_1 \prec_1 D_1'$. This means that there exists a leaf ℓ_1 of the oriented lamination $\widetilde{L}_1^u$ which crosses D_1 before crossing D_1'. As a consequence, if we endow $\widetilde{\phi}_{\mathrm{in}}(\ell_1)$ with the image under $\widetilde{\phi}_{\mathrm{in}}$ of the orientation of α_1, then $\widetilde{\phi}_{\mathrm{in}}(\ell_1)$ crosses $\widetilde{\phi}_{\mathrm{in}}(D_1)$ before crossing $\widetilde{\phi}_{\mathrm{in}}(D_1')$. Now recall that:

- $\widetilde{\phi}_{\mathrm{in}}(D_1)$ and $\widetilde{\phi}_{\mathrm{in}}(D_1')$ are strips bounded by leaves of the lamination $\widetilde{\phi}_{\mathrm{in}}(\widetilde{L}_1^s) = \widetilde{L}_2^s$.
- There exists an isotopy $(\widetilde{\phi}_t)$ joining $\widetilde{\phi}_{\mathrm{in}}$ to $\widetilde{\phi}_{\mathrm{out}}$ such that the lamination $\widetilde{\phi}_t(\widetilde{L}_1^u)$ is strongly transverse to the lamination $\widetilde{L}_2^s$ for every t.

We deduce that, if we endow $\widetilde{\phi}_{\mathrm{out}}(\ell_1)$ with the image under $\widetilde{\phi}_{\mathrm{out}}$ of the orientation of ℓ_1, then $\widetilde{\phi}_{\mathrm{out}}(\ell_1)$ crosses $\widetilde{\phi}_{\mathrm{in}}(D_1)$ before crossing $\widetilde{\phi}_{\mathrm{in}}(D_1')$. According to (4), this means that there is a leaf of the oriented lamination $\widetilde{L}_1^u$ which crosses $\widetilde{\phi}_{\mathrm{in}}(D_1)$ before crossing $\widetilde{\phi}_{\mathrm{in}}(D_1')$. By definition of the partial order $\prec_2$, this means that $\widetilde{\phi}_{\mathrm{in}}(D_1) \prec_2 \widetilde{\phi}_{\mathrm{in}}(D_1')$. □

Lemma 3.10 *Let D_1 be a connected component of $\widetilde{S}_1 \setminus \widetilde{L}^s_1$. Set $D_2 := \widetilde{\phi}_{\mathrm{in}}(D_1)$. Then the following are equivalent*:

1. *The map $\widetilde{\theta}_1$ restricted to the strip D_1 preserves the orientation of the foliation $\widetilde{F}^u_1$.*
2. *The map $\widetilde{\theta}_2$ restricted to the strip D_2 preserves the orientation of the foliation $\widetilde{F}^u_2$.*

Proof The proof is a bit intricate, because we need to introduce no fewer than six leaves and compare their orientations. Recall that we have chosen some orientations for the foliations $\widetilde{F}^u_1$ and $\widetilde{F}^u_2$. In the sequel, we will also consider the foliations $(\widetilde{\phi}_{\mathrm{in}})_*\widetilde{F}^u_1$, $(\widetilde{\phi}_{\mathrm{out}})_*\widetilde{F}^u_1$ and $(\widetilde{\phi}_t)_*\widetilde{F}^u_1$; we endow them with the images under $\widetilde{\phi}_{\mathrm{in}}$, $\widetilde{\phi}_{\mathrm{out}}$ and $\widetilde{\phi}_t$ of the orientation of $\widetilde{F}^u_1$.

We pick a leaf ℓ_1 of the lamination $\widetilde{L}^u_1$ so that $\ell_1 \cap D_1 \neq \varnothing$ (such a leaf always exists since the laminations $\widetilde{L}^s_1$ and $\widetilde{L}^u_1$ are strongly transverse). Then we set

$$\ell_2 := \widetilde{\phi}_{\mathrm{out}}(\ell_1), \quad \widehat{\ell}_2 := \widetilde{\phi}_{\mathrm{in}}(\ell_1),$$
$$\ell'_1 := \widetilde{\theta}_1(\ell_1 \cap D_1), \quad \ell'_2 := \widetilde{\theta}_2(\ell_2 \cap D_2), \quad \widehat{\ell}'_2 := \widetilde{\theta}_2(\widehat{\ell}_2 \cap D_2).$$

Observe that

$$\widehat{\ell}'_2 = \widetilde{\theta}_2(\widetilde{\phi}_{\mathrm{in}}(\ell_1) \cap D_2) = \widetilde{\theta}_2 \circ \widetilde{\phi}_{\mathrm{in}}(\ell_1 \cap D_1) = \widetilde{\phi}_{\mathrm{out}} \circ \widetilde{\theta}_1(\ell_1 \cap D_1) = \widetilde{\phi}_{\mathrm{out}}(\ell'_1) \tag{5}$$

(the third equality follows from Proposition 3.5). Now recall that, for $i = 1, 2$, both $\widetilde{L}^u_i$ and $(\widetilde{\theta}_i)_*(\widetilde{L}^u_i \cap D^s_i)$ are sublaminations of the foliation $\widetilde{\mathcal{F}}^u_i$. Also recall that $\widetilde{\phi}_{\mathrm{out}}(\widetilde{L}^u_1) = \widetilde{L}^u_2$. This provides some natural orientations on ℓ_1, ℓ'_1, ℓ_2, ℓ'_2, $\widehat{\ell}_2$ and $\widehat{\ell}'_2$:

- ℓ_1 and ℓ'_1 are leaves of the foliation $\widetilde{F}^u_1$, and hence inherit the orientation of $\widetilde{F}^u_1$.
- ℓ_2 and ℓ'_2 are leaves of the foliation $\widetilde{F}^u_2$, and hence inherit the orientation of $\widetilde{F}^u_2$; we endow them with the orientation of this foliation.
- $\widehat{\ell}_2$ is a leaf of the foliation $(\widetilde{\phi}_{\mathrm{in}})_*\widetilde{F}^u_1$, and hence inherits the orientation of $(\widetilde{\phi}_{\mathrm{in}})_*\widetilde{F}^u_1$;
- $\widehat{\ell}'_2$ is a leaf of the foliation $(\widetilde{\phi}_{\mathrm{out}})_*\widetilde{F}^u_1$, and hence inherits the orientation of $(\widetilde{\phi}_{\mathrm{out}})_*\widetilde{F}^u_1$.

By symmetry, it is enough to prove the implication (1) $\Longrightarrow$ (2). So we assume that the restriction of $\widetilde{\theta}_1$ to D^s_1 preserves the orientation of $\widetilde{F}^u_1$; in particular:

$$\widetilde{\theta}_1 \text{ maps the orientation of } \ell_1 \text{ to that of } \ell'_1. \tag{6}$$

According to (4):

$$\widetilde{\phi}_{\mathrm{out}} \text{ maps the orientation of } \ell_1 \text{ to that of } \ell_2. \tag{7}$$

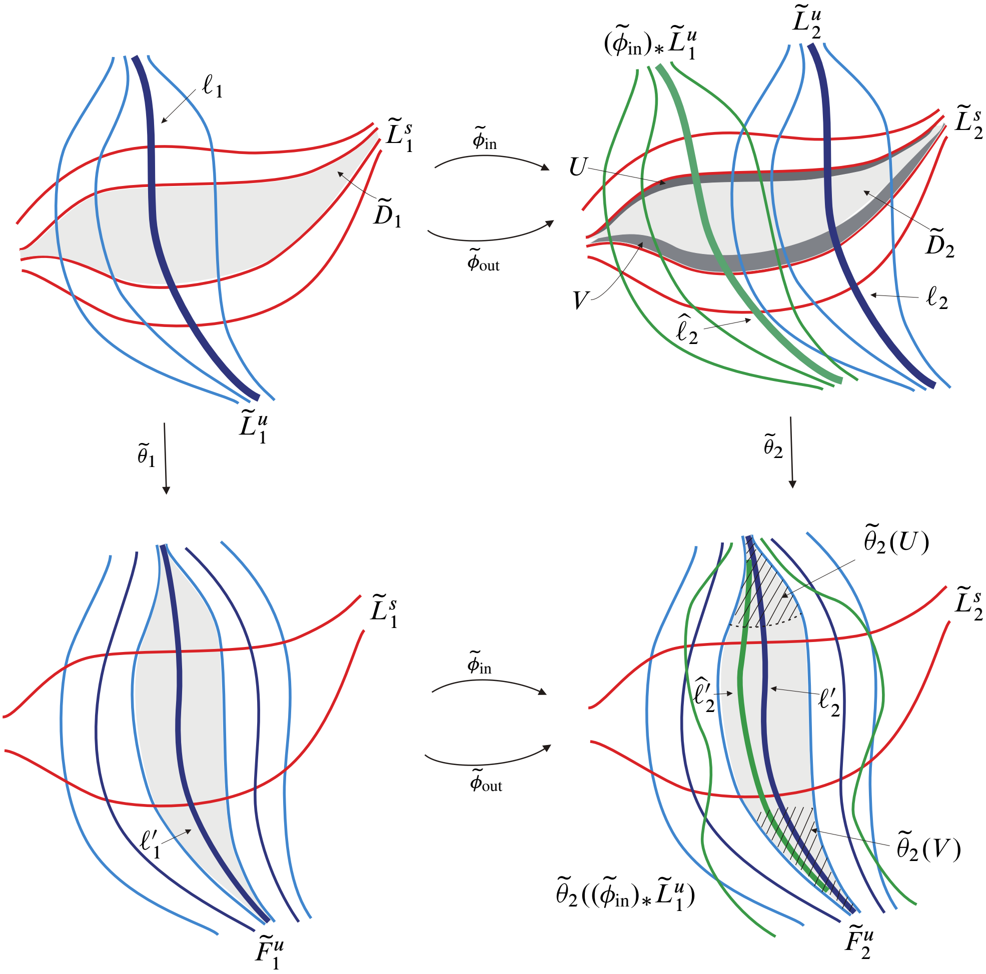

$(\widetilde{\phi}_{\mathrm{in}})_*\widetilde{L}^u_1$
$\widetilde{L}^u_2$
ℓ_1
$\widetilde{L}^s_1$
$\widetilde{\phi}_{\mathrm{in}}$
$\widetilde{L}^s_2$
U
$\widetilde{D}_1$
$\widetilde{D}_2$
$\widetilde{\phi}_{\mathrm{out}}$
V
ℓ_2
$\widehat{\ell}_2$
$\widetilde{L}^u_1$
$\widetilde{\theta}_1$
$\widetilde{\theta}_2$
$\widetilde{\theta}_2(U)$
$\widetilde{L}^s_1$
$\widetilde{L}^s_2$
$\widetilde{\phi}_{\mathrm{in}}$
$\widehat{\ell}'_2$
ℓ'_2
$\widetilde{\phi}_{\mathrm{out}}$
ℓ'_1
$\widetilde{\theta}_2(V)$
$\widetilde{\theta}_2((\widetilde{\phi}_{\mathrm{in}})_*\widetilde{L}^u_1)$
$\widetilde{F}^u_1$
$\widetilde{F}^u_2$

Figure 4: Proof of Lemma 3.10.

The orientations of ℓ_1, ℓ_2, $\widehat{\ell}_2$ and $\widehat{\ell}'_2$ are chosen in such a way that $\widetilde{\phi}_{\mathrm{in}}^{-1}$ maps the orientation of $\widehat{\ell}_2$ to that of ℓ_1, and $\widetilde{\phi}_{\mathrm{out}}$ maps the orientation of ℓ'_1 to that of $\widehat{\ell}'_2$. Putting this together with (6), we obtain that $\widetilde{\phi}_{\mathrm{out}} \circ \widetilde{\theta}_1 \circ \widetilde{\phi}_{\mathrm{in}}^{-1}$ maps the orientation of $\widehat{\ell}_2$ to that of $\widehat{\ell}'_2$. Using Proposition 3.5, we obtain:

$$\widetilde{\theta}_2 \text{ maps the orientation of } \widehat{\ell}_2 \text{ to that of } \widehat{\ell}'_2. \tag{8}$$

Our final goal is to prove that $\widetilde{\theta}_2$ maps the orientation of ℓ_2 to that of ℓ'_2. So, in view of (8), we need to compare the orientations of ℓ_2 and $\widehat{\ell}_2$ on the one hand, and the orientations ℓ'_2 and $\widehat{\ell}'_2$ on the other hand. We start with ℓ_2 and $\widehat{\ell}_2$.

Recall that D_2 is a strip in $\widetilde{S}_2$ bounded by two leaves of the stable lamination $\widetilde{L}^s_2$. We denote these two leaves by α and β in such a way that oriented unstable leaf ℓ_2 enters in D_2 by crossing α and exits D_2 by crossing β. According to (7), the orientation of $\ell_2 = (\widetilde{\phi}_{\mathrm{out}})_* \ell_1$ as a leaf of $\widetilde{L}^u_2 \subset \widetilde{F}^u_2$ coincides with the orientation as a leaf of $(\widetilde{\phi}_{\mathrm{out}})_* \widetilde{L}^u_1 \subset (\widetilde{\phi}_{\mathrm{out}})_* F^u_1$. Moreover, recall that there exists an isotopy $(\widetilde{\phi}_t)_{t\in[0,1]}$ joining $\widetilde{\phi}_0 = \widetilde{\phi}_{\mathrm{in}}$ to $\widetilde{\phi}_1 = \widetilde{\phi}_{\mathrm{out}}$ such that the lamination $\widetilde{\phi}_t(\widetilde{L}^u_1)$ is strongly transverse to the lamination $\widetilde{L}^u_2$ for every t. We deduce that $\widehat{\ell}_2 = (\widetilde{\phi}_{\mathrm{in}})_*(\ell_1)$ crosses D_2 in the same direction as $\ell_2 = (\widetilde{\phi}_{\mathrm{out}})_* \ell_1$. In other words:

(9) Both ℓ_2 and $\widehat{\ell}_2$ enter D_2 by crossing α and exit D_2 by crossing β.

Let U and V be some disjoint neighbourhoods of the stable leaves α and β in the strip D_2. Assertion (9) can be reformulated as follows:

(10) The arcs of oriented leaves $\ell_2 \cap D_2$ and $\widehat{\ell}_2 \cap D_2$ both go from U to V.

We are left to compare the orientations of ℓ'_2 and $\widehat{\ell}'_2$. First observe that $\widetilde{\theta}_2(D_2)$ is an open strip in $\widetilde{S}_2$, bounded by two leaves of the unstable lamination $\widetilde{L}^u_2 = (\widetilde{\phi}_{\mathrm{out}})_* \widetilde{L}^u_1$. The closure $\mathrm{Cl}(\widetilde{\theta}_2(D_2))$ of $\widetilde{\theta}_2(D_2)$ is the union of the open strip $\widetilde{\theta}_2(D_2)$ and its two boundary leaves. The boundary components of $\widetilde{\theta}_2(D_2)$ are leaves of both the foliations F^u_2 and $(\widetilde{\phi}_{\mathrm{out}})_* F^u_1$. Moreover, F^u_2 and $(\widetilde{\phi}_{\mathrm{out}})_* F^u_1$ induce two trivial oriented foliations on the closed strip $\mathrm{Cl}(\widetilde{\theta}_2(D_2))$. In particular, the leaves of F^u_2 and $(\widetilde{\phi}_{\mathrm{out}})_* F^u_1$ in $\mathrm{Cl}(\widetilde{\theta}_2(D_2))$ go from one end of $\mathrm{Cl}(\widetilde{\theta}_2(D_2))$ to the other end. In order to distinguish the two ends of the closed strip $\mathrm{Cl}(\widetilde{\theta}_2(D_2))$, we use the sets $\mathrm{Cl}(\widetilde{\theta}_2(U))$ and $\mathrm{Cl}(\widetilde{\theta}_2(V))$. These sets are disjoint neighbourhoods of the two ends of $\mathrm{Cl}(\widetilde{\theta}_2(D_2))$. So we just need to decide if the leaves go from $\mathrm{Cl}(\widetilde{\theta}_2(U))$ to $\mathrm{Cl}(\widetilde{\theta}_2(V))$, or the contrary. On the one hand, putting (8) and (10) together, we obtain that $\widehat{\ell}_2$ goes from $\mathrm{Cl}(\widetilde{\theta}_2(U))$ to $\mathrm{Cl}(\widetilde{\theta}_2(V))$. On the other hand, F^u_2 and $(\widetilde{\phi}_{\mathrm{out}})_* F^u_1$ are trivial oriented foliations on $\mathrm{Cl}(\widetilde{\theta}_2(D_2))$, and, according to (4), they induce the same orientation on the boundary leaves of D'_2. So we conclude that all the leaves of both the oriented foliations F^u_2 and $(\widetilde{\phi}_{\mathrm{out}})_* F^u_1$ go from $\mathrm{Cl}(\widetilde{\theta}_2(U))$ to $\mathrm{Cl}(\widetilde{\theta}_2(V))$. In particular:

(11) The oriented leaves ℓ'_2 and $\widehat{\ell}'_2$ go from $\widetilde{\theta}_2(U)$ to $\widetilde{\theta}_2(V)$.

From (10) and (11), we deduce that $\widetilde{\theta}_2|_{D_2}$ maps the orientation of ℓ_2 to that of ℓ'_2. By definition of the orientations of ℓ_2 and ℓ'_2, this means that the restriction of $\widetilde{\theta}_2$ to the strip D_2 preserves the orientation of the foliation $\widetilde{F}^u_2$. This completes the proof of the implication (1) $\Longrightarrow$ (2). □

Corollary 3.11 *Let $D_{1,0},\dots,D_{1,p_0-1}$ be connected components of $\widetilde{S}_1\setminus\widetilde{L}_1^s$ such that $\bigcap_{p=0}^{p_0-1}\widetilde{\theta}_1^p(D_{1,p})$ is nonempty. For $p=1,\dots,p_0-1$, let $D_{2,p}:=\widetilde{\phi}_{\mathrm{in}}(D_{1,p})$. Then the following are equivalent*:

1. *The map $\widetilde{\theta}_1^{p_0}$ restricted to $\bigcap_{p=0}^{p_0-1}\widetilde{\theta}_1^p(D_{1,p})$ preserves the orientation of the foliation $\widetilde{F}_1^u$.*
2. *The map $\widetilde{\theta}_2^{p_0}$ restricted to $\bigcap_{p=0}^{p_0-1}\widetilde{\theta}_2^p(D_{2,p})$ preserves the orientation of the foliation $\widetilde{F}_2^u$.*

Proof For $i=1,2$, consider the set $J_i\subset\{0,\dots,p_0-1\}$ so that the restriction of $\widetilde{\theta}_i$ to $D_{i,p}$ preserves the orientation of $\widetilde{F}_i^u$. On the one hand, Lemma 3.10 implies that the sets J_1 and J_2 coincide. On the other hand, it is clear that the restriction of $\widetilde{\theta}_i$ to $\bigcap_{j=0}^{p_0-1}\widetilde{\theta}_i^p(D_{i,p})$ preserves the orientation of the leaves of $\widetilde{F}_i^u$ if and only if the cardinality of J_i is even. □

Proof of Proposition 3.8 We consider two leaves γ_1 and γ_1' in $f_1^{s,\infty}$, we define $\gamma_2:=\Delta^s(\gamma_1)$ and $\gamma_2':=\Delta^s(\gamma_1')$, and we assume that $\gamma_1\prec_1\gamma_1'$. We aim to prove $\gamma_2\prec_2\gamma_2'$. Let $\chi_1^s(\widetilde{\gamma}_1)=(D_{1,p})_{p\geq 0}$, $\chi_1^s(\widetilde{\gamma}_1')=(D_{1,p}')_{p\geq 0}$, $\chi_2^s(\widetilde{\gamma}_2)=(D_{2,p})_{p\geq 0}$ and $\chi_2^s(\widetilde{\gamma}_2')=(D_{2,p}')_{p\geq 0}$. By definition of the map χ_i^s, this means that, for $i=1,2$,

$$\widetilde{\gamma}_i=\bigcap_{p\geq 0}\widetilde{\theta}_i^{-p}(D_{i,p})\quad\text{and}\quad\widetilde{\gamma}_i'=\bigcap_{p\geq 0}\widetilde{\theta}_i^{-p}(D_{i,p}').$$

And, since $\widetilde{\gamma}_2=\Delta^s(\widetilde{\gamma}_1)$ and $\widetilde{\gamma}_2'=\Delta^s(\widetilde{\gamma}_1')$, we have

$$D_{2,p}=\phi_{\mathrm{in}}(D_{1,p})\quad\text{and}\quad D_{2,p}'=\phi_{\mathrm{in}}(D_{1,p}')$$

for every $p\geq 0$. We denote by p_0 the smallest integer p such that $D_{1,p}\neq D_{1,p}'$.

Let us consider the case where the map $\widetilde{\theta}_1^{p_0}$ restricted to $\bigcap_{p=0}^{p_0-1}\widetilde{\theta}_1^{-p}(D_{1,p})$ preserves the orientation of the foliation $\widetilde{F}_1^u$.

- Proposition 2.29 implies that $D_{1,p_0}\prec_1 D_{1,p_0}'$.
- Since $\phi_{\mathrm{in}}\colon\mathcal{A}_1\to\mathcal{A}_2$ is order-preserving (Lemma 3.9), $D_{2,p_0}\prec_2 D_{2,p_0}'$.
- Corollary 3.11 implies that the map $\widetilde{\theta}_2^{p_0}$ restricted to $\bigcap_{p=0}^{p_0-1}\widetilde{\theta}_2^{-p}(D_{2,p})$ preserves the orientation of the foliation $\widetilde{F}_2^u$.
- Using again Proposition 2.29, we deduce from the two last items above that $\widetilde{\gamma}_2\prec_2\widetilde{\gamma}_2'$, as desired.

The case where the map $\widetilde{\theta}_1^{p_0}$ restricted to $\bigcap_{p=0}^{p_0-1}\widetilde{\theta}_1^{-p}(D_{1,p})$ reverses the orientation of the foliation $\widetilde{F}_1^u$ follows from the very same arguments. □

Corollary 3.12 *The map $\Delta^s\colon f_1^{s,\infty} \to f_2^{s,\infty}$ extends in a unique way to an order-preserving bijection $\Delta^s : f_1^s \to f_2^s$.*

Proof This is an immediate consequence of the following facts:

- $\Delta_s\colon f_1^{s,\infty} \to f_2^{s,\infty}$ is an order-preserving map (Proposition 3.8).
- For $i = 1, 2$, $f_i^{s,\infty}$ is a dense subset of the (nonseparated) one-dimensional manifold f_i^s (Proposition 2.8).
- For $i = 1, 2$, each leaf $\ell \in f_i^s$ has a neighbourhood U_ℓ in f_i^s such that the leaves in U_ℓ are totally ordered (Proposition 2.24). □

Of course, the stable and the unstable directions play some symmetric roles, so the same arguments as above allow one to prove the following analog of Corollary 3.12:

Corollary 3.13 *The map $\Delta^u\colon f_1^{u,\infty} \to f_2^{u,\infty}$ extends in a unique way to an order-preserving bijection $\widehat{\Delta}^u\colon f_1^u \to f_2^u$.*

3.6 Mating $\widehat{\Delta}^s$ and $\widehat{\Delta}^u$: construction of the map $\widehat{\Delta}$

Now, we will mate the maps $\widehat{\Delta}^s$ and $\widehat{\Delta}^u$ to obtain a $\widehat{\Delta}\colon \widetilde{S}_1 \to \widetilde{S}_2$. In view to that goal, we need the following lemma:

Lemma 3.14 *Consider a leaf ℓ_1^s of the stable foliation $\widetilde{\mathcal{F}}_1^s$ and a leaf ℓ_1^u of the unstable foliation $\widetilde{\mathcal{F}}_1^u$. Then ℓ_1^s intersects ℓ_1^u if and only if $\widehat{\Delta}^s(\ell_1^s)$ intersects $\widehat{\Delta}^u(\ell_1^u)$.*

Proof The case where the leaves ℓ_1^s and ℓ_1^u belong to $f_1^{s,\infty}$ and $f_1^{u,\infty}$ is a consequence of Proposition 3.6 (together with the definitions of the maps Δ^s, Δ^u and Δ): the leaves ℓ_1^s and ℓ_1^u intersect at x if and only if the leaves $\widehat{\Delta}^s(\ell_1^s) = \Delta^s(\ell_1^s)$ and $\widehat{\Delta}^u(\ell_1^u) = \Delta^u(\ell_1^u)$ intersect at $\Delta(x)$. The general case follows by density of $f_i^{s,\infty}$ and $f_i^{u,\infty}$ in $f_i^{s,\infty}$ and $f_i^{u,\infty}$. □

Now we define a map $\widehat{\Delta}\colon \widetilde{S}_1 \to \widetilde{S}_2$. Let $\tilde{x}$ be any point in $\widetilde{S}_1$. Denote by ℓ_1^s (resp. ℓ_1^u) the leaf of the stable foliation $\widetilde{\mathcal{F}}_1^s$ (resp. the unstable foliation $\widetilde{\mathcal{F}}_1^u$) passing through x. Recall that x is the unique intersection point of ℓ_1^s and ℓ_1^u. According to the preceding lemma, the stable leaf $\widehat{\Delta}^s(\ell_1^s)$ and the unstable leaf $\widehat{\Delta}^u(\ell_1^u)$ do intersect. According to Proposition 2.3, the intersection is a single point. We define $\widehat{\Delta}(\tilde{x})$ to be the unique intersection point of the leaves $\widehat{\Delta}^s(\ell_1^s)$ and $\widehat{\Delta}^u(\ell_1^u)$. In other words, $\widehat{\Delta}$ is defined by

$$\widehat{\Delta}(\ell_1^s \cap \ell_1^u) = \widehat{\Delta}^s(\ell_1^s) \cap \widehat{\Delta}^u(\ell_1^u). \tag{12}$$

By construction, the map $\widehat{\Delta}$ is bijective and maps the foliations $\widetilde{F}_1^s$ and $\widetilde{F}_1^u$ to the foliations $\widetilde{F}_2^s$ and $\widetilde{F}_2^u$, preserving the orders on the leaf spaces. Since the leaf spaces are locally totally ordered (Proposition 2.24), it follows that $\widehat{\Delta}$ is continuous. Hence, $\widehat{\Delta}$ is a homeomorphism.

Proposition 3.15 *The map* $\widehat{\Delta}\colon \widetilde{S}_1 \to \widetilde{S}_2$ *is equivariant with respect to the actions of the fundamental groups: for every* γ *of* $\pi_1(M_1)$,

$$\widehat{\Delta}\circ\gamma = (\widetilde{\Phi}_{\text{in}})_*(\gamma)\circ\widehat{\Delta}.$$

Proof This is a rather immediate consequence of the construction of $\widehat{\Delta}$. First recall that $\widehat{\Delta}$ is a continuous extension of the map $\Delta\colon \widetilde{S}_1^\infty \to \widetilde{S}_2^\infty$ and recall that $\widetilde{S}_1^\infty$ and $\widetilde{S}_2^\infty$ are dense subsets of $\widetilde{S}_1$ and $\widetilde{S}_2$. As a consequence, it is enough to prove that Δ is equivariant with respect to the actions of the fundamental groups. Now recall that Δ is defined as the composition of three maps:

$$\Delta = (\chi_2)^{-1}\circ(\widetilde{\phi}_{\text{in}})^{\otimes\mathbb{Z}}\circ\chi_1.$$

But we know that:

- The map χ_i commutes with the action of the fundamental group $\pi_1(M_i)$ for $i=1,2$ (Proposition 2.17).
- The map $\widetilde{\phi}_{\text{in}}$ satisfies $\widetilde{\phi}_{\text{in}}\circ\gamma = (\widetilde{\Phi}_{\text{in}})_*(\gamma)\circ\widetilde{\phi}_{\text{in}}$ (equation (2)).

This shows that the map Δ satisfies the equivariance relation $\Delta\circ\gamma = (\widetilde{\Phi}_{\text{in}})_*(\gamma)\circ\Delta$. □

Proposition 3.16 *The map* $\widehat{\Delta}\colon \widetilde{S}_1 \to \widetilde{S}_2$ *conjugates the Poincaré maps* $\widetilde{\theta}_1$ *and* $\widetilde{\theta}_2$; *that is*,

$$\widehat{\Delta}\circ\widetilde{\theta}_1 = \widetilde{\theta}_2\circ\widehat{\Delta}.$$

Proof On the one hand, for $i=1,2$, the coding map χ_i^s conjugates the Poincaré map $\widetilde{\theta}_i$ on $\widetilde{S}_i$ to the shift map on the symbolic space Σ_i^s (Proposition 2.18). On the other hand, the map $(\widetilde{\phi}_{\text{in}})^{\otimes\mathbb{Z}_{\geq 0}}$ obviously conjugates the shift map on Σ_1^s to the shift map on Σ_2^s. Hence, $\Delta^s = (\chi_2^s)^{-1}\circ(\widetilde{\phi}_{\text{in}})^{\otimes\mathbb{Z}_{\geq 0}}\circ\chi_1^s$ conjugates the action $\widetilde{\theta}_1$ on $f_1^{s,\infty}$ to the action of $\widetilde{\theta}_2$ on $f_2^{s,\infty}$. By density of $f_1^{s,\infty}$ in f_i^s, it follows that $\widehat{\Delta}^s$ conjugates the action $\widetilde{\theta}_1$ on f_1^s to the action of $\widetilde{\theta}_2$ on f_2^s. Similarly, $\widehat{\Delta}^u$ conjugates the action $\widetilde{\theta}_1$ on f_1^u to the action of $\widetilde{\theta}_2$ on f_2^u. Finally, since $\widetilde{\Delta}$ is defined by mating $\widehat{\Delta}^s$ and $\widehat{\Delta}^u$ (see (12)), this implies that $\widetilde{\Delta}$ conjugates $\widetilde{\theta}_1$ to $\widetilde{\theta}_2$. □

3.7 From the map $\widehat{\Delta}$ to the orbital equivalence

To conclude the proof of Theorem 1.2, we need to introduce the orbit spaces of the Anosov flows (Y_1^t) and (Y_2^t). The *orbit space* of (Y_i^t) is by definition the quotient of the manifold $\widetilde{M}_i$ by the action of the flow (Y_i^t). We denote it by O_i, and we denote by pr_i the natural projection of $\widetilde{M}_i$ on O_i. The action of the fundamental group $\pi_1(M_i)$ on $\widetilde{M}_i$ induces an action of this group on O_i. The two-dimensional foliations $\widetilde{\mathcal{F}}_i^s$ and $\widetilde{\mathcal{F}}_i^u$ are leafwise invariant under the flow (Y_i^t) and therefore can be projected in the orbit space O_i. They induce a pair (g_i^s, g_i^u) of transverse one-dimensional foliations on O_i.

The orbit space O_i by itself does not carry much information: indeed, O_i is always a separated manifold diffeomorphic to $\mathbb{R}^2$ (see [8, Proposition 2.1] or [2, Theorem 3.2]). The pair of transverse foliations (g_i^s, g_i^u) carries much more interesting information (see the work of Barbot and Fenley on the subject; good references are Barbot's habilitation memoir [3] and Barthelmé's lecture notes [4]). The action of $\pi_1(M_i)$ on O_i carries even richer dynamical information: actually, this action characterizes the flow (Y_i^t) up to topological equivalence (see Theorem 3.22 below).

Recall that Λ denotes the maximal invariant set of the initial hyperbolic plug (U, X), that Λ_i denotes the projection of Λ in the manifold $M_i = U/\psi_i$, and that $\widetilde{\Lambda}_i$ the complete lift of Λ_i in the universal cover $\widetilde{M}_i$. Now we denote by L_i the projection of the set $\widetilde{\Lambda}_i$ in O_i.

Lemma 3.17 *The projection* $\mathrm{pr}_i(\widetilde{S}_i)$ *of the surface* $\widetilde{S}_i$ *in the orbit space* O_i *is exactly the complement of the set* L_i *in* O_i*.*

Proof The set Λ is the union of the orbits of the vector field X which remain in U forever, ie which do not intersect ∂U. Hence, the set $\Lambda_i = \pi_i(\Lambda)$ is the union of the orbits of the vector field $Y_i = (\pi_i)_* X$ which do not intersect the surface $S_i = \pi_i(\partial U)$. As a further consequence, $\widetilde{\Lambda}_i$ is the union of the orbits of the vector field $\widetilde{Y}_i$ which do not intersect the surface $\widetilde{S}_i$. This means that the projection of $\widetilde{S}_i$ in the orbit space O_i is exactly the complement of the projection of the set $\widetilde{\Lambda}_i$. □

Proposition 3.16 can be rephrased as follows: two points $x, x' \in \widetilde{S}_1$ belong to the same orbit of the flow $(\widetilde{Y}_1^t)$ if and only if the points $\widehat{\Delta}(x)$ and $\widehat{\Delta}(x')$ belong to the same orbit of the flow $(\widetilde{Y}_2^t)$. As a consequence, the homeomorphism $\widehat{\Delta}\colon \widetilde{S}_1 \to \widetilde{S}_2$ induces a homeomorphism

$$\delta\colon \mathrm{pr}_1(\widetilde{S}_1) = O_1 \setminus L_1 \to \mathrm{pr}_2(\widetilde{S}_2) = O_2 \setminus L_2.$$

Since $\widehat{\Delta}$ is equivariant with respect to the actions of the fundamental groups (Proposition 3.15), the homeomorphism δ is also equivariant: for every $\gamma \in \pi_1(M_1)$,

$$\delta \circ \gamma = (\widetilde{\Phi}_{\text{in}})_*(\gamma) \circ \delta.$$

Our next step is to extend the map η to the whole orbit spaces.

Proposition 3.18 *The homeomorphism $\delta\colon \mathcal{O}_1 \setminus L_1 \to \mathcal{O}_2 \setminus L_2$ can be extended in a unique way to a homeomorphism $\overline{\delta}\colon \mathcal{O}_1 \to \mathcal{O}_2$ which is equivariant with respect to the actions of the fundamental groups of M_1 and M_2.*

We shall use the following general lemma of planar topology:

Lemma 3.19 *Let A and B be totally discontinuous subsets of $\mathbb{R}^2$ and $h\colon \mathbb{R}^2 \setminus A \to \mathbb{R}^2 \setminus B$. Assume that, for every compact subset K of $\mathbb{R}^2$, the set $h(K \setminus A)$ is relatively compact in $\mathbb{R}^2$. Then h can be extended to a homeomorphism of $\bar{h}\colon \mathbb{R}^2 \to \mathbb{R}^2$.*

This lemma is easy and certainly well known to people working in planar topology, but we were not able to find it in the literature. We provide a proof for sake of completeness.

Proof We proceed to the definition of $\bar{h}$. Let x be a point in A. We pick a decreasing sequence $(X_n)_{n\geq 0}$ of compact connected subsets of $\mathbb{R}^2$ so that $X_n \neq \{x\}$ for every n and so that $\bigcap_n X_n = \{x\}$. For every $n \geq 0$, let Y_n be the closure in $\mathbb{R}^2$ of the set $h(X_n \setminus A)$. Our assumptions imply that $(Y_n)_{n\geq 0}$ is a decreasing sequence of nonempty compact connected subsets of $\mathbb{R}^2$. As a consequence, the intersection $\bigcap_n Y_n$ must be a nonempty compact connected subset of $\mathbb{R}^2$. Moreover, since $\bigcap_n X_n = \{x\} \subset A$, the intersection $\bigcap_n Y_n$ must be included in B. Since B is totally disconnected, it follows that $\bigcap_n Y_n$ must be a singleton $\{y\}$. Standard arguments show that the point y does not depend on the choice of the sequence (X_n). We set $\bar{h}(x) := y$. Repeating the same procedure for each point $x \in A$, we get an extension $\bar{h}\colon \mathbb{R}^2 \to \mathbb{R}^2$ of h. The continuity of $\bar{h}$ follows easily from its definition.

Of course, the same procedure yields a continuous extension $\overline{h^{-1}}\colon \mathbb{R}^2 \to \mathbb{R}^2$ of the map $h^{-1}\colon \mathbb{R}^2 \setminus B \to \mathbb{R}^2 \setminus A$. Since $\mathbb{R}^2 \setminus A$ and $\mathbb{R}^2 \setminus B$ are dense in $\mathbb{R}^2$, the equalities $h \circ h^{-1} = \mathrm{Id}_{\mathbb{R}^2 \setminus B}$ and $h^{-1} \circ h = \mathrm{Id}_{\mathbb{R}^2 \setminus A}$ extend to $\bar{h} \circ \overline{h^{-1}} = \overline{h^{-1}} \circ \bar{h} = \mathrm{Id}_{\mathbb{R}^2}$. This shows that $\bar{h}$ is a homeomorphism. □

Lemma 3.20 *For $i = 1, 2$, the set L_i is totally discontinuous in $\mathcal{O}_i \simeq \mathbb{R}^2$.*

Let us introduce some terminology that will be used in the proof of Lemma 3.20. By a *local section* of a vector field Z on a three-manifold P, we mean a compact surface with boundary embedded in P and transverse to Z. A (Z^t)–invariant set $\Omega \subset P$ is said to be *transversally totally discontinuous* if $\Omega \cap \Sigma$ is totally discontinuous for every local section Σ of Z.

Proof By our assumptions, the maximal invariant set Λ_X of the hyperbolic plug (U, X) contains neither attractors nor repellers. Since Λ_X is a hyperbolic set, it follows that Λ_X is transversally totally discontinuous. Hence, the projection Λ_i of Λ_X in the manifold M_i is also transversally totally discontinuous (recall that Λ_X sits in the interior of U and that the projection $p_i\colon U \to M_i$ is a homeomorphism in restriction to the interior of U). As a further consequence, the complete lift $\widetilde{\Lambda}_i$ of Λ_i in the universal cover $\widetilde{M}_i$ is also transversally totally discontinuous.

Now recall that $(\widetilde{M}_i, \widetilde{Y}_i)$ is topologically equivalent to $\mathbb{R}^3$ equipped with the trivial vertical unit vector field. As a consequence, for every point $x \in \widetilde{M}_i$, we can find a local section Σ of $\widetilde{Y}_i$ such that $x \in \Sigma$ and no orbit of $\widetilde{Y}_i$ intersects Σ twice. This implies that the restriction to Σ of the projection $\mathrm{pr}\colon \widetilde{M}_i \to O_i$ is one-to-one, and hence a homeomorphism onto its image. Since $\widetilde{\Lambda}_i$ is transversally totally discontinuous, it follows that the set $L_i = \mathrm{pr}(\widetilde{\Lambda}_i)$ is totally discontinuous in O_i. □

Lemma 3.21 *For every compact set $K \subset O_1 \simeq \mathbb{R}^2$, the set $\eta(K \setminus L_1)$ has compact closure in $O_2 \simeq \mathbb{R}^2$.*

Proof For $i = 1, 2$, the surface $O_i \setminus L_i$ has infinitely many ends. One of them is the end of $O_i \simeq \mathbb{R}^2$, which we denote by ∞_i. The other ends are in one-to-one correspondence with the points of L_i (since L_i is totally discontinuous). Proving Lemma 3.21 is equivalent to proving that the homeomorphism $\eta\colon O_1 \setminus L_1 \to O_2 \setminus L_2$ maps the end ∞_1 to the end ∞_2.

From the viewpoint of the topology of the surface $O_i \setminus L_i$, nothing distinguishes ∞_i from the other ends. Hence, we need to introduce some dynamical invariants to prove that η necessarily maps ∞_1 to ∞_2.

For $i = 1, 2$, the foliation $\widetilde{\mathcal{F}}^s_i$ induces a one-dimensional foliation g^s_i on the space O_i. We denote by $g^s_{i,0}$ the restriction of the foliation g^s_i to $O_i \setminus L_i$. According to Lemma 3.17, $g^s_{i,0}$ can be obtained as the projection on O_i of the foliation $\widetilde{\mathcal{F}}^s_i \cap \widetilde{S}_i = \widetilde{F}^s_i$. As a consequence, η maps the foliation $g^s_{1,0}$ to the foliation $g^s_{2,0}$.

Since O_i is a plane, every leaf of the foliation g_i^s is a properly embedded line, going from ∞_i to ∞_i (recall that ∞_i is the unique end of O_i). The leaves of $g_i^s = (\mathrm{pr}_i)_* \widetilde{\mathcal{F}}_i^s$ that intersect $L_i = \mathrm{pr}_i(\Lambda_i)$ are the projections of the leaves of the lamination $W^s(\widetilde{\Lambda}_i)$. In particular, there exist leaves of g_i^s that do not intersect L_i. As a consequence, there exist leaves of $g_{i,0}^s$ going from ∞_i to ∞_i. On the other hand, if x is an end of $O_i \setminus L_i$ corresponding to a point of L_i, then there does not exist any leaf of $g_{i,0}^s$ going from x to x (because every leaf ℓ of $g_{i,0}^s$ is a connected component of $\widehat{\ell} \setminus L_i$, where $\widehat{\ell}$ is a line in O_i going from ∞_i to ∞_i). So the foliation $g_{i,0}^s$ allows us to distinguish ∞_i from the other ends of $O_i \setminus L_i$. Since η maps $g_{1,0}^s$ to $g_{2,0}^s$, it follows that η must map ∞_1 to ∞_2. Since ∞_i is the unique end of O_i, this exactly means that, for a compact set $K \subset O_1 \simeq \mathbb{R}^2$, the set $\eta(K \setminus L_1)$ has compact closure in $O_2 \simeq \mathbb{R}^2$. □

Proof of Proposition 3.18 Lemmas 3.20 and 3.21, together with the fact that O_1 and O_2 are homeomorphic to $\mathbb{R}^2$, show that we are exactly in the situation of Lemma 3.19. Applying this lemma, we get a homeomorphism $\overline{\delta}\colon O_1 \to O_2$ extending η. The equivariance of $\overline{\eta}$ follows from that of δ, by continuity and by density of $O_i \setminus L_i$ in O_i. □

We will now conclude the proof of Theorem 1.2 by using a result of Barbot.

Theorem 3.22 (see [2, Theorem 3.4] or [3, proposition 1.36 and corollaire 1.42]) *Two transitive Anosov flows are topologically equivalent if and only if there exists a homeomorphism between their orbit spaces which is equivariant with respect to the actions of the fundamental groups and which does not exchange the stable/unstable directions.*

Proof of Theorem 1.2 The theorem is an immediate consequence of Proposition 3.18 and Theorem 3.22. □

Acknowledgements Yu is supported by National Program for Support of Top-notch Young Professionals, the Fundamental Research Funds for the Central Universities and Shanghai Pilot Program for Basic Research.

References

[1] **M Aschenbrenner**, **S Friedl**, **H Wilton**, 3*–Manifold groups*, Eur. Math. Soc., Zürich (2015) MR Zbl

[2] **T Barbot**, *Generalizations of the Bonatti–Langevin example of Anosov flow and their classification up to topological equivalence*, Comm. Anal. Geom. 6 (1998) 749–798 MR Zbl

[3] **T Barbot**, *De l'hyperbolique au globalement hyperbolique*, habilitation, Université Claude Bernard – Lyon I (2005) `https://tel.archives-ouvertes.fr/tel-00011278`

[4] **T Barthelmé**, *Anosov flows in 3–manifolds*, lecture notes (2017) `https://sites.google.com/site/thomasbarthelme/research`

[5] **F Béguin**, **C Bonatti**, **B Yu**, *Building Anosov flows on 3–manifolds*, Geom. Topol. 21 (2017) 1837–1930 MR Zbl

[6] **F Béguin**, **B Yu**, *Coding periodic orbits for Anosov flows obtained by gluing hyperbolic plugs and applications*, in preparation

[7] **C Bonatti**, **R Langevin**, *Un exemple de flot d'Anosov transitif transverse à un tore et non conjugué à une suspension*, Ergodic Theory Dynam. Systems 14 (1994) 633–643 MR Zbl

[8] **S R Fenley**, *Quasigeodesic Anosov flows and homotopic properties of flow lines*, J. Differential Geom. 41 (1995) 479–514 MR Zbl

[9] **E Ghys**, *Flots d'Anosov sur les* 3*–variétés fibrées en cercles*, Ergodic Theory Dynam. Systems 4 (1984) 67–80 MR Zbl

LAGA, UMR 7539 du CNRS, Université Sorbonne Paris Nord
Villetaneuse, France

School of Mathematical Sciences, Tongji University
Shanghai, China

`beguin@math.univ-paris13.fr`, `binyu1980@gmail.com`

Received: 13 January 2021 Revised: 3 December 2021

Algebraic & Geometric Topology **23**:6 (2023) 2715–2733
DOI: 10.2140/agt.2023.23.2715
Published: 7 September 2023

Ribbon 2–knot groups of Coxeter type

JENS HARLANDER
STEPHAN ROSEBROCK

Wirtinger presentations of deficiency 1 appear in the context of knots, long virtual knots, and ribbon 2–knots. They are encoded by labeled oriented trees and, for that reason, are also called LOT presentations. These presentations are a well known and important testing ground for the validity (or failure) of Whitehead's asphericity conjecture. We define LOTs of Coxeter type and show that for every given n there exists a prime LOT of Coxeter type with group of rank n. We also show that label separated Coxeter LOTs are aspherical.

20F05, 20F06, 20F65; 57K20, 57K45

Dedicated to the memory of Stephen Pride

1 Introduction

Wirtinger presentations of deficiency 1 appear in the context of knots, long virtual knots, and ribbon 2–knots; see Harlander and Rosebrock [9]. They are encoded by labeled oriented trees and, for that reason, are also called LOT presentations. Adding a generator to the set of relators in a Wirtinger presentation P gives a balanced presentation of the trivial group. Thus the associated 2–complex $K(P)$ is a subcomplex of an aspherical (in fact contractible) 2–complex. Wirtinger presentations are a well-known and important testing ground for the validity (or failure) of Whitehead's asphericity conjecture, which states that a subcomplex of an aspherical 2–complex is aspherical. For more on the Whitehead conjecture see Bogley [3], Berrick and Hillman [1] and Rosebrock [18].

If P is a Wirtinger presentation and the group $G(P)$ defined by P is a 1–relator group, then $G(P)$ admits a 2–generator 1–relator presentation P' and the corresponding 2–complex $K(P')$ is aspherical. Since $K(P')$ and $K(P)$ have the same Euler characteristic and the same fundamental group, it follows (using Schanuel's lemma and Kaplansky's theorem, which states that finitely generated free $\mathbb{Z}G$–modules are Hopfian) that $K(P)$

is also aspherical. Thus, when investigating the asphericity of $K(P)$ for a given Wirtinger presentation P, the first thing to ask is if $G(P)$ is a 1–relator group.

Composite knot groups require more than two generators; see Norwood [17].[1] However, many knots have 2–generator 1–relator knot groups. Prime knots whose groups need more than two generators were known to Crowell and Fox in 1963. As one example, Crowell and Fox consider a certain prime 9–crossing knot, show that its Wirtinger presentation simplifies to

$$P = \langle x, y, z \mid y^{-1}xyx^{-1}y = x^{-1}zx^{-1}zxz^{-1}x,\ x^{-1}zxz^{-1}x = y^{-1}zyz^{-1}y\rangle,$$

and show that the length of the chain of elementary ideals for this knot group is 2. It follows that the rank (the minimal number of generators) of $G(P)$ is greater than 2 and therefore equal to 3. This can also be seen without the use of elementary ideals. We have an epimorphism

$$G(P) \to \Delta(3,3,3) = \langle x, y, z \mid x^2, y^2, z^2, (xy)^3, (xz)^3, (yz)^3\rangle$$

sending $x \to x$, $y \to y$ and $z \to z$. Since the rank of the Euclidean triangle group $\Delta(3,3,3)$ is 3 (see Klimenko and Sakuma [13]), we have $\operatorname{rank}(G(P)) = 3$.

This example motivates the first part of this article. It is easier to construct high-rank ribbon 2–knot groups than classical knot groups, because we do not have to verify that a given Wirtinger presentation can be read off a knot projection (a 4–regular planar graph). Below we define labeled oriented trees of Coxeter type and show that, given a Coxeter group W which abelianizes to $W_{\mathrm{ab}} = \mathbb{Z}_2$, there exists a Coxeter-type LOT group that maps onto W. Using this we give examples of prime LOT groups of arbitrarily high rank.

In the second part of the paper we investigate the question of asphericity of LOTs of Coxeter type. We show that label-separated LOTs of Coxeter type are aspherical. It turns out that the study of asphericity can be translated into questions concerning free subgroups of 1–relator LOT groups of dihedral type.

2 Groups defined by graphs

A *labeled oriented graph* (LOG) is an oriented finite graph Γ on vertices $\boldsymbol{x}$ and edges $\boldsymbol{e}$, where each oriented edge is labeled by a word in $\boldsymbol{x}^{\pm 1}$. Associated with a LOG Γ is

[1] The central Lemma 3 in Norwood's paper has a gap which can be filled, as was pointed out by Menasco and Reid [15, page 223, Remark 4] and also Bleiler [2].

the presentation

$$P(\Gamma) = \langle \boldsymbol{x} \mid \boldsymbol{r} = \{r_e \mid e \in \boldsymbol{e}\}\rangle,$$

where $r_e = xw(wy)^{-1}$ when $e = (x \xrightarrow{w} y)$ is the edge of Γ starting at x, ending at y, and labeled with the word w on letters in $\boldsymbol{x}^{\pm 1}$. We remark that what we call a labeled oriented graph is elsewhere called a *weakly labeled oriented graph* or *word-labeled oriented graph*. See Howie [12] and Harlander and Rosebrock [10].

Denote by $K(\Gamma)$ and $G(\Gamma)$ the standard 2–complex and the group defined by $P(\Gamma)$, respectively. The case where Γ is a tree, now called a *labeled oriented tree* (LOT), is special. It is known that the groups $G(\Gamma)$ where Γ is a LOT are precisely the ribbon 2–knot groups (see Yajima [22], Howie [12] and also Hillman [11, Section 1.7]), since, in that case, $G(\Gamma)$ is a group of weight 1 (normally generated by a single element, in fact by each generator) that has a deficiency 1 presentation $P(\Gamma)$. The 2–complexes $K(\Gamma)$ with Γ a LOT are of central importance to Whitehead's asphericity conjecture, since adding a generator to the set of relators in $P(\Gamma)$ gives a balanced presentation of the trivial group. So $K(\Gamma)$ is a subcomplex of a 2–dimensional contractible complex. A question that has been open for a long time asks if $K(\Gamma)$ is aspherical, ie $\pi_2(K(\Gamma)) = 0$. See Bogley [3], Berrick and Hillman [1] and Rosebrock [18].

A subtree $\Gamma_0 \subseteq \Gamma$ of a LOT is a *sub-LOT* if the label w of an edge in Γ_0 is a word in the vertices of Γ_0. A sub-LOT $\Gamma_0 \subseteq \Gamma$ is called *proper* if it has more than one vertex and is not all of Γ. A LOT is called *prime* if it does not contain proper sub-LOTs.

Let Υ be a simplicial graph on vertices $\boldsymbol{x}$, and suppose edges e are labeled with integers $m_e \geq 2$. Define

$$P(\Upsilon) = \langle \boldsymbol{x} \mid x^2 \text{ for } x \in \boldsymbol{x},\ (xy)^{m_e} \text{ if } e = \{x, y\} \text{ is an edge}\rangle.$$

The group $W = W(\Upsilon)$ defined by this presentation is called a *Coxeter group*. We refer to Υ as the *defining graph* for the Coxeter group. We remark that the graph Υ shows up in Davis [6, Example 7.1.6] (the Coxeter system associated to a labeled simplicial graph). It should not be confused with the Dynkin diagram, another labeled graph that appears in connection with Coxeter groups. Let $K = K(\Upsilon)$ be the 2–complex associated with $P(\Upsilon)$. Consider the universal covering $\widetilde{K}(\Upsilon)$. The 1–skeleton of $\widetilde{K}(\Upsilon)$ is the Cayley graph for $(W, \boldsymbol{x})$. All edges in $\widetilde{K}(\Upsilon)$ are double edges: for every $g \in W$ and $x \in \boldsymbol{x}$, we have an edge (g, x) connecting g to gx, and an edge (gx, x) connecting gx to $gxx = g$. Note that a double edge pair bounds two 2–cells in $\widetilde{K}(\Upsilon)$, coming from the relator x^2. We remove one and collapse the other one to an edge. This turns each double edge into a single unoriented edge. Every relator $(xy)^{m_e}$ gives rise to

$2m_e$ 2–cells with the same boundary. We remove all but one from this set. We denote the 2–complex obtained in this fashion by $\Sigma^{(2)}(\Upsilon)$. It is the 2–skeleton of the Davis complex $\Sigma(\Upsilon)$. See [6, Proposition 7.3.4]. We remark that the Davis complex is closely related to the Coxeter complex, but the complexes are not the same. For the definition of Coxeter complex, see [6, Example 5.2.7]. Under certain conditions, for example when the defining graph Υ is a tree, the Davis complex is 2–dimensional: $\Sigma(\Upsilon) = \Sigma^{(2)}(\Upsilon)$. See [6, Example 7.4.2].

Proposition 2.1 *Let Υ be a defining tree with associated Coxeter group $W(\Upsilon)$. Then:*

1. *For every edge $e = \{x, y\}$ of Υ we have a 2–cell κ_e in $\Sigma(\Upsilon)$ attached along a $2m_e$–gon whose edge labels read $(xy)^{m_e}$.*
2. *$\Sigma(\Upsilon)$ is the union of the 2–cells $w\kappa_e$ for $e \in \{\text{edges of}\,\Upsilon\}$ and $w \in W(\Upsilon)$. Furthermore, if $w_1\kappa_{e_1} \cap w_2\kappa_{e_2} \neq \varnothing$, then $e_1 \cap e_2 \neq \varnothing$; if $x = e_1 \cap e_2$, then the edge $w_1\kappa_{e_1} \cap w_2\kappa_{e_2}$ carries the label x.*
3. *$\Sigma(\Upsilon)$ is a tree of 2–cells: if we connect the barycenters of the 2–cells with the barycenters of their boundary edges, we obtain a tree. In particular, if M is a finite connected union of Coxeter 2–cells $w_i\kappa_{e_i}$ in $\Sigma(\Upsilon)$, then there exists a 2–cell $w\kappa_e$ in M that intersects with the rest of M in a single edge.*

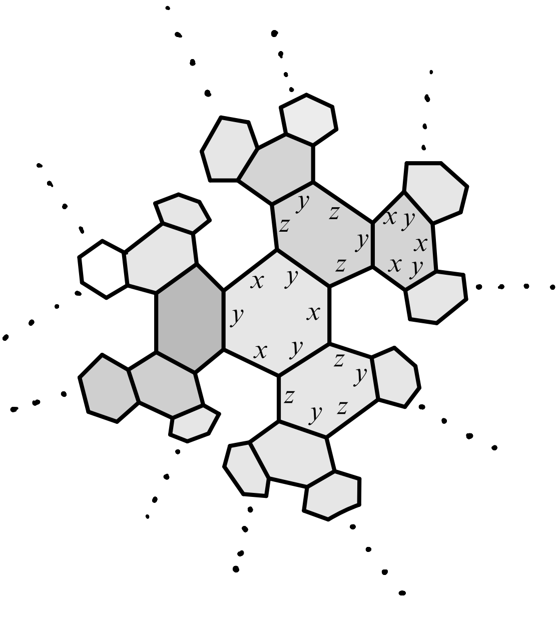


Figure 1: The Davis complex $\Sigma(\Upsilon)$ for $\Upsilon = x \overset{3}{—} y \overset{3}{—} z$. It is a tree of Coxeter cells.

Proof The statements (1) and (2) are clear from the construction of Υ. For an edge $e=\{x,y\}$, let $P(e)=\langle x,y \mid x^2,y^2,(xy)^{m_e}\rangle$. Let D_{m_e} be the dihedral group defined by $P(e)$. Since Υ is a tree, $W(\Upsilon)$ is an amalgamated product of the D_{m_e}. The associated Bass–Serre tree can be seen inside the Davis complex $\Sigma(\Upsilon)$. The vertices of that tree are the barycenters of the 2–cells and 1–cells, and the edges connect barycenters of 2–cells to the barycenters of the 1–cells in the boundary of that 2–cell. We can think of $\Sigma(\Upsilon)$ as a tree of Coxeter 2–cells. An example is shown in Figure 1.

Suppose $M=\bigcup_{i=0}^{k} D_i$ is a union of 2–cells. Let d_i be the barycenter of D_i. Let d_p be a vertex in the Bass–Serre tree furthest away from d_0 with $p\in\{0,\dots,k\}$. Consider a geodesic from d_0 to d_p and let d_q be the barycenter that is encountered just before getting to d_p when traveling along the geodesic. Then $\bigl(\bigcup_{i\neq p} D_i\bigr)\cap D_p = D_q\cap D_p$, which is a single edge. □

Lemma 2.2 *Let Γ be a LOT $e=(x\xrightarrow{w}y)$ an edge such that the word w contains letters and $z\neq x,y$ with even (positive or negative) exponent only. Then the relator $r_e=xw(wy)^{-1}$ reduces (up to cyclic permutation) to $\bar r_e=(yx)^{m_e}$, with $m_e\geq 1$ and odd, in $\langle \boldsymbol{x} \mid x^2$ for $x\in\boldsymbol{x}\rangle$.*

Proof The word w reduces to an alternating word $\overline{w}$ in the letters x and y. If $\overline{w}$ is the empty word, then $\bar r_e=xy$. There are four remaining cases to consider:

(1) $\overline{w}$ starts with x and has even length.

(2) $\overline{w}$ starts with x and has odd length.

(3) $\overline{w}$ starts with y and has even length.

(4) $\overline{w}$ starts with y and has odd length.

In case (1) we have $\overline{w}=xyxy$, say. So $x(xyxy)y(xyxy)=xxyxyyxyxy=xy$. In case (2) we have $\overline{w}=xyxyx$, say. So $x(xyxyx)y(xyxyx)=xxyxyxyxyxyx=(yx)^5$. In case (3) we have $\overline{w}=yxyx$, say. So $x(yxyx)y(yxyx)=xy$. In case (4) we have $\overline{w}=yxyxy$, say. So $x(yxyxy)y(yxyxy)=(xy)^5$. □

Definition 2.3 Let Γ be a LOT with vertex set $\boldsymbol{x}$. We say Γ is of *Coxeter type* if:

(1) For every edge $e=(x\xrightarrow{w}y)$, the word w contains letters $z\neq x,y$ with even (positive or negative) exponent only.

(2) For every edge $e=(x\xrightarrow{w}y)$, the relator $r_e=xw(wy)^{-1}$ reduces (up to cyclic permutation) to $\bar r_e=(yx)^{m_e}$, with $m_e\geq 2$, in $\langle \boldsymbol{x}\mid x^2$ for $x\in\boldsymbol{x}\rangle$.

Remark 2.4 Lemma 2.2 shows that, if Γ is a LOT of Coxeter type, then, for each edge e, $m_e \geq 3$ and is odd.

Let Γ be a LOT of Coxeter type. Define a tree Υ in the following way: erase orientations in Γ and, if $e = (x \xrightarrow{w} y)$ is an edge and the LOT relator r_e reduces to $\bar{r}_e = (yx)^{m_e}$ (up to cyclic permutation) in $\langle \boldsymbol{x} \mid x^2 \text{ for } x \in \boldsymbol{x} \rangle$, then label the (unoriented) edge e by m_e. Note that Υ is a defining tree for a Coxeter group. We have a map $P(\Gamma) \to P(\Upsilon)$ sending x to x which induces a group epimorphism $G(\Gamma) \to W(\Upsilon)$. This process can be reversed.

Lemma 2.5 *Let Υ be a defining tree for a Coxeter group where all m_e are odd. Then there exists a LOT of Coxeter type Γ such that the process described above produces Υ from Γ. In particular, $G(\Gamma)$ maps onto $W(\Upsilon)$.*

Proof Suppose $e = \{x, y\}$ is an edge in Υ. Orient it from x to y. Let $w = (yx)^{(m_e-1)/2}$. Let $e = (x \xrightarrow{w} y)$ be the corresponding edge in Γ. □

Note that the LOT Γ of Coxeter type constructed in the lemma is not prime. In fact, every edge is a sub-LOT. Note also that $G(\Gamma)$ is an Artin group. One can show that all Artin groups are LOG groups, but we will not pursue this here. Given a defining tree Υ, there are many LOTs of Coxeter type that give rise to Υ.

Lemma 2.6 *Let Υ be a defining tree where all m_e are odd. Suppose Γ is a LOT of Coxeter type such that the process just described produces Υ. Then there also exists a prime LOT that produces Υ.*

Proof Suppose Γ_0 is a proper sub-LOT of Γ. Let $e = (x \xrightarrow{w} y)$ be an edge in Γ_0 and z be a vertex not in Γ_0. Replace the label w by $z^2 w$ to obtain a new LOT Γ'. Then Γ_0 is not a sub-LOT of Γ', but Γ' still produces Υ. We can apply this procedure until we arrive at a LOT without proper sub-LOTs. □

3 LOT groups of high rank

Given two LOTs Γ_1 and Γ_2 and two valency-one vertices $x_i \in \Gamma_i$ for $i = 1, 2$, one can form a composite LOT $\Gamma = \Gamma_1 \cup_{x_1 = x_2} \Gamma_2$ by identifying the two vertices. The LOT group $G(\Gamma)$ is an amalgam $G(\Gamma_1) *_{\mathbb{Z}} G(\Gamma_2)$, and, avoiding trivial cases, the rank of $G(\Gamma)$ is greater than the rank of each $G(\Gamma_i)$ for $i = 1, 2$. This follows from a theorem of Karras and Solitar. See also [17]. However, Γ is not prime, and it is more difficult to provide lower bounds for the rank of prime LOT groups. This issue is already present in the classical knot world, as was discussed in the introduction. In this section we present a method for constructing prime LOTs with groups of arbitrarily high rank.

Theorem 3.1 (Carette and Weidmann [5]) *Let Υ be a defining graph with n vertices and assume that $m_e \geq 6 \cdot 2^n$ for each e. Then the rank of $W(\Upsilon)$ is n.*

Theorem 3.2 *Let $W = W(\Upsilon)$ be a Coxeter group such that $W_{\mathrm{ab}} = \mathbb{Z}_2$. There exists a prime labeled oriented tree Γ of Coxeter type such that $G = G(\Gamma)$ maps onto W.*

Proof Since $W_{\mathrm{ab}} = \mathbb{Z}_2$, the defining graph Υ is connected. In fact, the subgraph Υ_{odd} consisting of edges with odd label is connected, because an edge with an even label does not contribute a relation in W_{ab}, so $W(\Upsilon_{\mathrm{odd}})_{\mathrm{ab}} = W(\Upsilon)_{\mathrm{ab}}$. Thus Υ contains a maximal tree Υ_0 in which all labels m_e are odd. Then Υ and Υ_0 have the same set of vertices and we have an epimorphism $W(\Upsilon_0) \to W(\Upsilon)$. From Lemmas 2.5 and 2.6, we know that there is a prime LOT Γ of Coxeter type such that $G(\Gamma)$ maps onto $W(\Upsilon_0)$. □

Corollary 3.3 *For any given n there exists a prime labeled oriented tree Γ of Coxeter type with n vertices such that $G(\Gamma)$ has rank n. In particular, if $n \geq 3$, then $G(\Gamma)$ is not a 1–relator group.*

Proof This follows from Theorem 3.2 together with the Carette–Weidmann theorem, Theorem 3.1. □

Example 3.4 Let Γ be the prime LOT $x \xrightarrow{yz^2x} y \xrightarrow{zx^2y} z$. Note that $G(\Gamma)$ maps onto the amalgamated product $D_3 *_{\mathbb{Z}_2} D_3$, which cannot be generated by two elements. Thus the rank of $G(\Gamma)$ is 3 and it follows that this LOT group is not a 1–relator group.

Remark 3.5 If Γ is a LOT of Coxeter type and Υ is the associated defining tree, then $W(\Upsilon)$ is an amalgamated product of dihedral groups. A direct way to obtain upper bounds for the rank of $W(\Upsilon)$ without the full force of Theorem 3.1 is via Weidmann [21].

Remark 3.6 A *reorientation* of a LOT is obtained when changing signs on the exponents of letters that occur in the edge words, which has no effect on the quotient $W(\Upsilon)$. Thus, if $\mathrm{rk}(G(\Gamma)) = \mathrm{rk}(W(\Upsilon))$, then this equation holds also for all reorientations of Γ.

4 Largeness

A group is *large* if it has a subgroup of finite index that has a free quotient of rank ≥ 2. Large groups of deficiency 1 are studied in Button [4]. A list of properties can also be found there. If G is large, then:

(1) G contains free subgroups of rank ≥ 2.

(2) G is SQ-universal (every countable group is the subgroup of some quotient).

(3) G has finite-index subgroups with arbitrarily large first Betti number.

(4) G has uniformly exponential word growth.

(5) G has subgroup growth of strict type n^n (which is the largest possible growth for finitely generated groups).

(6) The word problem for G is solvable strongly generically in linear time.

Theorem 4.1 *Let Γ be a LOT of Coxeter type on at least three vertices. Then $G(\Gamma)$ is large.*

Proof The conditions imply that $W(\Upsilon)$ is an infinite group that is the fundamental group of a finite tree of groups where the vertex groups are either $\mathbb{Z}_2$ or dihedral groups D_m with $m \geq 3$ ($\mathbb{Z}_2$ vertex groups will appear when Υ has vertices of valency ≥ 3). Thus $W(\Upsilon)$ contains a free subgroup F of rank ≥ 2 of finite index (see Serre [20, Proposition 11, page 120)]. Let H be the preimage of F in $G(\Gamma)$. Then H is a subgroup of $G(\Gamma)$ of finite index that maps onto F. It follows that $G(\Gamma)$ is large. □

A characterization of virtual free Coxeter groups is given in Davis [6, Section 8.8]. When Υ is a tree, the characterization implies that $W(\Upsilon)$ is virtually free. This provides another proof for Theorem 4.1.

Example 4.2 As in Example 3.4 let Γ be the prime LOT $x \xrightarrow{yz^2x} y \xrightarrow{zx^2y} z$. We have $W(\Upsilon) = D_3 *_{\mathbb{Z}_2} D_3$. Let $\Delta(3,3,2)$ be the spherical triangle group (it is the symmetric group S_4) defined by $\langle x, y, z \mid x^2, y^2, z^2, (xy)^3, (yz)^3, (xz)^2\rangle$. We have an epimorphism $W(\Upsilon) \to \Delta(3,3,2)$ and we claim that the kernel V is free of rank ≥ 2. Indeed, since both D_3's of $W(\Upsilon)$ are also subgroups of $\Delta(3,3,2)$, it follows that V intersects both D_3's trivially and thus V acts freely on the Bass–Serre tree T for $W(\Upsilon) = D_3 *_{\mathbb{Z}_2} D_3$, and hence is free. Note that the valency of every vertex in T is equal to 3 (since the index of $\mathbb{Z}_2$ in the D_3's is 3), and so V cannot be cyclic. Here is why: Note that $V=\pi_1(X)$, where $X=T/V$ is a finite graph in which every vertex has valency 3. Let $v(X)$ and $e(X)$ denote the number of vertices and edges, respectively. We have $v(X)=\frac{2}{3}e(X)$ and we obtain $\chi(X)=v(X)-e(X)=\frac{2}{3}e(X)-e(X)<0$. Thus $\dim H_0(X)-\dim H_1(X)=1-\dim H_1(X)=\chi(X)<0$. So $\dim H_1(X)>1$ and hence $\dim V_{\mathrm{ab}}>1$. One can also check directly that $(xz)^2$ and $x(xz)^2x^{-1}=(zx)^2$ generate a free subgroup of V of rank 2.

5 The question of asphericity

Let Γ be a labeled oriented tree of Coxeter type and let Υ be the associated defining tree for the Coxeter group $W(\Upsilon)$. Let $\overline{K}(\Gamma)$ be the normal covering space with fundamental group the kernel of the epimorphism $G(\Gamma)\to W(\Upsilon)$. We will analyze the structure of $\overline{K}(\Gamma)$. We have maps

$$\overline{K}(\Gamma)\to\widetilde{K}(\Upsilon)\to\Sigma(\Upsilon),$$

and note that $\overline{K}(\Gamma)$ and $\widetilde{K}(\Upsilon)$ have the same 1–skeleton. Let $e=(x\xrightarrow{w}y)$ be an edge in Γ. Let $P_e=\langle \boldsymbol{x}_e \mid r_e\rangle$, where $\boldsymbol{x}_e\subseteq\boldsymbol{x}$ is the subset of the vertices of Γ that occur in r_e. Let $\boldsymbol{z}=\boldsymbol{x}_e-\{x,y\}$. Then $P_e=\langle x,y,\boldsymbol{z}\mid xw=wy\rangle$. The complex $K(P_e)$ is a subcomplex of $K(\Gamma)$. Consider the preimage of $K(P_e)$ under the covering projection $\overline{K}(\Gamma)\to K(\Gamma)$. It is a union of finite subcomplexes $w\overline{K}_e$ for $w\in W(\Upsilon)$, which we will now describe in detail. The 1–skeleton of $\overline{K}_e$ is a $2m_e$–gon with double edges labeled in an alternating way by x and y. At each of the $2m_e$ vertices we have a double edge for every $z\in\boldsymbol{z}$. The situation is depicted in Figure 2. We have $2m_e$ 2–cells, attached along the loop with label r_e, starting at every vertex. The dihedral group D_{m_e}, the stabilizer of the cell κ_e in $\Sigma(\Upsilon)$, acts freely on $\overline{K}_e$. It is convenient to replace $\overline{K}_e$ by a complex with a single D_{m_e} orbit of vertices. Let $\overline{L}_e$ be the 2–complex obtained

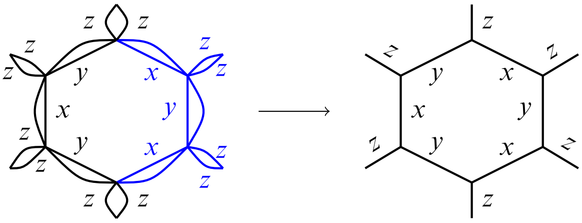


Figure 2: The complex $\overline{K}_e$ (on the left) in the case $e=(x\xrightarrow{w}y)\in\Gamma$ with corresponding edge $e=(x\xrightarrow{3}y)\in\Upsilon$, so the Coxeter relator is $(xy)^3$. On the right is the corresponding Coxeter cell κ_e together with z–edges. The blue part is a y–side in $\overline{K}_e$.

from $\overline{K}_e$ in the following way: at every vertex collapse one of the z–edges from the z–double edge for some $z \in \boldsymbol{z}$. The complex $\overline{L}_e$ is homotopy equivalent to $\overline{K}_e$. The 1–skeleton of $\overline{L}_e$ is a $2m_e$–gon with double edges labeled in an alternating way by x and y. At each of the $2m_e$ vertices we have a loop for every $z \in \boldsymbol{z}$. Let $\hat{r}_e$ be the word obtained from r_e by replacing every z^p for $z \in \boldsymbol{z}$ by $z^{p/2}$. Let $\widehat{P}_e = \langle x, y, \boldsymbol{z} \mid \hat{r}_e \rangle$. Note that the dihedral group D_{m_e} acts freely on $\overline{L}_e$ and we have a covering map $\overline{L}_e \to \overline{L}_e / D_{m_e} = K(\widehat{P}_e)$.

Lemma 5.1 *The* 2*–complex* $\overline{K}_e$ *is aspherical.*

Proof The complex $K(\widehat{P}_e)$ is aspherical because $\widehat{P}_e$ is a 1–relator presentation for which the relator is not a proper power. Thus $\overline{L}_e$ is aspherical, being a covering space of $K(\widehat{P}_e)$. Since $\overline{K}_e$ is homotopy equivalent to $\overline{L}_e$, it follows that $\overline{K}_e$ is aspherical. □

An *x–side* of $\overline{K}_e$ consists of a double edge with label x together with all the double edges connected to the two vertices of the x–double edge. A y–side is defined in the same way. See Figure 2, where the blue part on the left shows a y–side. Note that $\overline{K}_e$ has m_e x–sides and m_e y–sides. We refer to these as the *sides* of $\overline{K}_e$. We say $\overline{K}_e$ is *side injective* if the inclusion induced map $\pi_1(S) \to \pi_1(\overline{K}_e)$ is injective for every side S. An x–side in $\overline{L}_e$ is the image of an x–side under $\overline{K}_e \to \overline{L}_e$, etc.

Lemma 5.2 *The* 2*–complex* $\overline{K}_e$ *is x–side injective if and only if*

$$\langle x^2, y^2, z, xy^2x^{-1}, xzx^{-1} : z \in \boldsymbol{z} \rangle$$

is a free subgroup of $G(\widehat{P}_e)$ *on the given basis.*

Proof Recall that $m_e \geq 3$. An x–side S in $\overline{L}_e$ is an x–double edge, a y–double edge at each of the two vertices, and a loop for every $z \in \boldsymbol{z}$ at each of the two vertices. The image of $\pi_1(S)$ in $G(\widehat{P}_e)$ under the covering projection is the group in the statement of the lemma. □

Lemma 5.3 *If T is a subgraph of the* 1*–skeleton of $\overline{K}_e$ that does not involve every letter from $\boldsymbol{x}_e = \{x, y, \boldsymbol{z}\}$, then $\pi_1(T) \to \pi_1(\overline{K}_e)$ is injective.*

Proof We can argue with $\overline{L}_e$ instead of $\overline{K}_e$. A reduced loop γ in T gives a reduced word u in the generators of $\widehat{P}_e$ that does not involve all letters from $\boldsymbol{x}_e = \{x, y, \boldsymbol{z}\}$. The presentation $\widehat{P}_e$ has only one relator $\hat{r}_e$ that does involve all letters from the generating set $\boldsymbol{x}_e = \{x, y, \boldsymbol{z}\}$. The Freiheitssatz for 1–relator groups implies that u does not represent the trivial element in $G(\widehat{P})$. Thus γ is not trivial in $\pi_1(\overline{L}_e)$. □

We continue our analysis. The complex $\overline{K}(\Gamma)$ is a union of the complexes $w\overline{K}_e$ for $w \in W(\Upsilon)$ and $e \in$ edges of Γ. The maps

$$\overline{K}(\Gamma) \to \widetilde{K}(\Upsilon) \to \Sigma(\Upsilon)$$

give a one-to-one correspondence between the $w\overline{K}_e$ and Coxeter cells $w\kappa_e$. Since Υ is a tree, the Davis complex $\Sigma(\Upsilon)$ is a tree of Coxeter cells $w\kappa_e$ and so $\overline{K}(\Gamma)$ is a tree of complexes $w\overline{K}_e$. In complete analogy to Proposition 2.1, we have:

Proposition 5.4 *Consider $\overline{K}(\Gamma) = \bigcup w\overline{K}_e \to \Sigma(\Upsilon) = \bigcup w\kappa_e$.*

1. *$\overline{K}(\Gamma)$ is the union of the 2–complexes $w\overline{K}_e$ for $e \in \{$edges of $\Gamma\}$ and $w \in W(\Upsilon)$. Furthermore, if $w_1\overline{K}_{e_1} \cap w_2\overline{K}_{e_2} \neq \varnothing$, then $e_1 \cap e_2 \neq \varnothing$; if $x = e_1 \cap e_2$, then $w_1\overline{K}_{e_1} \cap w_2\overline{K}_{e_2} = T$, where T is the subgraph of an x–side S that carries the letters $\boldsymbol{x}_{e_1} \cap \boldsymbol{x}_{e_2}$.*
2. *$\overline{K}(\Gamma)$ is a tree of 2–complexes. In particular, if $\overline{M}$ is a finite connected union of 2–complexes $w_i\overline{K}_{e_i}$ in $\overline{K}(\Gamma)$, then there exists a 2–complex $w\overline{K}_e$ in $\overline{M}$ that intersects with the rest of $\overline{M}$ in a subgraph of a single side.*

Theorem 5.5 *Let Γ be a LOT of Coxeter type. Then $K(\Gamma)$ is aspherical if the $\overline{K}_e$ are side injective for every edge e in Γ.*

Proof We will show that $\overline{K}(\Gamma)$ is aspherical. It suffices to show that every finite connected union $\overline{M} = \bigcup_{i=1}^{n} w_i\overline{K}_{e_i}$ is aspherical. We first claim that the sides of the $w_i\overline{K}_{e_i}$ π_1–inject into the union $\overline{M}$. We do induction on n. If $n = 1$, the result follows from the hypothesis. Assume $n > 1$. Then, by Proposition 5.4(2), there exists a 2–complex $w\overline{K}_e$ in $\overline{M}$ that intersects with the rest of $\overline{M}$ in a subgraph T of a single side S (of course, T could be S). Now, by the induction hypothesis, the inclusion $S \subseteq \overline{M} - w\overline{K}_e = \overline{M}_0$ is π_1–injective, and the inclusion $S \subseteq w\overline{K}_e$ is π_1–injective by hypothesis. It follows that $\pi_1(\overline{M})$ is an amalgamated product $\pi_1(\overline{M}) = \pi_1(\overline{M}_0) *_{\pi_1(T)} \pi_1(w\overline{K}_e)$. Thus the inclusion $S \subseteq \overline{M}$ is π_1–injective. All other sides that occur in $\overline{M}$ are contained in either $\overline{M}_0$ or $w\overline{K}_e$. π_1–injectivity follows from the amalgamated product decomposition. Asphericity of $\overline{M}$ now follows from induction on n and the amalgamated product decomposition $\pi_1(\overline{M}) = \pi_1(\overline{M}_0) *_{\pi_1(T)} \pi_1(w\overline{K}_e)$. □

Remark The above proof shows more than asphericity. Since each $\pi_1(\overline{K}_e)$ is a finite-index subgroup of a 1–relator group, we see that $\pi_1(\overline{K})$ is a tree of groups, the vertex groups being finite-index subgroups of 1–relator groups, and the edge groups (over which we amalgamate) being finitely generated and free.

Definition 5.6 A labeled oriented tree Γ is called *label separated* if, for every pair of edges e_1 and e_2 that have a vertex in common, the intersection $\boldsymbol{x}_{e_1} \cap \boldsymbol{x}_{e_2}$ is a proper subset of both $\boldsymbol{x}_{e_1}$ and $\boldsymbol{x}_{e_2}$.

Theorem 5.7 *Let Γ be a label separated LOT of Coxeter type. Then $K(\Gamma)$ is aspherical.*

Proof The proof is very much the same as the proof of Theorem 5.5. Let $\overline{M} = \bigcup_{i=1}^{n} w_i \overline{K}_{e_i}$as before. Again it suffices to show that $\overline{M}$ is aspherical. If $n = 1$ then the proof is clear. It is instructive to look at the case $n = 2$. The intersection $w_1 \overline{K}_{e_1} \cap w_2 \overline{K}_{e_2} = T$ is the subgraph of a side that carries the letters $\boldsymbol{x}_{e_1} \cap \boldsymbol{x}_{e_2}$, which is a proper subset of both $\boldsymbol{x}_{e_1}$ and $\boldsymbol{x}_{e_2}$. Thus π_1–injectivity for the inclusions $T \subseteq w_i \overline{K}_{e_i}$ for $i = 1, 2$ follows from Lemma 5.3. We have $\pi_1(\overline{M}) = \pi_1(w_1 \overline{K}_{e_1}) *_{\pi_1(T)} \pi_1(w_2 \overline{K}_{e_2})$ and $\overline{M}$ is aspherical. For $n \geq 2$ we argue by induction and obtain (as in the proof of Theorem 5.5) a decomposition $\pi_1(\overline{M}) = \pi_1(\overline{M}_0) *_{\pi_1(T)} \pi_1(w \overline{K}_e)$, which proves asphericity of $\overline{M}$. □

6 Side injectivity

Let $P = \langle a, b, \boldsymbol{c} \mid r \rangle$, be a 1–relator group, where $\boldsymbol{c}$ is a finite set of letters (which could be empty). We assume that r is cyclically reduced and contains all generators. Assume further that $r = (ab)^m$ for some $m \geq 0$ modulo the relations $a^2 = b^2 = c = 1$ for $c \in \boldsymbol{c}$ and cyclic permutation. The number m is called *the dihedral type* of P.

Let $Q = \langle a, b, \boldsymbol{c} \mid (ab)^m, a^2, b^2 \text{ for } c \in \boldsymbol{c} \rangle$. We have an epimorphism $\phi\colon G(P) \to G(Q) = D_m$. Let $\overline{K}(P)$ be the covering of $K(P)$ associated with the kernel. Note that $\overline{K}(P)^{(1)} = \widetilde{K}(Q)^{(1)}$, which is the Cayley graph for D_m on the generating set $\{a, b, \boldsymbol{c}\}$. So $\overline{K}(P)^{(1)}$ is a $2m$–gon, consisting of double edges labeled in an alternating way with a and b, and at every vertex we have a c–loop for every $c \in \boldsymbol{c}$. An a–side of $\overline{K}(P)$ is a connected subgraph of the 1–skeleton that consists of a double edge with label a, together with all the b–double edges and c–loops connected to the two vertices of the a–double edge. A b–side is defined in an analogous way. We say P is *side injective* if the inclusion of any side $S \to \overline{K}(P)$ is π_1–injective.

Lemma 6.1 *Assume that P is of dihedral type $m \geq 3$. Suppose that, for every cyclically reduced word w in $\{a, b, \boldsymbol{c}\}^{\pm 1}$ which represents the trivial element in $G(P)$, some cyclic permutation of w contains a reduced subword u of the form*

$$a^{\pm 1} d_1 b^{\beta} d_2 a^{\alpha} d_3 b^{\pm 1} \quad \text{or} \quad b^{\pm 1} d_1 a^{\alpha} d_2 b^{\beta} d_3 a^{\pm 1},$$

where α and β are odd integers and the d_i are words in the generators containing a and b with even exponents (the d_i could be trivial). *Then P is side injective.*

Proof We begin with some notation. If w is a word in $\{a,b,\boldsymbol{c}\}^{\pm 1}$, then we denote by $\overline{w}$ the element of D_m that it represents. If $w = x_1 \dots x_n$ with $x_i \in \{a,b,\boldsymbol{c}\}^{\pm 1}$, then the lift $\gamma(w,V)$ of w into $\overline{K}(P)^{(1)}$, starting at a vertex V, is a path with vertices $V, \overline{x_1}\,V, \overline{x_1x_2}\,V, \dots, \overline{x_1 \dots x_n}\,V$. Now let w be a reduced word as in the statement. We assume without loss of generality that $w = w_1(a^{\pm 1}d_1b^{\beta}d_2a^{\alpha}d_3b^{\pm 1})w_2$. Consider $\gamma(w,V)$. Let $V' = \overline{w_1}\,V$. Among the vertices of $\gamma(w,V)$ we find V', $\bar{a}V'$, $\overline{ab}\,V'$, $\overline{aba}\,V'$, and $\overline{abab}\,V'$. These are five distinct vertices. A side of $\overline{K}(P)$ contains exactly four vertices. It follows that $\gamma(w,V)$ is not contained in a side. We conclude that P is side injective. □

Example 6.2 $P = \langle a,b \mid (ab)^m\rangle$ for $m \geq 3$ is side injective. This is because 1–relator presentations with torsion are Dehn presentations (in particular, $G(P)$ is hyperbolic). See Newman [16]. A word w that is trivial in the group contains a subword of length more than $\frac{1}{2}$ that of a cyclic permutation of the relator or its inverse; hence, it contains a cyclic permutation of $abab$, or its inverse. The result follows from Lemma 6.1.

Example 6.3 More generally, if $P = \langle a,b,\boldsymbol{c} \mid r(a,b,\boldsymbol{c})\rangle$ ($\boldsymbol{c}$ could be empty) is a Dehn presentation of dihedral type $m \geq 3$ such that more than half of a cyclic permutation of the relator or its inverse contains a subword u as in Lemma 6.1, then P is side injective. Recall that P is a Dehn presentation for instance if it satisfies the small cancellation condition $C'\big(\frac{1}{6}\big)$ or $C'\big(\frac{1}{4}\big) - T(4)$ (see Lyndon and Schupp [14, Chapter V, Theorem 4.4]). For example, if

$$r(a,b,\boldsymbol{c}) = a^{\alpha_1}d_1b^{\beta_1}d_2a^{\alpha_2}d_3b^{\beta_2}d_4a^{\alpha_3}d_5b^{\beta_3}d_6a^{\alpha_4}d_7b^{\beta_4}d_8,$$

where the α_i and β_i are odd integers satisfying $|\alpha_i| = |\alpha_j|$ and $|\beta_i| = |\beta_j|$ for all $i,j \leq 4$ and the d_i are words of the same length containing a and b with even exponents, and P satisfies the small cancellation condition $C'\big(\frac{1}{6}\big)$ or $C'\big(\frac{1}{4}\big) - T(4)$, then P is side injective. Concrete examples are

$$\langle a,b,c \mid (acbc^{-1}ac^{-1}bc)^2\rangle \quad \text{and} \quad \langle a,b,c \mid acbc^{-1}acbcac^{-1}bc^{-1}ac^{-1}bc\rangle,$$

which are $C'\big(\frac{1}{4}\big) - T(4)$, and

$$\langle a,b,c \mid acbca^{-1}cbc^{-1}a^{-1}c^{-1}bcac^{-1}bc^{-1}\rangle,$$

which is $C'\big(\frac{1}{6}\big)$. These presentations were checked with the help of GAP (see [7]) and the package SMALLCANCELLATION by Ivan Sadofschi Costa (see [19]).

Example 6.4 The Artin presentation $P=\langle a,b \mid \operatorname{prod}(a,b,m)=\operatorname{prod}(b,a,m)\rangle$ is not side injective for $m=3$, but is side injective for $m\geq 4$:

***m* = 3** We show that $P=\langle a,b \mid aba=bab\rangle$ is not side injective. We have $a^2(aba^2ba)a^{-2}=aba^2ba$ in $G(P)$ because $(aba)^2=aba^2ba$ is central. So

$$w=a^2ba^2ba^{-2}b^{-1}a^{-2}b^{-1}=1$$

in $G(P)$. Note that w lifts into a b–side of $\overline{K}(P)$.

***m* = 4** We show that $P=\langle a,b \mid abab=baba\rangle$ is side injective. Note that $x=abab$ is a central element. The quotient $G(P)/\langle x\rangle$ has a presentation $\langle a,b \mid (ab)^2\rangle$. Let $y=ba$; then the presentation rewrites to $\langle a,y \mid y^2\rangle$. In order to show that P is a–side injective, we have to show that a^2, b^2 and ab^2a^{-1} generate a free group of rank 3 in $G(P)$. We will do this by showing that $A=a^2$, $B=(ya^{-1})^2=ya^{-1}ya^{-1}$ and $C_0=a(ya^{-1})^2a^{-1}=aya^{-1}ya^{-1}a^{-1}$ generate a free group in the quotient presented by $Q=\langle a,y \mid y^2\rangle=\mathbb{Z}*\mathbb{Z}_2$. Let $C_1=BC_0$. We have

$$\begin{aligned}C_1&=ya^{-1}ya^{-1}aya^{-1}ya^{-1}a^{-1}=ya^{-1}yya^{-1}ya^{-1}a^{-1}=ya^{-1}a^{-1}ya^{-1}a^{-1}\\&=ya^{-2}ya^{-2}.\end{aligned}$$

And, finally, let $C=C_1A=ya^{-2}y$. In summary we have

$$A=a^2,\quad B=ya^{-1}ya^{-1},\quad C=ya^{-2}y.$$

The group $H=\langle A,B,C\rangle$ is a normal free subgroup of $G(Q)$ of rank 3 and index 4. Figure 3 shows a covering space $p\colon \overline{K}(Q)\to K(Q)$ such that $\pi_1(\overline{K}(Q))$ is free of rank 3 and $p_*\bigl(\pi_1(\overline{K}(Q))\bigr)=\langle A,B,C\rangle\leq\pi_1(K(Q))$. The argument for b–side injectivity is analogous.

***m* ≥ 6 and even** This case is easy. Let $x=\operatorname{prod}(a,b,m)$. The quotient $G(P)/\langle x\rangle$ is presented by $\langle a,b \mid (ab)^{m/2}\rangle$, which is a Dehn presentation, being a 1–relator presentation with torsion. Since $m\geq 6$, we have $\frac{1}{2}m\geq 3$. Side injectivity follows from Example 6.2.

***m* ≥ 5 and odd** Let $x=\operatorname{prod}(a,b,m)$ and $y=ba$. Note that $x=ay^{(m-1)/2}$. Using $a=xy^{(-m+1)/2}$ and $b=y^{(m+1)/2}x^{-1}$, the presentation P can be rewritten to $\langle x,y \mid x^2=y^m\rangle$. Thus $G(P)/\langle x^2\rangle$ is presented by $\langle x,y \mid x^2,y^m\rangle$, which is the hyperbolic group $\mathbb{Z}_2*\mathbb{Z}_m$. In the original generators, this is $\langle a,b \mid \operatorname{prod}(a,b,m)^2,(ba)^m\rangle$. If this were a Dehn presentation, we could proceed as in the previous case (at least for $m\geq 7$), but we do not know. Instead we argue as for $m=4$. For simplicity we assume $m=5$; the cases $m\geq 7$ go along the same lines. In order to show that P

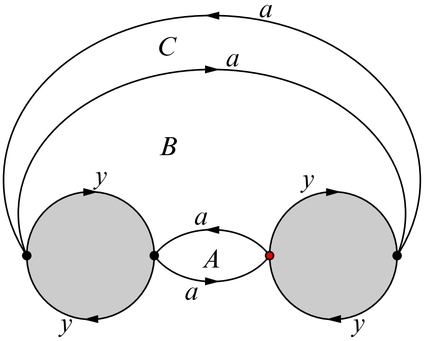


Figure 3: If $Q = \langle a, y \mid y^2 \rangle$ then the universal covering $\widetilde{K}(Q)$ is a tree with spheres attached. Here we see the intermediate covering $\overline{K}(Q)$ corresponding to the subgroup $H = \langle A, B, C \rangle$. The gray discs with boundary y^2 indicate 2–spheres.

is a–side injective we have to show that a^2, b^2 and ab^2a^{-1} generate a free group of rank 3 in $G(P)$. In terms of x and y, it suffices to show that $xy^{-2}xy^{-2}$, $y^3x^{-1}y^3x^{-1}$ and $(xy^{-2})y^3x^{-1}y^3x^{-1}(xy^{-2})^{-1}$ generate a free subgroup of rank 3 in the quotient presented by $Q = \langle x, y \mid x^2, y^5 \rangle$. Let

$$A = xy^3xy^3, \quad B = y^3xy^3x, \quad C_0 = xyxy^3xy^2x.$$

Let $C = C_0A = (xyxy^3xy^2x)(xy^3xy^3) = xyxy$.

Note that

$$C(y^{-1}Cy) = (xyxy)y^{-1}(xyxy)y = xy^2xy^2 = B^{-1}$$

and

$$C(y^{-1}Cy)(y^{-2}Cy^2) = xy^2xy^2y^{-2}xyxyy^2 = xy^3xy^3 = A.$$

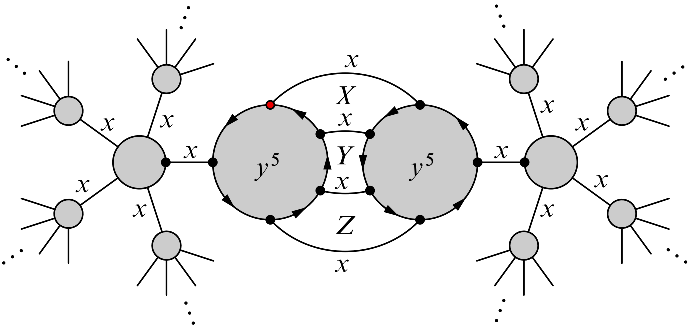


Figure 4: A rendering of the covering space $\overline{K}(Q)$. Each x–edge represents a double x–edge into which two discs with boundary x^2 are glued. Each gray disc represents five discs with boundary y^5.

So it suffices that to show that

$$X = C, \quad Y = y^{-1}Cy, \quad Z = y^{-2}Cy^2$$

generate a free subgroup of rank 3. Figure 4 shows a covering space $p\colon \overline{K}(Q) \to K(Q)$ such that $\pi_1(\overline{K}(Q))$ is free of rank 3 and $p_*\big(\pi_1(\overline{K}(Q))\big) = \langle X, Y, Z\rangle \leq \pi_1(K(Q))$. The argument for b–side injectivity is analogous.

Example 6.5 Let $P = \langle a, b, c \mid a(babcaba) = (babcaba)b\rangle$. Then P is side injective by Theorem 6.6.

Theorem 6.6 *Suppose P has dihedral type $m \geq 3$ and*

- $P = \langle a, b, \boldsymbol{c} \mid a(u_1c^{\epsilon}u_3) = (u_1c^{\epsilon}u_3)b\rangle$, *or*
- $P = \langle a, b, \boldsymbol{c} \mid a(u_1c^{\epsilon}u_2c^{\epsilon}u_3) = (u_1c^{\epsilon}u_2c^{\epsilon}u_3)b\rangle$,

where

(1) $c \in \boldsymbol{c}$ *and* $\epsilon = \pm 1$,

(2) *the words u_1 and u_3 do not contain c while u_2 is arbitrary, and*

(3) *both $u_1^{-1}a$ and u_3b^{-1} contain a subword u as in Lemma 6.1.*

Then P is side injective.

Proof We assume we are in the second case and $\epsilon = 1$. The first case is shown in an analogous way. Envision the relator disc placed in the plane as a rectangle, where the a on the very left of the equation and the b on the very right of the equation are horizontal edges, and the word $u_1cu_2cu_3$ is a vertical edge sequence. Connect the midpoints of c–edges on the left and right by horizontal red edges. See Figure 5. Suppose that w is a cyclically reduced word that represents the trivial element in $G(P)$. Let D be a reduced Van Kampen diagram with boundary w. We may assume that D is a topological disc.

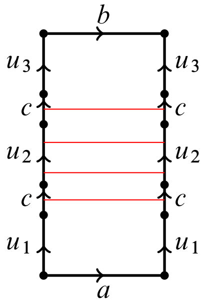


Figure 5: The relator disc drawn as a rectangle.

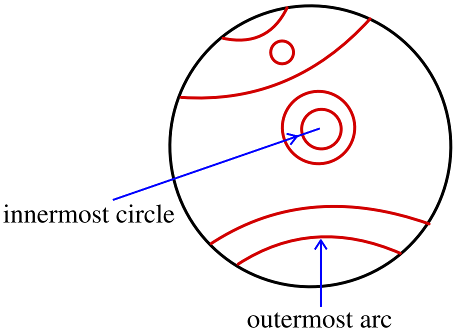


Figure 6: A disc with red arcs, indicating innermost circles and outermost arcs.

The red edges in our relator disc will form red circles and red arcs connecting points on the boundary of D. See Figure 6. Consider an innermost red circle. Going around the inside, we read off a word that freely reduces to $u_1^{-1}a^k u_1$ or $u_3 b^k u_3^{-1}$ for some $k \in \mathbb{Z}$. If $k = 0$, then D is not reduced. If $k \neq 0$, then $G(P)$ has torsion. Neither is the case; hence, there are no red circles in D. Consider an outermost red arc α. Let E be the component of $D - \alpha$ that does not contain anything red. Reading along the part of the boundary of D which belongs to E gives a reduced word (a subword of the reduced word w) equal to $u_1^{-1}a^k u_1$ or $u_3 b^k u_3^{-1}$. Because D is reduced, k cannot be zero. If k is positive then $u_1^{-1}a^k u_1$ contains $u_1^{-1}a$ and hence a word u as in Lemma 6.1. Also, $u_3 b^k u_3^{-1}$ contains bu_3^{-1}, and, since $u_3 b^{-1}$ contains a word u as in Lemma 6.1, so does $(u_3 b^{-1})^{-1} = bu_3^{-1}$. The case where k is negative goes the same way. It now follows from Lemma 6.1 that P is side injective. □

7 Last words about LOT applications

Theorem 7.1 *Let Γ be a LOT of Coxeter type. Suppose that, for every edge $e = (a \xrightarrow{w_e} b)$, the word w_e is of the form $u_1 c^\epsilon u_3$ or $u_1 c^\epsilon u_2 c^\epsilon u_3$ for some letter $c \neq a, b$, $\epsilon = \pm 1$, and words u_1, u_2 and u_3 as in Theorem 6.6. Then $K(\Gamma)$ is aspherical.*

Proof Each $\widehat{P}_e$ is side injective. This follows from Theorem 6.6. Thus each $\overline{K}_e$ is side injective. The result follows from Theorem 5.5. □

What if side injectivity fails?

Theorem 7.2 *Suppose Γ is a LOT of Coxeter type and e_1 and e_2 are two edges of Γ. Let $\overline{K}_{e_1} \cap \overline{K}_{e_2} = S$, which is a subgraph of a side of $\overline{K}_{e_1}$ and a side of $\overline{K}_{e_2}$. Let*

$N_1 = \ker(\pi_1(S) \to \pi_1(\overline{K}_{e_1}))$ and $N_2 = \ker(\pi_1(S) \to \pi_1(\overline{K}_{e_2}))$. Assume that

$$\frac{N_1 \cap N_2}{[N_1, N_2]} \neq 1.$$

Then Whitehead's asphericity conjecture is false.

Proof Suppose Whitehead's conjecture is true. Then $K(\Gamma)$ and hence $\overline{K}(\Gamma)$ is aspherical. Note that $\overline{K}_{e_1} \cup \overline{K}_{e_2}$ is a subcomplex of $\overline{K}(\Gamma)$. Let w be a reduced edge loop in S that represents a nontrivial element in the quotient $(N_1 \cap N_2)/[N_1, N_2]$. It is the boundary of a Van Kampen diagram D_1 for $\overline{K}_{e_1}$ and also the boundary of a Van Kampen diagram D_2 for $\overline{K}_{e_2}$. The two diagrams can be glued together to form a nontrivial element in $\pi_2(\overline{K}_{e_1} \cup \overline{K}_{e_2})$ (see Gutierrez and Ratcliffe [8]). This is a contradiction. □

References

[1] **A J Berrick**, **J A Hillman**, *Whitehead's asphericity question and its relation to other open problems*, from "Algebraic topology and related topics" (M Singh, Y Song, J Wu, editors), Springer (2019) 27–49 MR Zbl

[2] **S A Bleiler**, *Two-generator cable knots are tunnel one*, Proc. Amer. Math. Soc. 122 (1994) 1285–1287 MR Zbl

[3] **W A Bogley**, *J H C Whitehead's asphericity question*, from "Two-dimensional homotopy and combinatorial group theory" (C Hog-Angeloni, W Metzler, A J Sieradski, editors), London Math. Soc. Lecture Note Ser. 197, Cambridge Univ. Press (1993) 309–334 MR Zbl

[4] **J O Button**, *Large groups of deficiency* 1, Israel J. Math. 167 (2008) 111–140 MR Zbl

[5] **M Carette**, **R Weidmann**, *On the rank of Coxeter groups*, preprint (2009) arXiv 0910.4997

[6] **M W Davis**, *The geometry and topology of Coxeter groups*, London Mathematical Society Monographs Series 32, Princeton Univ. Press (2008) MR Zbl

[7] **The GAP Group**, *GAP*: *groups*, *algorithms*, *and programming*, software (2020) Version 4.11.0 Available at `https://www.gap-system.org`

[8] **M A Gutiérrez**, **J G Ratcliffe**, *On the second homotopy group*, Quart. J. Math. Oxford Ser. 32 (1981) 45–55 MR Zbl

[9] **J Harlander**, **S Rosebrock**, *Generalized knot complements and some aspherical ribbon disc complements*, J. Knot Theory Ramifications 12 (2003) 947–962 MR Zbl

[10] **J Harlander**, **S Rosebrock**, *Aspherical word labeled oriented graphs and cyclically presented groups*, J. Knot Theory Ramifications 24 (2015) art. id. 1550025 MR Zbl

[11] **J Hillman**, *Algebraic invariants of links*, 2nd edition, Series on Knots and Everything 52, World Sci., Hackensack, NJ (2012) MR Zbl

[12] **J Howie**, *On the asphericity of ribbon disc complements*, Trans. Amer. Math. Soc. 289 (1985) 281–302 MR Zbl

[13] **E Klimenko**, **M Sakuma**, *Two-generator discrete subgroups of* $\mathrm{Isom}(\mathbb{H}^2)$ *containing orientation-reversing elements*, Geom. Dedicata 72 (1998) 247–282 MR Zbl

[14] **R C Lyndon**, **P E Schupp**, *Combinatorial group theory*, Ergebnisse der Math. 89, Springer (1977) MR Zbl

[15] **W Menasco**, **A W Reid**, *Totally geodesic surfaces in hyperbolic link complements*, from "Topology '90" (B Apanasov, W D Neumann, A W Reid, L Siebenmann, editors), Ohio State Univ. Math. Res. Inst. Publ. 1, de Gruyter, Berlin (1992) 215–226 MR Zbl

[16] **B B Newman**, *Some results on one-relator groups*, Bull. Amer. Math. Soc. 74 (1968) 568–571 MR Zbl

[17] **F H Norwood**, *Every two-generator knot is prime*, Proc. Amer. Math. Soc. 86 (1982) 143–147 MR Zbl

[18] **S Rosebrock**, *Labelled oriented trees and the Whitehead-conjecture*, from "Advances in two-dimensional homotopy and combinatorial group theory" (W Metzler, S Rosebrock, editors), London Math. Soc. Lecture Note Ser. 446, Cambridge Univ. Press (2018) 72–102 MR Zbl

[19] **I Sadofschi Costa**, *The small cancellation package*, software (2018) Available at `https://github.com/isadofschi/smallcancellation`

[20] **J-P Serre**, *Trees*, Springer (2003) MR Zbl

[21] **R Weidmann**, *The rank problem for sufficiently large Fuchsian groups*, Proc. Lond. Math. Soc. 95 (2007) 609–652 MR Zbl

[22] **T Yajima**, *On a characterization of knot groups of some spheres in* R^4, Osaka Math. J. 6 (1969) 435–446 MR Zbl

Department of Mathematics, Boise State University
Boise, ID, United States

Pädagogische Hochschule Karlsruhe
Karlsruhe, Germany

`jensharlander@boisestate.edu, rosebrock@ph-karlsruhe.de`

Received: 5 March 2021 Revised: 21 December 2021

Algebraic & Geometric Topology **23**:6 (2023) 2735–2776
DOI: 10.2140/agt.2023.23.2735
Published: 7 September 2023

Weave-realizability for D–type

JAMES HUGHES

We study exact Lagrangian fillings of Legendrian links of D_n–type in the standard contact 3–sphere. The main result is the existence of a Lagrangian filling, represented by a weave, such that any algebraic quiver mutation of the associated intersection quiver can be realized as a geometric weave mutation. The method of proof is via Legendrian weave calculus and a construction of appropriate 1–cycles whose geometric intersections realize the required algebraic intersection numbers. In particular, we show that, in D–type, each cluster chart of the moduli of microlocal rank-1 sheaves is induced by at least one embedded exact Lagrangian filling. Hence, the Legendrian links of D_n–type have at least as many Hamiltonian isotopy classes of Lagrangian fillings as cluster seeds in the D_n–type cluster algebra, and their geometric exchange graph for Lagrangian disk surgeries contains the cluster exchange graph of D_n–type.

53D12; 57K33

1 Introduction

Legendrian links in contact 3–manifolds — see Bennequin [3] and Arnold [2] — are central to the study of 3–dimensional contact topology; see Ozbagci and Stipsicz [20] and Geiges [15]. Recent developments due to Casals, Gao, Ng and Zaslow [7; 5; 6] have revealed new phenomena regarding their Lagrangian fillings, including the existence of many Legendrian links $\Lambda \subseteq (\mathbb{S}^3, \xi_{\mathrm{st}})$ with infinitely many (smoothly isotopic) Lagrangian fillings in the Darboux 4–ball $(\mathbb{D}^4, \omega_{\mathrm{st}})$ which are not Hamiltonian isotopic. The relationship between cluster algebras and Lagrangian fillings — see Casals and Zaslow [7] and Gao, Shen and Weng [14] — has also led to new conjectures on the classification of Lagrangian fillings from Casals [4]. In particular, Conjecture 5.1 of [4] introduced a conjectural ADE classification of Lagrangian fillings. Here we study D–type and prove part of the conjectured classification.

The A–type was studied by Ekholm, Honda and Kálmán [9] and Pan [21] via Floer-theoretic methods, and by Shende, Treumann, Williams and Zaslow [23; 25] via

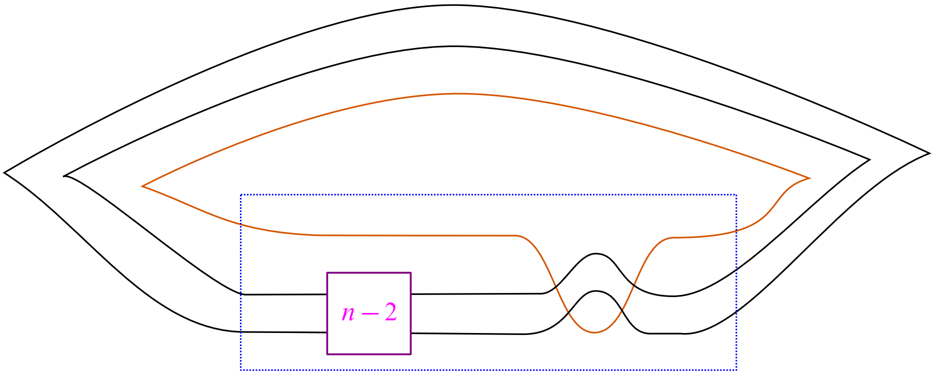


Figure 1: The front projection of $\lambda(D_n) \subseteq (\mathbb{S}^3, \xi_{\mathrm{st}})$. The box labeled with an $n-2$ represents $n-2$ positive crossings given by σ_1^{n-2}. When n is even, $\lambda(D_n)$ has three components, while when n is odd, $\lambda(D_n)$ only has two components.

microlocal sheaves. Their main result is that the A_n–Legendrian link $\lambda(A_n) \subseteq (\mathbb{S}^3, \xi_{\mathrm{st}})$, which is the max-tb representative of the $(2, n+1)$–torus link, has at least a Catalan number $C_{n+1} = \frac{1}{n+2}\binom{2n+2}{n+1}$ of embedded exact Lagrangian fillings, where C_{n+1} is precisely the number of cluster seeds in the finite-type A_n cluster algebra; see Fomin, Williams and Zelevinsky [13]. We will show that the same holds in D–type, namely that D_n–type Legendrian links have at least as many distinct Hamiltonian isotopy classes of Lagrangian fillings as there are cluster seeds in the D_n–type cluster algebra. This will be a consequence of a stronger geometric result, weave-realizability in D–type, which we discuss below.

By definition, the Legendrian link $\lambda(D_n) \subseteq (\mathbb{S}^3, \xi_{\mathrm{st}})$ with $n \geq 4$ of D_n–type is the standard satellite of the Legendrian link defined by the front projection given by the 3–stranded positive braid $\sigma_1^{n-2}(\sigma_2\sigma_1^2\sigma_2)(\sigma_1\sigma_2)^3$, where σ_1 and σ_2 are the Artin generators for the 3–stranded braid group. Figure 1 depicts a front diagram for $\lambda(D_n)$; note that the (-1)–framed closure of $\sigma_1^{n-2}(\sigma_2\sigma_1^2\sigma_2)(\sigma_1\sigma_2)^3$ is Legendrian isotopic to the rainbow closure of $\sigma_1^{n-2}(\sigma_2\sigma_1^2\sigma_2)$, the latter being depicted. The Legendrian link $\lambda(D_n)$ is also a max-tb representative of the smooth isotopy class of the link of the singularity $f(x, y) = y(x^2 + y^{n-2})$. Since these are algebraic links, the max-tb representative given above is unique — see eg [4, Proposition 2.2] — and has at least one exact Lagrangian filling; see Hayden and Sabloff [18].

The N–graph calculus developed by Casals and Zaslow [7] allows us to associate an exact Lagrangian filling of a (-1)–framed closure of a positive braid to a pair of trivalent

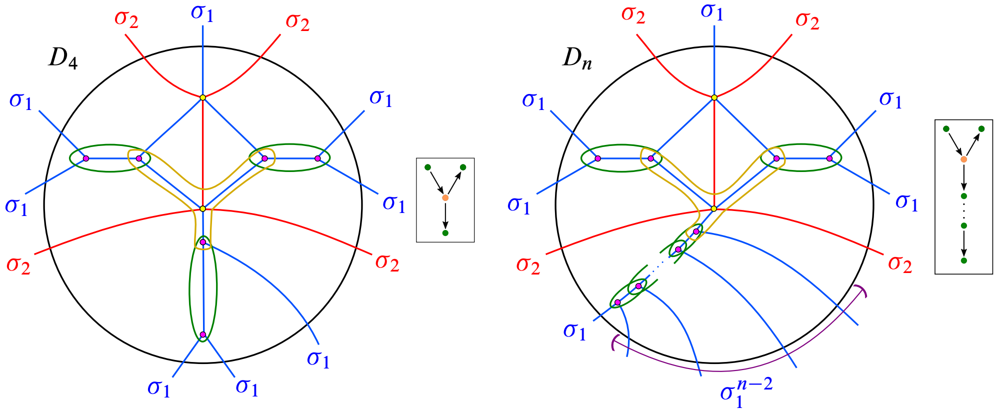


Figure 2: 3–Graphs $\Gamma_0(D_4)$ (left) and $\Gamma_0(D_n)$ (right), pictured with their associated intersection quivers $Q(\Gamma_0(D_4), \{\gamma_i^{(0)}\})$ and $Q(\Gamma_0(D_n), \{\gamma_i^{(0)}\})$. The basis $\{\gamma_i^{(0)}\}$ for $H_1\big(\Lambda(\Gamma_0(D_4)); \mathbb{Z}\big)$ is depicted by the dark green and orange and cycles drawn in the graph. Note that the quivers correspond to the D_4 and D_n Dynkin diagrams, usually depicted rotated 90° counterclockwise.

planar graphs satisfying certain properties. See Figure 2, left, for an example of a particular 3–graph, denoted by $\Gamma_0(D_4)$, associated to the Legendrian link $\lambda(D_4)$.[1] In Section 3, we will show that the 3–graph $\Gamma_0(D_4)$ generalizes to a family of 3–graphs $\Gamma_0(D_n)$, depicted in Figure 2, right, for any $n \geq 3$. In a nutshell, a 3–fold branched cover of $\mathbb{D}^2$, simply branched at the trivalent vertices of these 3–graphs, yields an exact Lagrangian surface in $(T^*\mathbb{D}^2, \omega_{\mathrm{st}})$, whose Legendrian lift is a Legendrian weave. One of the distinct advantages of the 3–graph calculus is that it combinatorializes an operation, known as Lagrangian disk surgery — see Polterovich [22] and Yau [27] — that modifies the weave in such a way as to yield additional — non-Hamiltonian isotopic — exact Lagrangian fillings of the link.

If we consider a 3–graph Γ and a basis $\{\gamma_i\}$ for the first homology of the weave $\Lambda(\Gamma)$ for $i \in [1, b_1(\Lambda(\Gamma))]$, we can define a quiver $Q(\Gamma, \{\gamma_i\})$ whose adjacency matrix is given by the intersection form in $H_1(\Lambda(\Gamma))$. Quivers come equipped with a involutive operation, known as quiver mutation, that produces new quivers; see Section 2.6 below or Fomin, Williams and Zelevinsky [12] for more on quivers. A key result of [7] tells us that Legendrian mutation of the weave induces a quiver mutation of the intersection quiver. Quivers related by a sequence of mutations are said to be mutation-equivalent,

[1] We use $\lambda(D_4)$, ie $n = 4$, as a first example because $n = 3$ would correspond to $\lambda(A_3)$, which has been studied previously [9; 21]. The study of $\lambda(D_4)$ is also the first instance where we require the machinery of 3–graphs rather than 2–graphs.

and the quivers that are of finite mutation type (ie the set of mutation-equivalent quivers is finite) have an ADE classification [13]. This classification parallels the naming convention for the D_n links described above: the intersection quiver associated to $\lambda(D_n)$ is a quiver in the mutation class of the D_n Dynkin diagram (the latter endowed with an appropriate orientation). See Figure 2 for examples of D_4 and D_n quivers. For our 3–graph $\Gamma_0(D_n)$ for $n \geq 3$, we will give an explicit basis $\{\gamma_i^{(0)}\} = \{\gamma_1^{(0)}, \dots, \gamma_n^{(0)}\}$ for $H_1\big(\Lambda(\Gamma_0(D_n)), \mathbb{Z}\big)$, whose intersection quiver $Q(\Gamma_0(D_n), \{\gamma_i^{(0)}\})$ is the standard D_n Dynkin diagram.

Let us introduce the following notion. By definition, a sequence of quiver mutations for $Q(\Gamma_0(D_n), \{\gamma_i^{(0)}\})$ is said to be *weave-realizable* if each quiver mutation in the sequence can be realized as a Legendrian weave mutation for a 3–graph. Our main result is the following theorem:

Theorem 1 *Any sequence of quiver mutations of* $Q(\Gamma_0(D_n), \gamma_1^{(0)}, \dots, \gamma_n^{(0)}\})$ *is weave-realizable.*

In other words, Theorem 1 states that, in D–type, any algebraic quiver mutation can actually be realized geometrically by a Legendrian weave mutation. Weave-realizability is of interest because it measures the difference between algebraic invariants — eg the cluster structure in the moduli of sheaves — and geometric objects, in this case Hamiltonian isotopy classes of exact Lagrangian fillings. In general, if any sequence of quiver mutations were weave-realizable, we would know that each cluster is inhabited by at least one embedded exact Lagrangian filling; this general statement remains open for an arbitrary Legendrian link. For instance, any link with an associated quiver that is not of finite mutation type satisfying the weave-realizability property would admit infinitely many Lagrangian fillings, distinguished by their quivers.[2] Note that weave-realizability was shown for A–type by Treumann and Zaslow [25], and beyond A– and D–types we currently do not know whether there are any other links satisfying the weave-realizability property.

We can further distinguish fillings by studying the cluster algebra structure on the moduli of microlocal rank-1 sheaves $\mathcal{C}(\Gamma)$ of a weave $\Lambda(\Gamma)$; see eg [7]. Specifically, sheaf quantization of each exact Lagrangian filling of $\lambda(D_n)$ induces a cluster chart on the coordinate ring of functions on $\mathcal{C}(\Gamma_0(D_n))$ via the microlocal monodromy functor, giving $\mathcal{C}(\Gamma_0(D_n))$ the structure of a cluster variety of D_n–type; see Shende,

[2] This would be independent of the cluster structure defined by the microlocal monodromy functor, which we actually must use for D–type.

Treumann, Williams and Zaslow [24; 23]. Describing a single cluster chart in this cluster variety requires the data of the quiver associated to the weave and the microlocal monodromy around each 1–cycle of the weave. Crucially, applying the Legendrian mutation operation to the weave induces a cluster transformation on the cluster chart, and the specific cluster chart defined by a Lagrangian fillings is a Hamiltonian isotopy invariant. Therefore, Theorem 1 has the following consequence:

Corollary 2 *Every cluster chart of the moduli of microlocal rank-*1 *sheaves* $\mathcal{C}(\Gamma_0(D_n))$ *is induced by at least one embedded exact Lagrangian filling of* $\lambda(D_n) \subset (\mathbb{S}^3, \xi_{st})$*. In particular, there exist at least* $(3n-2)C_{n-1}$ *exact Lagrangian fillings of the link* $\lambda(D_n)$ *up to Hamiltonian isotopy, where* C_n *denotes the* n^{th} *Catalan number.*

Moreover, weave-realizability implies a slightly stronger result. Specifically, we can consider the *filling exchange graph* associated to a link of D_n–type, where the vertices are Hamiltonian isotopy classes of embedded exact Lagrangians, and two vertices are connected by an edge if the two fillings are related by a Lagrangian disk surgery. Then weave-realizability implies that the filling exchange graph contains a subgraph isomorphic to the cluster exchange graph for the cluster algebra of D_n–type.

Remark As of yet, we have no way of determining whether our method produces all possible exact Lagrangian fillings of a type D_n link. This question remains open for A–type Legendrian links as well. In fact, the only known knot for which we have a complete nonempty classification of Lagrangian fillings is the Legendrian unknot, which has a unique filling; see Eliashberg and Polterovich [11].

In summary, our method for constructing exact Lagrangian fillings will be to represent them using the planar diagrammatic calculus of N–graphs developed in [7]. This diagrammatic calculus includes a mutation operation on the diagrams that yields additional fillings. We distinguish the resulting fillings up to Hamiltonian using a sheaf-theoretic invariant. From this data, we extract a cluster algebra structure and show that every mutation of the quiver associated to the cluster can be realized by applying our Legendrian mutation operation to the 3–graph, thus proving that there are at least as many distinct fillings as distinct cluster seeds of D_n–type. The main theorem will be proven in Section 3 after giving the necessary preliminaries in Section 2.

Added in proof

While writing this manuscript, we learned that recent independent work by Byung Hee An, Youngjin Bae and Eunjeong Lee [1] also produces at least as many exact

Lagrangian fillings as cluster seeds for links of ADE–type, providing an alternative proof to Corollary 2. To our understanding, they use an inductive argument that relies on the combinatorial properties of the finite-type generalized associahedron. Specifically, they leverage the fact that the Coxeter transformation in finite type is transitive if starting with a particular set of vertices by finding a weave pattern that realizes Coxeter mutations. While their initial 3–graph $\mathcal{G}(1,1,n)$ is the same as our $\Gamma_0(D_n)$, their method of computing a weave associated to an arbitrary sequence of quiver mutations requires concatenating some number of concordances corresponding to the Coxeter mutation before mutating. As a result, a 3–graph arising from a sequence of quiver mutations $\mu_1,\dots,\mu_i$ computed using this method is not explicitly shown to be related to a 3–graph arising from a sequence of quiver mutations $\mu_1,\dots,\mu_i,\mu_{i+1}$ by a single Legendrian mutation of the weave. In contrast, in our approach we are able to relate each 3–graph arising from a sequence of quiver mutations to the next by a single Legendrian mutation and a specific set of Legendrian Reidemeister moves. While both this manuscript and [1] use the framework of N–graphs to approach the problem of enumerating exact Lagrangian fillings, the proofs are different and independent, and our approach is able to give an explicit construction for realizing any sequence of quiver mutations via an explicit sequence of mutations in the 3–graph.

Acknowledgments

Many thanks to Roger Casals for his support and encouragement throughout this project. Thanks also to Youngjin Bae and Eric Zaslow for helpful conversations, and to the referee for insightful comments.

2 Preliminaries

In this section we introduce the ingredients required for the proof of Theorem 1 and Corollary 2. We first discuss the contact topology needed to understand weaves and their homology. We then discuss the sheaf-theoretic material related to distinguishing fillings via cluster algebraic methods.

2.1 Contact topology and exact Lagrangian fillings

A contact structure ξ on $\mathbb{R}^3$ is a 2–plane field given locally as the kernel of a 1–form $\alpha \in \Omega^1(\mathbb{R}^3)$ satisfying $\alpha \wedge d\alpha \neq 0$. The standard contact structure on $(\mathbb{R}^3, \xi_{\mathrm{st}})$ is given by the kernel of $\alpha = dz - y\,dx$. A Legendrian link λ in $(\mathbb{R}^3, \xi)$ is an embedding of a disjoint union of copies of $\mathbb{S}^1$ that is always tangent to ξ. By definition, the contact 3–sphere $(\mathbb{S}^3, \xi_{\mathrm{st}})$ is the one-point compactification of $(\mathbb{R}^3, \xi_{\mathrm{st}})$. Since a link in $\mathbb{S}^3$ can always be assumed to avoid a point, we will equivalently be considering Legendrian links in $(\mathbb{R}^3, \xi_{\mathrm{st}})$ and $(\mathbb{S}^3, \xi_{\mathrm{st}})$. By definition, the symplectization of $(\mathbb{R}^3, \xi_{\mathrm{st}})$ is given by $(\mathbb{R}^3 \times \mathbb{R}_t, d(e^t\alpha))$.

Given two Legendrian links λ_+ and λ_- in $(\mathbb{R}^3, \xi)$, an exact Lagrangian cobordism Σ from λ_- to λ_+ is an embedded compact orientable surface in the symplectization $(\mathbb{R}^3 \times \mathbb{R}_t, d(e^t\alpha))$ such that, for some $T > 0$,

- $\Sigma \cap \big(\mathbb{R}^3 \times [T, \infty)\big) = \lambda_+ \times [T, \infty)$,
- $\Sigma \cap (\mathbb{R}^3 \times (-\infty, -T)) = \lambda_- \times (-\infty, -T]$,
- Σ is an exact Lagrangian, ie $e^t\alpha = df$ for some function $f\colon \Sigma \to \mathbb{R}$.

The asymptotic behavior of Σ, as specified by the first two conditions, ensures that we can concatenate Lagrangian cobordisms. By definition, an exact Lagrangian filling of λ_+ is an exact Lagrangian cobordism from $\varnothing$ to λ_+.

We can also consider the Legendrian lift of an exact Lagrangian in the contactization $\big(\mathbb{R}_s \times \mathbb{R}^4, \ker\{ds - d(e^t\alpha)\}\big)$ of $(\mathbb{R}^4, d(e^t\alpha))$. Note that there exists a contactomorphism between $\big(\mathbb{R}_s \times \mathbb{R}^4, \ker\{ds - d(e^t\alpha)\}\big)$ and the standard contact Darboux structure $(\mathbb{R}^5, \xi_{\mathrm{st}})$, where $\xi_{\mathrm{st}} = \ker\{dz - y_1\,dx_1 - y_2\,dx_2\}$. We will often work with the Legendrian front projection $(\mathbb{R}^5, \xi_{\mathrm{st}}) \to \mathbb{R}^3_{x_1,x_2,z}$ for the latter. This will be a useful perspective for us, as it allows us to construct Lagrangian fillings by studying (wave)fronts in $\mathbb{R}^3 = \mathbb{R}^3_{x_1,x_2,z}$ of Legendrian surfaces in $(\mathbb{R}^5, \xi_{\mathrm{st}})$, and then projecting down to the standard symplectic Darboux chart $\mathbb{R}^4 = \mathbb{R}^4_{x_1,y_1,x_2,y_2}$. In this setting, the exact Lagrangian surface is embedded in $\mathbb{R}^4$ if and only if its Legendrian lift has no Reeb chords. The construction will be performed through the combinatorics of N–graphs, as we now explain.

2.2 3–Graphs and weaves

In this subsection, we discuss the diagrammatic method of constructing and manipulating exact Lagrangian fillings of links arising as the (-1)–framed closures of positive braids via the calculus of N–graphs. It will suffice here to take $N = 3$.

Definition 3 A 3–graph is a pair of embedded planar trivalent graphs $B, R \subseteq \mathbb{D}^2$ such that at any vertex $v \in B \cap R$ the six edges belonging to B and R incident to v alternate.

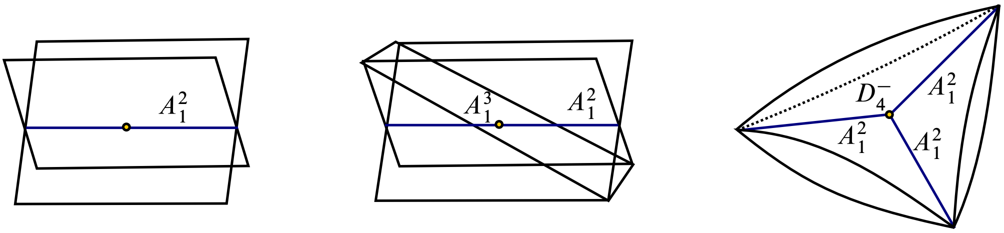


Figure 3: A_1^2 (left), A_1^3 (center) and D_4^- (right) singularities represented in the 3–graph by an edge, hexavalent vertex and trivalent vertex, respectively.

Equivalently, a 3–graph is an edge-bicolored graph with monochromatic trivalent vertices and interlacing hexavalent vertices. $\Gamma_0(D_4)$, depicted in Figure 2, left, contains two hexavalent vertices displaying the alternating behavior described in the definition.

Remark Casals and Zaslow [7] give a general framework for working with N–graphs, where $N-1$ is the number of embedded planar trivalent graphs. This allows for the study of fillings of Legendrian links associated to N–stranded positive braids. This can also be generalized to consider N–graphs in a surface other than $\mathbb{D}^2$. In our case, the family of links $\lambda(D_n)$ can be expressed as a family of 3–stranded braids, whence our choice to restrict N to 3 in $\mathbb{D}^2$.

Given a 3–graph $\Gamma \subseteq \mathbb{D}^2$, we describe how to associate a Legendrian surface $\Lambda(\Gamma) \subseteq (\mathbb{R}^5, \xi_{\mathrm{st}})$. To do so, we first describe certain singularities of $\Lambda(\Gamma)$ that arise under the Legendrian front projection $\pi\colon (\mathbb{R}^5, \xi_{\mathrm{st}}) \to (\mathbb{R}^3, \xi_{\mathrm{st}})$. In general, such singularities are known as Legendrian singularities or singularities of fronts. See [2] for a classification of such singularities. The three singularities we will be interested in are the A_1^2, A_1^3 and D_4^- singularities, pictured in Figure 3.

Before we describe our Legendrian surfaces, we must first discuss the ambient contact structure that they live in. For $\Gamma \subseteq \mathbb{D}^2$ we will take $\Lambda(\Gamma)$ to live in the first jet space $(J^1\mathbb{D}^2, \xi_{\mathrm{st}}) = (T^*\mathbb{D}^2 \times \mathbb{R}_z, \ker(dz - \theta_{\mathrm{st}}))$, where θ_{st} is the standard Liouville form on the cotangent bundle $T^*\mathbb{D}^2$. We can view $J^1\mathbb{D}^2$ as a certain local model for a contact structure, in the following way. If we take (Y, ξ) to be a contact 5–manifold, then, by the Weinstein neighborhood theorem, any Legendrian embedding $i\colon \mathbb{D}^2 \to (Y, \xi)$ extends to an embedding from $(J^1\mathbb{D}^2, \xi_{\mathrm{st}})$ to a small open neighborhood of $i(\mathbb{D}^2)$ with contact structure given by the restriction of ξ to that neighborhood. In particular, a Legendrian embedding of $i\colon \mathbb{S}^1 \to \mathbb{S}^3$ gives rise to a contact embedding $\tilde{\imath}\colon J^1\mathbb{S}^1 \to \mathrm{Op}(i(\mathbb{S}^1))$ into some open neighborhood $\mathrm{Op}(i(\mathbb{S}^1)) \subseteq \mathbb{S}^3$. Of particular note in our case is that, under

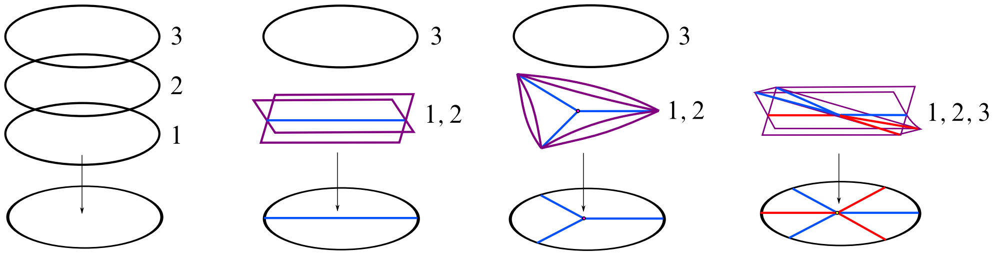


Figure 4: The weaving of the singularities pictured in Figure 3 along the edges of the N–graph. Gluing these local pictures together according to the 3–graph Γ yields the weave $\Lambda(\Gamma)$.

a Legendrian embedding $\mathbb{D}^2 \subseteq (\mathbb{R}^5, \xi_{\mathrm{st}})$, a Legendrian link λ in $J^1\partial\mathbb{D}^2$ is mapped to a Legendrian link in the contact boundary $(\mathbb{S}^3, \xi_{\mathrm{st}})$ of the symplectic $(\mathbb{R}^4, \omega_{\mathrm{st}} = d\theta_{\mathrm{st}})$ given as the codomain of the Lagrangian projection $(\mathbb{R}^5, \xi_{\mathrm{st}}) \to (\mathbb{R}^4, \omega_{\mathrm{st}})$. See [19] for a description of this Legendrian satellite operation.

To construct a Legendrian weave $\Lambda(\Gamma) \subseteq (J^1\mathbb{D}^2, \xi_{\mathrm{st}})$ from a 3–graph Γ, we glue together the local germs of singularities according to the edges of Γ. First consider three horizontal wavefronts $\mathbb{D}^2 \times \{1\} \sqcup \mathbb{D}^2 \times \{2\} \sqcup \mathbb{D}^2 \times \{3\} \subseteq \mathbb{D}^2 \times \mathbb{R}$ and a 3–graph $\Gamma \subseteq \mathbb{D}^2 \times \{0\}$. We construct the associated Legendrian weave $\Lambda(\Gamma)$ as follows:

- Above each blue (resp. red) edge, insert an A_1^2 crossing between the $\mathbb{D}^2 \times \{1\}$ and $\mathbb{D}^2 \times \{2\}$ sheets (resp. $\mathbb{D}^2 \times \{2\}$ and $\mathbb{D}^2 \times \{3\}$ sheets) so that the projection of the A_1^2 singular locus under $\pi : \mathbb{D}^2 \times \mathbb{R} \to \mathbb{D}^2 \times \{0\}$ agrees with the blue (resp. red) edge.
- At each blue (resp. red) trivalent vertex v, insert a D_4^- singularity between the sheets $\mathbb{D}^2 \times \{1\}$ and $\mathbb{D}^2 \times \{2\}$ (resp. $\mathbb{D}^2 \times \{2\}$ and $\mathbb{D}^2 \times \{3\}$) in such a way that the projection of the D_4^- singular locus agrees with v and the projection of the A_2^1 crossings agree with the edges incident to v.
- At each hexavalent vertex v, insert an A_1^3 singularity along the three sheets in such a way that the origin of the A_1^3 singular locus agrees with v and the A_1^2 crossings agree with the edges incident to v.

If we take an open cover $\{U_i\}_{i=1}^m$ of $\mathbb{D}^2 \times \{0\}$ by open disks, refined so that any disk contains at most one of these three features, we can glue together the resulting fronts according to the intersection of edges along the boundary of our disks. Specifically, if $U_i \cap U_j$ is nonempty, then we define $\Sigma(U_1 \cup U_2)$ to be the wavefront resulting from considering the union of wavefronts $\Sigma(U_1) \cup \Sigma(U_j)$ in $(U_1 \cup U_2) \times \mathbb{R}$. We define the Legendrian weave $\Lambda(\Gamma)$ as the Legendrian surface contained in $(J^1\mathbb{D}^2, \xi_{\mathrm{st}})$ with wavefront $\Sigma(\Gamma) = \Sigma\bigl(\bigcup_{i=1}^m U_i\bigr)$ given by gluing the local wavefronts of singularities together according to the 3–graph Γ [7, Section 2.3].

The smooth topology of a Legendrian weave $\Lambda(\Gamma)$ is given as a 3–fold branched cover over $\mathbb{D}^2$ with simple branched points corresponding to each of the trivalent vertices of Γ. The genus of $\Lambda(\Gamma)$ is then computed, using the Riemann–Hurwitz formula, to be

$$g(\Lambda(\Gamma)) = \tfrac{1}{2}\bigl(v(\Gamma) + 2 - 3\chi(\mathbb{D}^2) - |\partial\Lambda(\Gamma)|\bigr),$$

where $v(\Gamma)$ is the number of trivalent vertices of Γ and $|\partial\Lambda(\Gamma)|$ denotes the number of boundary components of Γ.

Example If we apply this formula to the 3–graph $\Gamma_0(D_4)$, pictured in Figure 2, we have six trivalent vertices and three link components, so the genus is computed as $g\bigl(\Lambda(\Gamma_0(D_4))\bigr) = \frac{1}{2}(6+2-3-3) = 1$.

For $\Gamma_0(D_n)$, we have three boundary components for even n and two boundary components for odd n. The number of trivalent vertices is $n+2$, so the genus $g\bigl(\Lambda(\Gamma_0(D_n))\bigr)$ is $\bigl\lfloor \frac{1}{2}(n-1)\bigr\rfloor$, assuming $n \geq 2$.

This computation tells us that $\Lambda(\Gamma_0(D_4))$ is smoothly a 3–punctured torus bounding the link $\lambda(D_4)$. Therefore, we can give a basis for $H_1(\Lambda(\Gamma_0(D_4)); \mathbb{Z})$ in terms of the four cycles pictured in Figure 2.

For $\Gamma_0(D_n)$, the corresponding weave $\Lambda(\Gamma_0(D_n))$ will be smoothly a genus $\bigl\lfloor \frac{1}{2}(n-1)\bigr\rfloor$ surface with a basis of $H_1(\Lambda(\Gamma); \mathbb{Z})$ given by n–cycles. Our computation of the genus in the example above agrees with a theorem of Chantraine [8] specifying the relationship between the Thurston–Bennequin invariant of $\lambda(D_n)$ and the genus of any exact Lagrangian filling Λ of $\lambda(D_n)$. In particular, $\mathrm{tb}(\lambda(D_n)) = n-1$ and therefore the Euler characteristic of Λ is $3-n$ when n is odd and $4-n$ when n is even. Thus, we recover the genus $g(\Lambda) = \bigl\lfloor \frac{1}{2}(n-1)\bigr\rfloor$ of any filling of $\lambda(D_n)$. In the next subsection, we describe a general method for giving a basis $\{\gamma_i^{(0)}\}$ for $i \in [1, n]$ of the first homology $H_1(\Lambda(\Gamma_0(D_n)); \mathbb{Z}) \cong \mathbb{Z}^n$.

2.3 Homology of weaves

We require a description of the first homology $H_1\bigl((\Lambda(\Gamma)); \mathbb{Z}\bigr)$ in order to apply the mutation operation to a 3–graph Γ. We first consider an edge connecting two trivalent vertices. Closely examining the sheets of our surface, we can see that each such edge

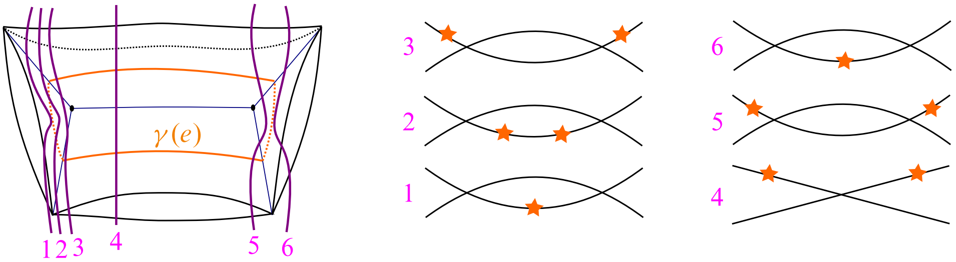


Figure 5: A short I–cycle $\gamma(e)$ for the edge $e \in G$ pictured in the wavefront $\Sigma(\Gamma)$ (left) and a vertical slicing of $\Sigma(\Gamma)$ (right).

corresponds to a 1–cycle, as pictured in Figure 5, left. We refer to such a 1–cycle as a short I–cycle. Similarly, any three edges of the same color that connect a single hexavalent vertex to three trivalent vertices correspond to a 1–cycle, as pictured in Figure 6, left. We refer to such a 1–cycle as a short Y–cycle. See Figures 5, right, and 6, right, for a diagram of these 1–cycles in the wavefront $\Sigma(\Gamma)$. We can also consider a sequence of edges starting and ending at trivalent vertices and passing directly through any number of hexavalent vertices, as pictured in Figure 7. Such a cycle is referred to as a long I–cycle. Finally, we can combine any number of I–cycles and short Y–cycles to describe an arbitrary 1–cycle as a tree with leaves on trivalent vertices and edges passing directly through hexavalent vertices.

In the proof of our main result, we will generally give a basis for $H_1(\Lambda(\Gamma); \mathbb{Z})$ in terms of short I–cycles and short Y–cycles. Indeed, Figure 8 gives a basis of $H_1\big(\Lambda(\Gamma_0(D_n)); \mathbb{Z}\big)$ consisting of $n-1$ short I–cycles and a single Y–cycle.

The intersection form $\langle \cdot, \cdot \rangle$ on $H_1(\Lambda(\Gamma))$ plays a key role in distinguishing our Legendrian weaves. If we consider a pair of 1–cycles $\gamma_1, \gamma_2 \in H_1(\Lambda(\Gamma))$ with nonempty geometric intersection in Γ, as pictured in Figure 9, we can see that the intersection of

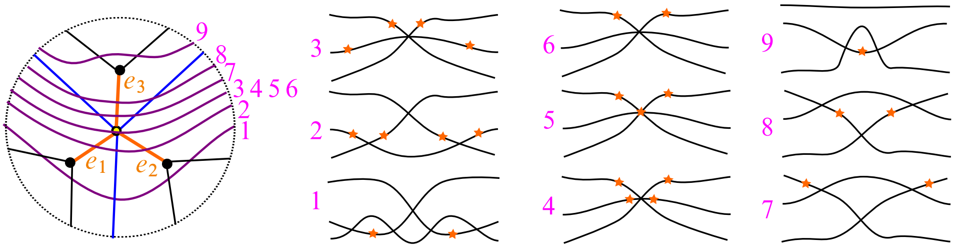


Figure 6: A short Y–cycle $\gamma(e)$ defined by the edges $e_1, e_2, e_3 \in G$ pictured in the wavefront $\Sigma(\Gamma)$ (left) and a vertical slicing of $\Sigma(\Gamma)$ (right).

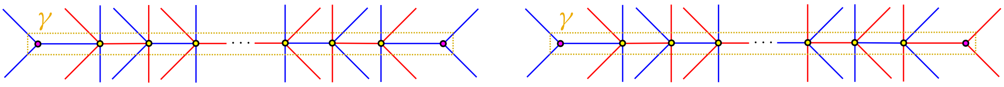


Figure 7: A pair of long I–cycles, both denoted by γ. The cycle on the left passes through an even number of hexavalent vertices, while the cycle on the right passes through an odd number.

their projection onto the 3–graph differs from the intersection in $\Lambda(\Gamma)$. Specifically, we can carefully examine the sheets that the 1–cycles cross in order to see that γ_1 and γ_2 intersect only in a single point of $\Lambda(\Gamma)$. If we fix an orientation on γ_1 and γ_2, then we can assign a sign to this intersection based on the convention given in Figure 9. We refer to the signed count of the intersection of γ_1 and γ_2 as their algebraic intersection and denote it by $\langle\gamma_1,\gamma_2\rangle$. We fix a counterclockwise orientation for all of our cycles and adopt the convention that any two cycles γ_1 and γ_2 intersecting at a trivalent vertex as in Figure 9 have algebraic intersection $\langle\gamma_1,\gamma_2\rangle = -1$.

Notation For the sake of visual clarity, we will represent an element of $H_1(\Lambda(\Gamma);\mathbb{Z})$ by a colored edge. This also ensures that the geometric intersection more accurately reflects the algebraic intersection. The original coloring of the blue or red edges can be readily obtained by examining Γ and its trivalent vertices.

In our correspondence between 3–graphs and weaves, we must consider how a Legendrian isotopy of the weave $\Lambda(\Gamma)$ affects the 3–graph Γ and its homology basis. We can

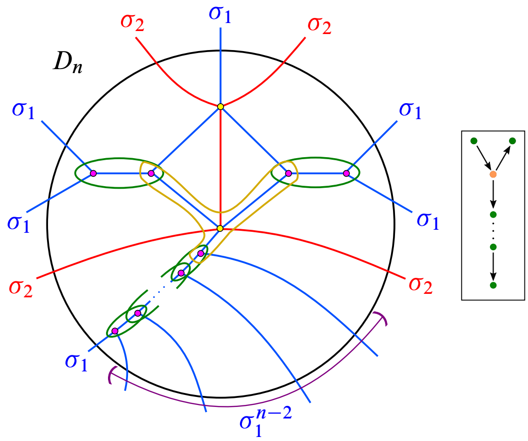


Figure 8: The 3–graph $\Gamma_0(D_n)$ and its associated intersection quiver. The basis $\{\gamma_i^{(0)}\}$ of $H_1\big(\Lambda(\Gamma_0(D_n));\mathbb{Z}\big)$ is given by the orange Y–cycle, the green I–cycles and the $n-3$ I–cycles not pictured.

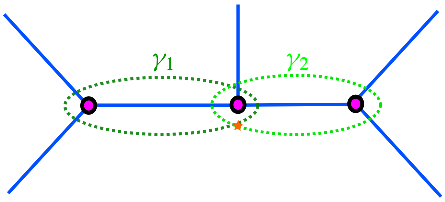


Figure 9: Intersection of two cycles, γ_1 and γ_2. The intersection point is indicated by an orange star. If we orient both cycles counterclockwise, then we will set $\langle\gamma_1,\gamma_2\rangle = -1$ as our convention.

restrict our attention to certain isotopies, referred to as Legendrian surface Reidemeister moves. These moves create specific changes in the Legendrian front $\Sigma(\Gamma)$, known as perestroikas or Reidemeister moves [2]. From [7], we have the following theorem, relating perestroikas of fronts to the corresponding 3–graphs:

Theorem 4 [7, Theorem 4.2] *Let Γ and Γ' be two 3–graphs related by one of the moves shown in Figure 10. Then the associated weaves $\Lambda(\Gamma)$ and $\Lambda(\Gamma')$ are Legendrian isotopic relative to their boundaries.* □

See Figure 11 for a description of the behavior of elements of $H_1(\Lambda(\Gamma);\mathbb{Z})$ under these Legendrian surface Reidemeister moves. In the pair of 3–graphs in Figure 11, top right, we have denoted a push-through by PT or PT^{-1} depending on whether we go from left to right or right to left. This helps us to specify the simplifications we make in the figures in the proof of Theorem 1, as this move is not as readily apparent as the

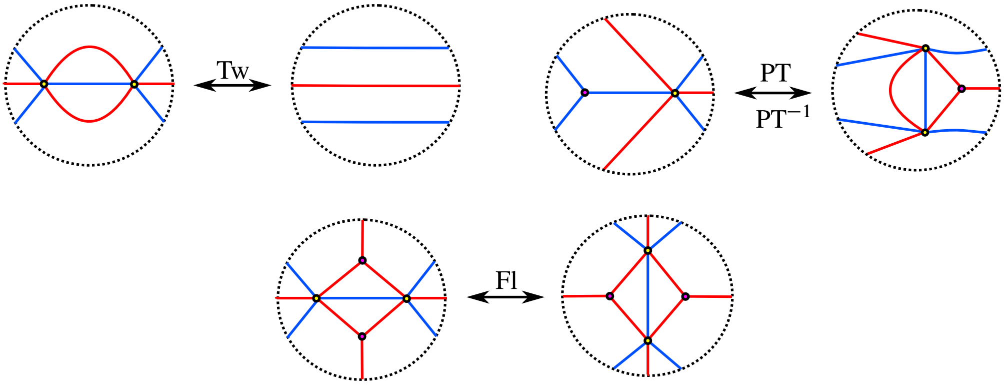


Figure 10: Legendrian surface Reidemeister moves for 3–graphs. From left to right: a candy twist, a push-through and a flop, denoted by Tw, PT and Fl, respectively.

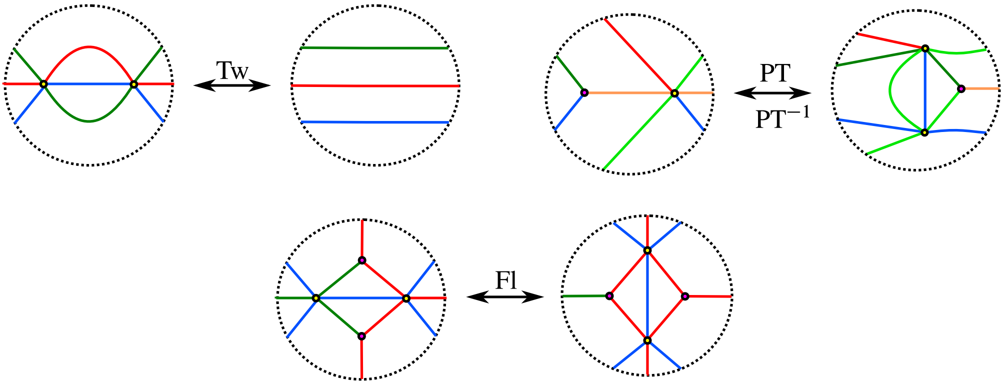


Figure 11: Behavior of certain homology cycles under Legendrian surface Reidemeister moves.

other two. We will refer to the PT^{-1} move as a reverse push-through. Note that an application of this move eliminates the geometric intersection between the light green and dark green cycles in Figure 11.

Remark It is also possible to verify the computations in Figure 11 by examining the relative homology of a cycle. Specifically, if we have a basis of the relative homology $H_1(\Lambda(\Gamma), \partial\Lambda(\Gamma); \mathbb{Z})$, then the intersection form on that basis allows us to determine a given cycle by Poincaré–Lefschetz duality.

2.4 Mutations of 3–graphs

We complete our discussion of general 3–graphs with a description of Legendrian mutation, which we will use to generate distinct exact Lagrangian fillings. Given a Legendrian weave $\Lambda(\Gamma)$ and a 1–cycle $\gamma \in H_1(\Lambda(\Gamma); \mathbb{Z})$, the Legendrian mutation $\mu_\gamma(\Lambda(\Gamma))$ outputs a 3–graph and a corresponding Legendrian weave smoothly isotopic to $\Lambda(\Gamma)$ but whose Lagrangian projection is generally not Hamiltonian isotopic to that of $\Lambda(\Gamma)$.

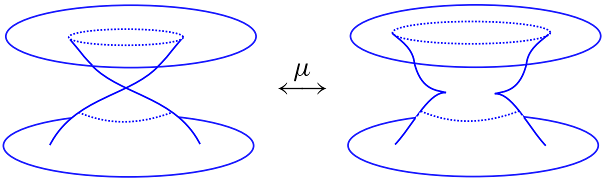


Figure 12: Local fronts for two Legendrian cylinders non-Legendrian isotopic relative to their boundaries.

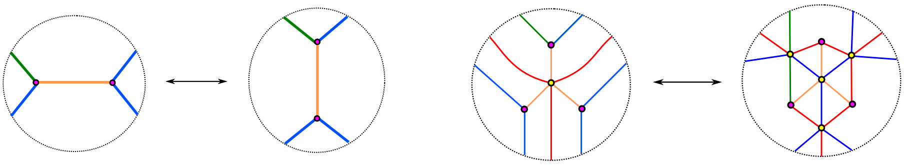


Figure 13: Mutations of a 3–graph. The pair of 3–graphs on the left depicts mutation at the orange I–cycle, while the pair of 3–graphs on the right depicts mutation at the orange Y–cycle. In both cases, the dark green edge depicts the effect of mutation on any cycle intersecting the orange cycle.

Definition 5 Legendrian surfaces $\Lambda_0, \Lambda_1 \subseteq (\mathbb{R}^5, \xi_{\mathrm{st}})$ with equal boundary $\partial\Lambda_0 = \partial\Lambda_1$ are mutation-equivalent if and only if there exists a compactly supported Legendrian isotopy $\{\widetilde{\Lambda}_t\}$ relative to the boundary, with $\widetilde{\Lambda}_0 = \Lambda_0$, and a Darboux ball (B, ξ_{st}) such that:

(i) Outside the Darboux ball, we have $\widetilde{\Lambda}_1|_{\mathbb{R}^5 \setminus B} = \Lambda_1|_{\mathbb{R}^5 \setminus B}$.

(ii) There exists a global front projection $\pi : \mathbb{R}^5 \to \mathbb{R}^3$ such that the pair of fronts $\pi|_{B \cap \widetilde{\Lambda}_1}$ and $\pi|_{B \cap \Lambda_1}$ coincides with the pair of fronts in Figure 12.

We briefly note that these two fronts lift to non-Legendrian isotopic Legendrian cylinders in $(\mathbb{R}^5, \xi_{\mathrm{st}})$, relative to the boundary, and that the 1–cycle we input for our operation is precisely the 1–cycle defined by the cylinder corresponding to Λ_0.

Combinatorially, we can describe mutation as certain manipulations of the edges of our graph. Figure 13, left, depicts mutation at a short I–cycle, while Figure 13, right,

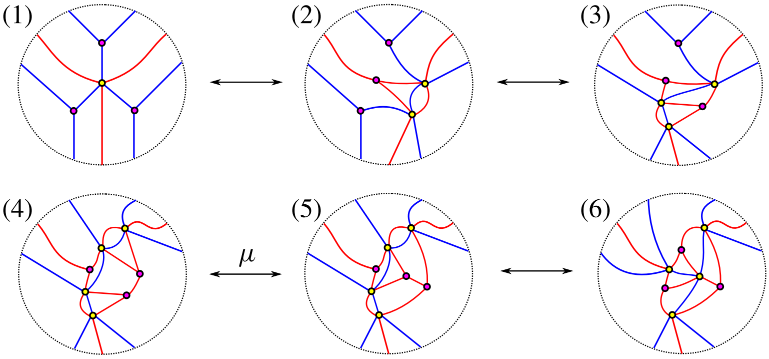


Figure 14: Mutation at a short Y–cycle given as a sequence of Legendrian surface Reidemeister moves and mutation at a short I–cycle. The Y–cycle in the initial 3–graph is given by the three blue edges that each intersect the yellow vertex in the center.

depicts mutation at a short Y–cycle. In the $N = 2$ setting, we can identify 2–graphs with triangulations of an n–gon, in which case mutation at a short I–cycle corresponds to a Whitehead move. In the 3–graph setting, in order to describe mutation at a short Y–cycle, we can first reduce the short Y–cycle case to a short I–cycle, as shown in Figure 14, before applying our mutation. See [7, Section 4.9] for a more general description of mutation at long I– and Y–cycles in the 3–graph.

The geometric operation above coincides with the combinatorial manipulation of the 3–graphs. Specifically, we have the following theorem:

Theorem 6 [7, Theorem 4.2.1] *Given two 3–graphs, Γ and Γ', related by either of the combinatorial moves described in Figure 13, the corresponding Legendrian weaves $\Lambda(\Gamma)$ and $\Lambda(\Gamma')$ are mutation-equivalent relative to their boundary.* □

2.5 Lagrangian fillings from weaves

We now describe in more detail how an exact Lagrangian filling of a Legendrian link arises from a Legendrian weave. If we label all edges of $\Gamma \subseteq \mathbb{D}^2$ colored blue by σ_1 and all edges colored red by σ_2, then the points in the intersection $\Gamma \cap \partial\mathbb{D}^2$ give us a braid word in the Artin generators σ_1 and σ_2 of the 3–stranded braid group. We can then view the corresponding link β as living in $(J^1\mathbb{S}^1, \xi_{\mathrm{st}})$. If we consider our Legendrian weave $\Lambda(\Gamma)$ as an embedded Legendrian surface in $(\mathbb{R}^5, \xi_{\mathrm{st}})$, then, according to our discussion above, it has boundary $\Lambda(\beta)$, where $\Lambda(\beta)$ is the Legendrian satellite of β with companion knot given by the standard unknot. In our local contact model, the projection $\pi\colon (J^1\mathbb{D}^2, \xi_{\mathrm{st}}) \to (T^*\mathbb{D}^2, \omega_{\mathrm{st}})$ gives an immersed exact Lagrangian surface with immersion points corresponding to Reeb chords of $\Lambda(\Gamma)$. If $\Lambda(\Gamma)$ has no Reeb chords, then π is an embedding and $\Lambda(\Gamma)$ is an exact Lagrangian filling of $\Lambda(\beta)$. Since $(\mathbb{S}^3, \xi_{\mathrm{st}})$ minus a point is contactomorphic to $(\mathbb{R}^3, \xi_{\mathrm{st}})$, an embedding of $\Lambda(\Gamma)$ into $(\mathbb{R}^5, \xi_{\mathrm{st}})$ gives an exact Lagrangian filling in $(\mathbb{R}^4, \xi_{\mathrm{st}})$ of $\Lambda(\beta) \subseteq (\mathbb{R}^3, \xi_{\mathrm{st}})$, as it can be assumed — after a Legendrian isotopy — to be disjoint from the point at infinity.

Remark We study embedded — rather than immersed — Lagrangian fillings due to the existence of an h–principle for immersed Lagrangian fillings [10, Theorem 16.3.2]. In particular, any pair of immersed exact Lagrangian fillings is connected by a one-parameter family of immersed exact Lagrangian fillings relative to the boundary. See also [16].

Our desire for embedded Lagrangians motivates the following definition:

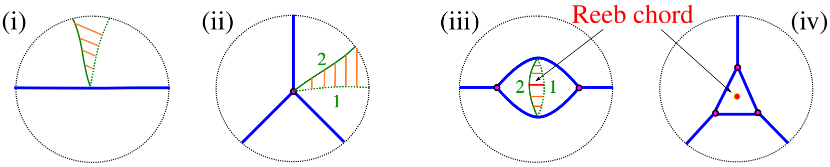


Figure 15: 2–Graphs with a choice of fronts illustrated by the green curves while the solid orange lines illustrate the difference in heights between sheets. A woven front for the pair of 2–graphs on the left can be chosen in such a way that the function giving the difference of heights between the two sheets of the front is 0 on G and increasing towards the boundary. Critical points of the difference function correspond to Reeb chords, so the pair of 2–graphs on the left are free. However, any difference function for the pair of 2–graphs on the right must have at least one critical point inside the face.

Definition 7 A 3–graph $\Gamma \subseteq \mathbb{D}^2$ is free if the associated Legendrian front $\Sigma(\Gamma)$ can be woven with no Reeb chords.

In the $N = 2$ setting, a 2–graph $\Gamma \subseteq \mathbb{D}^2$ is free if and only if G has no bounded faces contained in the interior of $\mathbb{D}^2$. See Figure 15 for examples illustrating this characterization. In the $N = 3$ setting, there is no such simple characterization, but many 3–graphs can be determined to be free by direct inspection, as done in [7, Section 7]. As an example, the 3–graph $\Gamma_0(D_n)$, depicted in Figure 8, is a free 3–graph of D_n–type. This can be verified by taking a woven front for $\Lambda_0(D_n)$ such that the functions giving the difference of heights between the three sheets take the value 0 on G and increase radially towards the boundary. Critical points of these difference functions correspond to Reeb chords. By construction, none of these difference functions have critical points, so $\Gamma_0(D_n)$ can be woven without Reeb chords and is a free 3–graph.

Crucially, the mutation operation described above preserves the free property of a 3–graph.

Lemma 8 [7, Lemma 7.4] *Let $\Gamma \subseteq \mathbb{D}^2$ be a free 3–graph. Then the 3–graph $\mu(\Gamma)$ obtained by mutating according to any of the Legendrian mutation operations given above is also a free 3–graph.* □

Therefore, starting with a free 3–graph and performing the Legendrian mutation operation gives us a method of creating additional embedded exact Lagrangian fillings.

At this stage, we have described the geometric and combinatorial ingredients needed for Theorem 1. The two subsequent subsections introduce the necessary algebraic invariants relating Legendrian weaves and 3–graphs to cluster algebras. These will be used to distinguish exact Lagrangian fillings.

2.6 Quivers from weaves

Before we describe the cluster algebra structure associated to a weave, we must first describe quivers and how they arise via the intersection form on $H_1(\Lambda(\Gamma);\mathbb{Z})$. A quiver is a directed graph without loops or directed 2–cycles. In the Legendrian weave setting, the data of a quiver can be extracted from a given weave and a basis of its first homology. The intersection quiver is defined as follows: each basis element $\gamma_i \in H_1(\Lambda(\Gamma);\mathbb{Z})$ defines a vertex v_i in the quiver and we have k arrows pointing from v_j to v_i if $\langle\gamma_i,\gamma_j\rangle = k$. We will only ever have k either 0 or 1 for quivers arising from fillings of $\lambda(D_n)$. See Figure 2, left, for an example of the quiver $Q\bigl(\Lambda(\Gamma_0(D_4)), \{\gamma_i^{(0)}\}\bigr)$ defined by $\Lambda(\Gamma_0(D_4))$ and the indicated basis for $H_1\bigl(\Lambda(\Gamma_0(D_4);\mathbb{Z})\bigr)$.

The combinatorial operation of quiver mutation at a vertex v is defined as follows; see eg [12]. First, for every pair of incoming edges and outgoing edges, we add an edge starting at the tail of the incoming edge and ending at the head of the outgoing edge. Next, we reverse the direction of all edges adjacent to v. Finally, we cancel any directed 2–cycles. If we start with the quiver Q, then we denote the quiver resulting from mutation at v by $\mu_v(Q)$. See Figure 16, bottom, for an example. Under this operation, we can naturally identify the vertices of Q with $\mu_v(Q)$, just as we can identify the homology bases of a weave before and after Legendrian mutation.

Remark The crucial difference between algebraic and geometric intersections is captured in the step canceling directed 2–cycles. This cancellation is implemented by default in a quiver mutation, as the arrows of the quiver *only* capture algebraic intersections. In contrast, the geometric intersection of homology cycles after a Legendrian mutation will, in general, not coincide with the algebraic intersection. This dissonance will be explored in detail in Section 3.

The following theorem relates the two operations of quiver mutation and Legendrian mutation:

Theorem 9 [7, Section 7.3] *Given a 3–graph Γ, Legendrian mutation at an embedded cycle γ induces a quiver mutation for the associated intersection quivers, taking $Q(\Gamma,\{\gamma_i\})$ to $\mu_\gamma(Q(\Gamma,\{\gamma_i\}))$.* □

See Figure 16 for an example showing the quiver mutation of $Q(\Gamma_0(D_4), \{\gamma_i^{(0)}\})$ for $i \in [1,4]$, corresponding to Legendrian mutation applied to $\Lambda(\Gamma_0(D_4))$.

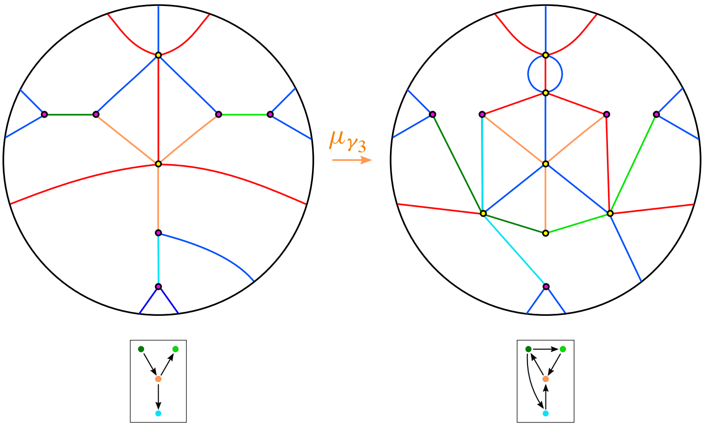


Figure 16: Mutation of $\Gamma_0(D_4)$ and its associated intersection quiver at the short Y–cycle colored in orange. Note that the sign of the intersection between the dark green I–cycle and the orange Y–cycle changes from negative to positive, reflecting the reversal of the arrow in the quiver under mutation.

2.7 Microlocal sheaves and clusters

To introduce the cluster structure mentioned above, we need to define a sheaf-theoretic invariant. We first consider the dg category of complexes of sheaves of $\mathbb{C}$–modules on $\mathbb{D}^2 \times \mathbb{R}$ with constructible cohomology sheaves. For a given 3–graph Γ and its associated Legendrian $\Lambda(\Gamma)$, we denote by $\mathcal{C}(\Gamma) := \mathrm{Sh}^1_{\Lambda(\Gamma)}(\mathbb{D}^2 \times \mathbb{R})_0$ the subcategory of microlocal rank-one sheaves with microlocal support along $\Lambda(\Gamma)$, which we require to be zero in a neighborhood of $\mathbb{D}^2 \times \{-\infty\}$. Here we identify the unit cotangent bundle $T^{\infty,-}(\mathbb{D}^2 \times \mathbb{R})$ with the first jet space $J^1(\mathbb{D}^2)$. With this identification, the sheaves of $\mathcal{C}(\Gamma)$ are constructible with respect to the stratification given by the Legendrian front $\Sigma(\Gamma)$. Work of Guillermou, Kashiwara and Schapira implies that $\mathcal{C}(\Gamma)$ is an invariant under Hamiltonian isotopy [17].

As described in [7, Section 5.3], this category has a combinatorial description. Given a 3–graph Γ, the data of the moduli space of microlocal rank-one sheaves is equivalent to the following:

(i) Assign to each face F (connected component of $\mathbb{D}^2 \setminus G$) of a flag $\mathcal{F}^\bullet(F)$ in the vector space $\mathbb{C}^3$.

(ii) For each pair F_1 and F_2 of adjacent faces sharing an edge labeled by σ_i, we require that the corresponding flags satisfy

$$\mathcal{F}^j(F_1) = \mathcal{F}^j(F_2), \quad 0 \le j \le 3,\ j \ne i, \qquad \text{and} \qquad \mathcal{F}^i(F_1) \ne \mathcal{F}^i(F_2).$$

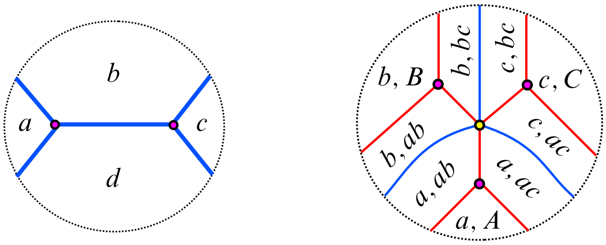


Figure 17: The data of the flag moduli space given in the neighborhood of a short I–cycle (left) and a short Y–cycle (right). Lines are represented by lowercase letters, while planes are written in uppercase. The intersection of the two lines a and b is written as ab.

Finally, we consider the moduli space of flags satisfying (i) and (ii) modulo the diagonal action of GL_n on $\mathcal{F}^\bullet$. The precise statement [7, Theorem 5.3] is that the flag moduli space, denoted by $\mathcal{M}(\Gamma)$, is isomorphic to the space of microlocal rank-one sheaves $\mathcal{C}(\Gamma)$. Since $\mathcal{C}(\Gamma)$ is an invariant of $\Lambda(\Gamma)$ up to Hamiltonian isotopy, it follows that $\mathcal{M}(\Gamma)$ is an invariant as well. In the I–cycle case, when the edges are labeled by σ_1, the moduli space is determined by four lines $a \neq b \neq c \neq d \neq a$, as pictured in Figure 17, left. If the edges are labeled by σ_2, then the data is given by four planes $A \neq B \neq C \neq D \neq A$. Around a short Y–cycle, the data of the flag moduli space is given by three distinct planes $A \neq B \neq C \neq A$ contained in $\mathbb{C}^3$ and three distinct lines $a \subsetneq A$, $b \subsetneq B$ and $c \subsetneq C$ with $a \neq b \neq c \neq a$, as pictured in Figure 17, right.

To describe the cluster algebra structure on $\mathcal{C}(\Gamma)$, we need to specify the cluster seed associated to the quiver $Q(\Lambda(\Gamma), \{\gamma_i\})$ via the microlocal monodromy functor μ_{mon}, which is a functor from the category $\mathcal{C}(\Gamma)$ to the category of rank one local systems on $\Lambda(\Gamma)$. As described in [24; 23], the functor μ_{mon} takes a 1–cycle as input and outputs the isomorphism of sheaves given by the monodromy about the cycle. Since it is locally defined, we can compute the microlocal monodromy about an I–cycle or Y–cycle using the data of the flag moduli space in a neighborhood of the cycle. If we have a short I–cycle γ with flag moduli space described by the four lines a, b, c and d, as in Figure 17, left, then the microlocal monodromy about γ is given by the cross ratio

$$\frac{a \wedge b}{b \wedge c} \frac{c \wedge d}{d \wedge a}.$$

Similarly, for a short Y–cycle with flag moduli space given as in Figure 17, right, the microlocal monodromy is given by the triple ratio

$$\frac{B(a)C(b)A(c)}{B(c)C(a)A(b)}.$$

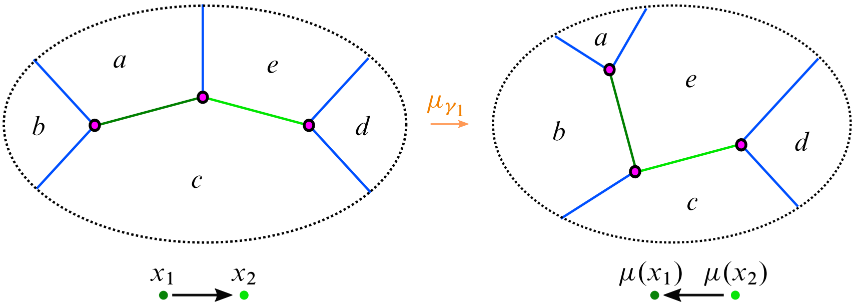


Figure 18: Prior to mutating at γ_1, we have $\langle\gamma_1,\gamma_2\rangle = -1$. Computing the cross ratios for γ_1 and $\mu_1(\gamma_1)$, we can see that the cross ratio transforms as $\mu_1(\gamma_1) = (b\wedge c/c\wedge e)(e\wedge a/a\wedge b) = x_1^{-1}$ under mutation. Similarly, computing the cross ratios for γ_1 and $\mu_1(\gamma_2)$ and applying the relation $e\wedge b\cdot a\wedge c = b\wedge c\cdot e\wedge a + a\wedge b\cdot c\wedge e$, we have $\mu_1(x_2) = (e\wedge a/a\wedge c)(c\wedge d/d\wedge e)\bigl(1+(a\wedge b/b\wedge c)(c\wedge e/e\wedge a)\bigr)$.

As described in [7, Section 7.2], the microlocal monodromy about a 1–cycle gives rise to an X–cluster variable at the corresponding vertex in the quiver. Under mutation of the 3–graph, the cross ratio and triple ratio transform as cluster X–coordinates. Specifically, if we start with a 3–graph with cluster variables x_j, then the cluster variables x_j' of the 3–graph after mutating at γ_i are given by

$$x_j' = \begin{cases} x_j^{-1} & \text{if } i = j, \\ x_j(1+x_i^{-1})^{-\langle\gamma_i,\gamma_j\rangle} & \text{if } \langle\gamma_i,\gamma_j\rangle > 0, \\ x_j(1+x_i)^{-\langle\gamma_i,\gamma_j\rangle} & \text{if } \langle\gamma_i,\gamma_j\rangle < 0. \end{cases}$$

See Figure 18 for an example.

The goal of the next section will be to realize each possible mutation of the D_n quiver as a mutation of the corresponding 3–graph. This will imply that there are at least as many exact Lagrangian fillings as cluster seeds of D_n–type. There exists a complete classification of all finite mutation type cluster algebras and, in fact, the number of cluster seeds of D_n–type is $(3n-2)C_{n-1}$ [13].

Remark Other than Legendrian weaves, it is not known whether methods of generating exact Lagrangian fillings of $\lambda(D_n)$ access all possible cluster seeds of D_n–type. When constructing fillings of D_4 by opening crossings, as in [9; 21], experimental evidence suggests that it is only possible to access at most 46 out of the possible 50 cluster seeds by varying the order of the crossings chosen. Of note in the combinatorial setting, we also contrast the 3–graphs $\Gamma(D_4)$ with double wiring diagrams for the torus link $T(3,3)$, which is the smooth type of $\lambda(D_4)$. The moduli of sheaves $\mathcal{C}(\Gamma(D_4))$

for $\Gamma(D_4)$ embeds as an open positroid cell into the Grassmannian $\mathrm{Gr}(3,6)$ [5], so we can identify some cluster charts with double wiring diagrams. The double wiring diagrams associated to $\mathrm{Gr}(3,6)$ only access 34 out of 50 distinct cluster seeds via local moves applied to an initial double wiring diagram [12].

3 Proof of the main results

In this section, we state and prove Theorem 11, which implies Theorem 1. The following definitions relate the algebraic intersections of cycles to geometric intersections in the context of 3–graphs.

Definition 10 A 3–graph Γ with associated basis $\{\gamma_i\}$ for $i \in [1, b_1(\Lambda(\Gamma))]$ of $H_1(\Lambda(\Gamma);\mathbb{Z})$ is *sharp at a cycle* γ_j if, for any other cycle $\gamma_k \in \{\gamma_i\}$, the geometric intersection number of γ_j with γ_k is equal to the algebraic intersection $\langle \gamma_j, \gamma_k \rangle$.

Γ is *locally sharp* if, for any cycle $\gamma \in \{\gamma_i\}$, there exists a sequence of Legendrian surface Reidemeister moves taking Γ to some other 3–graph Γ' such that Γ' is sharp at the corresponding cycle $\gamma' \in H_1(\Lambda(\Gamma');\mathbb{Z})$.

A 3–graph Γ with a set of cycles Γ is *sharp* if Γ is sharp at all $\gamma_i \in \{\gamma_i\}$.

For 3–graphs that are not sharp, it is possible that a sequence of mutations will cause a cycle to become immersed. This is the only obstruction to weave-realizability. Therefore, sharpness is a desirable property for our 3–graphs, as it simplifies our computations and helps us avoid creating immersed cycles. We will not be able to ensure sharpness for all $\Gamma(D_n)$ that arise as part of our computations (eg see the Type III.i normal form in Figure 20), but we will be able to ensure that each of our 3–graphs is locally sharp.

3.1 Proof of Theorem 1

The following result is slightly stronger than the statement of Theorem 1, as we are able to show that each 3–graph in our sequence of mutations is locally sharp:

Theorem 11 *Let $\mu_{v_1}, \dots, \mu_{v_k}$ be a sequence of quiver mutations, with initial quiver $Q(\Gamma_0(D_n), \{\gamma_i^{(0)}\})$. Then there exists a sequence $\Gamma_0(D_n), \dots, \Gamma_k(D_n)$ of 3–graphs such that:*

(i) *$\Gamma_{j-1}(D_n)$ is related to $\Gamma_j(D_n)$ by mutation at a cycle γ_j and by Legendrian surface Reidemeister moves I, II and III. The cycle γ_j represents the vertex v_j*

in the intersection quiver and it is given by one of the cycles in the initial basis $\{\gamma_i^{(0)}\}$ after mutation and Reidemeister moves.

(ii) *$\Gamma_j(D_n)$ is sharp at γ_j.*

(iii) *$\Gamma_j(D_n)$ is locally sharp.*

(iv) *The basis of cycles for $\Gamma_j(D_n)$, obtained from the initial basis $\{\gamma_i^{(0)}\}$ by mutation and Reidemeister moves, consists entirely of short* Y*–cycles and short* I*–cycles.*

The conditions (ii)–(iv) allow us to continue to iterate mutations after applying a small number of simplifications at each step. Theorem 1 thus follows from Theorem 11.

Proof We proceed by organizing the 3–graphs arising from any sequence of mutations of $\Gamma_0(D_n)$ into four types, in line with the organization scheme introduced by Vatne for quivers of D_n–type [26]. Vatne's classification of quivers in the mutation class of D_n–type uses the configuration of a certain subquiver to define the different types. Outside of that subquiver, there are a number of disjoint subquivers of A_n–type that are referred to as A_n tail subquivers. We will refer to the corresponding cycles in the 3–graph as A_n tail subgraphs, or simply A_n tails when it is clear from context whether we are referring to the quiver or the 3–graph. For each type, Vatne describes the results of quiver mutation at different vertices, which can depend on the existence of A_n tail subquivers. See Figures 21, 27, 31 and 35 for the four types and their mutations.

Notation As mentioned in the previous section, cycles are pictured as colored edges for the sake of visual clarity. Throughout this section, we denote all of the dark green cycles by γ_1, light green cycles by γ_2, orange cycles by γ_3, light blue cycles by γ_4, pink cycles by γ_5, purple cycles by γ_6, and olive cycles by γ_7. With this notation, γ_i will correspond to the vertex labeled by v_i in the quivers given below.

***A_n* tails** We briefly describe the behavior of the A_n tail subquivers, as given in [26], in terms of weaves. Any of the n vertices in an A_n tail subquiver can have valence between 0 and 4. Cycles in the quiver are oriented with length 3. If a vertex v has valence 3, then two of the edges form part of a 3–cycle, while the third edge is not part of any 3–cycle. If v has valence 4, then two of the edges belong to one 3–cycle and the remaining two edges belong to a separate 3–cycle.

Any A_n tail of the quiver can be represented by a sharp configuration of n I–cycles in the 3–graph. See Figure 19 for an identification of I–cycles with quiver vertices of a given valence. Mutation at any vertex v_i in the quiver corresponds to mutation at the I–cycle γ_i in the 3–graph, so it is readily verified that mutation preserves the

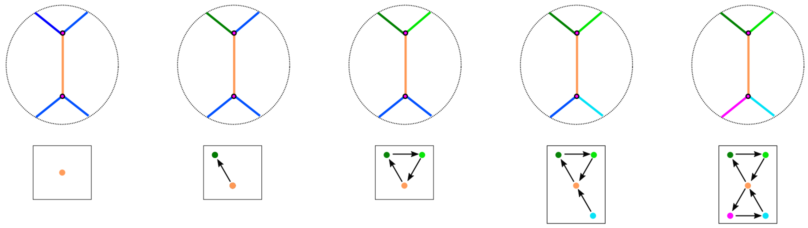


Figure 19: All possible arrangements of I–cycles in an A_n tail of the 3–graph corresponding to a given vertex in the A_n tail subquiver of valence between 0 and 4.

number of I–cycles and requires no application of Legendrian surface Reidemeister moves to simplify. The sequences of mutations given in the remainder of the proof As a consequence, any sequence of A_n tail mutations is weave-realizable, and a sharp 3–graph remains sharp after mutation at A_n tail I–cycles that only intersect other A_n tail I–cycles.

Normal forms For each of the four types of D_n quivers described in [26], we give a set of specific subgraphs of $\Gamma(D_n)$, which we refer to as normal forms. These normal

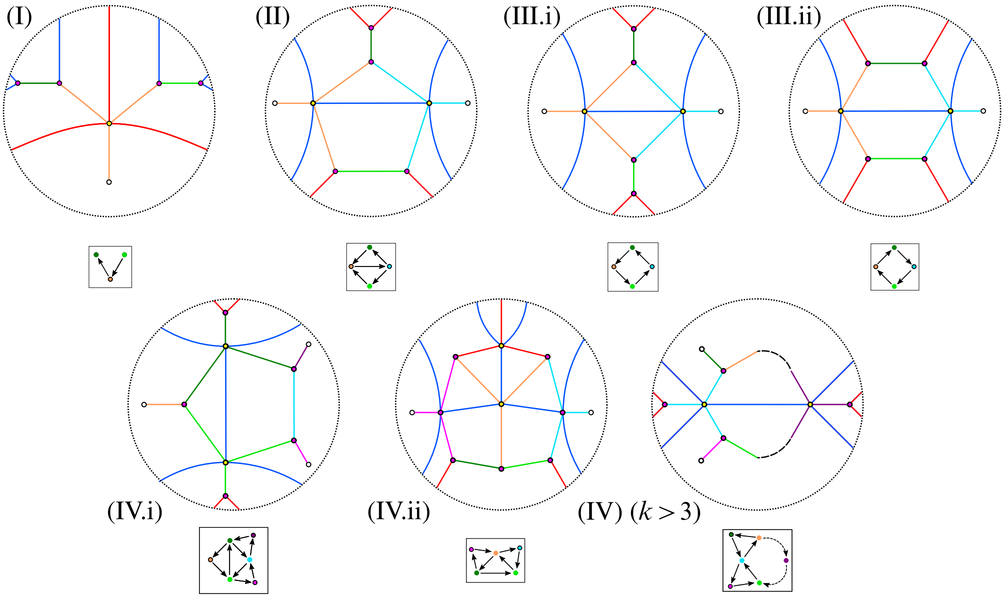


Figure 20: Normal forms labeled by their type. The possible addition of I–cycles corresponding to A_n tails of the quiver are represented by unfilled circles appended to the end of edges that do not intersect the boundary.

forms are pictured in Figure 20. We indicate the possible existence of A_n tail subgraphs by an unfilled circle. In our discussion below, we will say that an edge of the 3–graph carries a cycle if it is part of a homology cycle. We will generally use this terminology to specify which edges cannot carry a cycle.

For each possible quiver mutation, we describe the possible mutations of the 3–graph and show that the result matches the quiver type and retains the properties listed in Theorem 11 above. In addition, the Legendrian surface Reidemeister moves we describe ensure that the A_n tail subgraphs continue to consist solely of short I–cycles. If the mutation results in a long I–cycle or pair of long I–cycles connecting our A_n tail to the rest of the 3–graph, we can simplify by applying a sequence of n push-throughs to ensure that these are all short I–cycles. It is readily verified that we can always do this and that no other simplifications of the A_n tails are required following any other mutations. We include A_n tail cycles only where relevant to the specific mutation. In our computations below, we generally omit the final steps of applying a series of push-throughs to make any long I– or Y–cycles into short I– or Y–cycles. Figure 26 provides an example where these push-throughs are shown for both an I–cycle and a Y–cycle.

In order to simplify the overall presentation of the normal forms and the computations below, we allow for the following variations in the Type I and Type IV cases. In the Type I case, mutating at either of the short I–cycles γ_1 or γ_2 in the Type I normal form produces one of four possible configurations of the cycles γ_1, γ_2 and γ_3 in a 3–graph corresponding to a Type I quiver. Since these mutations are readily computed, we simplify our presentation by giving a single normal form rather than four, and describing the relevant mutations of two of the four possible 3–graphs in Figures 22, 23, 24 and 25. The remaining cases can be seen by swapping the cycle(s) to the left of the short Y–cycle with the cycle(s) to the right of it. This symmetry corresponds to reversing all of the arrows in the quiver. In general, we will implicitly appeal to similar symmetries of the normal form 3–graphs to reduce the number of cases we must consider. In the Type IV case, the edge(s) corresponding to γ_3, γ_5 or γ_6 need not carry a cycle. See the discussion of Type IV quiver mutations below for a more detailed description.

Type I We start with 3–graphs, always endowed with a homology basis, whose associated intersection quivers are a Type I quiver. See Figure 21 for the relevant quiver mutations.

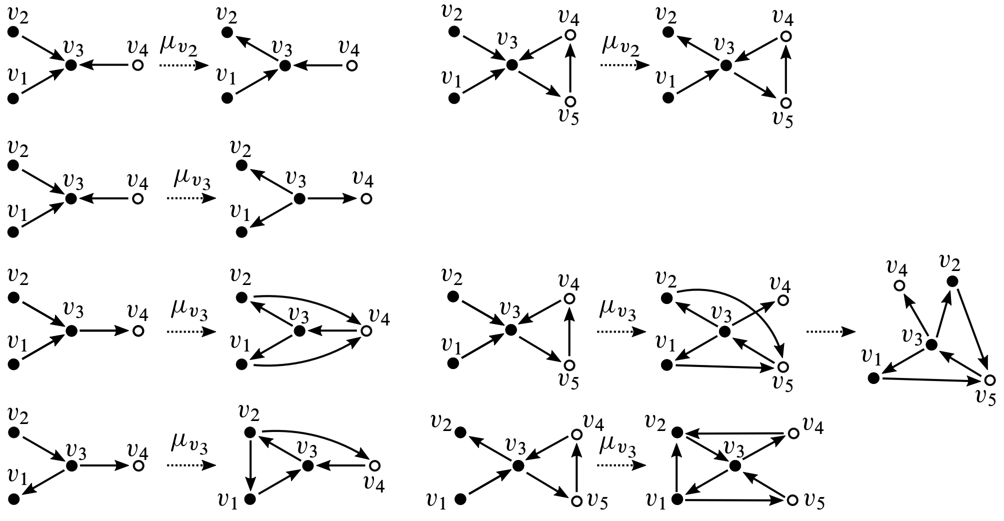


Figure 21: From top to bottom: two Type I to Type I quiver mutations, Type I to Type II quiver mutations, and Type I to Type IV quiver mutations. The arrow labeled by μ_{v_i} indicates mutation at the vertex v_i. Unfilled circles represent potential A_n tails. In each line, the first quiver mutation shows the case where v_3 is only adjacent to one A_n tail vertex, while the second quiver mutation shows the case where v_3 is adjacent to two A_n tail vertices. Note that reversing the direction of all of the arrows simultaneously before mutating gives additional possible quiver mutations of the same type.

Type I to Type I There are two possible Type I to Type I mutations of 3–graphs, depicted in Figure 22. As shown in Figure 22, left, mutation at γ_1 only affects the sign of the intersection of γ_1 with the γ_3. This reflects the fact that the corresponding quiver mutation has only reversed the orientation of the edge between v_1 and v_3. Mutating at any other I–cycle is equally straightforward and yields a Type I to Type I mutation as well.

Type I to Type I For the second possible Type I to Type I mutation, we proceed as pictured in Figure 22, right. Mutation at γ_3 does not create any new additional geometric or algebraic intersections. Instead, it takes positive intersections to negative intersections and vice versa. This is reflected in the quivers pictured underneath the 3–graphs, as the orientation of edges has reversed under the mutation. As explained above, we could simplify the resulting 3–graph by applying a push-through move to each of the long I–cycles to get a sharp 3–graph where the homology cycles are made up of a single short Y–cycle and some number of short I–cycles.

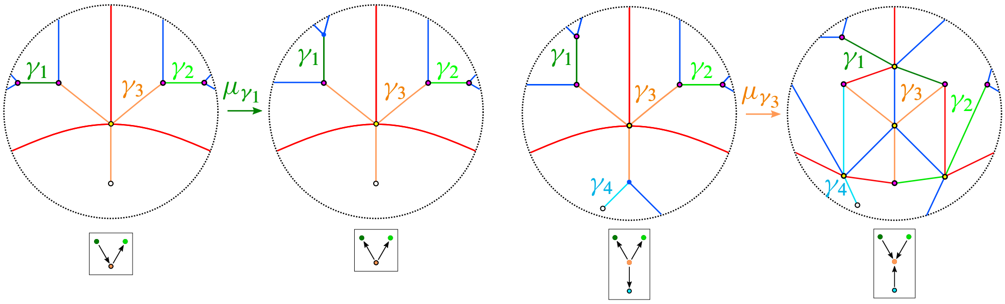


Figure 22: Type I to Type I mutation. Arrows labeled by μ indicate mutation at a cycle of the same color.

Type I to Type II In Figure 23 we consider the cases where the Y–cycle γ_3 intersects one I–cycle (top) or two I–cycles (bottom) in the A_n tail subgraph. Mutation at γ_3

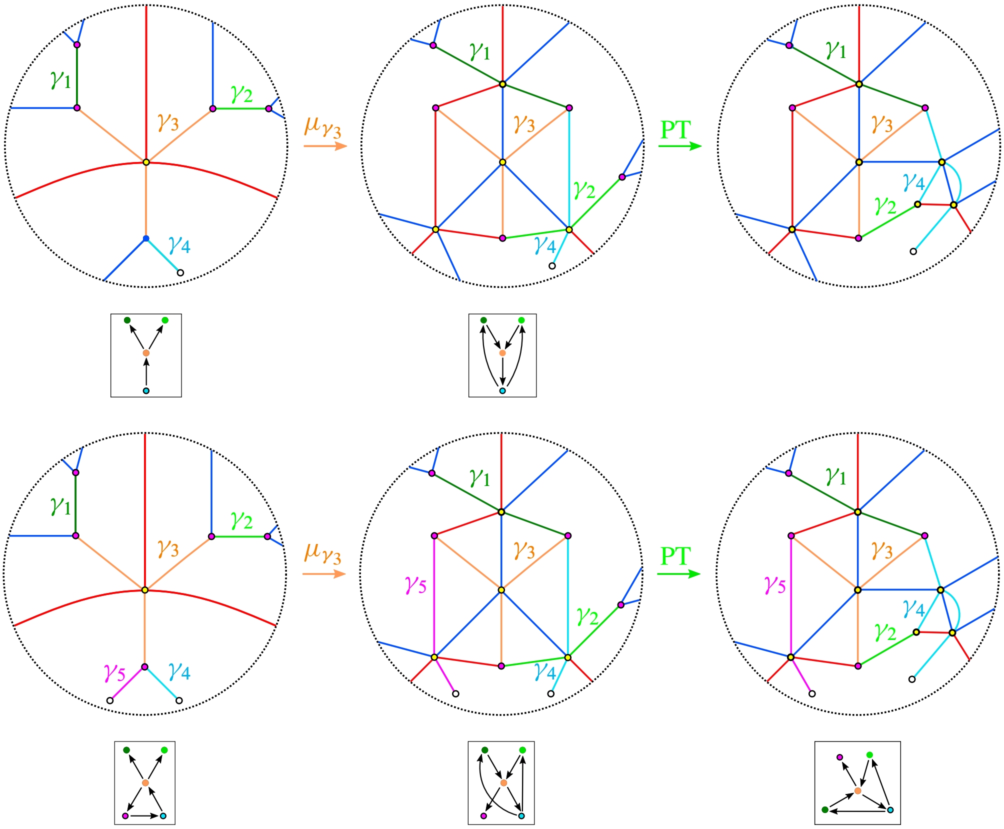


Figure 23: Type I to Type II mutations. Legendrian surface Reidemeister are moves labeled as in Theorem 4 and Figure 10.

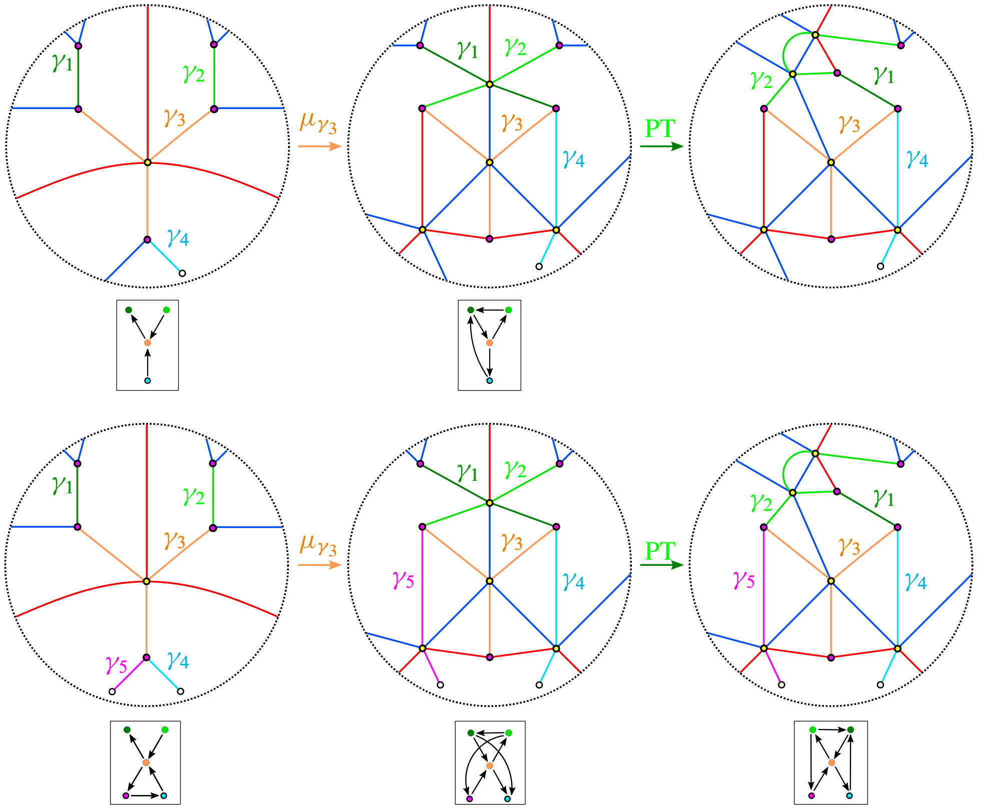


Figure 24: Type I to Type IV.i mutations.

introduces an intersection between γ_2 and γ_4 that causes the second 3–graph in of each mutation sequences to no longer be sharp. Applying a push-through to γ_2 resolves this intersection so that the geometric intersection between γ_2 and γ_4 matches their algebraic intersection. This simplification ensures that the result of μ_{γ_3} is a sharp 3–graph that matches the Type II normal form. If we compare the mutations in the top and bottom sequences, we can see that the presence of the A_n tail cycle γ_5 does not affect the computation.

Type I to Type IV.i We now consider the first of two Type I to Type IV mutations, shown in Figure 24. Starting with the configuration of cycles at the left of each sequence and mutating at γ_3 causes γ_1 and γ_2 to cross. Applying a push-through to γ_1 or to γ_2 (not pictured) simplifies the resulting intersection and yields a Type IV.i normal form made up of the cycles γ_1, γ_2, γ_3 and γ_4. The sequences on the top and bottom of Figure 24 differ only by the presence of the A_n tail cycle γ_5.

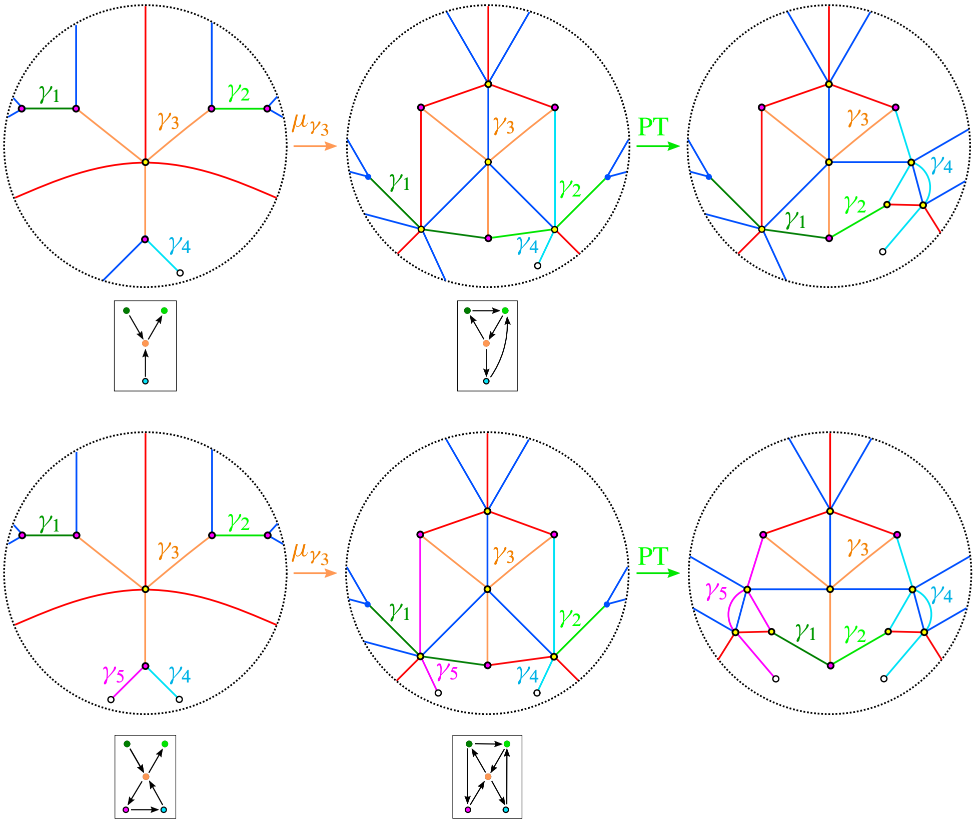


Figure 25: Type I to Type IV.ii mutations.

Type I to Type IV.ii In Figure 25, we consider the cases where γ_1 intersects one I–cycle (top) or two I–cycles (bottom) in the A_n tail subgraph, as we did in the Type I to Type II case. As in the Type I to Type II case, we must apply a push-through to resolve the new intersections in between that cause the second 3–graph in each sequence to fail to be sharp. When we include both γ_4 and γ_5 in the sequence on the right, we get two new intersections after mutating, and therefore require two push-throughs. Note that, in the Type IV.ii case, we must first apply the push-through to γ_1 and γ_2 in order to ensure that we can apply a push-through to any additional cycles in the A_n tail subgraph. This causes the Y–cycles of the graph to correspond to different vertices in the quiver than in the Type IV.i normal form, which is the main reason we distinguish between the normal forms for Type IV.i and Type IV.ii.

The above cases describe all possible mutations of the Type 1 normal form. Each of these mutations yields a sharp 3–graph with short I–cycles and Y–cycles, as desired.

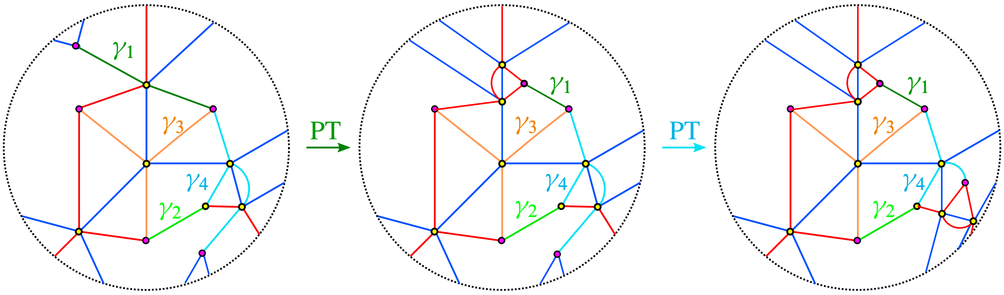


Figure 26: Push-through examples. The first push-through move simplifies the long I–cycle labeled γ_1, while the second simplifies the long Y–cycle labeled γ_4.

In Figure 26, we show how to apply push-throughs to completely simplify the long I– and Y–cycles pictured in the Type I to Type IV.ii graph. As mentioned above, these push-throughs are identical to any other computation required to simplify our resulting 3–graphs to a set of short I– and Y–cycles.

Type II We now consider mutations of our Type II normal form. See Figure 27 for the relevant quivers. As shown in the figure, performing a quiver mutation at the 2–valent vertices labeled by v_1 or v_2 yields a Type III quiver, while a quiver mutation at the vertices labeled v_3 or v_4 yields either another Type II quiver or a Type I quiver, depending on the intersection of v_3 or v_4 with any A_n tail subquivers.

Type II to Type I We first consider the sequence of 3–graphs pictured in Figure 28. Mutation at γ_4 results in a new geometric intersection between γ_2 and γ_3 even though

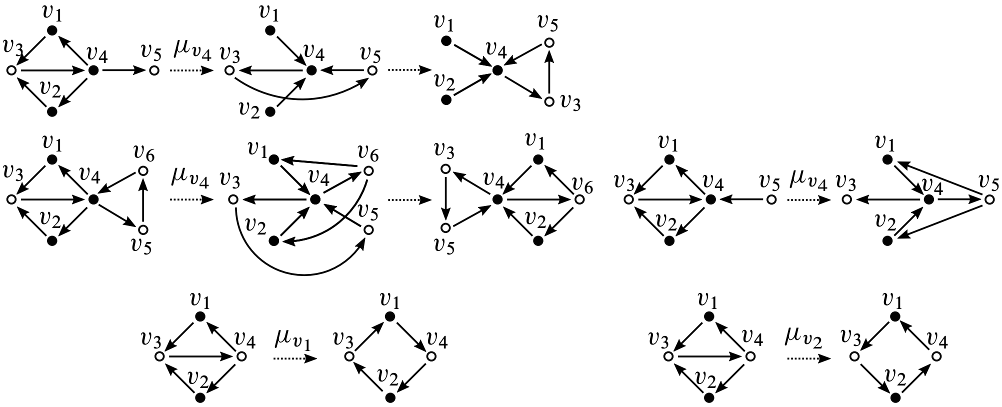


Figure 27: From top to bottom: Type II to Type I, Type II to Type II and Type II to Type III quiver mutations.

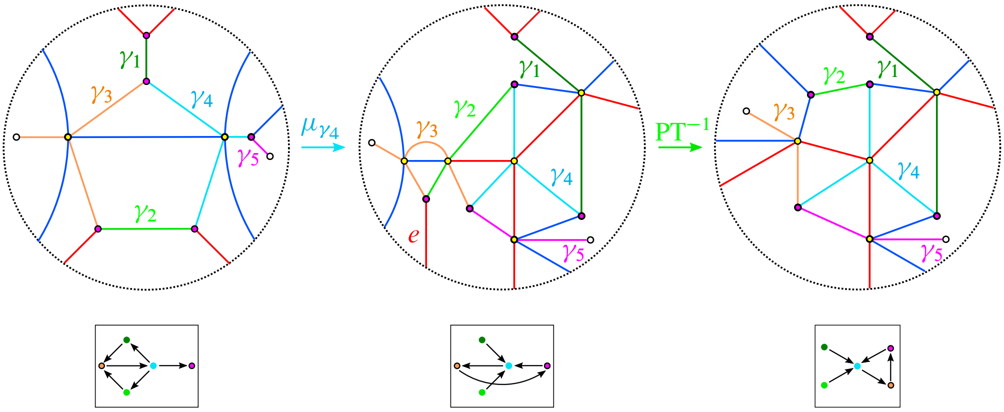


Figure 28: Type II to Type I mutations. The red e labels an edge in the 3–graph that does not carry a cycle.

their algebraic intersection is zero. We can resolve this by applying a reverse push-through at the trivalent vertex where γ_2 and γ_3 meet. The resulting 3–graph is sharp, as γ_2 and γ_3 no longer have any geometric intersection. This computation is identical if γ_3 were to intersect a single A_n tail cycle and we mutated at γ_3 instead. Note that here we require the adjacent red edge labeled e to not carry a cycle, as specified by our normal form.

Type II to Type II We now consider the sequence shown in Figure 29. After mutating at γ_4, we have the same intersection between γ_2 and γ_3 as in the previous case. We again resolve this intersection by applying a reverse push-through at the same trivalent vertex. In this case, we also have an intersection between γ_1 and γ_6, which we resolve via push-through of γ_1. As a result, γ_6 becomes a Y–cycle, and the Type II normal form is now made up of the cycles γ_1, γ_2, γ_4 and γ_6, while γ_3 becomes an A_n tail cycle.

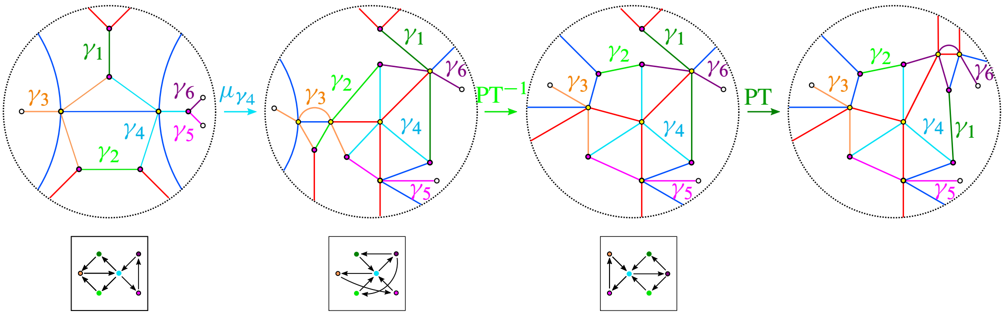


Figure 29: Type II to Type II mutations.

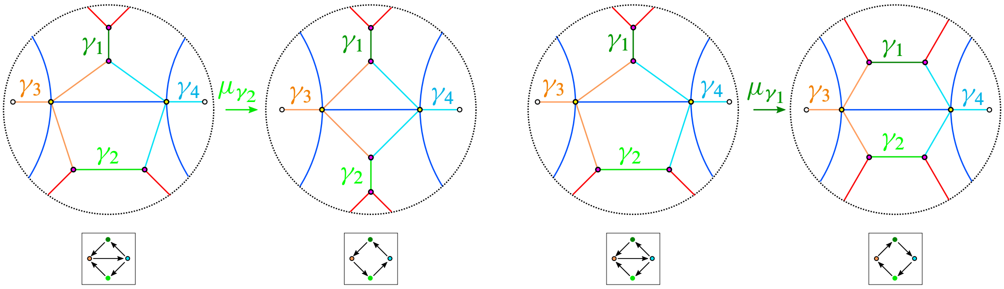


Figure 30: Type II to Type III mutations.

Type II to Type III.i Mutation at γ_1 or γ_2 in the Type II normal form yields either of the Type III normal forms. In the sequence in Figure 30, left, mutation at γ_2 leads to a geometric intersection between γ_3 and γ_4 at two trivalent vertices. Since the signs of these two intersections differ, the algebraic intersection $\langle \gamma_3, \gamma_4 \rangle$ is zero, so the resulting 3–graph is not sharp. However, it is sharp at γ_1 and γ_2, and applying a flop to the 3–graph removes the geometric intersection between γ_3 and γ_4 at the cost of introducing the same intersection between γ_1 and γ_2. Therefore, applying the flop does not make the 3–graph sharp, but it does show that the 3–graph resulting from our mutation is locally sharp at every cycle.

Type II to Type III.ii In the sequence in Figure 30, right, mutation at γ_1 yields a sharp 3–graph that matches the Type III.ii normal form.

Type III Figure 31 illustrates the Type III quiver mutations. Figures 32, 33 and 34 depict the corresponding Legendrian mutations of the Type III normal forms.

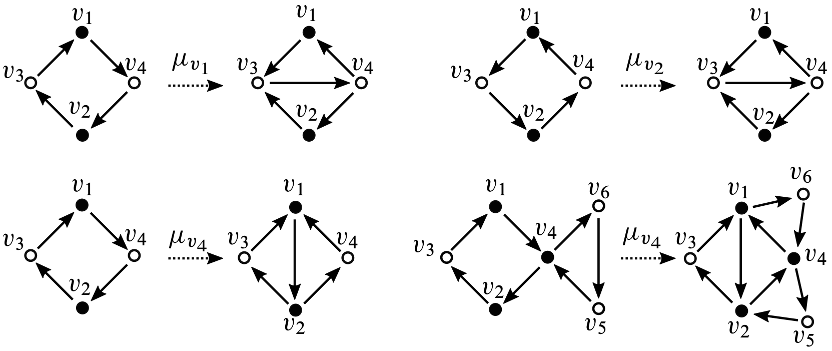


Figure 31: Type III to Type II quiver mutations (top) and Type III to Type IV quiver mutations (bottom).

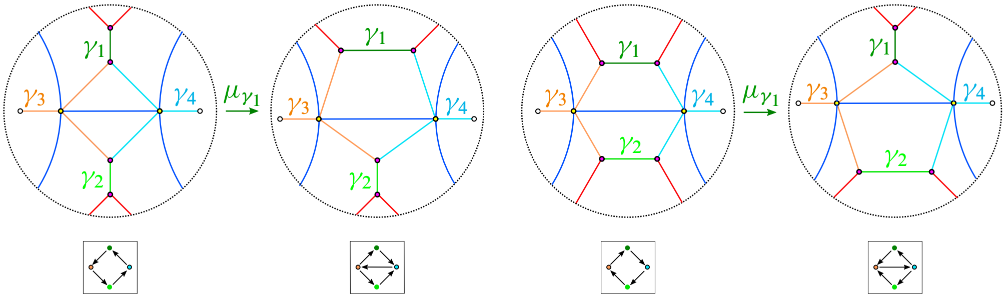


Figure 32: Type III.i to Type II mutations (left) and Type III.ii to Type II mutations (right).

Type III.i to Type II We first consider the sequence of 3–graphs in Figure 32, left. Mutating at γ_1 or γ_2 immediately yields a Type II normal form. Mutating at γ_1 and γ_2 in succession yields a Type III.ii normal form. Note that, if the 3–graph were not sharp at γ_1 or γ_2, we would first need to apply a flop. We can always apply this move because the 3–graph is locally sharp at each of its cycles. See the Type III.i to Type IV.i subcase below for an example where we demonstrate this move.

Type III.ii to Type II In the sequence in Figure 32, right, mutation at either γ_1 or γ_2 yields a Type II normal form. Mutation at γ_1 and γ_2 in succession yields a Type III.i normal form. Therefore, applying these two moves in succession can take us between both of our Type III normal forms.

Type III.i to Type IV We now consider the sequence of 3–graphs in Figure 33. Since the initial 3–graph is not sharp at γ_4, we must first apply a flop before mutating. After

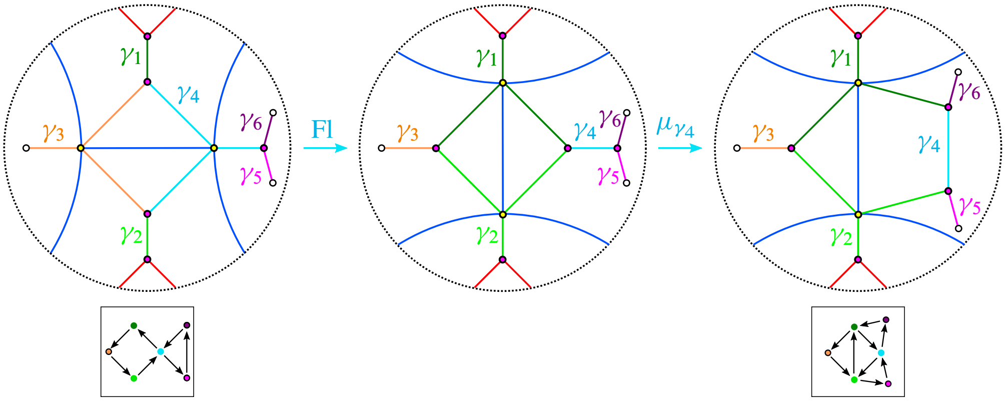


Figure 33: Type III.i to Type IV mutations.

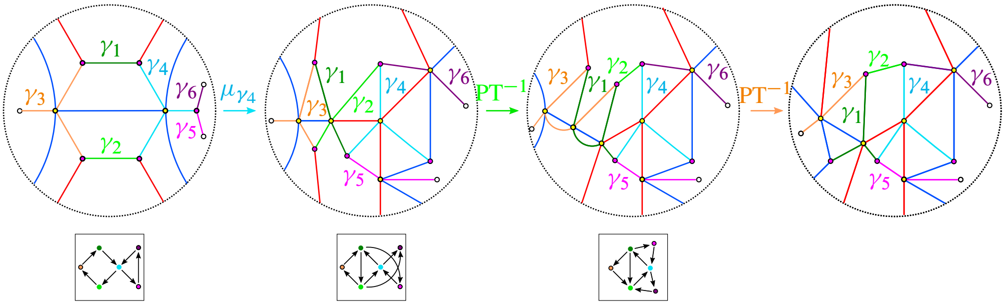


Figure 34: Type III.ii to Type IV mutations.

applying this flop, γ_4 is a short I–cycle and the 3–graph is sharp at γ_4. Mutating at γ_4 then yields a Type IV.i normal form. The short I–cycles γ_5 and γ_6 are included to indicate where any A_n tail cycles would be sent under this mutation.

Type III.ii to Type IV In Figure 34, mutation at γ_4 causes γ_1 and γ_2 to cross while still intersecting γ_3 and γ_4 at either end. We resolve this by first applying a push-through to γ_2 and then applying a reverse push-through to the trivalent vertex where γ_1 and γ_3 intersect a red edge. This results in a sharp 3–graph with γ_1, γ_2, γ_3 and γ_4 making up the Type IV normal form. We again include γ_5 and γ_6 as cycles belonging to a potential A_n tail subgraph in order to show where the A_n tail cycles are sent under this mutation.

Type IV Figure 35 illustrates all of the relevant Type IV quivers and their mutations. In general, the edges of a Type IV quiver have the form of a single k–cycle with the possible existence of 3–cycles or outward-pointing "spikes" at any of the edges along the k–cycle. At the tip of each of these spikes is a possible A_n tail subquiver. We will refer to a vertex at the tip of any of the spikes (eg the vertex v_3 in Figure 35) as a spike vertex and any vertex along the k–cycle will be referred to as a k–cycle vertex. A homology cycle corresponding to a spike vertex will be referred to as a spike cycle. Mutating at a spike vertex increases the length of the internal k–cycle by one, while mutating at a k–cycle vertex decreases the length by 1, so long as $k > 3$. Figures 36, 37, 38 and 39 illustrate the corresponding mutations of 3–graphs for Type IV to Type I and Type IV to Type III when $k = 3$.

Type IV.i to Type I We first consider the sequence of 3–graphs in Figure 36. Mutation at γ_1 causes γ_2 and γ_4 to cross. Application of a reverse push-through at the trivalent vertex where γ_2 and γ_4 intersect a red edge removes this crossing and yields a Type I normal form where γ_1 is the sole Y–cycle.

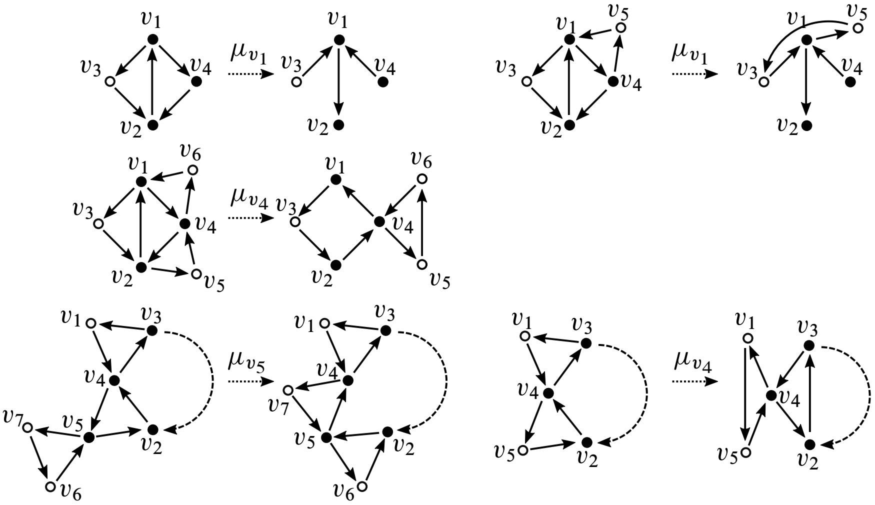


Figure 35: From top to bottom: Type IV to Type I, Type IV to Type III, and Type IV spike vertex (left) and cycle vertex (right) quiver mutations. The presence or absence of the A_n tail vertices v_6 and v_7 in the quiver mutation depicted in the first column, third row corresponds to the presence or absence of spikes appearing in the resulting quiver.

Type IV.ii to Type I Mutation at γ_3 in Figure 37 yields a 3–graph with geometric intersections between γ_1 and γ_5 and between γ_2 and γ_4. The application of reverse push-throughs at the trivalent vertex intersections of γ_1 with γ_5 and γ_2 with γ_4 removes these geometric intersections, resulting in a Type I normal form where γ_1 is the sole Y–cycle. We also apply a candy twist (Legendrian surface Reidemeister move I) to simplify the intersection at the top of the resulting 3–graph.

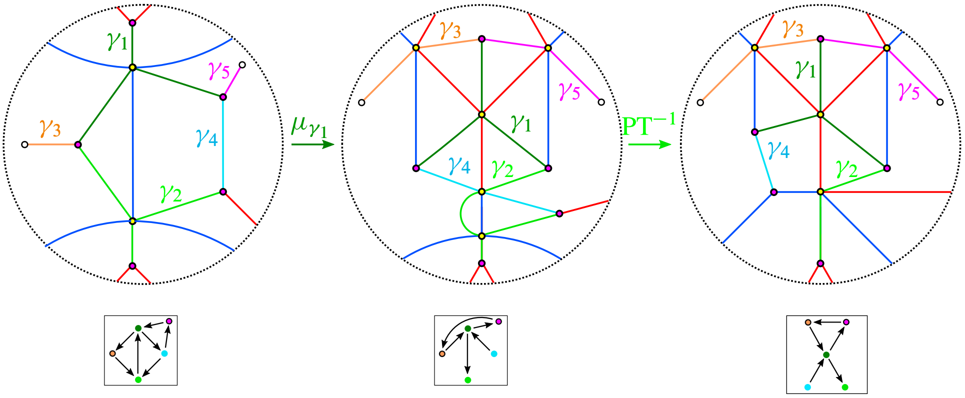


Figure 36: Type IV.i to Type I mutations.

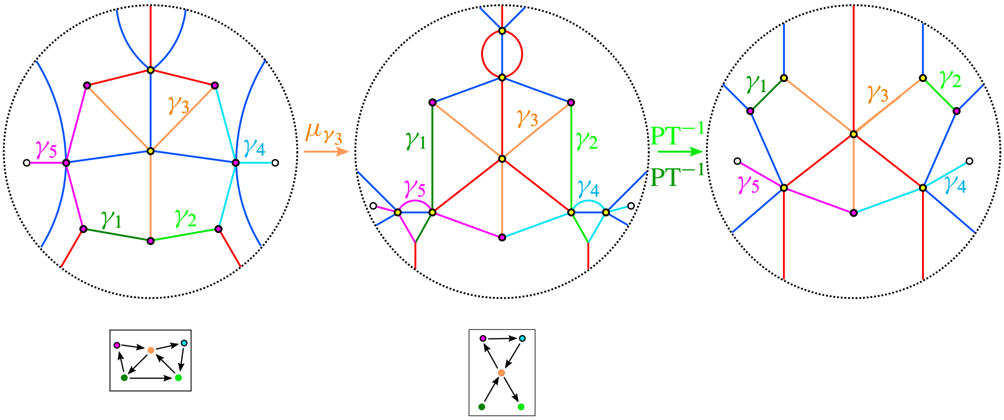


Figure 37: Type IV.ii to Type I mutations.

Type IV.i to Type III We now consider the two sequences of 3–graphs in Figure 38. Mutation at any of γ_1, γ_2, γ_3 or γ_4 in the Type IV.i normal form yields a Type III normal form. Specifically, mutation at γ_4 yields a Type III.i normal form that requires no simplification, while mutation at γ_3 (not pictured) yields a Type III.ii normal form that also requires no simplification. The computation for mutation at γ_1 is pictured in

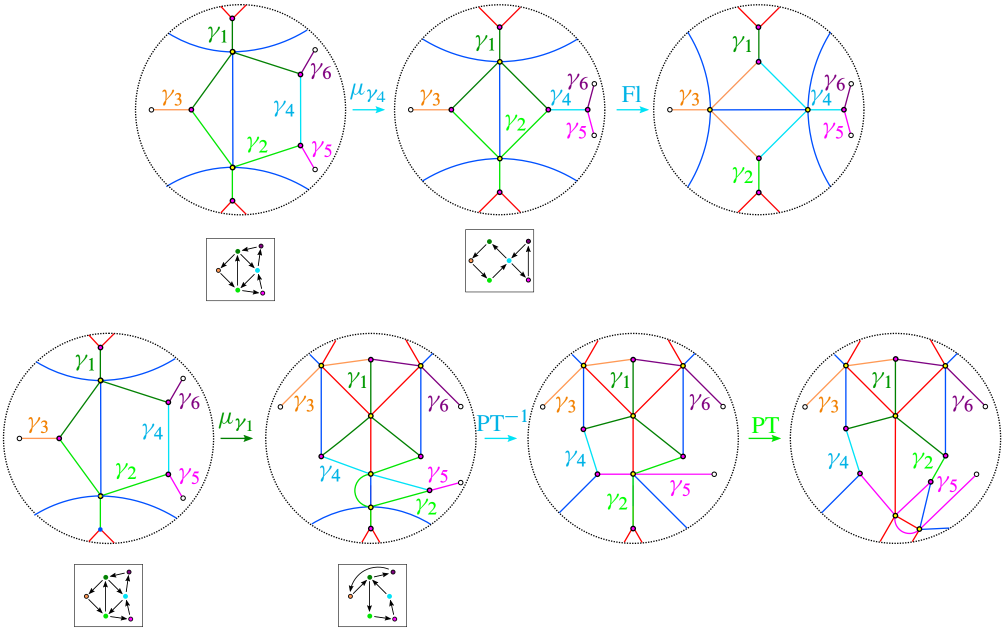


Figure 38: Type IV.i to Type III mutations.

the sequence on the right and is identical to the computation for mutation at γ_2. The first step of the simplification is the same as the Type IV.i to Type I subcase described above. However, we require the application of an additional push-through to remove the geometric intersection between γ_2 and γ_5. This makes γ_5 into a Y–cycle and results in a Type III normal form.

Type IV.ii to Type III Mutation at γ_1 in our Type IV.ii normal form, depicted in Figure 39, results in a pair of geometric intersections between γ_3 and γ_5. Application of a flop removes these geometric intersections and results in a sharp 3–graph with Y–cycles γ_1 and γ_4, which matches our Type III.ii normal form. Note that the computations for mutations involving a Type IV.ii 3–graph with a single spike cycle are identical.

The remaining three subcases are all Type IV to Type IV mutations.

Type IV.ii to Type IV Figure 40 depicts mutation of a Type IV.ii normal form at a spike cycle. Mutating at γ_5 results in an additional geometric intersection between γ_1 and γ_3. We first apply a reverse push-through at the trivalent vertex where γ_1, γ_2 and γ_3 meet. This introduces an additional geometric intersection between γ_2 and γ_3, which we resolve by applying a push-through to γ_3. Application of a reverse push-through to the trivalent vertex where γ_1 and γ_5 intersect a red edge resolves the final geometric intersection between γ_1 and γ_5. The Y–cycles of the resulting 3–graph correspond to k–cycle vertices of the quiver. As shown below, none of the other Type IV to Type IV mutations result in Y–cycles corresponding to spike vertices. Therefore, assuming we have simplified after each of our mutations in the manner described above, the only possible way a Type IV.ii 3–graph arises is by mutating from the initial Type I graphs in Figure 25. Hence, all other Type IV 3–graphs only have Y–cycles corresponding to k–cycle vertices in the quiver. The computations involving a Type IV.ii 3–graph with a single spike cycle are again identical.

Type IV to Type IV Figure 41 depicts Type IV to Type IV mutations when the length of the quiver k–cycle is greater than 3. When mutating at a homology cycle corresponding to a k–cycle vertex of the quiver, we have two possibilities. Figure 41, top, shows the case where γ_4 intersects another Y–cycle γ_2, which corresponds to a k–cycle vertex in the quiver. Figure 41, bottom, considers the case where γ_4 only intersects I–cycles. In both of these cases we must apply a reverse push-through to the trivalent vertex where γ_3 and γ_4 intersect a red edge in order to simplify the 3–graph. This particular simplification requires that neither of the two edges adjacent to the leftmost edge of γ_4 carries a cycle before we mutate. A similar computation (not

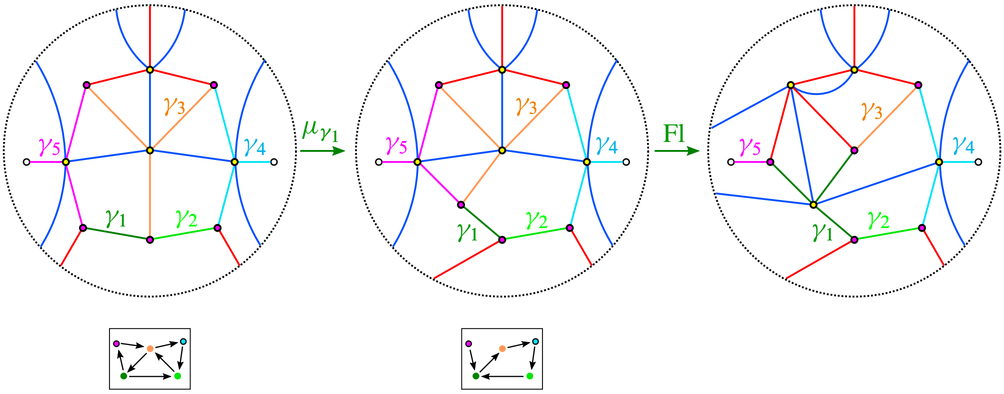


Figure 39: Type IV.ii to Type III mutations.

pictured) involving the Y–cycle γ_2 would also require that neither of the two edges adjacent to the bottommost edge of γ_2 carry a cycle. Crucially, our computations show that Type IV to Type IV mutation preserve this property, ie that both of the Y–cycles have an edge that is adjacent to a pair of edges which do not carry a cycle. When $k=4$, the resulting 3–graph resulting from the computations in the top line will have a short

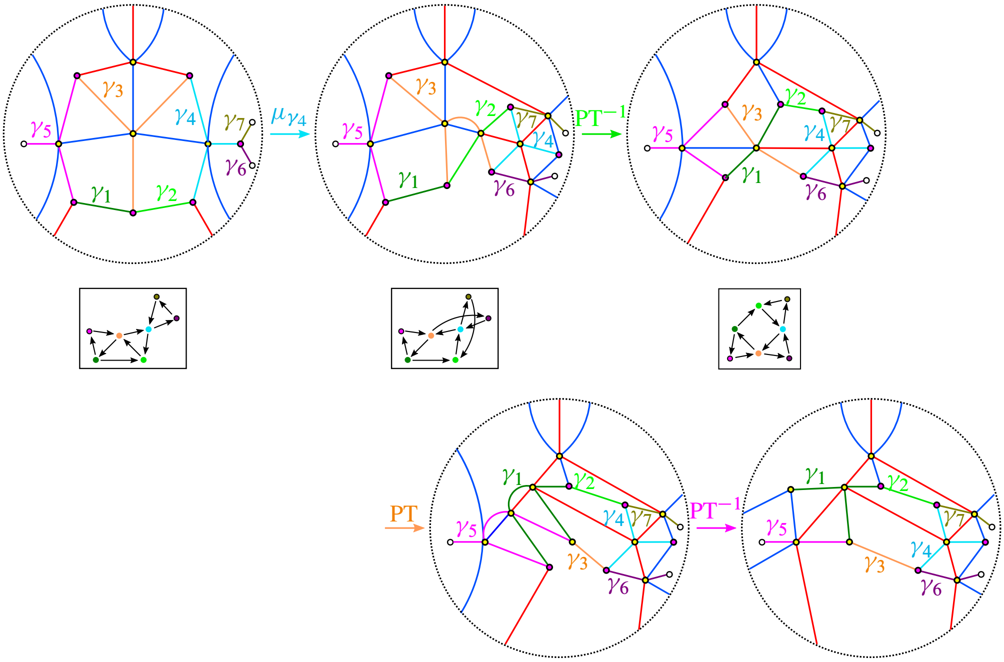


Figure 40: Type IV.ii graph mutation at a spike cycle.

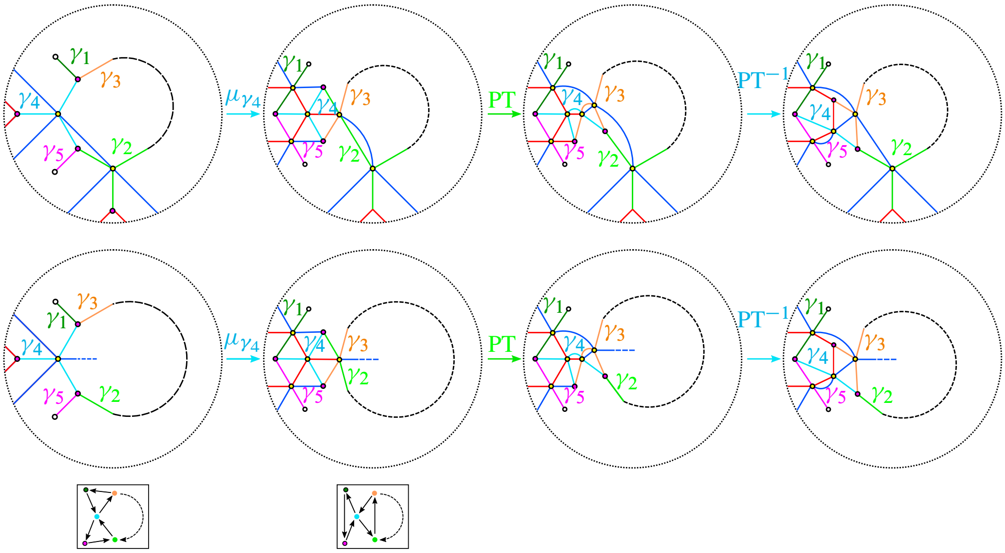


Figure 41: Type IV to Type IV mutations at homology cycles corresponding to k–cycle vertices in the quiver. Mutating at γ_2, γ_3 or γ_4 (corresponding to k–cycle vertices in the quiver) in the 3–graphs on the left decreases the length of the k–cycle in the quiver by 1.

I–cycle adjacent to γ_2 and γ_3, while the 3–graph resulting from the computations in the bottom line will have a short Y–cycle adjacent to γ_2 and γ_3.

Type IV to Type IV Figure 42 depicts mutation at a spike cycle. Since we have already discussed the Type IV.ii spike cycle subcase above, we need only consider the case where each of the spike cycles is a short I–cycle. The cycles γ_7 and γ_6 are included to help indicate where A_n tail cycles are sent under this mutation. The computation for mutating at a spike edge for Type IV.i (ie the $k = 3$ case) is identical to the $k > 3$ case. We have omitted the case where each of the cycles involved in our mutation is an I–cycle, but the computation is again a straightforward mutation of a single I–cycle that requires no simplification.

In each of the Type IV to Type IV subcases above, mutating at a Y–cycle or an I–cycle and applying the simplifications as shown preserves the number of Y–cycles in our graph. Therefore, our computations match the normal form we gave in Figure 20 with $k-2$ short I–cycles in the normal form 3–graph not belonging to any A_n tail subgraphs.

This completes our classification of the mutations of normal forms. In each case, we have produced a 3–graph of the correct normal form that is locally sharp and made up of

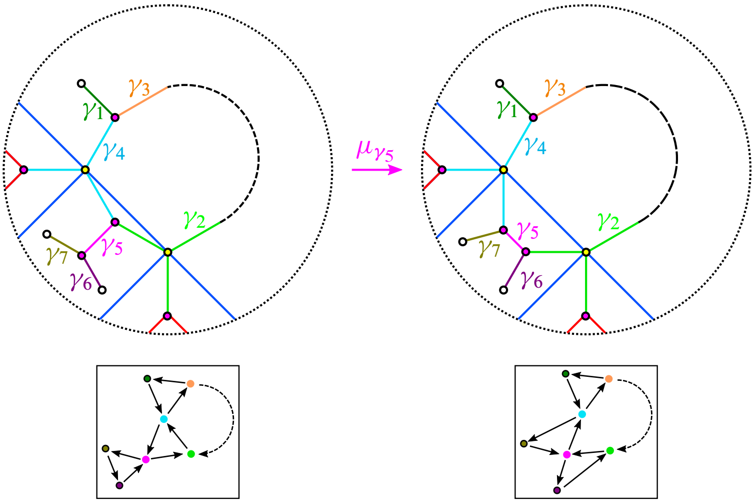


Figure 42: Type IV to Type IV mutations at spike cycles. Mutating at the spike cycles γ_1 or γ_5 in the 3–graphs on the left increases the length of the k–cycle in the intersection quiver by 1.

short Y–cycles and I–cycles. Thus, any sequence of quiver mutations for the intersection quiver $Q(\Gamma_0(D_n), \{\gamma_i^{(0)}\})$ of our initial $\Gamma_0(D_n)$ is weave-realizable. Hence, given any sequence of quiver mutations, we can apply a sequence of Legendrian mutations to our original 3–graph to arrive at a 3–graph with intersection quiver given by applying that sequence of quiver mutations to $Q(\Gamma_0(D_n), \{\gamma_i^{(0)}\})$, as desired. □

Having proven weave-realizability for $\Gamma_0(D_n)$, we conclude with a proof of Corollary 2.

3.2 Proof of Corollary 2

We take our initial cluster seed in $\mathcal{C}(\Gamma)$ to be the cluster seed associated to $\Gamma_0(D_n)$. The cluster variables in this initial seed exactly correspond to the microlocal monodromies along each of the homology cycles of the initial basis $\{\gamma_i^{(0)}\}$. The intersection quiver $Q(\Gamma_0(D_n), \{\gamma_i^0\})$ is the D_n Dynkin diagram and thus the cluster seed is D_n–type. By definition, any other cluster seed in the D_n–type cluster algebra is obtained by a sequence of quiver mutations starting with the quiver $Q(\Gamma_0(D_n), \{\gamma_i^0\})$ and its associated cluster variables. Theorem 1 implies that any quiver mutation of $Q(\Gamma_0(D_n), \{\gamma_i^0\})$ can be realized by a Legendrian mutation in $\Lambda(\Gamma_0(D_n))$, so we have proven the first part of the corollary. The remaining part of the corollary follows from the fact that the D_n–type cluster algebra is known to be of finite mutation type with $(3n-2)C_{n-1}$ distinct cluster seeds. □

3.3 Further study

While a classification of E–type quivers is not yet known, it seems likely that the techniques in this manuscript could be used to show weave-realizability for Lagrangian fillings arising from $\lambda(E_6)$, $\lambda(E_7)$ and $\lambda(E_8)$. Identifying normal forms for the expected weave fillings [4, Conjecture 5.1] could even aid in such a classification of E–type quivers. More generally, it is possible that the methods used here may be adapted to show weave-realizability for any positive braid.

References

[1] **B H An**, **Y Bae**, **E Lee**, *Lagrangian fillings for Legendrian links of finite type*, preprint (2021) arXiv 2101.01943

[2] **V I Arnold**, *Singularities of caustics and wave fronts*, Mathematics and its Applications (Soviet Series) 62, Kluwer Academic, Dordrecht (1990) MR Zbl

[3] **D Bennequin**, *Entrelacements et équations de Pfaff*, from "Third Schnepfenried geometry conference, I" (D Bernard, T Hangan, R Lutz, editors), Astérisque 107, Soc. Math. France, Paris (1983) 87–161 MR Zbl

[4] **R Casals**, *Lagrangian skeleta and plane curve singularities*, J. Fixed Point Theory Appl. 24 (2022) art. id. 34 MR Zbl

[5] **R Casals**, **H Gao**, *Infinitely many Lagrangian fillings*, Ann. of Math. 195 (2022) 207–249 MR Zbl

[6] **R Casals**, **L Ng**, *Braid loops with infinite monodromy on the Legendrian contact DGA*, J. Topol. 15 (2022) 1927–2016 MR

[7] **R Casals**, **E Zaslow**, *Legendrian weaves: N–graph calculus, flag moduli and applications*, Geom. Topol. 26 (2022) 3589–3745 MR Zbl

[8] **B Chantraine**, *Lagrangian concordance of Legendrian knots*, Algebr. Geom. Topol. 10 (2010) 63–85 MR Zbl

[9] **T Ekholm**, **K Honda**, **T Kálmán**, *Legendrian knots and exact Lagrangian cobordisms*, J. Eur. Math. Soc. 18 (2016) 2627–2689 MR Zbl

[10] **Y Eliashberg**, **N Mishachev**, *Introduction to the h–principle*, Graduate Studies in Math. 48, Amer. Math. Soc., Providence, RI (2002) MR Zbl

[11] **Y Eliashberg**, **L Polterovich**, *Local Lagrangian 2–knots are trivial*, Ann. of Math. 144 (1996) 61–76 MR Zbl

[12] **S Fomin**, **L Williams**, **A Zelevinsky**, *Introduction to cluster algebras*: *Chapters* 1–3, book project (2016) arXiv 1608.05735

[13] **S Fomin**, **L Williams**, **A Zelevinsky**, *Introduction to cluster algebras*: *Chapters* 4–5, book project (2017) arXiv 1707.07190

[14] **H Gao**, **L Shen**, **D Weng**, *Augmentations*, *fillings*, *and clusters*, preprint (2020) arXiv 2008.10793

[15] **H Geiges**, *An introduction to contact topology*, Cambridge Studies in Advanced Mathematics 109, Cambridge Univ. Press (2008) MR Zbl

[16] **M Gromov**, *Partial differential relations*, Ergebnisse der Math. (3) 9, Springer (1986) MR Zbl

[17] **S Guillermou**, **M Kashiwara**, **P Schapira**, *Sheaf quantization of Hamiltonian isotopies and applications to nondisplaceability problems*, Duke Math. J. 161 (2012) 201–245 MR Zbl

[18] **K Hayden**, **J M Sabloff**, *Positive knots and Lagrangian fillability*, Proc. Amer. Math. Soc. 143 (2015) 1813–1821 MR Zbl

[19] **L Ng**, **D Rutherford**, *Satellites of Legendrian knots and representations of the Chekanov–Eliashberg algebra*, Algebr. Geom. Topol. 13 (2013) 3047–3097 MR Zbl

[20] **B Ozbagci**, **A I Stipsicz**, *Surgery on contact* 3*–manifolds and Stein surfaces*, Bolyai Society Mathematical Studies 13, Springer (2004) MR Zbl

[21] **Y Pan**, *Exact Lagrangian fillings of Legendrian* $(2,n)$ *torus links*, Pacific J. Math. 289 (2017) 417–441 MR Zbl

[22] **L Polterovich**, *The surgery of Lagrange submanifolds*, Geom. Funct. Anal. 1 (1991) 198–210 MR Zbl

[23] **V Shende**, **D Treumann**, **H Williams**, **E Zaslow**, *Cluster varieties from Legendrian knots*, Duke Math. J. 168 (2019) 2801–2871 MR Zbl

[24] **V Shende**, **D Treumann**, **E Zaslow**, *Legendrian knots and constructible sheaves*, Invent. Math. 207 (2017) 1031–1133 MR Zbl

[25] **D Treumann**, **E Zaslow**, *Cubic planar graphs and Legendrian surface theory*, Adv. Theor. Math. Phys. 22 (2018) 1289–1345 MR Zbl

[26] **D F Vatne**, *The mutation class of* D_n *quivers*, Comm. Algebra 38 (2010) 1137–1146 MR Zbl

[27] **M-L Yau**, *Surgery and isotopy of Lagrangian surfaces*, from "Proceedings of the Sixth International Congress of Chinese Mathematicians, II" (C-S Lin, L Yang, S-T Yau, J Yu, editors), Adv. Lect. Math. 37, International, Somerville, MA (2017) 143–162 MR Zbl

Department of Mathematics, University of California, Davis
Davis, CA, United States

jmhughes@math.ucdavis.edu

Received: 12 March 2021 Revised: 12 October 2021

Algebraic & Geometric Topology **23**:6 (2023) 2777–2821
DOI: 10.2140/agt.2023.23.2777
Published: 7 September 2023

Mapping class groups of surfaces with noncompact boundary components

RYAN DICKMANN

We show that the pure mapping class group is uniformly perfect for a certain class of infinite-type surfaces with noncompact boundary components. We then combine this result with recent work in the remaining cases to give a complete classification of the perfect and uniformly perfect pure mapping class groups for infinite-type surfaces. We also develop a method to cut a general surface into simpler surfaces and extend some mapping class group results to the general case.

57K20, 57S05; 57M07

1 Introduction

Let S be a connected, orientable and second-countable surface, possibly with boundary. The *mapping class group* $\mathrm{Map}(S)$ is the group of all isotopy classes relative to the boundary of S of orientation-preserving homeomorphisms of S. The elements in this group are considered up to isotopy relative to the boundary. A finite-type surface refers to a surface with $\pi_1(S)$ finitely generated, and otherwise we say a surface is of infinite type. The $\mathrm{Map}(S)$ for infinite-type surfaces are commonly referred to as *big mapping class groups*. These groups have been the recent focus of many papers, but the case of noncompact boundary components has been largely untouched with only a single paper of Fabel [10] known to the author considering such groups.

The *pure mapping class group* $\mathrm{PMap}(S)$ is the subgroup of $\mathrm{Map}(S)$ consisting of elements that fix the ends of S, and $\mathrm{PMap}_c(S)$ is the subgroup of compactly supported elements. We equip these groups with the natural compact–open topology. Recently George Domat (and the author in one case) showed the following:

Theorem 1.1 [8] *Let S be any infinite-type surface with only compact boundary components. Then $\overline{\mathrm{PMap}_c(S)}$ and $\mathrm{PMap}(S)$ are not perfect.*

This partially answered Problem 8 of Aramayona, Patel and Vlamis [2]. In the finite-type case, it is a well-known result of Powell that pure mapping class groups are perfect for genus at least 3 [18]. Surprisingly, a new phenomenon occurs when we also consider surfaces with noncompact boundary components, and, even though the general case seems extremely complicated at first glance, it turns out that it is possible to completely classify the surfaces with perfect or uniformly perfect pure mapping class groups. A *disk with handles* will refer to a surface which can be constructed from a disk by removing points from the boundary and then attaching infinitely many handles accumulating to some subset of these points. We say compact boundary components are added to a surface when we delete open balls with disjoint closures from the interior. We say punctures are added when we remove isolated interior points.

Theorem A *Let S be an infinite-type surface. Then*:

- $\overline{\mathrm{PMap}_c(S)}$ *is uniformly perfect if and only if S is a disk with handles.*
- $\overline{\mathrm{PMap}_c(S)}$ *is perfect if and only if S is a connected sum of finitely many disks with handles with possibly finitely many punctures or compact boundary components added.*

In [2], it was shown for surfaces with only compact boundary components that $\mathrm{PMap}(S) = \overline{\mathrm{PMap}_c(S)}$ if and only if S has at most one end accumulated by genus, and otherwise $\mathrm{PMap}(S)$ factors as a semidirect product of $\overline{\mathrm{PMap}_c(S)}$ with some $\mathbb{Z}^n$, where n is possibly infinite. See Theorem 6.1 for a precise statement. Once we extend this result to the general case, we immediately get a classification of the perfect $\mathrm{PMap}(S)$. A disk with handles with exactly one end will be called a *sliced Loch Ness monster*.[1]

[1] This name was chosen because the interior of such a surface is often referred to as the *Loch Ness monster*. The author apologizes for adding to the already out of hand terminology.

Roughly speaking, a degenerate end refers to a end which is the result of deleting an embedded closed subset of the Cantor set from the boundary of a surface (see Definition 3.10). For the following theorem, we throw out surfaces with degenerate ends to give a classification which better fits the chosen definition of a sliced Loch Ness monster.

Theorem B *Let S be an infinite-type surface without degenerate ends. Then*

- $\mathrm{PMap}(S)$ *is uniformly perfect if and only if S is a sliced Loch Ness monster.*
- $\mathrm{PMap}(S)$ *is perfect if and only if S is a sliced Loch Ness monster with possibly finitely many punctures or compact boundary components added.*

Since a sliced Loch Ness monster has a single end, the pure mapping class group and the mapping class group coincide. Therefore, this also gives new examples of surfaces with uniformly perfect mapping class groups. These results show there is an interesting distinction between these mapping class groups and the previously studied cases. In particular, the results of Powell and Domat demonstrate a consistent behavior for pure mapping class groups of surfaces without noncompact boundary components, but the cases we study demonstrate a more complicated behavior. Also, many of the tools from the other cases do not easily extend as one would hope, so new techniques need to be discovered.

Disks with handles and sliced Loch Ness monsters will be an essential part of this paper. In Section 4 we will show how to cut a disk with handles into a collection of sliced Loch Ness monsters, so we can use these simpler surfaces as building blocks for a general argument. We can summarize the decomposition results with the following theorem, which is partially inspired by a result in [2]. See Section 3 for some of the terminology.

Theorem C *Every disk with handles without planar ends can be cut along a collection of disjoint essential arcs into sliced Loch Ness monsters.*

Furthermore, any infinite-type surface with infinite genus and no planar ends can be cut along disjoint essential simple closed curves into components which are either

(i) *Loch Ness monsters with $k \in \mathbb{N} \cup \{\infty\}$ compact boundary components added, possibly accumulating to the single end;*[2] *or*

(ii) *disks with handles with $k \in \mathbb{N} \cup \{\infty\}$ compact boundary components added possibly accumulating to some subset of the ends.*

[2] Here we are using $\mathbb{N} = \{0, 1, 2, \dots\}$.

Acknowledgements

The author is indebted to his advisor, Mladen Bestvina, for helping him throughout this project. Thanks to George Domat for many helpful conversations and for writing up the first version of the fragmentation lemma which is critical to this paper. Thanks to Dan Margalit for introducing the author to mapping class groups. Thanks to the referee for suggesting numerous improvements and corrections.

2 Outline

In Section 3, we discuss the necessary background including the classification of surfaces for orientable noncompact surfaces. The case of compact boundary was done by Kerékjártó [13] and Richards [20]. The general case was done by Brown and Messer [6]. We also give examples of surfaces which demonstrate the interesting new phenomena that occur for surfaces with noncompact boundary. Some understanding of the general classification and the possible cases may be useful to the usual infinite-type surface researcher, especially when considering arguments involving cutting a surface along noncompact objects, such as a union of infinitely many curves or a union of lines or rays.

In Section 4, we prove Theorem C, and also define the *boundary chains* of a surface with noncompact boundary components (see Definition 4.2). Intuitively speaking, a boundary chain can be thought of as a collection of noncompact boundary components which can be realized in the surface as a circle with points removed.

In Section 5, we prove Theorem A. The proof that $\overline{\mathrm{PMap}_c(S)}$ is uniformly perfect for a disk with handles uses standard tricks for writing elements as commutators (see for example the proof that the symmetric group on a countably infinite set is uniformly perfect [16]). First we use a fragmentation lemma (see Lemma 5.3) to decompose a map in $\overline{\mathrm{PMap}_c(S)}$ into a product of two simpler maps. Then, after decomposing the surface into simpler pieces using Theorem C, we can apply a standard trick to write each of the simpler maps as a single commutator.

In Section 6, we discuss how to extend the work of [2] to the general case (see Theorem 6.13) and then prove Theorem B. The main proof in Section 6 involves a natural way to turn a surface with noncompact boundary components into one without them via capping the boundary chains (see Construction 6.12). We first extend the

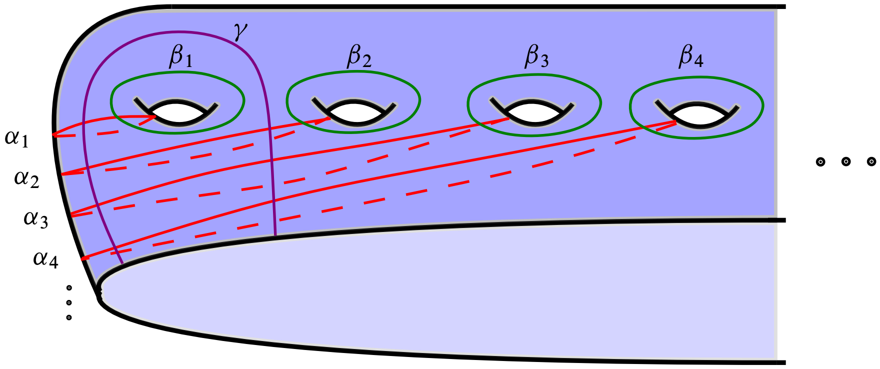


Figure 1: A sliced Loch Ness monster with two infinite collections of curves. The collection $\{\beta_i\}$ eventually leaves every compact subsurface, but every curve in the collection $\{\alpha_i\}$ intersects an arc γ.

Alexander method to the general case (see Theorem 6.4) using a doubling trick. We also extend some well-known facts to the general case (see Lemma 6.2 and Theorem 6.8).

One natural question that immediately comes to mind is whether the mapping class groups of surfaces with noncompact boundary are even different at all from the compact boundary counterparts. Is every one of these mapping class groups just naturally isomorphic to some mapping class group for a surface with (possibly empty) compact boundary? To the contrary, the following example shows that the mapping class group for a surface with noncompact boundary can correspond to a proper subgroup of the mapping class group for the interior surface. Consider the surface with infinite genus, one end, one noncompact boundary component, and no compact boundary components. It follows from the classification of surfaces in Section 3 that there is a unique surface with these properties. This is the 1–*sliced Loch Ness monster*, which we denote by L_s. If we take an infinite collection of curves $\{\alpha_i\}$ accumulating to the boundary as in Figure 1, then the infinite product of Dehn twists $\cdots T_{\alpha_3} T_{\alpha_2} T_{\alpha_1}$ does not correspond to a homeomorphism of L_s. To see this, take another infinite collection of curves $\{\beta_i\}$ and an arc γ as shown in the figure. If we let L be the interior of L_s, then the infinite product of twists corresponds to a well-defined homeomorphism $T = \prod_{i=1}^{\infty} T_{\alpha_i} \in \mathrm{Map}(L)$. Restricting maps on L_s to the interior induces a homomorphism

$$i : \mathrm{Map}(L_s) \to \mathrm{Map}(L),$$

but T is not in the image. Assume otherwise, and conflate T with a homeomorphism on L_s which restricts to T on L. Note $T(\gamma)$ intersects all of the β_i, so it follows the image is not compact, a contradiction. This follows a similar argument as Proposition 7.1 of [17]. We extend this type of argument to a more general setting in Theorem 6.9.

It will follow from Lemma 6.2 that i is injective. Since we have just shown that i is not surjective, we see that $\mathrm{Map}(L_s)$ truly corresponds to a proper subgroup of $\mathrm{Map}(L)$. Note more work must be done to show that $\mathrm{Map}(L_s)$ and $\mathrm{Map}(L)$ are not abstractly isomorphic. Once we are done though, this will follow from Theorem A.

The above example also partially motivated some of this work. In [8], Domat shows that certain multitwists (a product of powers of Dehn twists about disjoint curves) cannot be written as a product of commutators in $\overline{\mathrm{PMap}_c(S)}$. These multitwists involve a collection of curves similar to the α_i in Figure 1. The hope was that a natural subgroup without these types of multitwists would be a perfect group.

3 Background

3.1 Classification of surfaces

3.1.1 Compact boundary Here we summarize the classification theorems from [20; 6], starting with the case of compact boundary. We briefly review the necessary terminology. We always let a surface refer to a connected, orientable and second-countable 2–manifold. We will assume subsurfaces are connected unless stated otherwise. A *complementary domain* of a surface S is a subsurface which is the closure of some component of $S \setminus K$ for a compact subsurface K.

Definition 3.1 An *exiting sequence* for a surface S is a sequence of subsurfaces $\{U_i\}$ such that the following properties hold:

- $U_{i+1} \subset U_i$ for all i.
- $\bigcap_{i=1}^{\infty} U_i = \varnothing$.
- Each U_i is a complementary domain.

Two such sequences $\{U_i\}$ and $\{U_i'\}$ are considered equivalent if for any i there exists a j with $U_j \subset U_i'$, and conversely. This defines an equivalence relation on the set of exiting sequences, and an equivalence class is referred to as an *end* of the surface. The *ends space* of S is the collection of all equivalence classes, denoted by $E(S)$. Note that for a given compact exhaustion the complementary domains of the compact subsurfaces can be used to build exiting sequences. The ends space is an invariant which does not depend on the choice of a compact exhaustion here.

For a given subsurface U, let $U^\star$ be the set of ends such that there is a representative sequence eventually contained within U. We now equip $E(S)$ with a basis generated by sets of the form $U^\star$ ranging over all subsurfaces U such that U is a complementary domain. This basis gives a topology on $E(S)$ which is totally disconnected, second-countable, compact and Hausdorff (see [1]). Topological spaces with these properties are always homeomorphic to a closed subset of the Cantor set.

We say an end is *accumulated by genus* if there is a representative sequence $\{U_i\}$ such that every U_i has infinite genus. We denote the set of ends accumulated by genus by $E_\infty(S)$. An end is *planar* if there is a representative sequence in which some U_i is homeomorphic to a subset of the plane. The space of planar ends is exactly $E(S) \setminus E_\infty(S)$. We say an end is *isolated* if it is isolated in the topology on the space of ends. Isolated planar ends are referred to as *punctures*.

When we consider surfaces with compact boundary, there is the following classification theorem:

Theorem 3.2 (classification of surfaces with compact boundary [20]) *Two surfaces with compact boundary are homeomorphic if and only if they have homeomorphic pairs $(E(S), E_\infty(S))$, the same genus and the same number of compact boundary components.*

3.1.2 Noncompact boundary Now we summarize the ideas for the general case, following [6]. The previous definitions all apply to a general surface without adaptation, but we need more information to capture all the new possibilities. Note compact or, more generally, finite-type exhaustions for a surface S with noncompact boundary components must include those subsurfaces whose boundary intersects the noncompact boundary components of S in a union of intervals.

For a surface with infinitely many compact boundary components, we must record the ends which are accumulated by these components. We refer to these as *ends accumulated by compact boundary*, and we denote the space of these ends by $E_\partial(S)$. This can be precisely defined in a similar manner to accumulated by genus.

Let $\overline{\partial}S$ be the disjoint union of the noncompact boundary components of a surface S. Let $E(\overline{\partial}S)$ be the set of ends of $\overline{\partial}S$. This is just a discrete space with two points

$$\begin{tikzcd}
\pi_0(\overline{\partial}S) & E(\overline{\partial}S) \arrow[l, "e"'] \arrow[r, "v"] & E(S) & E_\infty(S) \arrow[l] \\
 & \mathcal{O} \arrow[u] & E_\partial(S) \arrow[u] &
\end{tikzcd}$$

Figure 2: A surface diagram.

associated to each component. Let $v\colon E(\overline{\partial}S) \to E(S)$ be the function that takes an end of a noncompact boundary component to the end of the surface to which it corresponds. Note it is possible that both ends of a noncompact boundary component get mapped by v to the same end of S, as is the case for the 1–sliced Loch Ness monster from Figure 1.

Let $e\colon E(\overline{\partial}S) \to \pi_0(\overline{\partial}S)$ be the map that takes an end to the corresponding noncompact boundary component. Here $\pi_0(\overline{\partial}S)$ denotes the discrete set of noncompact boundary components of S. If we fix an orientation on S, then, for an arbitrary component $p \in \pi_0(\overline{\partial}S)$, we may distinguish the right and left ends of $e^{-1}(p)$. An *orientation* of $E(\overline{\partial}S)$ is a subset $\mathcal{O} \subset E(\overline{\partial}S)$ that contains exactly the right ends for the given orientation. We collect all of this information in Figure 2.

The unlabeled arrows are the inclusion maps. We will refer to this as the *surface diagram* for the surface S. See [6] for the construction of a surface from a given *abstract surface diagram*, which is a diagram of the above form consisting of topological spaces and maps satisfying various technical conditions. The abstract surface diagram provides a bundle of data whose homeomorphism types are in correspondence with the homeomorphism types of surfaces. Here we consider diagrams to be homeomorphic when there are homeomorphisms between each of the sets which commute with the arrows. We will not use abstract surface diagrams in this paper, so we leave it to the reader to review this definition if desired. One should also note that for the nonorientable case there is extra data to consider which is not represented in Figure 2.

Theorem 3.3 (classification of surfaces [6]) *Two surfaces are homeomorphic if and only if they have homeomorphic surface diagrams, the same genus and the same number of compact boundary components.*

Since the general case is vastly more complicated, we give a few illustrative examples, some of which were discussed in the introduction of [6].

Example 3.4 See Figure 3. The two surfaces shown have homeomorphic ends spaces $E(S) = E_\infty(S) = \omega \cdot 2 + 1$. Notice the doubles of these surfaces are homeomorphic. Here the double of a surface with boundary is constructed by taking two copies and

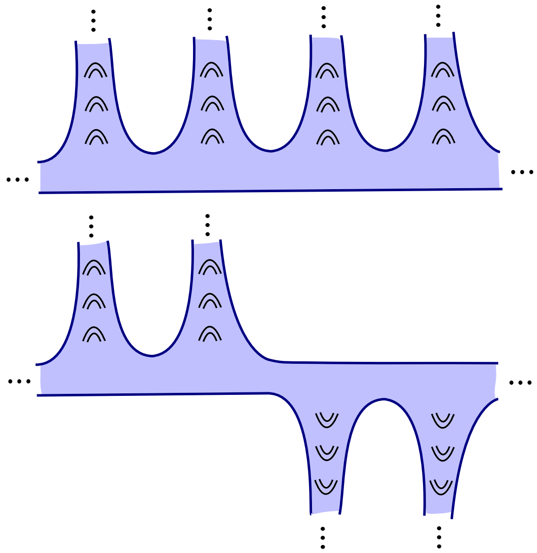


Figure 3: Nonhomeomorphic surfaces with homeomorphic doubles. The boundary components are represented by the blue lines.

gluing along the boundary by the identity. However, the surfaces themselves are not homeomorphic since they have nonhomeomorphic diagrams. To see this, note that the upper surface has a noncompact boundary component such that both ends get sent by υ to accumulation points of $E(S)$, but the lower surface does not. It follows that there cannot be homeomorphisms between their $E(\overline{\partial} S)$ and $E(S)$ sets which commute with the υ maps.

Example 3.5 Take an annulus and from each boundary component remove a point and a sequence accumulating to the point monotonically. There is a choice whether both sequences converge in the same direction or not, and this gives two nonhomeomorphic surfaces. These surfaces have homeomorphic end spaces $E(S) = \omega \cdot 2 + 1$, and even the top rows of their surface diagrams are homeomorphic. The full diagrams are not homeomorphic, however, because the orientations disagree. When the sequences go in the same direction, either $\mathcal{O}$ or $E(\overline{\partial} S) \setminus \mathcal{O}$ contains (the preimages of) both of the accumulation points, but, when the sequences go in opposite directions, $\mathcal{O}$ contains exactly one of the accumulation points. Similar reasoning gives another explanation why the surfaces in Example 3.4 are nonhomeomorphic.

This example highlights an interesting distinction from the compact boundary case. A connected sum of surfaces with compact boundary always gives a unique surface, up to homeomorphism, but for general surfaces there may be at most two homeomorphism types, depending on the orientation of the attaching map. The above two surfaces are each the connected sum of the same disks with boundary points removed. For orientable surfaces, a connected sum determines a unique surface if and only if at least one of the surfaces has an orientation-reversing self-homeomorphism.

Now we define a class of surfaces essential to this paper. By attaching a handle or tube to a surface we mean removing two open balls with disjoint closures and then identifying the resulting boundary components by an orientation-reversing map of degree -1.

Definition 3.6 (disk with handles) A *disk with handles* is a surface which can be constructed by taking a disk, removing a closed embedded subset $\mathcal{P}$ of the Cantor set from the boundary, and then attaching infinitely many handles accumulating to some subset of $\mathcal{P}$. The choice of infinitely many handles was chosen to simplify the statement of the theorems and to remove finite-type surfaces.

Remark 3.7 Let D be a disk with boundary points removed, and S a disk with handles constructed from D. When we attach a sequence of handles to D, it is possible the two corresponding sequences of open balls accumulate to different points in $E(D)$. This joins these ends into a single end of $E(S)$. This is highlighted in Construction 3.9 and Figures 4 and 5. Due to this phenomenon, a general disk with handles is much more complicated than one might first expect.

If we assume this type of handle attaching does not occur, then the possible disks with handles are classified by homeomorphism types of the pair $(E(S), E_\infty(S))$ with the additional structure of a cyclic ordering. Note that this gives another way to distinguish the surfaces in Example 3.4. A more complicated type of ordering, allowing repeats, is required to classify general disks with handles. A major part of the Brown–Messer construction for a surface from a given diagram involves the delicate construction of such orderings [6].

Now we also want to consider a more specific class of surfaces.

Definition 3.8 (Loch Ness monsters) A *Loch Ness monster* refers to the unique surface with one end, infinite genus and empty boundary. A *sliced Loch Ness monster* is any of the surfaces with one end, infinite genus, no compact boundary components, but at least one noncompact boundary component. Equivalently, a sliced Loch Ness monster is a disk with handles with one end. By the classification of surfaces, a surface

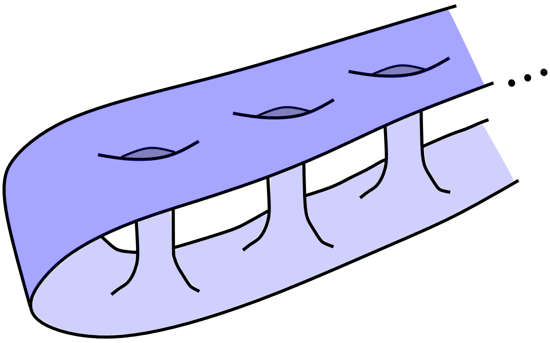


Figure 4: The 2–sliced Loch Ness monster.

with these properties is determined by the possibly infinite number of noncompact boundary components. We sometimes refer to an n–sliced Loch Ness monster to emphasize the number of boundary components.

In order to help visualize these surfaces, we give the following construction:

Construction 3.9 Take a strip $\mathbb{R}\times[-1,1]$ and remove small disjoint open balls centered around the points $(n,0)$ for $n\in\mathbb{Z}\setminus\{0\}$. Now identify pairs of boundary components centered at $(\pm n,0)$ via horizontal reflection. Equivalently, we may view this process as attaching tubes to the strip. The resulting manifold is the 2–sliced Loch Ness monster. See Figure 4. Similarly, we can construct the n–sliced Loch Ness monster for any finite n by taking a disk with n points removed from the boundary and attaching tubes to join all of the ends. To get the ∞–sliced Loch Ness monster, we can take a disk with any infinite embedded closed subset of the Cantor set removed from the boundary and attach tubes to join all of the ends as before. By the classification of surfaces, no matter what infinite set of points we remove in this construction we always get the same surface. This is somewhat unintuitive, but it is better understood once we realize that any interesting topology in the original ends space is collapsed when we attach tubes to get a surface with a single end.

The choice to define sliced Loch Ness monsters independently of the number of boundary components simplifies the statement of the main theorems. In particular, for the first statement of Theorem C it will be simpler to include sliced Loch Ness monsters which have any number of noncompact boundary components going out the single end. See Figure 5 for an example of a disk with handles which suggests that we should include 2–sliced Loch monsters in the list of building blocks.

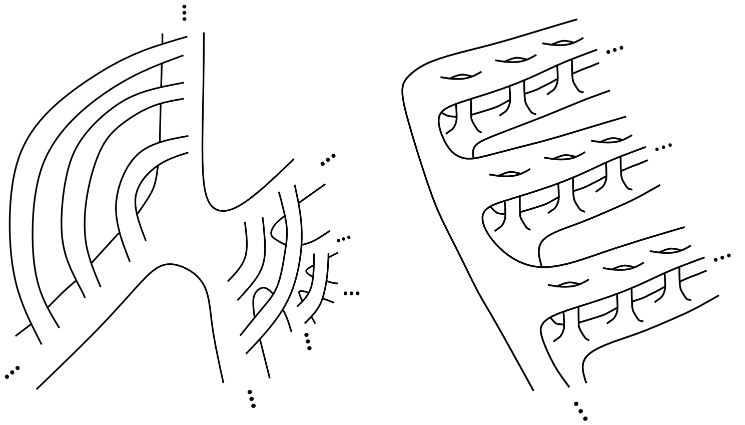


Figure 5: Two visualizations of the same disk with handles. This surface can be cut along arcs into infinitely many 2–sliced Loch Ness monsters.

According to Theorem C, an infinite-type surface with every end accumulated by genus can be cut along curves into Loch Ness monsters and disks with handles (without planar ends), each possibly with compact boundary components added. Therefore, this class of surfaces corresponds to the set of all surfaces which result from a possibly infinite procedure of connected sum operations with these building blocks. In Remark 4.14 we discuss a possible extension of Theorem C to general surfaces possibly with finite genus and planar ends. In this case, we must allow more building blocks, in particular disks with boundary points removed and possibly compact boundary components added or finitely many handles attached. Many basic examples one should consider involve inductive procedures of connected sum operations with these building blocks. Note that Theorem C or any extension thereof can only tell us that some procedure exists for connecting together building blocks to create a general surface, but this procedure may not necessarily be describable in an inductive manner.

3.2 Big mapping class groups

Let $\operatorname{Homeo}^+_\partial(S)$ be the group of orientation-preserving homeomorphisms of a surface S which fix the boundary pointwise. The *mapping class group* $\operatorname{Map}(S)$ is defined to be

$$\operatorname{Map}(S) = \operatorname{Homeo}^+_\partial(S)/{\sim},$$

where two homeomorphisms are equivalent if they are isotopic relative to the boundary of S. We will often conflate a mapping class group element with a representative homeomorphism. We equip $\operatorname{Homeo}^+_\partial(S)$ with the compact–open topology, which induces the quotient topology on $\mathrm{Map}(S)$. We equip subgroups of $\mathrm{Map}(S)$ with the subspace topology. The mapping class group of a subsurface will correspond to the subgroup of elements which have a representative supported in the subsurface. The *pure mapping class group* $\mathrm{PMap}(S)$ is the subgroup of $\mathrm{Map}(S)$ consisting of elements which fix the ends of S.

We say $f \in \mathrm{Map}(S)$ is *compactly supported* if f has a representative that is the identity outside of a compact subsurface of S. The subgroup consisting of compactly supported mapping classes is denoted by $\mathrm{PMap}_c(S)$. Note any compactly supported mapping class is in the subgroup $\mathrm{PMap}(S)$.

Definition 3.10 (degenerate ends) Notice removing an embedded closed subset of the Cantor set from the boundary of a surface does not change the underlying mapping class group. We refer to the resulting ends as *degenerate*. More generally, this will refer to ends with a representative sequence $\{U_i\}$ such that some U_i is homeomorphic to a disk with boundary points removed. It may be convenient in some cases to only work with homeomorphism types of surfaces up to filling in the degenerate ends. We will allow these ends except when stated otherwise. Note that given the definition of a finite-type surface from the introduction, there can be finite-type surfaces with infinitely many degenerate ends.

Now we review the definition of a handle shift from [2] which will be used throughout Section 6. Let Σ be the surface obtained by gluing handles onto $\mathbb{R} \times [-1,1]$ periodically with respect to the map $(x,y) \mapsto (x+1,y)$. We refer to this surface as a *strip with genus*. For some $\epsilon > 0$, let $\sigma\colon \mathbb{R} \times [-1,1] \to \mathbb{R} \times [-1,1]$ be the map determined by setting

$$\sigma(x,y) = \begin{cases} (x+1,y) & \text{for } (x,y) \in \mathbb{R} \times [-1+\epsilon, 1-\epsilon], \\ (x,y) & \text{for } (x,y) \in \mathbb{R} \times \{-1,1\}, \end{cases}$$

and interpolating continuously on $\mathbb{R} \times [-1,-1+\epsilon] \cup \mathbb{R} \times [1-\epsilon,1]$. By extending this map to the attached handles, we get a homeomorphism on Σ, which we conflate with σ. A homeomorphism $h\colon S \to S$ is a *handle shift* if there exists a proper embedding $\iota\colon \Sigma \to S$ such that

$$h(x) = \begin{cases} (\iota \circ \sigma \circ \iota^{-1})(x) & \text{if } x \in \iota(\Sigma), \\ x & \text{otherwise.} \end{cases}$$

The embedding ι is required to be proper, so it induces a map $\hat{\iota}\colon E(\Sigma) \to E_\infty(S)$. A handle shift h then has an attracting and a repelling end, denoted by h^+ and h^-, respectively. In general, the attracting and repelling ends can be the same, though the handle shifts used in Section 6 will have different attracting and repelling ends.

3.3 Curves and arcs

A simple closed curve in a surface S is the image of a topological embedding $\mathbb{S}^1 \hookrightarrow S$. A simple closed curve is trivial if it is isotopic to a point; it is peripheral if it is either isotopic to a boundary component or bounds a once-punctured disk. We will often refer to a simple closed curve as just a curve.

An arc in S is a topological embedding $\alpha\colon I \hookrightarrow S$, where I is the closed unit interval, with $\alpha(\partial I) \subset \partial S$. We consider all isotopies between arcs to be relative to ∂I; ie, the isotopies are not allowed to move the endpoints. An arc is trivial if it is isotopic to an arc whose image is completely contained in ∂S; it is peripheral if it bounds a disk with a single point removed from the boundary. This last definition is the only nonstandard one, and we include it since it aligns with the definition of a peripheral curve. It may be useful in some cases to extend the definition of trivial/peripheral to include arcs or curves which are trivial/peripheral in the surface after degenerate ends are filled in.

A curve or arc is essential when it is not trivial nor peripheral; it is separating if its complement is disconnected and nonseparating otherwise. We will often conflate a curve or arc with its isotopy class. All curves and arcs will be assumed to be essential unless stated otherwise. We say curves or arcs intersect if they cannot be isotoped to be disjoint, and we say they are in minimal position when they are isotoped to have the smallest number of intersections. We say a subsurface is essential if the inclusion of the subsurface induces an injective map of fundamental groups.

By cutting along a collection of curves or arcs, we mean removing disjoint open regular neighborhoods of each of the curves or arcs. Throughout this paper, we will conflate the complement of a curve or arc with this cut surface. We will also occasionally conflate the complement of a subsurface with its closure.

4 Decomposing an infinite-type surface

4.1 Outline

In this section, we prove Theorem C along with several other decomposition results. This is crucial for the proof of the main theorems, since a general surface can be extremely complicated. We also want an approachable method for visualizing a surface from the surface diagram data. Our work builds off the Brown–Messer classification [6] with some inspiration from [19]. The classification theorem from the latter paper is incorrect as stated: it cannot distinguish the pairs of surfaces from Examples 3.4 and 3.5. On the other hand, the argument given there does provide a more intuitive approach. We precisely define some of the ideas from [19].

The main idea is to study what happens when we remove the boundary of a surface S. Deleting a compact boundary component leaves a puncture, which corresponds to an isolated end of S°, the interior of S. Deleting the noncompact boundary components is more complicated as there could be several ends corresponding to these boundary components which get sent to a single end.

We show that we can think of the noncompact boundary components and their ends as being grouped together into chains, and that removing the boundary components from a chain sends all of the corresponding ends to a single end of S°. An important tool will be Lemma 4.12, which allows us to cut a surface along curves so each resulting component has at most one boundary chain. After we discuss the types of surfaces which have a single boundary chain (see Lemma 4.8 and the remarks at the end of its proof), we can apply Lemma 4.12 to prove the "furthermore" statement of Theorem C by representing the components with boundary chains as disks with handles possibly with compact boundary components added. The boundary chains will then correspond to the boundaries of the disks with handles.

4.2 Boundary ends and chains

First we want to precisely define the map on ends spaces induced by deleting the boundary. Consider the inclusion of S in $S' = S \cup_{\partial S} \big(\partial S \times [0,\infty)\big)$. Notice that S' is homeomorphic to S°. Consider a compact exhaustion $\{S'_i\}$ of S'. Let $S_i = S'_i \cap S$, so $\{S_i\}$ is a compact exhaustion of S. Choose an end in $E(S)$ and let $\{U_i\}$ be an exiting sequence representative for this end consisting of complementary domains of the S_i. By replacing components of the $S \setminus S_i$ with components of the $S' \setminus S'_i$, we can get an exiting sequence in S'. It follows that we have a well-defined canonical map

$$\pi \colon E(S) \to E(S^{\circ}).$$

Proposition 4.1 *The map π is continuous.*

Proof Let $U^{\star} \subseteq E(S^{\circ})$ be a basis element defined by some complementary domain U in S°. This gives a complementary domain U_S in S after adding in the boundary, and so it defines a basis element $U_S^{\star} \subseteq E(S)$. We are done after noting that $\pi^{-1}(U^{\star}) = U_S^{\star}$. □

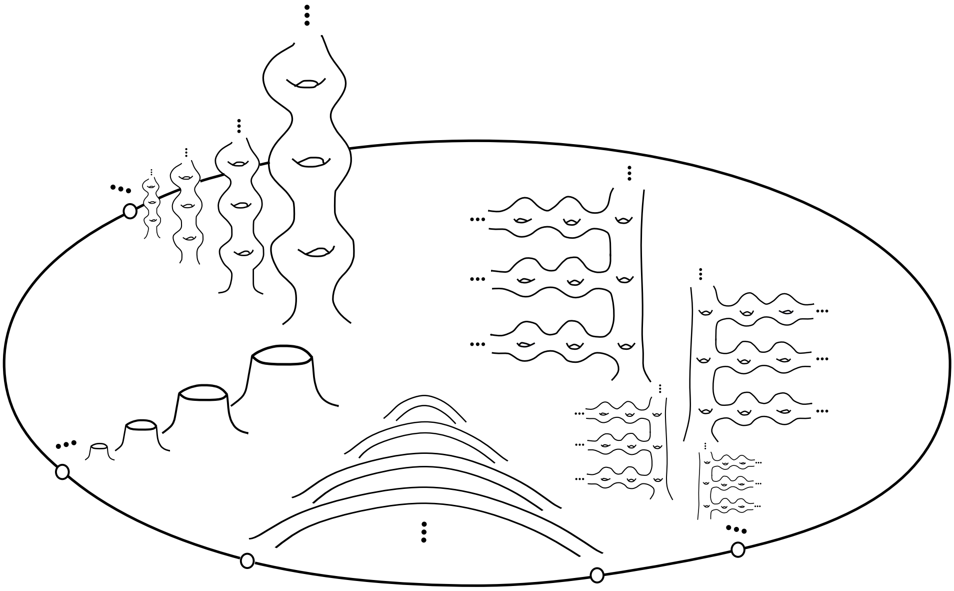


Figure 6: A surface with a single boundary chain.

Now recall the definition of a surface diagram from Section 3 (see Figure 2). Let V be the image of v, the map which sends ends of noncompact boundary components to ends of S. Now we use the map π to define a boundary chain. Intuitively, this can be thought of as a set of boundary ends and boundary components which can be realized in the surface as a circle with points removed.

Definition 4.2 (boundary chains) A *boundary chain* of a surface S is a subset of $E(S)$ of the form $\pi^{-1}(p)$, where $p \in \pi(V)$. The collection of all such sets is denoted by $C(S)$ and is referred to as the *set of boundary chains* for S. Occasionally, we will conflate definitions and use boundary chain to refer to the union of noncompact boundary components with ends in the chain.

Now we define the set of boundary ends for a surface.

Definition 4.3 (boundary ends) Let $B(S)$ be the union of the boundary chains. This will be referred to as the *set of boundary ends*, and any element of $B(S)$ is a *boundary end*.

An end in $E(S)$ is said to be an *interior end* if it is not in $B(S)$. If a boundary end in S is isolated from the other ends, then we refer to it as a *boundary puncture*. Note that

$B(S)$ contains V, but it is possible that $B(S)$ contains additional ends. The definitions above were specifically chosen to include additional ends, such as the ones from the following example:

Example 4.4 Consider a disk with a Cantor set removed from the boundary. We want every end of this surface to be considered a boundary end, but there are some ends which are not in the image of v. These correspond to points in the Cantor set which are not the endpoint of any interval that is removed during the usual middle thirds construction.

We can use π to define an equivalence relation on $B(S)$ for which $C(S)$ is the resulting quotient. After equipping $B(S)$ with the subspace topology, $C(S)$ inherits the quotient topology. Note π is injective on $E(S) \setminus B(S)$. The set of boundary chains exactly records the noninjectivity of π on $B(S)$.

Remark 4.5 Since there are countably many boundary components in a surface, $\pi(B(S))$ is countable.

Remark 4.6 The subset $B(S) \subseteq E(S)$ is not necessarily closed. For example, take a once-punctured sphere, remove infinitely many open balls with disjoint closures accumulating to the puncture, and then remove a single point from each of the resulting boundary components. It is not necessarily open either, as in the case of a disk with a point removed from the boundary with interior punctures added accumulating to the boundary end. By Proposition 4.1, each boundary chain is a closed subset. Then, by Remark 4.5, $B(S)$ is the countable union of closed subsets.

Example 4.7 Consider any disk with handles S. The interior of S is the Loch Ness monster, since it corresponds to a once-punctured sphere with handles attached accumulating to the puncture. Every boundary end of S gets sent by π to the single end of the Loch Ness monster, so any disk with handles has a single boundary chain.

This last statement has a partial converse, which provides a more intuitive way to think about a boundary chain:

Lemma 4.8 *Every surface S with infinite genus, one boundary chain and only boundary ends is a disk with handles possibly with compact boundary components added.*

Before we prove Lemma 4.8, we first need a few facts.

Proposition 4.9 *Let S be a noncompact surface without boundary. Then the following are equivalent*:

(i) *There exists a compact exhaustion $\{S_i\}$ of S such that each ∂S_i has a single component.*

(ii) *S has exactly one end.*

Proof Since the complementary regions of a compact exhaustion can be used to build exiting sequences, and the ends space is independent of this choice of a compact exhaustion, the first condition immediately implies the second. Assuming the second condition, S is either a finite-type surface with one puncture, or the Loch Ness monster. In either case, we can directly construct the desired exhaustion. □

Proposition 4.10 *Let S be a noncompact surface with no compact boundary components and no interior ends. Then the following are equivalent*:

(i) *There exists a compact exhaustion $\{S_i\}$ of S such that each ∂S_i has a single component.*

(ii) *S has exactly one boundary chain.*

Proof Suppose the first condition holds. To get the second condition, it suffices to show that the interior of S has a single end. Remove open regular neighborhoods of the boundary from each S_i, shrinking the neighborhoods as we increase i so we get a compact exhaustion for the interior. Each subsurface in this exhaustion has one boundary component, so we are done by Proposition 4.9.

Now suppose the second condition holds, so the interior of S has a single end by the definition of a boundary chain. Let $\{K_i\}$ be a compact exhaustion of the interior of S given by Proposition 4.9 such that each ∂K_i has a single boundary component. Let N be an open regular neighborhood of the boundary chain. Note that, if we set $S_i = K_i \cap (S \setminus N)$, then we get a compact exhaustion $\{S_i\}$ for $S \setminus N$.

We want to modify the K_i so the resulting S_i each have a single boundary component. First remove subsurfaces from $\{K_i\}$ so $K_1 \cap N \neq \varnothing$. Now isotope ∂K_1 so it is transverse to $\partial \overline{N}$ and each component of $K_1 \cap \overline{N}$ is a bigon. Now we proceed inductively. Remove some subsurfaces from the exhaustion so K_i contains the previously modified K_{i-1}, and isotope ∂K_i in $S \setminus K_{i-1}$ so its position with $\overline{N}$ is as above. We can ensure the bigons exhaust the interior of N, so the modified sequence $\{K_i\}$ is an exhaustion of

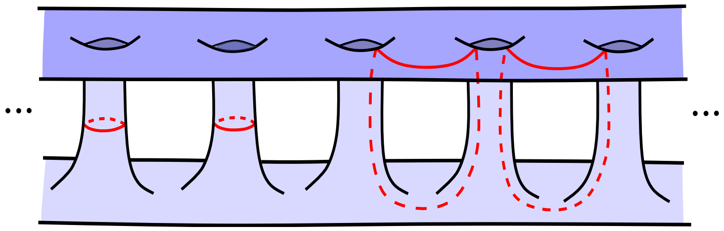


Figure 7: A surface satisfying the conditions of Lemma 4.8 with a collection of curves that cuts it into a surface with zero genus and one boundary chain.

the interior of S. Now, since S_i is the result of removing disjoint bigons from K_i, we conclude that each S_i has one boundary component. We are now done since $S \setminus N$ is homeomorphic to S. □

Now we are ready to prove Lemma 4.8.

Proof of Lemma 4.8 Throughout this proof, we modify S also calling the new surface at each step S. First cap any compact boundary components of S with disks. Since S has no interior ends, one boundary chain and now no compact boundary components, Proposition 4.10 gives us a compact exhaustion $\{S_i\}$ of S such that each ∂S_i has one component. Now we want to find an infinite sequence of nonseparating curves such that cutting S along the curves gives a surface with no genus and one boundary chain. See Figure 7 for an example. We must be careful since cutting the surface from this figure along a sequence of horizontal curves about each tube similar to the two leftmost curves gives two surfaces each with one boundary chain. If we exclude the curve about the middle tube, then cutting gives a surface with two boundary chains.

To get the desired collection of curves, note we can find a finite collection of curves in each S_i which cut it into a surface with no genus and one boundary component (after capping the compact boundary components resulting from cutting), and we can ensure each collection extends to subsequent collections. The desired collection of curves is then the increasing union of these collections. Cut S along these curves, and cap the resulting boundary components with disks. Now S has no genus, and, by applying Proposition 4.10 to a modified compact exhaustion, we see S has one boundary chain.

As in the first paragraph of the proof of Proposition 4.10, by removing open regular neighborhoods from the boundary of the S_i, we can get a compact exhaustion of the interior of S satisfying the first condition of Proposition 4.9. Then, by Proposition 4.9

and classification of surfaces, the interior of S is homeomorphic to a once-punctured sphere. Therefore, if we fill in the boundary ends of S, we get a compact surface which must be homeomorphic to a disk. We are then done after reversing the above steps, since this will correspond to deleting points from the boundary and then attaching handles as in the definition of a disk with handles. Finally, if there were initially any compact boundary components then reversing the capping corresponds to adding back in these components.

We should mention that a version of this lemma holds if we allow surfaces with finite genus. In this case, our surface will be homeomorphic to a disk with boundary points removed with finitely many (possibly zero) handles attached and possibly compact boundary components added. We could also allow interior ends, and then we would need to allow a final step where we delete interior points from the modified disk and then possibly attach handles or compact boundary components accumulating to any of the ends. The overall takeaway of this lemma is that a surface with a single boundary chain is homeomorphic to a modified disk. □

4.3 Decomposition results

Lemma 4.11 *Every infinite-type surface S without boundary and without planar ends can be cut along a collection of disjoint essential simple closed curves into Loch Ness monsters with $k \in \mathbb{N} \cup \{\infty\}$ compact boundary components added.*

Proof This was first shown in [2] as a tool to prove Theorem 6.1. See Section 6 for this argument. We provide a different proof, which gives us more control over the ends of the components in the cut surface. Recall that the ends space $E(S)$ is homeomorphic to a closed subset of the Cantor set. Let T be some locally finite tree with $E(T)$ homeomorphic to $E(S)$.[3] We can think of S as a thickened version of T with genus added accumulating to every end. For simplicity, we will assume T has no vertices of valence one.

We may write T as a union of rays $\{R_i\}$ where, for each distinct R_i and R_j, $R_i \cap R_j$ is empty or a single vertex. To see this, enumerate a countable dense subset $\{x_i\}$ of $E(T)$ and fix some basepoint vertex v. Begin by letting R_1 be the ray from v to x_1, then let R_2 be the ray from v to x_2 with the interior of the overlap with R_1 deleted. Continue

[3] The ends space of a tree is defined analogously to the ends space of a surface. For locally finite trees, the ends space is always homeomorphic to a closed subset of the Cantor set.

in this manner to build the desired collection $\{R_i\}$. Since T has no vertices of valence one, this will exhaust the entire tree.

Associate each R_i with a Loch Ness monster L_i. Let $n_i \in \mathbb{N} \cup \{\infty\}$ be the number of vertices in $R_i \cap \bigcup_{j \neq i} R_j$. For each i, remove n_i open balls with disjoint closures from L_i with the balls accumulating to the single end when n_i is infinite. Associate each boundary component of L_i with a vertex in $R_i \cap \bigcup_{j \neq i} R_j$, and attach the boundary components of distinct L_i and L_j when these components correspond to the same vertex of T. Let S' be the resulting surface and let $\{\alpha_i\}$ be the collection of curves in S' corresponding to the attached boundary components.

Now $E(T)$ and $E(S')$ are homeomorphic. This requires showing a correspondence between a compact exhaustion of T and an exhaustion of S'. One approach is to subdivide T, then write it as a union of stars of the vertices from the original tree. Then associate each star with an n–holed torus, where n is the number of edges in the star. The stars and the tori can then be attached to build compact exhaustions for T and S', respectively. By the classification of surfaces, S' is homeomorphic to S. Cutting S' along the α_i gives components which are Loch Ness monsters with compact boundary components added, so we are done. □

The argument from this lemma will be referenced often in the following proofs. By decomposing a tree T into rays we mean writing T as a union of rays which are either disjoint or intersect one another at a single vertex.

Lemma 4.12 *Every surface S can be cut along a collection of disjoint simple closed curves into components with at most one boundary chain. Additionally, we may assume the components with boundary chains have only boundary ends.*

Proof We can assume S has noncompact boundary components, since otherwise the lemma holds trivially. Recall that, by the definition of a boundary chain, two boundary ends $p, q \in B(S)$ are in the same boundary chain if and only if $\pi(q) = \pi(p)$. Suppose we can cut S° along curves into one-ended components so that $\pi(B(S))$ is contained in the dense subset of $E(S^\circ)$ corresponding to these components. Now, when we cut S along these same curves, each component of the cut surface has at most one boundary chain. Since π maps interior ends outside of $\pi(B(S))$, we get the last statement of the lemma. Therefore, it suffices to decompose S° in this manner.

Following the proof of Lemma 4.11, represent S° by a tree T with no valence one vertices. Fix a base vertex v and let T' be the union of rays from v to an end in $\pi(B(S))$.

By Remark 4.5, $\pi(B(S))$ is countable, so enumerate the elements of $\pi(B(S))$ as a sequence $\{x_i\}$. As in the proof of Lemma 4.11, we can use an inductive process to decompose T' into a collection of rays $\{R_i\}$ where each element of $\{x_i\}$ corresponds to the end of one of the rays. Then we can decompose the remainder of T into rays. Now we follow the proof of Lemma 4.11 to decompose S° as desired. Note for this last step we need to allow one-ended pieces with finite genus into our decomposition since we are not assuming S has only ends accumulated by genus (see Remark 4.14). We may also need to allow nonessential curves. □

Lemma 4.13 *Every disk with handles S without planar ends can be cut along a collection of disjoint essential arcs into sliced Loch Ness monsters.*

Proof Let D be a disk with points removed from the boundary used to construct S. Note we may realize D as a closed neighborhood of a tree T properly embedded in $\mathbb{C}$.[4] As before, we will assume this tree has no valence one vertices. Recall from Remark 3.7 that the handles may be attached in a way that joins ends of D together. By similar reasoning to Proposition 4.1 and the preceding remarks, the process of attaching handles determines a well-defined continuous quotient map

$$q\colon E(D) \to E(S).$$

By classification of surfaces, two disks with handles without planar ends are homeomorphic when there is a homeomorphism between the base disks which respects the quotient maps induced by attaching the handles.

We argue by analogy to the proof of Lemma 4.11. First suppose that q is injective. Decompose T into rays $\{R_i\}$ and then associate each ray with a 1–sliced Loch Ness monster. Attach these surfaces along intervals on their boundaries according to the incidences of the R_i in T. This attaching procedure is analogous to the procedure from Lemma 4.11 with boundary connected sum operations in place of the connected sum operations. This gives a disk with handles with a base disk homeomorphic to D and an injective quotient map, so it is homeomorphic to S. It then follows that we can cut S into 1–sliced Loch Ness monsters. See Figure 8, left, for an example of a disk with handles constructed from a thickened binary tree being cut into 1–sliced Loch Ness monsters. However, if q is not injective, we need to be a little more careful. See for example Figure 8, right. If we choose to cut this surface along arcs similar to the ones

[4] One approach is to take a triangulation of D and then build T from a spanning tree of the dual 1–skeleton.

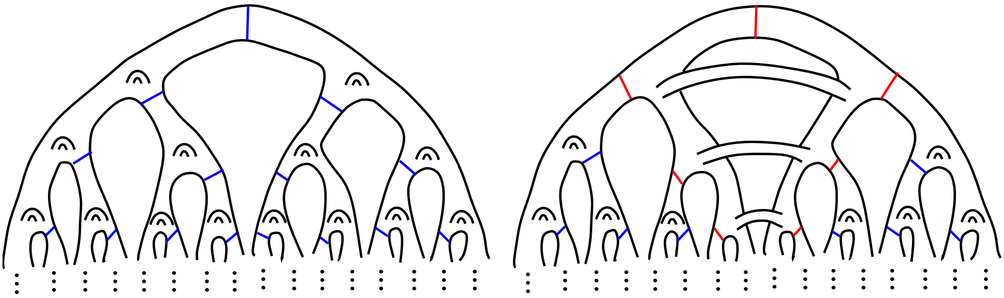


Figure 8: Left: a disk with handles gets cut along blue arcs into 1–sliced Loch Ness monsters. Right: a more complicated disk with handles gets cut along the red and blue arcs into a 2–sliced Loch Ness monster and 1–sliced Loch Ness monsters. The surface bounded by the red arcs corresponds to two rays chosen to exhaust the subset $q^{-1}(x_1)$, where x_1 is the single element of U.

used for the left surface, then we will have components in the cut surface which are not sliced Loch Ness monsters. Let

$$U = \{p \in E(S) : |q^{-1}(p)| \geq 2\}.$$

Enumerate a countable dense subset $\{x_i\}$ of U. Now, when we decompose T into rays, first choose rays that exhaust a dense subset of each $q^{-1}(x_i)$. Here we are conflating the ends space of D with the ends space of T. Then decompose the remainder of T to exhaust a dense subset of the entire ends space. Let $\{R_i\}$ be the resulting rays. Similar to before, associate each R_i with a disk with one boundary puncture D_i, and attach the D_i along intervals on their boundaries according to the incidences of the R_i to get a base disk homeomorphic to D.

Choose some x_i and consider the subset of rays with an end corresponding to an element of $q^{-1}(x_i)$. Attach infinitely many handles to the union of the respective D_j in order to join the boundary ends of the D_j into a single end. Similar to Construction 3.9, this gives n–sliced Loch Ness monsters where $n \geq 2$. Repeat this process for every x_i. Now, for the remaining rays, attach handles to the corresponding disks to get 1–sliced Loch Ness monsters. Attaching handles in this manner to the base disk gives an equivalence relation on $E(D)$ which agrees with the equivalence relation given by q on a dense subset. Therefore, by continuity and the above remarks, this construction gives a surface homeomorphic to S. Now we are done, since cutting this surface along the α_i gives sliced Loch Ness monsters. □

Now we combine everything thus far.

Proof of Theorem C The first statement of this theorem is Lemma 4.13. Let S be an infinite-type surface with infinite genus and no planar ends. Apply Lemma 4.12 to cut S into components with at most one boundary chain, where the components with a boundary chain have only boundary ends. We can assume each component has infinite genus since S has only ends accumulated by genus. Then, by Lemma 4.8, the components with a boundary chain are disk with handles possibly with compact boundary components added. The other components are Loch Ness monsters possibly with compact boundary components added. □

Remark 4.14 If we allow planar ends then similar decomposition results hold, where we have to allow other one-ended building blocks. For example, when decomposing a disk with handles with planar ends similar to Lemma 4.13, we need to include disks with one boundary puncture. We could also allow finite genus. For example, when decomposing a surface without boundary similar to Lemma 4.11, we have to allow one-ended surfaces with finite genus and possibly with infinitely many compact boundary components added. In these cases we may need to allow cutting along peripheral curves and arcs.

One possible extension of Theorem C to general surfaces with noncompact boundary involves using Lemma 4.12 and the extension of Lemma 4.8 mentioned in the final remarks of its proof. In this case, we must add the modified disks discussed in these remarks to our building blocks.

5 Main results

5.1 Background

Domat has shown for surfaces with compact boundary components and at least two ends accumulated by genus that $\overline{\mathrm{PMap}_c(S)}$ is not perfect [8]. In the appendix of that paper, the author and Domat use the Birman exact sequence to extend this to the case with one end accumulated by genus. On the other hand, Calegari has shown that the mapping class group of the sphere minus a Cantor set is uniformly perfect [7]. Now we want to show many surfaces with noncompact boundary components have uniformly perfect mapping class groups.

First we need to extend a result of Patel and Vlamis to the general case, since we will use this implicitly throughout the proof of Theorem A.

Theorem 5.1 [17] *For any infinite-type surface S with only compact boundary components and at most one end accumulated by genus*, $\overline{\mathrm{PMap}_c(S)} = \mathrm{PMap}(S)$.

This result was originally stated for compact boundary, but the proof in [17] also works when there are infinitely many compact boundary components. The argument uses pants decompositions which we can construct without adaptation when there are only compact boundary components. Pants decompositions seem more tedious to use in the general case, so we instead give a slightly modified proof using a more general exhaustion. To simplify our arguments we will assume surfaces do not have degenerate ends (Definition 3.10) for the entirety of Section 5. Note this will not affect the proof of Theorem A, since filling in degenerate ends does not change the mapping class group.

Theorem 5.2 *For any infinite-type surface S with at most one end accumulated by genus*, $\overline{\mathrm{PMap}_c(S)} = \mathrm{PMap}(S)$.

Proof Let $f \in \mathrm{PMap}(S)$ be an arbitrary element. We want to find a sequence $\{f_i\}$ of elements of $\mathrm{PMap}_c(S)$ such that $f_i \to f$ in the compact–open topology. Let $\{S_i\}$ be an exhaustion of S by essential finite-type surfaces. It will suffice to show that there is always some compactly supported f_i which agrees with f on S_i. We can assume that the complementary domains of each S_i are of infinite type.

Note the orbit of any curve in S under $\mathrm{PMap}(S)$ is determined, up to isotopy, by the partition it determines on $E(S)$, the partition it determines on the compact boundary components of S, and the topological type of the complementary domains. The orbit of an arc, up to isotopy, is determined by the same properties and the endpoints of the arc. This is also true for curves and arcs in any surface.

Fix some S_i and let n be large enough that $f(S_i) \subset S_n$. Let $\{\alpha_k\}$ be the components of $\partial S_i \setminus \partial S$. First suppose α_k is a separating curve or arc. Since S has at most one end accumulated by genus, $S \setminus \alpha_k$ has one component U with finite genus. Increase n if necessary so that $S_n \cap U$ contains all of this genus. Note $f(\alpha_k)$ and α_k determine the same partition on $E(S_n)$ and the same partition on the compact boundary components of S_n.

Let $V = S_n \cap U$ and $W = S_n \cap f(U)$. Since S_n contains all the genus of U, we must have that V and W have the same genus. It follows that α_k and $f(\alpha_k)$ have homeomorphic complementary domains. Now, if α_k is nonseparating, then α_k and $f(\alpha_k)$ are both nonseparating in S_n, and so the complementary domains are homeomorphic in this case too. Therefore, we can find some $g \in \mathrm{PMap}(S_n)$ which takes $f(\alpha_k)$ to α_k. We can also require gf to fix the orientation of α_k when it is a nonseparating curve.

Now we build a compactly supported element f_i which approximates f on S_i. Start by finding some $g_1 \in \mathrm{PMap}(S_n)$ which takes $f(\alpha_1)$ to α_1. Now find some $g_2 \in \mathrm{PMap}(S_n \setminus \alpha_1)$ which takes $g_1 f(\alpha_2)$ to α_2. Repeat this process to find a sequence of compactly supported elements $\{g_k\}$ such that $g = \cdots g_3 g_2 g_1$ sends each $f(\alpha_k)$ to α_k. Also choose the g_k so that gf fixes the orientation of each α_k. Now, finally, $g \in \mathrm{PMap}_c(S)$ sends $f(S_i)$ to S_i. Then let h_i be equal to gf in S_i and the identity outside S_i, so $f_i = g^{-1} h_i$ agrees with f on S_i. □

5.2 Fragmentation

The main tool for the proof of Theorem A is a fragmentation lemma that allows us to write a map in $\overline{\mathrm{PMap}_c(S)}$ as a product of two simpler maps. This is based on fragmentation results from [9; 14], and was originally formulated by Domat in the case of no boundary. Here we provide a proof that works in the general case.

Lemma 5.3 (fragmentation) *Let S be any infinite-type surface and $f \in \overline{\mathrm{PMap}_c(S)}$. There exist two sequences of compact subsurfaces $\{K_i\}$ and $\{C_i\}$, with each sequence consisting of pairwise disjoint surfaces, and $g, h \in \overline{\mathrm{PMap}_c(S)}$ such that*

(i) $\mathrm{supp}(g) \subseteq \bigcup_i C_i$ *and* $\mathrm{supp}(h) \subseteq \bigcup_i K_i$,

(ii) $f = hg$.

Proof Fix a compact exhaustion $\{S_i\}$ of S by essential subsurfaces, and begin by setting $K_1' = S_1$. Choose some n large enough that $f(K_1') \subset S_n$, and then set $K_1 = S_n$. Now there exists some $\phi_1 \in \mathrm{PMap}(K_1)$ such that $\phi_1 f$ fixes $\partial K_1'$. Let

$$\psi_1 = \phi_1 f|_{K_1'} \in \mathrm{PMap}(K_1').$$

Then $\psi_1^{-1}\phi_1 \in \mathrm{PMap}(K_1)$ and $\psi_1^{-1}\phi_1 f$ fixes K_1'. Let $g_1 = \psi_1^{-1}\phi_1$. Next let $K_2', \dots, K_j'$ be the components of some $S_n \setminus S_{n-1}$, where n is large enough that $f(K_i')$ is disjoint from K_1 for each $2 \le i \le j$.

Now we run the same argument as before to get elements $\phi_2, \dots, \phi_j$ contained in some $\mathrm{PMap}(K_2), \dots, \mathrm{PMap}(K_j)$, respectively, with all of the K_i pairwise disjoint and such that $K_i' \subseteq K_i$ and each $\phi_i f$ fixes $\partial K_i'$. Our choices for the new K_i will be the components of some $S_n \setminus S_m$, where n and m are any numbers such that $f(K_i') \subset S_n \setminus S_m$ for each $2 \le i \le j$, and $K_1 \subset S_m$. Then let $\psi_i = \phi_i f|_{K_i'}$ and $g_i = \psi_i^{-1}\phi_i$, so that each $g_i f$ fixes K_i'.

Continue this process to obtain an infinite sequence of elements g_i and compact subsurfaces $K_i' \subseteq K_i$ such that $g_i \in \mathrm{PMap}(K_i)$, each $g_i f$ fixes K_i', and the K_i are pairwise disjoint. The g_i are compactly supported and have pairwise disjoint supports, so the product $\cdots g_3 g_2 g_1$ converges to $\bar{g} \in \overline{\mathrm{PMap}_c(S)}$. Set $g = \bar{g} f$, so that g fixes every K_i'. Now let $\{C_i\}$ be the complementary domains of $\bigcup_i K_i'$ in S, and note the C_i are compact, since each is contained in some $S_n \setminus S_m$. Note that in general the C_i are allowed to intersect the K_i. Let $h = \bar{g}^{-1}$, so that $f = hg$. Now $\mathrm{supp}(h) \subseteq \bigcup_i K_i$, as desired. Also $\bigcup_i C_i = S \setminus \bigcup_i K_i'$ and $g = \bar{g} f$ fixes each of the K_i', which shows that $\mathrm{supp}(g) \subseteq \bigcup_i C_i$.

There is one subtlety we should mention. It will often be the case that a homeomorphism supported in some K_i or C_i will be trivial in $\overline{\mathrm{PMap}_c(S)}$, so we should throw these subsurfaces out of our collections. For example, if the surface has any interior punctures, then the K_i and C_i will contain annuli bounding that puncture, and any map supported in the union of these annuli is trivial in $\overline{\mathrm{PMap}_c(S)}$. Note the above proof would also work if we were to instead work with the subgroup of $\mathrm{Homeo}_\partial^+(S)$ corresponding to homeomorphisms which can be approximated by compactly supported homeomorphisms. In this case, we would not throw out any of the subsurfaces. We could also relax the infinite-type assumption if desired. □

Remark 5.4 A critical observation is that some of the compact subsurfaces we get from fragmentation can be modified. Say K is a compact subsurface whose boundary is composed of alternating essential arcs in S and arcs in ∂S. Let $f \in \mathrm{Map}(K)$ and conflate f with a representative homeomorphism. Since f fixes ∂K, we can assume after an isotopy that f is the identity in an open regular neighborhood N of ∂K, so $f \in \mathrm{Map}(K')$, where $K' = K \setminus N$. The boundary of K' is then a union of essential simple closed curves in S.

Modifying the subsurfaces in this manner may turn a surface which separates into one that does not. For example, the rightmost two subsurfaces shown in Figure 9 can be modified to be nonseparating. This idea can be extended as follows:

Lemma 5.5 *Suppose S is a disk with handles. Let g and h be maps given by fragmentation on some $f \in \overline{\mathrm{PMap}_c(S)}$, and let $\{C_i\}$ and $\{K_i\}$ be the respective sequences of compact subsurfaces. We can assume the following*:

(i) *Each ∂K_i and ∂C_i is a single essential simple closed curve.*

(ii) *$S \setminus \bigcup_i K_i$ and $S \setminus \bigcup_i C_i$ are homeomorphic to S with compact boundary components added accumulating to some subset of the ends.*

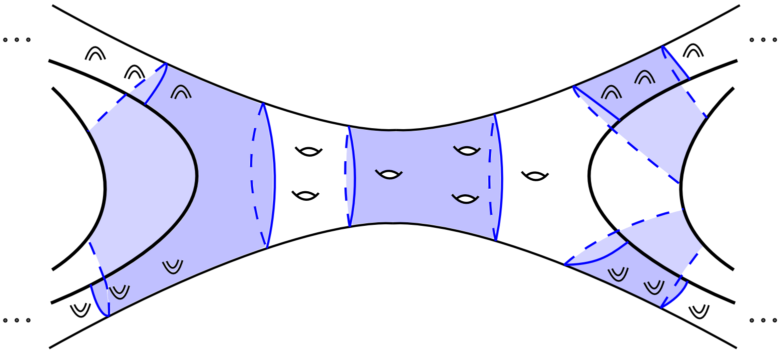


Figure 9: Example of one of the maps produced via fragmentation on a surface with two boundary chains (bold lines). The blue shaded regions represent the K_i before we modify them.

Proof Recall fragmentation depends on a given choice of a compact exhaustion $\{S_i\}$. By Proposition 4.10, we can choose our exhaustion so each ∂S_i has one component composed of alternating essential arcs in S and arcs in ∂S. From the proof of Lemma 5.3, each C_i and K_i is either some S_n or a component of some $S_n \setminus S_m$. We now show we can assume the desired conditions for h and the K_i, and the proof for the other map is similar. Since each component of ∂K_i intersects ∂S, we can modify the K_i as in Remark 5.4 so that ∂K_i is a union of essential simple closed curves. See Figure 10 for an example. In the case of fragmentation on the 2–sliced Loch Ness monster (see Figure 4 and Construction 3.9), this process will often give K_i with two boundary components, and in general this can give any finite number of boundary components.

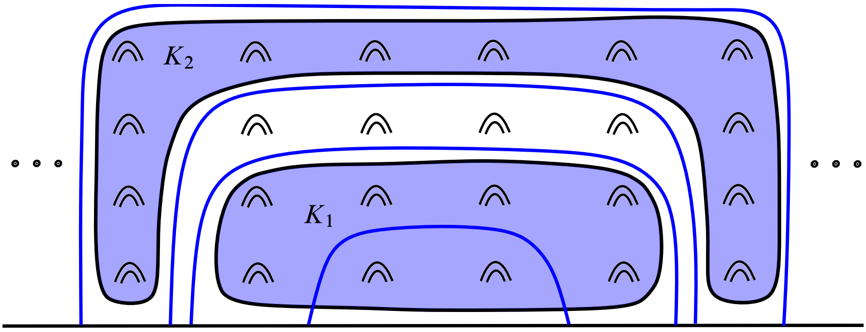


Figure 10: Example of fragmentation on a 1–sliced Loch Ness. The blue arcs correspond to the compact exhaustion used for the fragmentation. The K_i correspond to the modified subsurfaces containing the support of one of the maps from fragmentation.

Now note that none of the K_i bounds a common subsurface. By selecting the compact subsurfaces carefully in the proof of Lemma 5.3, we can assume that $S \setminus \bigcup_i K_i$ has infinite genus. It follows that $S \setminus \bigcup_i K_i$ is a disk with handles with compact boundary components added. We can also assume an end in $S \setminus \bigcup_i K_i$ is accumulated by genus if and only if the corresponding end of S is accumulated by genus. Therefore, we have the second condition of the lemma.

For any K_i with n compact boundary components where $n > 1$, connect all the components together with $n-1$ disjoint arcs $\{\alpha_k\}_{k=1}^{n-1}$ in $S \setminus \bigcup_i K_i$. Now enlarge K_i by adding in a small closed regular neighborhood of $\partial K_i \cup \bigcup_k \alpha_k$. Repeat this for every K_i, making sure the new subsurfaces are all disjoint. Now we have the first condition of the lemma. In order to maintain the second condition, we must also assume that only finitely many of the new K_i intersect any given compact subsurface. This is possible by choosing the arcs at each stage carefully. In particular, at each step let S_{j_i} be the largest subsurface in the original compact exhaustion which does not intersect K_i and choose the arcs to be outside of S_{j_i}. □

Lemma 5.6 *Suppose S is a connected sum of finitely many disks with handles. Let g and h be maps given by fragmentation on some $f \in \overline{\mathrm{PMap}_c(S)}$, and let $\{C_i\}$ and $\{K_i\}$ be the respective sequences of compact subsurfaces. We can assume the following*:

(i) *$S \setminus K_1$ and $S \setminus C_1$ are disks with handles with one compact boundary component added.*

(ii) *For the remaining C_i and K_i, each ∂K_i and ∂C_i is a single essential simple closed curve.*

(iii) *Each component of $S \setminus \bigcup_i K_i$ and $S \setminus \bigcup_i C_i$ is a disk with handles with compact boundary components added accumulating to some subset of the ends.*

Proof Suppose S is a connected sum of n disks with handles. By piecing together compact exhaustions of the disks with handles and using Proposition 4.10, we can choose our exhaustion $\{S_i\}$ of S for fragmentation so each ∂S_i has n components, each corresponding to one of the boundary chains, composed of alternating essential arcs in S and arcs in ∂S. For the h map, K_1 is equal to some S_n. Then modifying K_1 as in Remark 5.4 gives the first condition of the lemma. We get the remaining conditions for this map by following the proof of Lemma 5.5 for each component of $S \setminus K_1$. For the g map, enlarge its C_1 to be some S_n which contains K_1 and any of the C_i which intersect K_1. Then we get the desired conditions for this map by the same argument.

Note we could have stated a version of this lemma with different conditions for this second map, but that will not be necessary for the following proofs. □

5.3 Proof of Theorem A

First we use fragmentation along with standard commutator tricks to show every element of $\overline{\mathrm{PMap}_c(S)}$ can be written as a product of two commutators when S is a sliced Loch Ness monster. Then we will show the same for any disk with handles by applying Lemma 4.13. Finally we extend to the remaining cases using Lemma 4.12. During the upcoming proofs, we are implicitly using the fact that

$$\overline{\mathrm{PMap}_c(S)} = \mathrm{PMap}(S) = \mathrm{Map}(S)$$

when S is a sliced Loch Ness monster.

Lemma 5.7 $\overline{\mathrm{PMap}_c(S)}$ *is uniformly perfect when* S *is a disk with handles.*

Proof Let g be any of the two maps given by fragmentation on a general $f \in \overline{\mathrm{PMap}_c(S)}$ and let $\{C_i\}$ be the corresponding sequence of compact subsurfaces. First consider the case when our surface is the 1–sliced Loch Ness monster, L_s. By Lemma 5.5, we may assume each ∂C_i has one component and the complement of $\bigcup_i C_i$ is homeomorphic to L_s with infinitely many compact boundary components added accumulating to the single end.

Realize L_s as the closed upper half-plane with a handle attached inside an ϵ–ball at every integer point, and let ψ be the map $(x, y) \to (x+1, y)$ extended to the attached handles and isotoped in a neighborhood of the boundary to be the identity. Now we can assume, using the change-of-coordinates principle or by replacing g with a conjugate, that the C_i are contained inside the vertical strip bounded by the lines $x = \pm\frac{1}{2}$ and also the support of ψ. Letting $a = \prod_{k\geq 0} \psi^k g \psi^{-k}$, we can now write

$$g = \psi a^{-1} \psi^{-1} a = [\psi, a^{-1}].$$

See Figure 11. It now follows that we can write any $f \in \overline{\mathrm{PMap}_c(S)}$ as the product of two commutators.

Next we extend this to any n–sliced Loch Ness monster, L_s^n. First we need a model of this surface that works with the above method. Take a copy of L_s with the above half-plane model, and denote it by T. Now take the disjoint union with $n-1$ new copies of L_s realized in any way. Attach handles from T to each additional copy of L_s to join all the ends into a single end. When removing open balls from T in the process

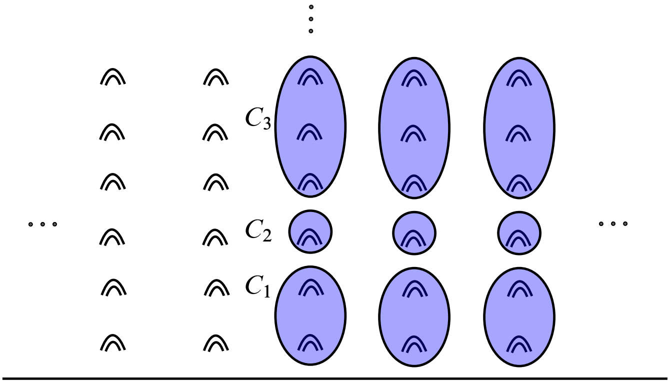


Figure 11: The last step for showing $\overline{\mathrm{PMap}_c(L_s)}$ is uniformly perfect. The support of the element a is shown in blue.

of attaching these handles, choose the open balls to be below the line $y = \frac{1}{2}$. Similar to Construction 3.9, this yields a surface homeomorphic to L_s^n, which we use as our model. See Figure 12 for an example when $n = 3$. Let $T' \subset L_s^n$ be the subsurface corresponding to the area of T above the line $y = \frac{1}{2}$. Now we can use Lemma 5.5 and the change-of-coordinates principle as before to assume that the C_i are contained above the attached handles within a vertical strip of T'. We then let ψ be the map which acts as the previous shift map on T' and fixes the remainder of L_s^n. Proceed as before to show $g = [\psi, a^{-1}]$.

Now suppose S is any disk with handles. After applying Lemma 5.5, we can assume $S \setminus \bigcup_i C_i$ is homeomorphic to S with infinitely many compact boundary components

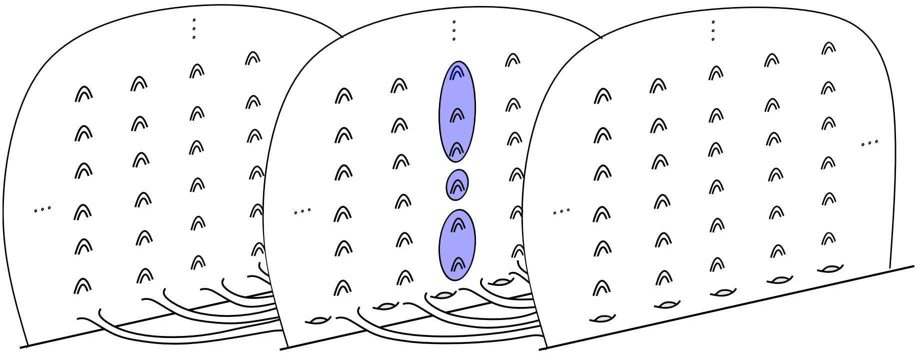


Figure 12: A model of the 3–sliced Loch Ness monster used in the proof of Lemma 5.7 with the surfaces C_i shown in blue in a vertical strip in the middle piece.

added accumulating to some subset of the ends. Using a slight variation of Lemma 4.13, where we allow the disks with handles to have compact boundary components added, we can cut S along a collection of disjoint arcs $\{\alpha_j\}$ which miss the C_i, so the components of the cut surface are sliced Loch Ness monsters. When we then cut out the C_i, we get sliced Loch Ness monsters with compact boundary components added. Give the components the models discussed in the previous paragraphs, and apply the change-of-coordinates principle argument to each component to assume each C_i is contained in a vertical strip within its respective component. Let $\{\psi_i\}$ be the collection of plane shift maps for each component analogous to the previous paragraphs. Since the supports of the ψ_i are disjoint, we have a well-defined product $\psi = \prod_i \psi_i$, and then we can show $g = [\psi, a^{-1}]$ as before. □

Lemma 5.8 $\overline{\mathrm{PMap}_c(S)}$ *is perfect when* S *is a connected sum of finitely many disks with handles with possibly finitely many punctures or compact boundary components added.*

Proof Let g be a map given by fragmentation on a general $f \in \overline{\mathrm{PMap}_c(S)}$, and let $\{C_i\}$ be the corresponding compact subsurfaces. First suppose S has no punctures or compact boundary components. When fragmenting in this case, we get supports with boundary components that are curves which separate ends (see the two leftmost subsurfaces in Figure 9). If a map is supported within one of these subsurfaces, then we cannot move the support off of itself as we did in the other cases. This is commonly referred to as a nondisplaceable subsurface (see [15, Definition 1.8]).

Apply Lemma 5.6, so we can assume the C_i have the desired properties. We can assume C_1 has genus at least 3 by replacing it with a connected compact surface containing C_1 and more of the C_i. Now $g = g_1 g_2$, where $g_1 \in \mathrm{PMap}(C_1)$ and $g_2 \in \mathrm{PMap}(S \setminus C_1)$. The classic result of Powell [18] tells us we can write g_1 as a product of commutators. By the method in the previous lemma, we can write g_2 as a single commutator. It follows that every element in $\overline{\mathrm{PMap}_c(S)}$ can be written as a product of commutators.

The cases with finitely many punctures and compact boundary components are done similarly. To consider the cases with punctures, we can slightly modify the fragmentation process by replacing a compact exhaustion with an exhaustion of finite-type surfaces. Then, depending on the number of boundary chains, we use a modification of either Lemma 5.5 or Lemma 5.6 such that C_1 includes the boundary components and punctures. □

These lemmas complete the reverse implications from Theorem A, so now we discuss why the other directions hold. For all infinite-type surfaces with only compact boundary components, $\overline{\mathrm{PMap}_c(S)}$ is not perfect, by the work of Domat. His proof relies on finding a particular sequence of disjoint essential annuli. Then he shows some multitwist about the core curves of these annuli cannot be written as a product of commutators. His work can be summarized by the following theorem. For the statement of this theorem, a nondisplaceable surface in S refers to an essential subsurface K disjoint from the noncompact boundary components of S such that $f(K) \cap K \neq \varnothing$ for all $f \in \overline{\mathrm{PMap}_c(S)}$. Note a subsurface K is nondisplaceable if it separates ends, ie if $S \setminus K$ is disconnected and induces a partition of $E(S)$ into two sets. A subsurface is also nondisplaceable if some component of $S \setminus K$ is a finite-type subsurface containing a compact boundary component of ∂S.

Theorem 5.9 [8] *Let S be an infinite-type surface such that there exists an infinite sequence of disjoint nondisplaceable essential annuli that eventually leaves every compact subsurface. Then $\overline{\mathrm{PMap}_c(S)}$ is not perfect.*

The hypothesis of Theorem 5.9 holds whenever there are interior ends of S accumulated by genus, except in the case of the Loch Ness monster, which was handled separately in the appendix of [8]. It also holds if there are infinitely many planar ends or infinitely many compact boundary components. By using Lemma 4.12, we see the hypothesis of Theorem 5.9 holds whenever there are infinitely many boundary chains as well. The only cases that remain are exactly the surfaces from Lemmas 5.7 and 5.8. This proves the forward direction of the second bullet point in Theorem A.

Finally, in order to show the forward direction of the first bullet point, we must explain why $\overline{\mathrm{PMap}_c(S)}$ is not uniformly perfect when S has more than one boundary chain, any planar ends or any compact boundary components. We will only sketch the details, since the main ideas here are taken from [8]. The issue in these cases is that there is some essential curve α which is nondisplaceable under the action of $\overline{\mathrm{PMap}_c(S)}$. Take a curve which either separates ends or bounds a finite-type subsurface containing a compact component of ∂S. The orbit of α can then be used to build a Bestvina–Bromberg–Fujiwara projection complex (see [4]) on which $\overline{\mathrm{PMap}_c(S)}$ acts by isometries. This complex is quasi-isometric to a tree, and the Dehn twist about α is a WWPD element (in the language of Bestvina, Bromberg and Fujiwara; see [3]). An adaptation of a construction of Brooks [5] then gives a quasimorphism from $\overline{\mathrm{PMap}_c(S)}$ to $\mathbb{R}$ which is unbounded on $\{T_\alpha^n\}_{n=1}^\infty$. Combining this with the fact that homogeneous quasimorphisms are bounded on commutators, we see $\overline{\mathrm{PMap}_c(S)}$ cannot be uniformly perfect.

6 Extending results

6.1 Background

In the case of surfaces with only compact boundary components, it is known that $\mathrm{PMap}(S)$ factors as a semidirect product containing $\overline{\mathrm{PMap}_c(S)}$ as one of the factors.

Theorem 6.1 (Aramayona, Patel and Vlamis [2, Corollary 6]) *Let S be an infinite-type surface with compact boundary components. Then*

$$\mathrm{PMap}(S) = \overline{\mathrm{PMap}_c(S)} \rtimes H,$$

where $H \cong \mathbb{Z}^{n-1}$ when there is a finite number $n > 1$ of ends of S accumulated by genus, $H \cong \mathbb{Z}^{\infty}$ when there are infinitely many ends accumulated by genus, and H is trivial otherwise. Furthermore, H is generated by pairwise commuting handle shifts.

Here $\mathbb{Z}^{\infty}$ refers to the direct product of a countably infinite number of copies of $\mathbb{Z}$. Although many of the results of Aramayona, Patel and Vlamis are stated for the case of compact boundary, the proofs all apply to surfaces with only compact boundary components.

In order to extend this result, we will also need to extend a well-known fact about when the inclusion of a subsurface induces an injective map between mapping class groups. Recall the definition of a degenerate end (Definition 3.10). We say a boundary chain of a surface is degenerate when every end in the chain is degenerate. After filling in degenerate ends, degenerate chains become compact boundary components. Similar to a Dehn twist about a compact boundary component, we can also speak of a Dehn twist about a degenerate chain.

Lemma 6.2 *Let S be any surface and Σ a closed essential subsurface. The natural homomorphism $i : \mathrm{PMap}(\Sigma) \to \mathrm{PMap}(S)$ is injective when the following holds:*

(i) *No compact component of $\partial\Sigma$ bounds a disk with a single interior puncture.*

(ii) *No two compact components of $\partial\Sigma$ bound an annulus.*

(iii) *There are no degenerate chains in Σ such that each boundary component of the chain bounds an upper half-plane.*

The proof will rely on the Alexander method for infinite-type surfaces. The case of compact boundary components was done in [12]. We will use a slight modification of the standard definition for a stable Alexander system.

Definition 6.3 A stable Alexander system for a surface without degenerate ends is a locally finite collection of essential simple closed curves and essential arcs Γ in a surface S such that the following properties hold:

- The elements in Γ are in pairwise minimal position.
- For distinct $\alpha_i, \alpha_j \in \Gamma$, we have that α_i is not isotopic to α_j.
- For all distinct $\alpha_i, \alpha_j, \alpha_k \in \Gamma$, at least one of the sets: $\alpha_i \cap \alpha_j$, $\alpha_i \cap \alpha_k$ or $\alpha_j \cap \alpha_k$ is empty.
- The collection Γ fills S; ie each complementary component is a disk or a disk with a single interior puncture.
- Every $f \in \mathrm{Homeo}^+_\partial(S)$ that preserves the isotopy class of each element of Γ is isotopic to the identity.

We say Γ is a stable Alexander system for a surface with degenerate ends if it becomes a stable Alexander system when the degenerate ends are filled in.

Theorem 6.4 (Alexander method) *For any infinite-type surface S, there exists a stable Alexander system Γ.*

Proof We will assume the compact boundary case from [12]. First suppose S has noncompact boundary and no degenerate ends. Embed it in the natural way inside the double, dS. Let Γ be a stable Alexander system for dS.

For an arbitrary $\gamma \in \Gamma$, isotope it to be transverse and in minimal position with ∂S so that $\gamma \cap S$ is either a curve or a union of arcs in S. Let Γ' be the collection of all curves and arcs formed in this manner. After possibly removing repeated occurrences of isotopy classes, Γ' is a stable Alexander system for S.

Now suppose S has degenerate ends. Apply the argument to S with the degenerate ends filled in, then isotope the arcs along the boundary if necessary so they descend to arcs in S. □

The proof of Lemma 6.2 will also rely on some facts about arcs. Note, given the current definition of an essential arc, in a surface with degenerate ends there may be essential arcs which bound a disk with boundary points removed. These arcs can be isotoped to be disjoint from any curve. In fact, we have the following:

Proposition 6.5 *Let S be a surface which contains essential simple closed curves. An essential arc α in S can be isotoped to be disjoint from any curve if and only if it bounds a disk with boundary points removed.*

Proof The reverse direction is clear, so suppose some essential arc α can be isotoped to be disjoint from any curve. Let $\{S_i\}$ be an exhaustion of S by compact essential subsurfaces. For any S_i large enough to contain α, we must have that α bounds a disk in S_i. Otherwise, we could construct a curve in S which cannot be isotoped away from α. It follows that α is separating and a component of $S \setminus \alpha$ has a compact exhaustion composed of only disks. This component cannot be compact, since then α would be trivial, and it cannot contain compact boundary components or interior ends, since then we could construct a curve which cannot be isotoped away from α. By Proposition 4.10, this component has a single boundary chain. Since it has no genus, no compact boundary components and no interior ends, it must be a disk with boundary points removed. □

For the following proposition and its proof, we allow all isotopies of arcs to move the endpoints along the boundary.

Proposition 6.6 *Let S be an infinite-type surface with nonempty boundary, and α an essential arc in S. There exists a collection of curves Γ disjoint from α such that the following holds: if β is an arc with endpoints on the same boundary components as α, and β can be isotoped to be disjoint from any curve in Γ, then α and β are isotopic.*

Proof First suppose S has no degenerate ends. Let $\{S_i\}$ be a compact exhaustion of S. Delete the first few subsurfaces in the exhaustion so that each $S_i \setminus \alpha$ is complex enough to contain essential simple closed curves. First suppose α is nonseparating in S_i. Then let Γ_i be a finite collection of curves in minimal position which fills the interior of $S_i \setminus \alpha$, so each complementary component of Γ_i in S_i is a disk or an annulus. When α is separating in S_i, it is possible it bounds an annulus or a pair of pants. Then let Γ_i be a collection which fills the interior of the other component. If the compact component is a pair of pants, add the curve bounding the two boundary components not containing α to Γ_i. For all other cases, we just let Γ_i be a collection which fills the interiors of both components of $S_i \setminus \alpha$. Let $\Gamma = \bigcup_i \Gamma_i$.

Suppose β is any arc as in the statement of the lemma. Choose some i large enough that S_i contains α and β and both these arcs have endpoints on the same boundary components of S_i. Now isotope β to be disjoint from every curve in Γ_i. Let A be the complementary component of Γ_i in S_i which contains α and β. Note that the complementary components of Γ_i in $S \setminus \alpha$ which intersect α are annuli. Therefore, A is the result of gluing two annuli together along a pair of arcs on their boundaries or by gluing a single annulus to itself along two arcs on the boundary. These arcs all correspond to α in A after the gluing. The single annulus case only occurs when α is an arc between two different compact boundary components of S_i, and in this case the annulus gets glued to itself by arcs on the same boundary component. It follows that A is a pair of pants. It is standard fact that there is a unique arc, up to isotopy, between any two boundary components of a pair of pants (see [11, Proposition 2.2]). It follows that β must be isotopic to α in S_i. Since this holds for all sufficiently large i, we see that β is isotopic to α in S.

Now suppose the surface has degenerate ends. If α does not become trivial after these ends are filled in, then we can apply the above argument to the filled-in surface to get the desired collection of curves. Otherwise, let Γ be the collection of all curves in S. By Proposition 6.5, if β can be isotoped to be disjoint from every curve, it must be an arc which bounds a disk with boundary points removed. The arc α also has this property. Now, since α and β have endpoints on the same boundary components, they induce the same partition of the ends space and have homeomorphic complementary components, so it follows that α and β are isotopic. □

Proof of Lemma 6.2 The last condition is similar to the first condition in the sense that it prevents Dehn twists from being in the kernel, in this case Dehn twists about degenerate chains. For example, consider any compact surface with one boundary component and then delete an embedded closed subset of the Cantor set from the boundary to form a degenerate chain. Attaching closed upper half-planes to each boundary component in the degenerate chain yields a surface with a single puncture, and the Dehn twist about the chain becomes trivial in the mapping class group of the new surface. We give a proof following Farb and Margalit [11].

Let $f \in \mathrm{PMap}(\Sigma)$ be in the kernel and conflate it with a representative homeomorphism. We extend f by the identity to a homeomorphism which represents $i(f)$. Let Γ be a stable Alexander system for Σ.

Let α be any essential simple closed curve in Σ. Since $i(f)$ is isotopic to the identity and $i(f)$ agrees with f on Σ, we have that $f(\alpha)$ is isotopic to α in S. Let $K \subset S$ be a compact essential subsurface which contains this isotopy. If K can be isotoped to be contained within Σ then we are done, so assume otherwise. Now, after isotoping ∂K and $\partial\Sigma$ to be transverse and in minimal position, $K \cap \partial\Sigma$ is a union of arcs in K. Since $f(\alpha)$ and α are contained in the interior of Σ, they are disjoint from these arcs, and it follows from a standard fact of isotopies in the compact case that there is an isotopy in Σ from $f(\alpha)$ to α missing the arcs. See for example [11, Lemma 3.16]. Although the lemma here is stated for curves instead of arcs, the same proof extends to our setting with minor changes. Therefore, f fixes the isotopy class of every curve in Σ.

Let α be an arbitrary arc in Γ. By Proposition 6.6, we can find a collection of curves in Σ such that $f(\alpha)$ is isotopic to α, by an isotopy possibly moving the endpoints, if it can be isotoped to miss each curve in the collection. This last condition holds since f fixes the isotopy class of every curve. Now we can assume by an isotopy not moving the endpoints that $f(\alpha)$ agrees with α outside of an open collar neighborhood N of the boundary components. Since Γ descends to a stable Alexander system for $S \setminus N$, we can apply the Alexander method to $S \setminus N$ to show f is supported in N. The components of N are annuli and strips $\mathbb{R} \times [-1,1]$. Since the mapping class groups of the latter components are trivial, it follows that f is a possibly infinite product of Dehn twists supported in the annuli. By the given conditions, we must now have that f is isotopic to the identity, since otherwise $i(f)$ would be nontrivial. □

Remark 6.7 Deleting a noncompact boundary component is topologically the same as attaching an upper half-plane to the component. Therefore, we can extend the above proof to show that homomorphisms such as the one discussed in Section 2 are injective. In particular, we will still have injectivity as long as we do not delete any degenerate chains or compact boundary components.

As an application of Lemma 6.2, we mention a potentially useful theorem:

Theorem 6.8 *Let S be an infinite-type surface with no compact boundary components and no degenerate chains, and suppose $f \in \operatorname{Map}(S)$ fixes the isotopy class of every curve. Then f must be the identity.*

Proof The conditions on S are necessary since otherwise a Dehn twist about a compact boundary component or degenerate chain would provide a counterexample.

Let $S' = S \cup_{\partial S} (\partial S \times [0,\infty))$ and let i be the map from $\operatorname{Map}(S)$ to $\operatorname{Map}(S')$ induced by the inclusion of S into S'. Since the conditions of Lemma 6.2 are satisfied by this inclusion, i must be injective. Curves in S' can always be isotoped by an innermost bigon argument to be inside of S, so $i(f)$ must fix every curve in S' up to isotopy. By the Alexander method for surfaces without boundary, $i(f)$ must be the identity, and so f must be as well by injectivity of i. □

Now we will prove a theorem which is a direct extension to the result shown in Section 2.

Theorem 6.9 *Let S be an infinite-type surface with at least one nondegenerate boundary chain. Then the map $i:\mathrm{Map}(S)\to\mathrm{Map}(S^{\circ})$ given by restricting a mapping class to the interior is not surjective.*

Proof By Lemma 4.12, we can cut S along curves so that each component of the cut surface has at most one boundary chain. Consider one of the components A which has a nondegenerate boundary chain. By Lemma 4.12, we can assume A has no interior ends. Now A must have a boundary end which is either accumulated by genus or compact boundary components. Cap all the compact boundary components with disks, and then apply Proposition 4.10 to get a compact exhaustion $\{A_i\}$ of A such that each ∂A_i has one component. Isotope each ∂A_i into the interior of A to get a curve α_i. Note we can assume after isotopies that $\{\alpha_i\}$ is a pairwise disjoint collection and each α_i is disjoint from the disks used to cap the compact boundary components.

Undo the capping of the compact boundary components, and then note each α_i bounds a compact subsurface and these subsurfaces form a compact exhaustion of $A\setminus C$, where C is the union of noncompact boundary components of A. Observe that $\{\alpha_i\}$ contains infinitely many nonisotopic curves. Otherwise, α_{i+1} and α_i would bound an annulus for all sufficiently large i. Then, by considering the compact exhaustion of $A\setminus C$ given by the α_i, we see that $A\setminus C$, and therefore A, has finite genus and finitely many compact boundary components. However, this is not possible by assumption. Now throw away any repeated occurrences of isotopy classes from $\{\alpha_i\}$.

Now we want to show that $T=\prod_{i=1}^{\infty}T_{\alpha_i}\in\mathrm{Map}(S^{\circ})$ is not in the image of i. Let γ be any essential arc in $A\subseteq S$ with endpoints on the noncompact boundary components such that γ does not bound a disk with boundary points removed. Now we use the same approach from Section 2 to show that, if T were in the image of i, then there would be a homeomorphism on S which sends γ to something noncompact, a contradiction. Conflate T with a homeomorphism on S which restricts to T on the interior. By the construction of the α_i and γ, for all sufficiently large i we have that γ cannot be isotoped to be disjoint from α_i. Note here we are implicitly using Proposition 6.5 applied to γ. Now we can find an infinite collection of curves $\{\beta_i\}$ which eventually leaves every compact subsurface of S such that each α_i intersects β_i, and therefore $T(\gamma)$ intersects each β_i. We are then done since it follows that $T(\gamma)$ is noncompact. One approach for finding the β_i is to consider a compact exhaustion $\{S_i\}$ of S and choose each β_i in some $S_{n_i}\setminus S_{m_i}$, where n_i and m_i go to infinity as i does. □

6.2 Extending Aramayona–Patel–Vlamis

First we will give a proof of Theorem 6.1, and then explain how to extend it to the general case. We say a handle shift h cuts a curve α when h^+ and h^- are on opposite sides of α. Let S be a surface with only compact boundary components. A *principal exhaustion* of S is an exhaustion of S by finite-type subsurfaces such that the following conditions hold for all i:

(i) Each complementary domain of S_i is an infinite-type surface.

(ii) Each component of ∂S_i is separating.

Now we state a few results from [2] which we will assume for the following proofs. Let $H_1^{\text{sep}}(S, \mathbb{Z})$ denote the subgroup of the first homology of a surface generated by classes that can be represented by separating curves on the surface.

Lemma 6.10 [2, Lemma 4.2] *Let S be a surface with only compact boundary components. Given a principal exhaustion $\{S_i\}$ of S there exists a basis of $H_1^{\text{sep}}(S, \mathbb{Z})$ composed of curves in the boundary of the S_i.*

Lemma 6.11 [2, Proposition 3.3] *Suppose S is a surface with only compact boundary components. Then we have the following*:

(1) *There is an injection ϕ from $H_1^{\text{sep}}(S, \mathbb{Z})$ to $H^1(\mathrm{PMap}(S), \mathbb{Z})$, thought of as the group of all homomorphisms from $\mathrm{PMap}(S)$ to $\mathbb{Z}$.*

(2) *Let α be a curve representing an element in $H_1^{\text{sep}}(S, \mathbb{Z})$. The homomorphism $\phi(\alpha)\colon \mathrm{PMap}(S) \to \mathbb{Z}$ sends a handle shift h to a nonzero element if and only if it cuts α, and it sends any map in $\overline{\mathrm{PMap}_c(S)}$ to 0. We can assume $\phi(\alpha)$ sends a given handle shift cutting α to 1.*

Proof of Theorem 6.1 First assume S has no planar ends or compact boundary components. The case of at most one end accumulated by genus was done in [17], so assume S has at least two ends accumulated by genus. Let $\{\alpha_i\}$ be a collection of curves forming a basis for $H_1^{\text{sep}}(S, \mathbb{Z})$, which exists by Lemma 6.10 and the fact that principal exhaustions always exist for surfaces with only compact boundary components. Now cut S along each of the α_i. Each separating curve in the cut surface bounds a compact surface, since otherwise the collection of curves above would not form a basis. Since any infinite-type surface with more than one end has separating curves which do not bound a compact subsurface, it follows that each component of the cut surface is a

Loch Ness monster with $k \in \mathbb{N} \cup \{\infty\}$ compact boundary components added. Note this gives another proof of Lemma 4.11, and the collection of curves given by this lemma will provide an example of a basis for $H_1^{\mathrm{sep}}(S, \mathbb{Z})$.

Each component Z of the cut surface can be modeled as $\mathbb{R}^2$ with k open disks removed along the horizontal axis and handles attached periodically and vertically above each removed disk. Let Y be the surface obtained from $[-1,1] \times [0,\infty) \subset \mathbb{R}^2$ by attaching a handle inside a small neighborhood about each interior integer point. We can properly embed k disjoint copies of Y into Z so that each copy of $[-1,1] \times \{0\} \subset Y$ is mapped to a different boundary component of Z.

Now we paste all of the components back together to form the original surface S. We can choose the embeddings of Y above so the union of their images is a collection of disjoint strips with genus. This then gives a collection of handle shifts $\{h_i\}$, where each h_i cuts only α_i. By Lemma 6.11, we have homomorphisms $\phi(\alpha_i)\colon \mathrm{PMap}(S) \to \mathbb{Z}$ such that $\phi(\alpha_i)$ sends h_i to 1 and every other h_j to 0. Let H be the subgroup topologically generated by the $\{h_i\}$. Since all of the h_i commute, H is a direct product of countably many copies of $\mathbb{Z}$. The product map $\phi = \prod_{i=1}^{n} \phi(\alpha_i)$ gives a homomorphism from $\mathrm{PMap}(S)$ to H. Then, by Lemma 6.11, we have a split exact sequence

$$\begin{tikzcd} 1 \arrow[r] & \overline{\mathrm{PMap}_c(S)} \arrow[r] & \mathrm{PMap}(S) \arrow[r, "\phi"] & H \arrow[r] \arrow[l, bend left, "s"] & 1, \end{tikzcd}$$

where s is inclusion. The cases of surfaces with planar ends and compact boundary components are done similarly. When there are planar ends, we choose handle shifts which miss the planar ends. Then we get the desired semidirect product. □

The general case is a corollary of this result using Lemma 6.2 along with a new version of the usual capping trick.

Construction 6.12 (capping boundary chains) Let S be a surface with noncompact boundary components. Using Lemma 4.12, we can cut S along curves so that the components of the cut surface each have at most one boundary chain. Let $\{S_i\}$ be the collection of components with exactly one boundary chain. By the final remarks in the proof of Lemma 4.8, we can build each S_i by adding topology to a disk with boundary points removed, which we will call D_i. Now we cap the boundary chains of S by attaching a copy of each D_i to the boundary of $S_i \subseteq S$ via the identity. We will denote the resulting surface by $\overline{S}$.

As an example, capping the boundary chain of any sliced Loch Ness monster gives the Loch Ness monster. Capping the boundary chain of a strip with genus gives the unique surface with empty boundary and exactly two ends, both of which are accumulated by genus (often referred to as the ladder surface). This construction was chosen because the inclusion of a surface into the capped-off surface induces a map on the ends spaces which preserves ends accumulated by genus and planar ends. Note there is a natural homomorphism

$$i\colon \mathrm{PMap}(S) \to \mathrm{PMap}(\overline{S}) \tag{1}$$

induced by inclusion, and i is injective by Lemma 6.2.

Theorem 6.13 *Let S be any infinite-type surface. Then*

$$\mathrm{PMap}(S) = \overline{\mathrm{PMap}_c(S)} \rtimes H,$$

where $H \cong \mathbb{Z}^{n-1}$ when there is a finite number $n > 1$ of ends of S accumulated by genus, $H \cong \mathbb{Z}^{\infty}$ when there are infinitely many ends accumulated by genus, and H is trivial otherwise. Furthermore, H is generated by pairwise commuting handle shifts.

Proof Recall that the case of at most one end accumulated by genus was done in Theorem 5.2. Assume S is a surface with noncompact boundary components, without planar ends or compact boundary components, and with at least two ends accumulated by genus. Let $\overline{S}$ be the capped surface given by Construction 6.12 and let i be the homomorphism between pure mapping class groups from (1) above. Note $\overline{S}$ has the same number of ends accumulated by genus as S. By Theorem 6.1, there is a split exact sequence as above with $\overline{S}$ in the place of S. Recall H is the subgroup topologically generated by disjoint handle shifts $\{h_i\}$ and s is the inclusion map. It suffices to show each of the h_i can be chosen to be inside $i(\mathrm{PMap}(S))$, because then by injectivity of i we get a split exact sequence

$$\begin{tikzcd} 1 \arrow[r] & \overline{\mathrm{PMap}_c(S)} \arrow[r] & \mathrm{PMap}(S) \arrow[r, "\phi \circ i"] & H \arrow[r] \arrow[l, bend left, "i^{-1} \circ s"] & 1. \end{tikzcd}$$

Apply Lemma 4.12 to cut S along a collection of curves so that each component of the cut surface has at most one boundary chain. As in Construction 6.12, each of the components with boundary chains can be represented as disks with boundary points removed with additional topology added. In fact, by the assumption that there are no planar ends, these components are disks with handles possibly with compact boundary

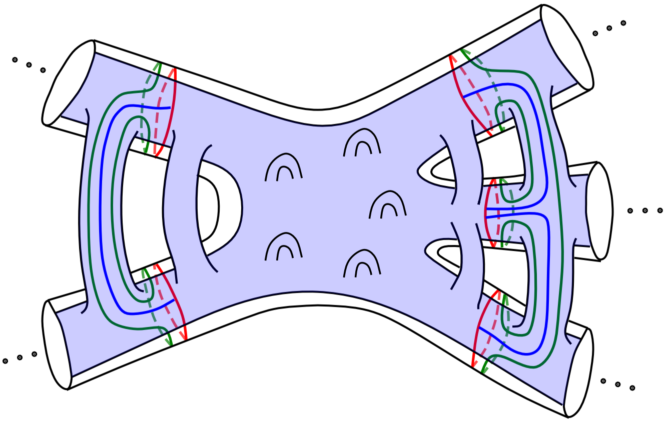


Figure 13: A disk with handles shaded blue embedded in the capped-off surface. The red curves are created by closing up arcs in the disk with handles. The blue arcs are used to replace the red curves with the green curves.

components added. We can piece together compact exhaustions on the components to get an exhaustion $\{S_i\}$ for S, and, using Proposition 4.10, we can choose the exhaustion so $\partial S_i \setminus \partial S$ is always composed of separating curves and arcs with endpoints on boundary components of the same chain. Also we can assume the exhaustion satisfies the first condition in the definition of a principal exhaustion.

Now we modify this exhaustion to get a principal exhaustion of $\overline{S}$. For every arc β_k in $\partial S_i \setminus \partial S$, there is a corresponding arc β'_k in the attached disk which, together with β_k, closes up to a curve γ_k. The γ_k together with the curves in $\partial S_i \setminus \partial S$ bound a compact subsurface $K_i \subset \overline{S}$. Then $\{K_i\}$ is a compact exhaustion for $\overline{S}$ which is not necessarily principal, but we can modify it so it becomes principal. Let U be any complementary domain of K_1 such that ∂U has $n > 1$ components. Connect each component of ∂U together with $n-1$ disjoint arcs in $U \cap S$. Now enlarge K_1 by adding a closed regular neighborhood in U of the arcs and the boundary components, then repeat this for each complementary domain with more than one boundary component. See Figure 13 for an example. Now remove some subsurfaces from the exhaustion so that $K_1 \subset K_2$, and then repeat the above process for K_2. Continue in this manner to get a principal exhaustion.

Now we sketch the final details. Find a homology basis $\{\alpha_i\}$ of $H_1^{\mathrm{sep}}(\overline{S}, \mathbb{Z})$ composed of curves that are boundary components for surfaces in the above principal exhaustion. Then we cut $\overline{S}$ along these curves and we get components which are Loch Ness monsters with compact boundary components added. Next we build the subgroup H by taking the group topologically generated by disjoint handle shifts h_i, where each h_i cuts α_i

and no other curve in the basis. In this part of the proof there is a great deal of choice for how to embed these strips; in particular, we can assume the strips are contained in S. The remaining cases are done similarly to the proof of Theorem 6.1. □

Now we show why Theorems 6.13 and A imply Theorem B.

Proof of Theorem B The reverse directions of Theorem B are immediate from Theorem A. Now notice that the commutator subgroup of

$$\mathrm{PMap}(S) = \overline{\mathrm{PMap}_c(S)} \rtimes H$$

is contained in $\overline{\mathrm{PMap}_c(S)}$ since H is abelian. Therefore, $\mathrm{PMap}(S)$ cannot be perfect when S has more than one end accumulated by genus. Since $\mathrm{PMap}(S) = \overline{\mathrm{PMap}_c(S)}$ when S has one end accumulated by genus, we get the forward implications of Theorem B from the forward implications of Theorem A and the above remark. □

References

[1] **L V Ahlfors**, **L Sario**, *Riemann surfaces*, Princeton Math. Series 26, Princeton Univ. Press (1960) MR Zbl

[2] **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 Zbl

[3] **M Bestvina**, **K Bromberg**, **K Fujiwara**, *Constructing group actions on quasi-trees and applications to mapping class groups*, Publ. Math. Inst. Hautes Études Sci. 122 (2015) 1–64 MR Zbl

[4] **M Bestvina**, **K Bromberg**, **K Fujiwara**, *Proper actions on finite products of quasi-trees*, Ann. H. Lebesgue 4 (2021) 685–709 MR Zbl

[5] **R Brooks**, *Some remarks on bounded cohomology*, from "Riemann surfaces and related topics: proceedings of the 1978 Stony Brook conference" (I Kra, B Maskit, editors), Ann. of Math. Stud. 97, Princeton Univ. Press (1981) 53–63 MR Zbl

[6] **E M Brown**, **R Messer**, *The classification of two-dimensional manifolds*, Trans. Amer. Math. Soc. 255 (1979) 377–402 MR Zbl

[7] **D Calegari**, *Big mapping class groups and dynamics*, blog post (2009) Available at `http://tinyurl.com/calegari-blog`

[8] **G Domat**, *Big pure mapping class groups are never perfect*, Math. Res. Lett. 29 (2022) 691–726 MR Zbl With an appendix joint with R Dickmann

[9] **R D Edwards**, **R C Kirby**, *Deformations of spaces of imbeddings*, Ann. of Math. 93 (1971) 63–88 MR Zbl

[10] **P Fabel**, *The mapping class group of a disk with infinitely many holes*, J. Knot Theory Ramifications 15 (2006) 21–29 MR Zbl

[11] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Math. Series 49, Princeton Univ. Press (2012) MR Zbl

[12] **J Hernández Hernández**, **I Morales**, **F Valdez**, *The Alexander method for infinite-type surfaces*, Michigan Math. J. 68 (2019) 743–753 MR Zbl

[13] **B Kerékjártó**, *Vorlesungen uber Topologie*, Grundl. Math. Wissen. 8, Springer (1923) Zbl

[14] **K Mann**, *Automatic continuity for homeomorphism groups and applications*, Geom. Topol. 20 (2016) 3033–3056 MR Zbl

[15] **K Mann**, **K Rafi**, *Large scale geometry of big mapping class groups*, preprint (2019) arXiv 1912.10914

[16] **O Ore**, *Some remarks on commutators*, Proc. Amer. Math. Soc. 2 (1951) 307–341 Zbl

[17] **P Patel**, **N G Vlamis**, *Algebraic and topological properties of big mapping class groups*, Algebr. Geom. Topol. 18 (2018) 4109–4142 MR Zbl

[18] **J Powell**, *Two theorems on the mapping class group of a surface*, Proc. Amer. Math. Soc. 68 (1978) 347–350 MR Zbl

[19] **A O Prishlyak**, **K I Mischenko**, *Classification of noncompact surfaces with boundary*, Methods Funct. Anal. Topology 13 (2007) 62–66 MR Zbl

[20] **I Richards**, *On the classification of noncompact surfaces*, Trans. Amer. Math. Soc. 106 (1963) 259–269 MR Zbl

Department of Mathematics, University of Utah
Salt Lake City, UT, United States

Current address: *Department of Mathematics, Georgia Institute of Technology*
Atlanta, GA, United States

rdickmann3@gatech.edu

Received: 2 April 2021 Revised: 6 March 2022

Algebraic & Geometric Topology **23**:6 (2023) 2823–2856
DOI: 10.2140/agt.2023.23.2823
Published: 7 September 2023

Pseudo-Anosov homeomorphisms of punctured nonorientable surfaces with small stretch factor

SAYANTAN KHAN
CALEB PARTIN
REBECCA R WINARSKI

We prove that in the nonorientable setting, the minimal stretch factor of a pseudo-Anosov homeomorphism of a surface of genus g with a fixed number of punctures is asymptotically on the order of $1/g$. Our result adapts the work of Yazdi to nonorientable surfaces. We include the details of Thurston's theory of fibered faces for nonorientable 3–manifolds.

37E30

1 Introduction

Let $S_{g,n}$ be a surface of genus g with n punctures. The mapping class group of $S_{g,n}$ consists of homotopy classes of orientation-preserving homeomorphisms of $S_{g,n}$. The Nielsen–Thurston classification of mapping classes (elements of the mapping class group) says that each mapping class is periodic, preserves some multicurve, or has a representative that is pseudo-Anosov. For each pseudo-Anosov homeomorphism $\varphi\colon S_{g,n} \to S_{g,n}$, the stretch factor $\lambda(\varphi)$ is an algebraic integer that describes the amount by which φ changes the length of curves. Arnoux and Yoccoz [4] and Ivanov [15] prove that the set

$$\mathrm{Spec}(S_{g,n}) = \{\log(\lambda(\varphi)) \mid \varphi \text{ is a pseudo-Anosov homeomorphism of } S_{g,n}\}$$

is a closed discrete subset of $(0,\infty)$. The minimum of $\mathrm{Spec}(S_{g,n})$,

$$\ell_{g,n} = \min\{\log(\lambda(\varphi)) \mid \varphi \text{ is a pseudo-Anosov homeomorphism of } S_{g,n}\},$$

quantitatively describes both the dynamics of the mapping class group of $S_{g,n}$ and the geometry of the moduli space of $S_{g,n}$.

Penner [26] showed that for orientable surfaces,

$$\ell_{g,0} \asymp \frac{1}{g}.$$

Penner conjectured that $\ell_{g,n}$ will have the same asymptotic behavior for $n \geq 0$ punctures. Following Penner, substantial attention has been given to finding bounds for $\ell_{g,n}$ — see Aaber and Dunfield [1], Bauer [5], Hironaka, Hirose, Kin and Takasawa [12; 13; 14; 17], Loving [21], and Minakawa [24] — calculating $\ell_{g,n}$ for specific values of (g,n) — see Cho, Ham and Song [7; 11], Lanneau and Thiffeault [18], and Song, Ko and Los [27] — and finding asymptotic behavior of $\ell_{g,n}$ for *orientable* surfaces with $n \geq 0$ — see Kin and Takasawa [17], Tsai [30], Valdivia [31], and Yazdi [33]. We adapt a result of Yazdi [33] to nonorientable surfaces.

Theorem 1.1 *Let $\mathcal{N}_{g,n}$ be a nonorientable surface of genus g with n punctures, and let $\ell'_{g,n}$ be the logarithm of the minimum stretch factor of the pseudo-Anosov mapping classes acting on $\mathcal{N}_{g,n}$. Then, for any fixed $n \in \mathbb{N}$, there are positive constants $B'_1 = B'_1(n)$ and $B'_2 = B'_2(n)$ such that, for any $g \geq 3$, the quantity $\ell'_{g,n}$ satisfies*

$$\frac{B'_1}{g} \leq \ell'_{g,n} \leq \frac{B'_2}{g}.$$

Pseudo-Anosov homeomorphisms Let S be a (possibly nonorientable) surface of finite type. A homeomorphism $\varphi\colon S \to S$ is said to be *pseudo-Anosov* if there exist a pair of transverse measured singular foliations $\mathcal{F}_s$ and $\mathcal{F}_u$ and a real number λ such that

$$\varphi(\mathcal{F}_s) = \lambda^{-1} \cdot \mathcal{F}_s \quad \text{and} \quad \varphi(\mathcal{F}_u) = \lambda \cdot \mathcal{F}_u.$$

The *stretch factor* of φ is the algebraic integer $\lambda = \lambda(\varphi)$.

Endow S with a Riemannian metric. The stretch factor $\lambda(\varphi)$ measures the growth rate of the length of geodesic representatives of a simple closed curve S under iteration of φ; see Fathi, Laudenbach and Poénaru [8, Proposition 9.21]. Moreover, $\log(\lambda(\varphi))$ is the minimal topological entropy of any homeomorphism of S that is isotopic to φ [8, Exposé 10].

Geometry of moduli space Let $\mathcal{T}_{g,n}$ denote the Teichmüller space of $S_{g,n}$; that is, the space of isotopy classes of hyperbolic metrics on $S_{g,n}$. When endowed with the Teichmüller metric, the mapping class group of $S_{g,n}$ acts properly discontinuously on $\mathcal{T}_{g,n}$ by isometries. The quotient of this action is the *moduli space* of $S_{g,n}$. The set $\operatorname{Spec}(S_{g,n})$ is the length spectrum of geodesics in the moduli space of $S_{g,n}$. Therefore the quantity $\ell_{g,n}$ is the length of the shortest geodesic in the moduli space of $S_{g,n}$.

Explicit bounds In his foundational work, Penner found $(\log 2)/(12g-12+4n)$ to be a lower bound for $\ell_{g,n}$ for orientable surfaces [26]. He also determined $(\log 11)/g$ to be an upper bound for $\ell_{g,0}$. Penner's work proves that $\ell_{g,0} \asymp 1/g$. McMullen [23] later asked:

Question 1.2 (McMullen) Does $\lim_{g\to\infty} g \cdot \ell_{g,0}$ exist, and, if so, what does it converge to?

To this end, Bauer [5] strengthened the upper bound for $\lim_{g\to\infty} g \cdot \ell_{g,0}$ to $\log 6$, and Minakawa [24] and Hironaka and Kin [13] further sharpened the upper bounds for $\lim_{g\to\infty} g\cdot\ell_{g,0}$ and $\lim_{g\to\infty} g\cdot\ell_{0,2g+1}$ to $\log(2+\sqrt{3})$. Later, Aaber and Dunfield [1], Hironaka [12] and Kin and Takasawa [16] determined that $\log((3+\sqrt{5})/2)$ is an upper bound for $\lim_{g\to\infty} g \cdot \ell_{g,0}$ and conjectured it is the supremum of $\lim_{g\to\infty} g \cdot \ell_{g,0}$.

Asymptotic behavior of punctured surfaces Tsai initiated the study of asymptotic behavior of $\ell_{g,n}$ along lines in the (g,n)–plane [30]. In particular, Tsai determined that for orientable surfaces of fixed genus $g \geq 2$, the asymptotic behavior in n is

$$\ell_{g,n} \asymp \frac{\log n}{n}.$$

Further, she showed that $\ell_{0,n} \asymp 1/n$. Later, Yazdi [33] determined that for an orientable surface with a fixed number of punctures $n \geq 0$, the asymptotic behavior in g is

$$\ell_{g,n} \asymp \frac{1}{g},$$

confirming the conjecture of Penner.

Nonorientable surfaces Let $\mathcal{N}_{g,n}$ be a nonorientable surface of genus g with n punctures. As above, let $\ell'_{g,n}$ denote the minimum stretch factor of pseudo-Anosov homeomorphisms of $\mathcal{N}_{g,n}$. For any $n \geq 0$ and $g \geq 1$, $\ell_{g-1,2n}$ is a lower bound for $\ell'_{g,n}$, which can be seen by passing to the orientation double cover of $\mathcal{N}_{g,n}$ (note that the definition of genus is different for orientable and nonorientable surfaces). Because the upper bounds for $\ell_{g,n}$ are constructed by example, upper bounds for $\ell'_{g,n}$ do not follow immediately from passing to the orientation double cover. Recently, Liechti and Strenner determined $\ell'_{g,0}$ for $g \in \{4,5,6,7,8,10,12,14,16,18,20\}$ [20]. Our work captures the asymptotic behavior for the punctured case.

Techniques To prove Theorem 1.1, we adapt the strategy of Yazdi [33] to nonorientable surfaces with punctures. The lower bound of $\ell'_{g,n}$ is found by lifting to the orientation double cover of $\mathcal{N}_{g,n}$. The upper bound (as in all prior work) is constructive. Fix $n \geq 0$: the desired number of punctures. Yazdi's construction is as follows. For a sequence of genera $g_{n,k}$, where k goes from 3 to ∞ and $g_{n,k} = (14k-2)n+2$, use the Penner construction [25] to obtain a homeomorphism $f_{n,k}$ of $S_{g_{n,k},n}$ that is pseudo-Anosov and has low stretch factor. In order to find pseudo-Anosov homeomorphisms of $S_{g,n}$ with small stretch factor for all g (not just those in the sequence above), construct a mapping torus for each $f_{n,k}$. To do this, Yazdi's appeals to a technique involving the use of Thurston's theory of fibered faces.

Thurston norm for nonorientable 3–manifolds In Thurston's development of what is now called the Thurston norm for 3–manifolds [28], his definitions and theorems required that all surfaces were orientable. Thurston said that the theorems should still be true for nonorientable surfaces, but there are some subtleties that have not been addressed elsewhere in the literature. In this paper, we write the details of Thurston's theory of fibered faces to nonorientable 3–manifolds. In particular, for orientable 3–manifolds, the Thurston norm is a norm on the second homology of a 3–manifold that measures the minimum complexity of an embedded (orientable) surface; it will need to be adjusted in nonorientable 3–manifolds. Specifically, the Thurston norm does not recognize embedded nonorientable surfaces in the second homology of a nonorientable 3–manifold. To address this limitation, we instead calculate the Thurston norm on the first cohomology of a nonorientable manifold. We develop a (weak) version of Poincaré duality in Theorem 2.7 that suffices to define a Thurston norm on $H^1(M;R)$ for a nonorientable 3–manifold M.

Fibered faces A special case of Thurston's hyperbolization theorem says that the monodromy of any fibration of a hyperbolic 3–manifold over S^1 is a pseudo-Anosov homeomorphism. Therefore by finding other fibrations of the same 3–manifold, one obtains additional pseudo-Anosov homeomorphism. Work of Fried [9; 10], Matsumoto [22], and Agol, Leininger and Margalit [2] can be used to bound the stretch factors of certain pseudo-Anosov homeomorphisms obtained in this way.

Outline In Section 2 we state Thurston's theory of fibered faces and adapt it to the nonorientable setting. In Section 3 we show how Thurston's theory of fibered faces can be used to construct pseudo-Anosov homeomorphisms of low stretch factor for nonorientable surfaces. Specifically, we state and prove the Nielsen–Thurston classification for nonorientable surfaces. Then we adapt the results of Fried [9; 10], Matsumoto [22], and Agol, Leininger and Margalit [2] used to construct pseudo-Anosov homeomorphisms with low stretch factor of orientable surfaces to the nonorientable setting. In Section 4, we prove Theorem 1.1, following the strategy of Yazdi.

Acknowledgements This work is the result of an REU at the University of Michigan in summer 2020. We are grateful to Alex Wright for organizing the REU and to Livio Liechti for suggesting the project. We are thankful to Eriko Hironaka, Dan Margalit, and the referees for their comments. We are also grateful to the other co-organizers, mentors and participants in the REU: Paul Apisa, Chaya Norton, Christopher Zhang, Bradley Zykoski, Anne Larsen and Rafael Saavedra. The REU was partially funded by NSF grant DMS 185615. Winarski acknowledges support of the NSF through grant 2002951.

2 Thurston norm for nonorientable 3–manifolds

Thurston defined a norm on $H_2(M;\mathbb{R})$ where M is an orientable 3–manifold [28], and this norm is now called the Thurston norm. In his manuscript, Thurston wrote: "Most of this paper works also for nonorientable manifolds but for simplicity we only deal with the orientable case". However, the details are not explained in Thurston's work or in subsequent literature. Therefore the goal of this section is to write the details of the Thurston norm for nonorientable 3–manifolds. We recall the Thurston norm for orientable manifolds in Section 2.1. In Section 2.2 we describe the challenge of defining the Thurston norm on $H_2(M;\mathbb{R})$ if M is nonorientable and present the solution of defining the Thurston norm instead on $H^1(M;\mathbb{R})$. However, Poincaré duality does not hold for nonorientable manifolds. We therefore define a condition — *relative orientability* — on a pair consisting of a manifold and an embedded surface. A surface that is relatively orientable in a nonorientable 3–manifold M will have a corresponding cohomology class in $H^1(M;\mathbb{Z})$, giving a version of Poincaré duality for nonorientable 3–manifolds as stated in Theorem 2.7. Finally, in Section 2.4, we define the oriented sum for relatively oriented embedded surfaces in nonorientable manifolds.

2.1 Thurston norm and mapping tori

In this section we recall the Thurston norm for orientable surfaces and how it detects when a 3–manifold fibers over a circle.

Mapping tori Let S be a surface and $\varphi\colon S \to S$ be a homeomorphism. A *mapping torus* of S by φ is the 3–manifold M_φ given by the identification

$$M_\varphi := \frac{S \times [0,1]}{(x,1) \sim (\varphi(x),0)}.$$

A mapping torus is a *fibration over* S^1, denoted by $S \to M_\varphi \to S^1$. A fibration defines a flow on M, called the *suspension flow*, where, for any $x_0 \in S$ and $t_0 \in S^1$, the pair

(x_0, t_0) is sent to $(x_0, t_0 + t)$. The fiber of a fibration is the preimage of any point $\theta \in S^1$ under the projection map $M_\varphi \to S^1$. If we do not specify θ, the fiber as a subset of M_φ is only well defined up to isotopy. The homology class of the fiber in $H_2(M_\varphi; \mathbb{R})$ is well defined.

A natural inverse question is to determine when a 3–manifold fibers over a circle, and the possible fibers. To this end, Thurston established a correspondence between second homology of 3–manifolds and surfaces embedded in 3–manifolds.

Complexity of an embedded surface Let M be a compact orientable closed 3–manifold. Let S be a connected surface embedded in M. The complexity of S is $\chi_-(S) = \max\{-\chi(S), 0\}$. If the surface S has multiple components $S_1, \dots, S_m$ then $\chi_-(S) = \sum_{i=1}^{m} \chi_-(S_i)$. The elements in $H_2(M; \mathbb{Z})$ can be represented by embedded surfaces inside of M [28, Lemma 1].

Thurston norm Let a be a homology class in $H_2(M; \mathbb{Z})$. Define the integer valued norm $x\colon H_2(M; \mathbb{Z}) \to \mathbb{Z}$ as

$$x(a) = \min\{\chi_-(S) \mid [S] = a \text{ and } S \text{ is compact, properly embedded and oriented}\}.$$

We then linearly extend x to $H_2(M; \mathbb{Q})$. The *Thurston norm* is the unique continuous $\mathbb{R}$–valued function that is an extension of x to $H_2(M; \mathbb{R})$. The unit ball for the Thurston norm is a convex polyhedron in $H_2(M; \mathbb{R})$.

The following remarkable theorem of Thurston [28] determines all possible fibrations of an oriented 3–manifold over the circle. We use the restatement of Yazdi [33].

Theorem 2.1 (Thurston) *Let M be an orientable 3–manifold. Let $\mathcal{F}$ be the set of homology classes in $H_2(M; \mathbb{R})$ that are representable by fibers of fibrations of M over the circle.*

(i) *Elements of $\mathcal{F}$ are in one-to-one correspondence with (nonzero) lattice points inside some union of cones over open faces of the unit ball in the Thurston norm.*

(ii) *If a surface F is transverse to the suspension flow associated to some fibration of $M \to S^1$ then $[F]$ lies in the closure of the corresponding cone in $H_2(M; \mathbb{R})$.*

The class $[F]$ has orientation such that the positive flow direction is pointing outwards relative to the surface. An open face of the unit ball is said to be a *fibered face* if the cone over the face contains the fibers of a fibration.

The goal for the rest of this section is to prove a version of Theorem 2.1 for compact nonorientable 3–manifolds. Most of the work in the proof will involve reducing the version for nonorientable 3–manifolds to the orientable version by passing to the double cover.

2.2 Thurston norm on cohomology of nonorientable mapping tori

Let $\mathcal{N}$ be a compact nonorientable surface. A naïve first attempt at defining the Thurston norm would be to define it on the $H_2(\mathcal{N};\mathbb{R})$, like in the orientable case. However, if the norm is defined on $H_2(\mathcal{N};\mathbb{R})$, the nonorientable version of Theorem 2.1 will not be true. Let $\varphi\colon \mathcal{N} \to \mathcal{N}$ be a homeomorphism and let N_φ be associated mapping torus. Clearly, N_φ fibers over S^1, and $\mathcal{N}$ is the fiber of this fibration. However, the homology class associated to $\mathcal{N}$ is the zero homology class, since the top-dimensional homology of nonorientable compact surfaces is 0–dimensional.

Our workaround for this problem will be to define a norm on the first cohomology $H^1(N_\varphi)$ rather than the second homology $H_2(N_\varphi)$. By Poincaré duality they are isomorphic for orientable 3–manifolds, but that is not true for nonorientable 3–manifolds.

Poincaré duality To see why Poincaré duality fails for nonorientable 3–manifolds, we will work through the construction of the isomorphism between first cohomology and second homology for orientable 3–manifolds. Let M be a 3–manifold. To define the Poincaré dual of $H^1(M;\mathbb{Z})$, we first define a homotopy class of maps $M \to S^1$. Then we construct an element of $H_2(M;\mathbb{Z})$. Let α be a 1–form on M and $[\alpha]$ its class in $H^1(M;\mathbb{Z})$. Fix a basepoint $y_0 \in M$. The associated map $f_\alpha\colon M \to S^1$ is given by

$$f_\alpha(y) := \int_{y_0}^{y} \alpha \bmod \mathbb{Z}. \tag{1}$$

The choice of basepoint does not affect the homotopy class of f_α (see [6, Section 5.1] for the details).

Now let $\theta \in S^1$ be a regular value so that $S = f_\alpha^{-1}(\theta)$ is a surface. To construct the associated element of $H_2(M;\mathbb{Z})$, we choose an orientation on S by assigning positive values of α to the outward-pointing normal vectors on S. Then S inherits an orientation from the orientation on M, and we have defined a fundamental class $[S] \in H_2(M;\mathbb{Z})$. We claim that $[S]$ is the Poincaré dual to α.

Lemma 2.2 *Let θ and θ' be two regular values of the function f_α and let $S = f_\alpha^{-1}(\theta)$ and $S' = f_\alpha^{-1}(\theta')$. Then, for any closed 2–form ω on M,*

(i) $\int_S \omega = \int_{S'} \omega$, *and*

(ii) $\int_S \omega = \int_M \alpha \wedge \omega$.

In particular, the homology class of S is Poincaré dual to α.

Proof To see (i), observe that S and S' are homologous, ie $f_\alpha^{-1}([\theta,\theta'])$ is a singular 3–chain that has S and S' as boundaries. By Stokes' theorem,

$$\int_{S-S'} \omega = \int_{f_\alpha^{-1}([\theta,\theta'])} d\omega = 0.$$

To prove (ii), observe that because α is the pullback of $d\xi$ along the map f_α we can write the right-hand side as

$$\int_M \alpha \wedge \omega = \int_{S^1} \left(\int_{f_\alpha^{-1}(\xi)} \omega \right) d\xi.$$

By Sard's theorem, almost every $\xi \in [0,1]$ is a regular value. Therefore the right-hand side is well defined. By (i), the inner integral is a constant function, as we vary over the ξ which are regular values of f_α. Then the integral of $d\xi$ over S^1 is 1, giving us the identity

$$\int_M \alpha \wedge \omega = \int_S \omega. \qquad \square$$

What we have here is an explicit formula for the Poincaré duality map from $H^1(M;\mathbb{R})$ to $H_2(M;\mathbb{R})$. For orientable 3–manifolds, this is an isomorphism.

Theorem 2.3 (Poincaré duality for orientable 3–manifolds) *Let M be an orientable 3–manifold, and let S be an orientable embedded surface. Then there exists a 1–form α and a regular value $\theta \in S^1$ such that S and $f_\alpha^{-1}(\theta)$ are homologous surfaces.*

Let N be a nonorientable 3–manifold. The map above from $H^1(N;\mathbb{R})$ to $H_2(N;\mathbb{R})$ is still well defined. However the map from $H^1(N;\mathbb{Z})$ to $H_2(N;\mathbb{Z})$ has a nontrivial kernel. For example, when N_φ is the mapping torus of a nonorientable surface $\mathcal{N}$, as above, the fiber is trivial in $H_2(N;\mathbb{Z})$.

Nonorientable manifolds Let N be a nonorientable 3–manifold. Let $\widetilde{N}$ and the covering map $p\colon \widetilde{N} \to N$ be the orientation double covering space of N. Let ι be the orientation-reversing deck transformation of $\widetilde{N}$. Let $N = N_\varphi$ be the mapping torus of the nonorientable surface $\mathcal{N}$ by a homeomorphism $\varphi\colon \mathcal{N} \to \mathcal{N}$. Then $\widetilde{N}$ is the mapping torus of $(\mathcal{S}, \widetilde{\varphi})$, where $\mathcal{S}$ is the orientation double cover of $\mathcal{N}$, and $\widetilde{\varphi}$ is the orientation-preserving lift of φ.

Defining the Thurston norm on cohomology In order to define the Thurston norm on $H^1(N;\mathbb{Z})$, we first need to relate $H^1(N;\mathbb{R})$ and $H^1(\widetilde{N};\mathbb{R})$. We do so by pulling back $H^1(N;\mathbb{R})$ to $H^1(\widetilde{N};\mathbb{R})$ via p. We also state the following lemma without proof (the proof is elementary).

Lemma 2.4 *The pullback $p^*\colon H^1(N;\mathbb{R}) \to H^1(\widetilde{N};\mathbb{R})$ maps $H^1(N;\mathbb{R})$ bijectively to the ι^*–invariant subspace of $H^1(\widetilde{N};\mathbb{R})$.*

Next we use Lemma 2.4 to define the Thurston norm on $H^1(N;\mathbb{R})$.

Thurston norm for nonorientable 3–manifolds Let $\alpha \in H^1(N;\mathbb{R})$ and let $\tilde{x}$ be the Thurston norm on $H^1(\widetilde{N};\mathbb{R}) \cong H_2(\widetilde{N};\mathbb{R})$. The *Thurston norm on* $H^1(N;\mathbb{R})$, is the norm $x\colon H^1(N;\mathbb{R}) \to \mathbb{R}$ defined by

$$x(\alpha) := \tilde{x}(p^*\alpha).$$

Note that defining the Thurston norm on $H^1(N;\mathbb{R})$ rather than $H_2(N;\mathbb{R})$ is not quite satisfactory. In particular, fibers of fibrations are embedded surfaces in N. In the orientable case, the embedded surfaces define the Thurston norm. In Section 2.3, we develop a (weak) version of Poincaré duality for nonorientable 3–manifolds.

2.3 Weak inverse to the Poincaré duality map

We state and prove a weak version of Poincaré duality for *relatively oriented* (nonorientable) surfaces embedded in 3–manifolds as Theorem 2.7.

Relative oriented surfaces Let M be a 3–manifold, and S an embedded surface in M. The surface S is said to be *relatively oriented with respect to M* if there is a nowhere-vanishing vector field on S that is transverse to the tangent plane of S. Two such vector fields are said to induce the same orientation if they induce the same local orientation after choosing a local frame for S. A surface S is *relatively oriented* in M if both S and the choice of positive normal vector field are specified.

If S and M are orientable, then S is relatively oriented with respect to M. But even if M is nonorientable, a nonorientable embedded surface S may be relatively oriented in M. In particular, we have the following lemma.

Lemma 2.5 *Let $\mathcal{N}$ be the fiber of a fibration $f\colon N \to S^1$. Then $\mathcal{N}$ is relatively oriented in N.*

Proof Consider a nonzero tangent vector v pointing in the positive direction at a point $\theta \in S^1$. One can pull back the tangent vector v to a nowhere-vanishing vector field over $f^{-1}(\theta) = \mathcal{N}$ because f is a fibration, ie a submersion. The pulled-back vector field defines a relative orientation for $\mathcal{N}$ in N. □

Orientable manifolds Now let M be an orientable 3–manifold, and let S be an orientable embedded surface. If S is relatively oriented with respect to M, then a choice of orientation on S determines an orientation on M and vice versa. We also need to define the notion of *incompressible surfaces* to state our version of Poincaré duality.

Incompressible surfaces Let S be a surface with positive genus embedded in a 3–manifold M. The surface S is said to be *incompressible* if there does not exist an embedded disc D in M such that D intersects S transversely and $D \cap S = \partial D$. The following result of Thurston demonstrates the link between incompressible surfaces and fibers of fibrations.

Theorem 2.6 [28, Theorem 4] *Let M be an oriented 3–manifold that fibers over S^1. Let S be an incompressible surface embedded in M. If S is homologous to a fiber, then S is isotopic to the fiber.*

In the remainder of the section, we will be working with a nonorientable 3–manifold N and an embedded nonorientable surface $\mathcal{N}$. Let $\tilde{N}$ and the covering map $p\colon \tilde{N} \to N$ be the orientation double covering space of N. Let $\tilde{\mathcal{N}}$ be the preimage of $\mathcal{N}$ under p. Let $\iota\colon \tilde{N} \to \tilde{N}$ be the orientation-reversing deck transformation of p. We will initiate N and $\mathcal{N}$ in each result below, but we suppress the initiation of the orientation double cover.

Theorem 2.7 (Poincaré duality for nonorientable 3–manifolds) *Let N be a compact nonorientable 3–manifold, and let $\mathcal{N}$ be a relatively oriented incompressible surface embedded in N. Then there exists $[\alpha] \in H^1(N;\mathbb{Z})$ such that the pullback of $[\alpha]$ to $\tilde{N}$ is the Poincaré dual of $\tilde{\mathcal{N}}$ in $\tilde{N}$. If $[\alpha]$ has a 1–form representative α that vanishes nowhere on N, then $\mathcal{N}$ is homeomorphic to $f_\alpha^{-1}(\theta)$ for all $\theta \in S^1$.*

We will refer to the 1–form α given in Theorem 2.7 as the *Poincaré dual* of the nonorientable surface $\mathcal{N}$. Before proving Theorem 2.7, we need three lemmas.

Lemma 2.8 *Let N be a nonorientable 3–manifold. Let $\mathcal{N}$ be a relatively oriented embedded surface in N, and let $\tilde{\mathcal{N}} = p^{-1}(\mathcal{N})$ in $\tilde{N}$. Then the Poincaré dual to $[\tilde{\mathcal{N}}]$ is ι^*–invariant.*

Proof A positive vector field on $\mathcal{N}$ that is transverse to its tangent plane in N lifts to a relative orientation of $\tilde{\mathcal{N}}$ in $\tilde{N}$. Since $\tilde{\mathcal{N}}$ and $\tilde{N}$ are orientable, the relative orientation of $\tilde{\mathcal{N}}$ defines an orientation of $\tilde{\mathcal{N}}$, and thus the homology class $[\tilde{\mathcal{N}}]$ in $H_2(\tilde{N};\mathbb{R})$ is well defined.

Next we show that ι reverses the orientation of $\tilde{\mathcal{N}}$. To do so, we first observe that because $\mathcal{N}$ is relatively oriented in N, the outward-pointing transverse vector field on $\mathcal{N}$ must lift to an outward-pointing transverse vector field on $\widetilde{\mathcal{N}}$. In particular, for any outward-pointing vector $\tilde{v}$ on $\mathcal{N}$, the vector $\iota(\tilde{v})$ is also outward-pointing.

Lift an outward-pointing transverse vector field on $\mathcal{N}$ to an outward-pointing transverse vector field $\widetilde{V}$ on $\widetilde{\mathcal{N}}$. Let (v_1, v_2, v_3) be a local frame for some point in $\widetilde{\mathcal{N}}$ such that v_3 is in $\widetilde{V}$. Since ι reverses the orientation of $\widetilde{N}$ but preserves the direction of v_3, ι must reverse the orientation of the pair (v_1, v_2). In particular, that means ι reverses the orientation of $\widetilde{\mathcal{N}}$.

Therefore $[\widetilde{\mathcal{N}}]$ is in the (-1)–eigenspace of the ι_* action on $H_2(\widetilde{N};\mathbb{R})$. Let the cohomology class $[\widetilde{\alpha}]$ be the Poincaré dual to $[\widetilde{\mathcal{N}}]$. Let $\widetilde{\alpha}$ be a representative 1–form $\widetilde{\alpha}$ of $[\widetilde{\alpha}]$ (that need not be ι^*–invariant). We use the fact that $\iota^2 = \mathrm{id}$ in the first and third equalities:

$$\begin{aligned}\int_{\iota_*\widetilde{\mathcal{N}}} \omega &= \int_{\widetilde{\mathcal{N}}} \iota^*\omega && \text{(by a change of variables)}\\ &= \int_{\widetilde{N}} \widetilde{\alpha} \wedge \iota^*\omega && \text{(Poincaré duality)}\\ &= \int_{\widetilde{N}} \iota^*(\iota^*\widetilde{\alpha} \wedge \omega)\\ &= \int_{\widetilde{N}} -(\iota^*\widetilde{\alpha} \wedge \omega) && (\iota \text{ is orientation-reversing}).\end{aligned}$$

Because $\iota_*[\widetilde{\mathcal{N}}] = -[\widetilde{\mathcal{N}}]$,

$$\int_{\iota_*\widetilde{\mathcal{N}}} \omega = -\int_{\widetilde{\mathcal{N}}} \omega = -\int_{\widetilde{N}} \widetilde{\alpha} \wedge \omega.$$

Since

$$\int_{\widetilde{N}} \widetilde{\alpha} \wedge \omega = \int_{\widetilde{N}} \iota^*\widetilde{\alpha} \wedge \omega$$

for all ω, it follows that $\widetilde{\alpha}$ and $\iota^*\widetilde{\alpha}$ differ by an exact form, and therefore the cohomology class $[\widetilde{\alpha}]$ is ι^*–invariant. □

As above, we will denote the Poincaré dual to $[\widetilde{\mathcal{N}}]$ by $[\widetilde{\alpha}]$. The class $[\widetilde{\alpha}]$ is an ι^*–invariant element of $H^1(\widetilde{N};\mathbb{Z})$, but it is not clear that $[\widetilde{\alpha}]$ is the pullback of an element of $H^1(N;\mathbb{Z})$ under p. In the next lemma, we show that is indeed the case.

Lemma 2.9 *Let N be a nonorientable 3–manifold. Let $[\widetilde{\alpha}] \in H^1(\widetilde{N}, Z)$ and let $\widetilde{S}$ be the Poincaré dual of $[\widetilde{\alpha}]$ in $\widetilde{N}$. There exists $[\alpha] \in H^1(N;\mathbb{Z})$ such that $\widetilde{\alpha} = p^*\alpha$.*

Proof It will suffice to show that for any simple closed curve γ in N, the integral of $\widetilde{\alpha}$ along any path lift of γ is an integer. Let $x_0 \in N$ be a basepoint of γ. Note that γ has two (path) lifts, $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ under p in $\tilde{N}$, one based at each element of $p^{-1}(x_0)$. Either $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are both simple closed curves based at each of the two preimages $p^{-1}(x_0)$, or $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are both arcs between the two points of $p^{-1}(x_0)$. If each lift, $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$, of γ is a closed curve in $\tilde{N}$, the integral $\int_{\tilde{\gamma}_i} \tilde{\alpha}$ will be an integer since $[\tilde{\alpha}] \in H^1(\tilde{N};\mathbb{Z})$ for $i = 1, 2$.

If each lift, $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$, of γ is an arc between the two preimages of $p^{-1}(x_0)$, we consider the simple closed curve $\tilde{\gamma} = \tilde{\gamma}_1 \cup \tilde{\gamma}_2$. We note that $\iota(\tilde{\gamma}) = \tilde{\gamma}$. By Lemma 2.8, $\tilde{\alpha}$ is ι^*–invariant. Therefore we have that $\int_{\tilde{\gamma}_1} \tilde{\alpha} = \int_{\tilde{\gamma}_2} \tilde{\alpha}$, so

$$\int_{\tilde{\gamma}} \tilde{\alpha} = 2 \int_{\tilde{\gamma}_1} \tilde{\alpha}.$$

It will suffice to show that $\int_{\tilde{\gamma}} \tilde{\alpha}$ is an even integer. Without loss of generality, we can assume all intersections of the simple closed curve $\tilde{\gamma}$ with the surface $\tilde{\mathcal{S}}$ are transverse. Since $\tilde{\alpha}$ is a representative of the Poincaré dual to $[\tilde{\mathcal{S}}]$, the integral of $\tilde{\alpha}$ along $\tilde{\gamma}$ is the signed intersection number of $\tilde{\gamma}$ with $\tilde{\mathcal{S}}$. The intersection number must be even, for if $\tilde{\gamma}$ and $\tilde{\mathcal{S}}$ intersect at a point y, then they also intersect at $\iota(y)$. □

The last lemma we need is that lifts of incompressible surfaces are incompressible.

Lemma 2.10 *Let N be a nonorientable 3–manifold. If $\mathcal{N}$ is a relatively oriented incompressible surface in N, then $\tilde{\mathcal{N}} = p^{-1}(\mathcal{N})$ is incompressible in $\tilde{N}$.*

Proof Because $\mathcal{N}$ is incompressible in N, the map on fundamental groups induced by inclusion $\mathcal{N} \to N$ is injective. Since $p_* : \pi_1(\tilde{N}) \to \pi_1(N)$ is injective, the induced map $\pi_1(\tilde{\mathcal{N}}) \to \pi_1(\tilde{N})$ must also be injective. An injective induced map on fundamental groups is equivalent to the orientable surface $\tilde{\mathcal{N}}$ being incompressible. □

We now have everything we need to finish proving Theorem 2.7.

Proof of Theorem 2.7 Let $\tilde{\mathcal{N}} = p^{-1}(\mathcal{N})$. The relative orientation of $\tilde{\mathcal{N}}$ determines a homology class $[\tilde{\mathcal{N}}] \in H_2(N;\mathbb{Z})$. Let the 1–form $\tilde{\alpha}$ be the Poincaré dual to $[\tilde{\mathcal{N}}]$ in $\tilde{N}$. By Lemma 2.9, there exists a 1–form $\alpha \in H^1(N;\mathbb{Z})$ such that $\tilde{\alpha} = p^*\alpha$.

We define the map $f_\alpha : N \to S^1$ according to (1). Because α is nonvanishing, f_α has full rank everywhere. Therefore f_α is a fibration. The map $f_\alpha \circ p$ is a lift of f_α to $\tilde{N}$ under p, and is therefore also a fibration. By Lemma 2.10, $\tilde{\mathcal{N}}$ is incompressible. It follows from the orientable version of Poincaré duality that $\tilde{\mathcal{N}}$ and $p^{-1}(f_\alpha^{-1}(\theta))$ are homologous surfaces in $\widetilde{N}$. Theorem 2.6 then tells us $\widetilde{\mathcal{N}}$ must be isotopic to a fiber of $f_\alpha \circ p$. The restriction of p to the homeomorphic surfaces $\widetilde{\mathcal{N}}$ and $p^{-1}(f_\alpha^{-1})(\theta)$ determines two equivalent 2–fold covering maps $\widetilde{\mathcal{N}} \to \mathcal{N}$ and $p^{-1}(f_\alpha^{-1}(\theta)) \to f_\alpha^{-1}(\theta)$. Therefore the image surfaces $\mathcal{N}$ and $f_\alpha^{-1}(\theta)$ must also be homeomorphic. □

Note that the above proof does not tell us that $\mathcal{N}$ and $f_\alpha^{-1}(\theta)$ are isotopic. Isotopy of the fibers of N requires the isotopy between $\widetilde{\mathcal{N}}$ and $p^{-1}(f_\alpha^{-1}(\theta))$ to be ι^*–equivariant. However, the theorem is sufficient for our application.

We conclude the section with a nonorientable version of Theorem 2.1.

Theorem 2.11 *Let N be a compact nonorientable 3–manifold, and let $\mathcal{F}$ be the elements of $H^1(N;\mathbb{Z})$ corresponding to fibrations of N over S^1.*

(i) *Elements of $\mathcal{F}$ are in one-to-one correspondence with (nonzero) lattice points — ie points of $H^1(N;\mathbb{Z})$ — inside some union of cones over open faces of the unit ball in the Thurston norm.*

(ii) *Let $\mathcal{N}$ be relatively oriented surface in N that transverse to the suspension flow associated to some fibration $f\colon N \to S^1$. Let $[\alpha]$ be the Poincaré dual $[\alpha]$ to $\mathcal{N}$. Then $[\alpha]$ lies in the closure of the cone in $H^1(N;\mathbb{R})$ containing the 1–form corresponding to f.*

Proof For (i), we observe that by Theorem 2.1 the fibrations of $\widetilde{N}$ are in one-to-one correspondence with lattice points inside a union of cones over open faces of the unit ball in $H_2(\widetilde{N};\mathbb{R})$. Let $\widetilde{\mathcal{K}}$ be the union of cones in $H_2(\widetilde{N};\mathbb{R})$. By Poincaré duality, $\widetilde{\mathcal{K}}$ is in one-to-one correspondence to a union of cones in $H^1(\widetilde{N};\mathbb{R})$, which we will call $\widetilde{\mathcal{K}}^*$.

Because $H^1(N;\mathbb{R})$ is isomorphic to a subspace of $H^1(\widetilde{N};\mathbb{R})$, we can construct a union of cones in $H^1(N;\mathbb{R})$ that map to the intersection of $p^*(H^1(N;\mathbb{R}))$ with $\widetilde{\mathcal{K}}^*$. Indeed, every lattice point in $\widetilde{\mathcal{K}}^*$ corresponds to a fibration $f\colon N \to S^1$, since the pullback of f to $H^1(\widetilde{N};\mathbb{Z})$ corresponds to a fibration of $\widetilde{N}$. Conversely, every fibration of $f\colon N \to S^1$ must correspond to an element of $\widetilde{\mathcal{K}}^*$, since the composition $f \circ p$ is a fibration of $\widetilde{N} \to S^1$.

For (ii), assume that the surface $\mathcal{N}$ is transverse to the suspension flow of a fibration $f\colon N \to S^1$. Then $\widetilde{\mathcal{N}}$ is transverse to the suspension flow $p \circ f\colon \widetilde{N} \to S^1$. Let $\widetilde{\alpha}$ be the pullback of α under p. Then $\widetilde{\alpha}$ is the Poincaré dual of $\widetilde{\mathcal{N}}$. By Theorem 2.1, the 1–form $\widetilde{\alpha}$ lies in the closure of a component of $\widetilde{\mathcal{K}}^*$ that contains the 1–form corresponding to $f \circ p$. Let $\widetilde{K}$ be this component. Let $K \subset H^1(N;\mathbb{R})$ be the preimage of $\widetilde{K}$ under p^*. The cone K contains both α and the 1–form corresponding to f, as desired. □

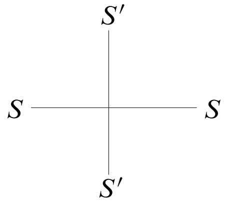


Figure 1: Cross section of intersection of S and S'.

2.4 Oriented sums

The next step in studying embedded nonorientable surfaces will be to describe *oriented sums*. Let M be a 3–manifold. The oriented sum of two embedded surfaces in M is additive in both the Euler characteristic and $H^1(M;\mathbb{R})$. This operation is well known in the case of orientable 3–manifolds (along with orientable embedded surfaces), but we will sketch the relevant details. We then extend the construction to relatively oriented embedded surfaces.

Oriented sum for oriented manifolds Let M be an orientable manifold. Let S and S' be orientable embedded surfaces in M. Assume that S and S' intersect transversally. Thus $S \cap S'$ is a disjoint union of copies of S^1. For each component of $S \cap S'$, take a tubular neighborhood that has cross section as in Figure 1.

We then perform a surgery on the leaves of S and S' so that the outward-pointing normal vector fields match as in Figure 2.

By performing this surgery at all the intersections, we get a new submanifold S'' of M (which may have multiple components). This new submanifold S'' is called the *oriented sum* of S and S'. The operation of taking oriented sums is additive on Euler

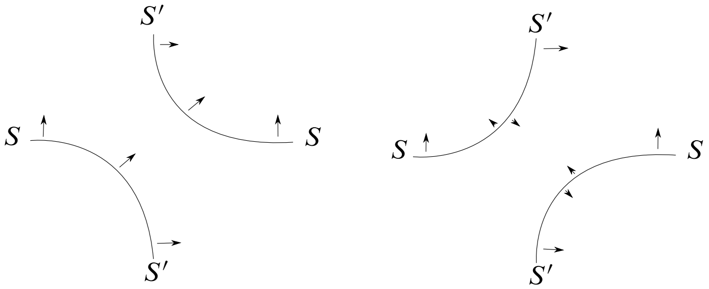


Figure 2: On the left, the normal vectors on S and S' are consistent. On the right, they are not.

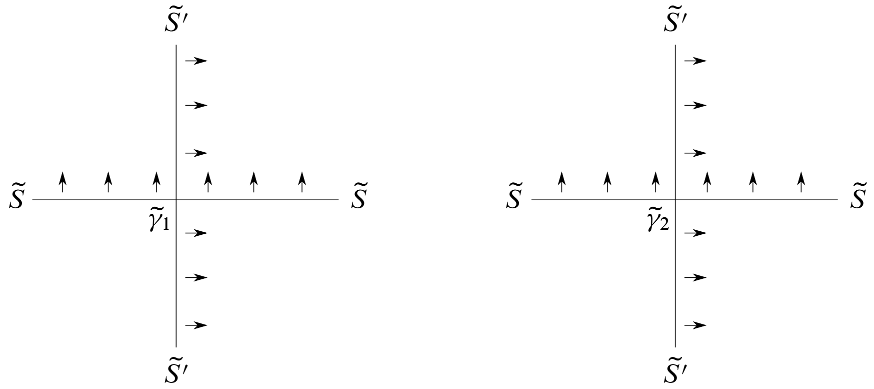


Figure 3: Neighborhoods of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, with the outward-pointing normal vector field.

characteristic, as well as the homology classes (and thus the cohomology classes of their Poincaré duals),

$$\chi(S'') = \chi(S) + \chi(S'), \quad [S''] = [S] + [S'].$$

Oriented sum for nonorientable manifolds Let N be a nonorientable 3–manifold and let $\mathcal{N}$ and $\mathcal{N}'$ be embedded surfaces in N that are relatively oriented. We define the oriented sum on $\mathcal{N}$ and $\mathcal{N}'$ as follows. As above, let $p\colon \widetilde{N} \to N$ be the orientation double cover and let ι be the orientation-reversing deck transformation of $\widetilde{N}$. Let $\widetilde{\mathcal{N}} = p^{-1}(\mathcal{N})$ and $\widetilde{\mathcal{N}}' = p^{-1}(\mathcal{N}')$, which are embedded oriented surfaces in $\widetilde{N}$. The oriented sum of $\mathcal{N}$ and $\mathcal{N}'$ is the image under p of the oriented sum of $\widetilde{\mathcal{N}}$ and $\widetilde{\mathcal{N}}'$.

To see that the operation is well defined, we recall that ι preserves the relative orientation of $\widetilde{\mathcal{N}}$ and $\widetilde{\mathcal{N}}'$. Therefore ι leaves the outward normal vector fields on $\widetilde{\mathcal{N}}$ and $\widetilde{\mathcal{N}}'$ invariant (see the proof of Lemma 2.8). Thus a leaf L of $\widetilde{\mathcal{N}}$ is surgered with a leaf of L' of $\widetilde{\mathcal{N}}'$ if and only if $\iota(L)$ and $\iota(L')$ are surgered. Therefore surgery factors through p and $[\mathcal{N}] + [\mathcal{N}']$ is well defined for nonorientable surfaces.

Example 2.12 Let γ be a component of $\mathcal{N} \cap \mathcal{N}'$ and $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ be the path lifts of γ. One possible orientation of $\widetilde{S}$ and $\widetilde{S}'$ is given in Figure 3. The outward-pointing normal vectors to $\widetilde{\mathcal{N}}$ and $\widetilde{\mathcal{N}}'$ determine which leaves are glued together along $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$.

To preserve the normal vector field, glue the left $\widetilde{\mathcal{N}}$ leaf to the bottom $\widetilde{\mathcal{N}}'$ leaf near $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$. Since $\iota(\widetilde{\gamma}_1) = \widetilde{\gamma}_2$, the outward-pointing normal vector fields point the same (relative) directions.

Additivity By the consistency of the oriented sum in N and $\widetilde{N}$, it easily follows that the oriented sum is additive in Euler characteristic, as well as in terms of Poincaré dual, since the Poincaré dual was also defined by passing to the orientation double cover.

3 Mapping classes with small stretch factors

In this section, we provide a strategy to compute pseudo-Anosov homeomorphisms with small stretch factors.

3.1 Mapping class groups of nonorientable surfaces

Let $\mathcal{N}$ be a nonorientable surface and let $\widetilde{\mathcal{N}}$ and the covering map $p\colon \widetilde{\mathcal{N}} \to \mathcal{N}$ be its orientation double covering space. Every homeomorphism $\varphi\colon \mathcal{N} \to \mathcal{N}$, has a unique orientation-preserving lift $\widetilde{\varphi}\colon \widetilde{\mathcal{N}} \to \widetilde{\mathcal{N}}$.

A consequence is that lifting homeomorphisms induces a monomorphism between homeomorphisms of $\mathcal{N}$ and orientation-preserving homeomorphisms of $\widetilde{\mathcal{N}}$. Every homotopy of $\mathcal{N}$ lifts to a homotopy of $\widetilde{\mathcal{N}}$. Therefore there is an inclusion from the mapping class group of $\mathcal{N}$ to the (orientation-preserving) mapping class group of $\widetilde{\mathcal{N}}$. This inclusion also respects the Nielsen–Thurston classification of mapping classes, both qualitatively, and quantitatively, as the following proposition shows.

Proposition 3.1 *Let $\varphi\colon \mathcal{N} \to \mathcal{N}$ be a homeomorphism and let $\widetilde{\varphi}\colon \widetilde{\mathcal{N}} \to \widetilde{\mathcal{N}}$ be the orientation-preserving lift of φ. Then*

(i) *φ is periodic if and only if $\widetilde{\varphi}$ is periodic*,

(ii) *φ is reducible if and only if $\widetilde{\varphi}$ is reducible*, *and*

(iii) *φ is pseudo-Anosov if and only if $\widetilde{\varphi}$ is pseudo-Anosov. Moreover*, *if φ has stretch factor λ*, *then $\widetilde{\varphi}$ also has stretch factor λ.*

Proof The fact that the map from $\operatorname{Mod}(\mathcal{N})$ to $\operatorname{Mod}(\widetilde{\mathcal{N}})$ is type-preserving follows from Aramayona, Leininger and Souto [3, Lemma 10] (while the statement of the lemma is for orientable surfaces, the argument, which we will skip, is identical for nonorientable surfaces).

Suppose now that $\varphi\colon \mathcal{N} \to \mathcal{N}$ is a pseudo-Anosov homeomorphism with stretch factor λ and stable and unstable foliations $\mathcal{F}_s$ and $\mathcal{F}_u$ respectively. Let $\widetilde{\mathcal{F}}_s$ and $\widetilde{\mathcal{F}}_u$ denote the lifts of the stable and unstable foliations to the orientation double cover. Let γ be a simple closed curve in $\widetilde{\mathcal{N}}$. We need to show that the following identities hold for all γ

(see [8, Exposé 5] for the definition of intersection number with measured foliations; the fact that these identities suffice follows from [8, Lemma 9.15]):

$$i(\gamma, \tilde{\varphi}(\tilde{\mathcal{F}}_u)) = \lambda \cdot i(\gamma, \tilde{\mathcal{F}}_u), \tag{2}$$

$$i(\gamma, \tilde{\varphi}(\tilde{\mathcal{F}}_s)) = \frac{1}{\lambda} \cdot i(\gamma, \tilde{\mathcal{F}}_s). \tag{3}$$

To see that (2) holds, we partition γ into short arcs $\{\gamma_i\}$ such that the restriction of the covering map p to a neighborhood of each arc is a homeomorphism. Then

$$i(\gamma_i, \tilde{\mathcal{F}}_u) = i(p(\gamma_i), \mathcal{F}_u), \tag{4}$$

$$i(\gamma_i, \tilde{\varphi}(\tilde{\mathcal{F}}_u)) = i(p(\gamma_i), \varphi(\mathcal{F}_u)). \tag{5}$$

Since we know that $\mathcal{F}_u$ is the unstable foliation for φ with stretch factor λ, we can compute the ratio of the right-hand side of (4) and (5),

$$i(p(\gamma_i), \varphi(\mathcal{F}_u)) = \lambda \cdot i(p(\gamma_i), \mathcal{F}_u). \tag{6}$$

Combining (4), (5) and (6) and summing over all γ_i gives us that (2) holds. A similar argument also proves that (3) holds. □

3.2 Constructing pseudo-Anosov maps using oriented sums

The goal of this section is to prove that the stretch factor of any pseudo-Anosov homeomorphism provides an asymptotic upper bound for the minimum stretch factor. We do this in Proposition 3.2.

Proposition 3.2 *Let $\mathcal{N}_g$ be a nonorientable surface of genus g and let $\varphi\colon \mathcal{N}_g \to \mathcal{N}_g$ be a pseudo-Anosov homeomorphism with stretch factor λ. Let N_φ be the mapping torus of $\mathcal{N}_g$ by φ. Let $\mathcal{N}_{g'}$ be a genus 3 nonorientable relatively orientable surface embedded in N_φ that is transverse to the suspension flow associated to φ. Then, for all $k \in \mathbb{Z}^+$, there is a pseudo-Anosov homeomorphism of the oriented sum $\mathcal{N}_g + k\mathcal{N}_{g'}$ with stretch factor at most λ.*

Our strategy for proving Proposition 3.2 is to find fibrations of N_φ over S^1 that have fiber $\mathcal{N}_g + k\mathcal{N}_{g'}$. We then apply a special case of Thurston's hyperbolization theorem, which says that the mapping torus of an orientable surface S by a homeomorphism φ is hyperbolic if and only if φ is pseudo-Anosov [29, Theorem 0.1]. In particular, Thurston's theorem implies that if $M = M_\varphi$ fibers over S^1 in two ways, either both monodromies are pseudo-Anosov or neither monodromy is pseudo-Anosov. Finally, we adapt theorems of Fried and Matsumoto (Theorem 3.4) and Agol, Leininger and Margalit (Theorem 3.5) to work for mapping tori with nonorientable fibers.

We will use the following two facts for orientable surfaces and hyperbolic 3–manifolds that fiber over S^1:

(1) [6, Lemma 5.7] A Thurston norm-minimizing surface S is incompressible.

(2) [28, Corollary 2] The fiber of any fibration over S^1 minimizes the Thurston norm in its homology class.

Proposition 3.3 *Let $\mathcal{N}'$ be a genus 3 nonorientable relatively orientable surface embedded in N that is transverse to the suspension flow associated to φ. Let α be the Poincaré dual of $\mathcal{N}$ and α' the Poincaré dual of $\mathcal{N}'$. If the oriented sum of $\mathcal{N}$ and $\mathcal{N}'$ is connected, then $\mathcal{N} + \mathcal{N}'$ is homeomorphic to the fiber of the fibration given by $\alpha + \alpha'$.*

Proof We first need to show that $\mathcal{N}'$ is incompressible to consider its Poincaré dual. This follows from the fact that the preimage $\widetilde{\mathcal{N}}'$ in the orientation double cover is a genus 2 surface, and minimizes the Thurston norm in its homology class. If it did not minimize the Thurston norm in the homology class, then the norm minimizing surface in its homology class would have to be a torus or a sphere, but that would contradict the fact that the 3–manifold is the mapping torus of a pseudo-Anosov map. By Calegari [6, Lemma 5.7], we have that $\widetilde{\mathcal{N}}'$ incompressible, and therefore $\mathcal{N}'$ is incompressible.

Let $p\colon \widetilde{N} \to N$ be the orientation double cover of N. The surface $\mathcal{N}$ minimizes the Thurston norm because it is a fiber of f. Similarly, $p^{-1}(\mathcal{N})$ also minimizes the Thurston norm. Thus the Thurston norm of α is $2\chi_-(\mathcal{N})$. Likewise, the Thurston norm of α' is $2\chi_-(\mathcal{N}')$.

By Theorem 2.11(ii), α' lies in the same cone in $H^1(N;\mathbb{Z})$ as α. The Thurston norm x on $H^1(N;\mathbb{Z})$ is a linear function on that cone. Since the Thurston norm is also linear on oriented sums of $\mathcal{N}$ and $\mathcal{N}'$,

$$x(\alpha+\alpha') = x(\alpha) + x(\alpha') = 2\chi_-(\mathcal{N}) + 2\chi_-(\mathcal{N}') = 2\chi_-(\mathcal{N}+\mathcal{N}').$$

Because $2\chi_-(\mathcal{N}+\mathcal{N}')$ achieves the Thurston norm of $\alpha+\alpha'$, the preimage $p^{-1}(\mathcal{N}+\mathcal{N}')$ achieves the Thurston norm of the pullback of $\alpha+\alpha'$ under p. Therefore $p^{-1}(\mathcal{N}+\mathcal{N}')$ is incompressible. Thus $\mathcal{N}+\mathcal{N}'$ is also incompressible.

By Theorem 2.11(i), $\alpha+\alpha'$ corresponds to some other fibration $f''\colon N \to S^1$. By Theorem 2.7, the fiber of f'' must be homeomorphic to $\mathcal{N}+\mathcal{N}'$. □

In the proof of Proposition 3.2, we will use Proposition 3.3 along with a theorem of Thurston to obtain a pseudo-Anosov homeomorphism φ_k of the surface of genus $g+kg'$. We then use Theorems 3.4 and 3.5 to obtain a upper bound on the stretch factor of φ_k.

Theorem 3.4 (Fried [9; 10] and Matsumoto [22]) *Let M be an orientable hyperbolic 3–manifold and let $\mathcal{K}$ be the union of cones in $H^1(M;\mathbb{R})$ whose lattice points correspond to fibrations over S^1. There exists a strictly convex function $h\colon \mathcal{K} \to \mathbb{R}$ satisfying the following properties*:

(i) $h(cu) = \frac{1}{c}h(u)$ *for all* $c > 0$ *and* $u \in \mathcal{K}$.

(ii) *For every primitive lattice point* $u \in \mathcal{K}$, $h(u) = \log(\lambda)$, *where* λ *is the stretch factor of the pseudo-Anosov map associated to this lattice point.*

(iii) $h(u)$ *goes to* ∞ *as* u *approaches* $\partial\mathcal{K}$.

Theorem 3.5 (Agol, Leininger and Margalit) *Let $\mathcal{K}$ be a fibered cone for a mapping torus M and let $\overline{\mathcal{K}}$ be its closure in $H^1(M;\mathbb{R})$. If $u \in \mathcal{K}$ and $v \in \overline{\mathcal{K}}$, then $h(u+v) < h(u)$.*

Proof of Proposition 3.2 The oriented sum

$$\mathcal{S} = \mathcal{N}_g + k\mathcal{N}_{g'}$$

constructed in Proposition 3.3 is a surface of genus $g + kg'$, and $\mathcal{S}$ is homeomorphic to a fiber of N_φ given by $\alpha + k\alpha'$. Let $\varphi_k : \mathcal{S} \to \mathcal{S}$ be the monodromy of N_φ over $\mathcal{S}$. By Thurston's theorem, φ_k is pseudo-Anosov. We claim that φ_k has stretch factor at most λ.

Let $p\colon \widetilde{N} \to N_\varphi$ be the orientation double cover of N_φ. Let $h|_N$ be the restriction of h to the pullback $p^*(H^1(N_\varphi;\mathbb{R}))$ in $H^1(\widetilde{N};\mathbb{R})$. The restriction $h|_N$ satisfies all the properties of Theorems 3.4 and 3.5.

Let $\tilde{\varphi}$ be the orientation-preserving lift of φ to $p^{-1}(\mathcal{N})$. Since $\tilde{\alpha}$ is the pullback of α, the map $\tilde{\varphi}$ is the pseudo-Anosov homeomorphism associated to $\tilde{\alpha}$. By Proposition 3.1, the stretch factor of $\tilde{\varphi}$ is λ.

Let $\mathcal{K}$ be the cone in $H^1(N_\varphi;\mathbb{R})$ that contains α. Since $\mathcal{N}_{g'}$ is transverse to the suspension flow in the direction of φ, we have that α' is in the closure of $\mathcal{K}$ in $H^1(N;\mathbb{R})$. Let $\tilde{\alpha}$ be the pullback of α under p and let $\tilde{\alpha}'$ be the pullback of α' under p. Then $h|_N(\tilde{\alpha} + \tilde{\alpha}') < h|_N(\tilde{\alpha})$. By Theorem 3.4, $h(\tilde{\alpha})$ is equal to the stretch factor of the pseudo-Anosov homeomorphism associated to $\tilde{\alpha}$. Therefore $h|_N(\tilde{\alpha} + \tilde{\alpha}') < \log(\lambda)$. It follows that the stretch factor of φ_k is less than λ. □

4 Minimal stretch factors for nonorientable surfaces with marked points

In this section we will use Theorem 2.11 and Proposition 3.2 to adapt the methods of Yazdi [33] to nonorientable surfaces. We recall the statement of the main theorem:

Theorem 1.1 *Let $\mathcal{N}_{g,n}$ be a nonorientable surface of genus g with n punctures, and let $\ell'_{g,n}$ be the logarithm of the minimum stretch factor of the pseudo-Anosov mapping classes acting on $\mathcal{N}_{g,n}$. Then, for any fixed $n \in \mathbb{N}$, there are positive constants $B'_1 = B'_1(n)$ and $B'_2 = B'_2(n)$ such that, for any $g \geq 3$, the quantity $\ell'_{g,n}$ satisfies*

$$\frac{B'_1}{g} \leq \ell'_{g,n} \leq \frac{B'_2}{g}.$$

Observe that the lower bound for the nonorientable case follows easily from the lower bound for the orientable case. Indeed, let φ be a pseudo-Anosov map with the minimal stretch factor on $\mathcal{N}_{g,n}$. The orientation double cover of $\mathcal{N}_{g,n}$ is $\mathcal{S}_{G,2n}$, where $G = g-1$. Note that in the nonorientable case we measure genus as the number of copies of the projective plane attached to S^2 via a connect sum and in the orientable case we measure genus as the number of copies of the torus attached to S^2 via a connected sum. Let $\tilde{\varphi}\colon \mathcal{S}_{G,2n} \to \mathcal{S}_{G,2n}$ be the orientation-preserving lift of φ. By Proposition 3.1, $\tilde{\varphi}$ has the same stretch factor as φ. The logarithm of the former is bounded below by B_1/G, where B_1 is given by Yazdi [33], and thus the stretch factor of φ is bounded below as well. The more challenging part of the proof is showing that the upper bound holds.

We will closely follow Yazdi's construction, which proceeds in five steps, though we will reorder them for clarity. In Steps 1 and 2, we construct a family of pseudo-Anosov homeomorphisms of $\mathcal{N}_{g_i,n}$, where $\{g_i\}$ is an unbounded increasing sequence. However the sequence $\{g_i\}$ does not contain all natural numbers. In Step 3 we give an upper bound to the stretch factor of the previously constructed homeomorphisms. In Steps 4 and 5, we construct pseudo-Anosov maps on surfaces of genera that do not belong to the sequence $\{g_i\}$. It is in Steps 4 and 5 that we use Thurston's fibered face theory. We have adapted each of Yazdi's five steps to work for nonorientable surfaces.

Step 1: constructing the surfaces

We begin by defining a family of surfaces $P_{n,k}$. Let S be an orientable surface of genus 5 with three boundary components. Call the boundary components c, d and e. Choose an orientation of S and let c, d and e inherit the induced orientations. Let p and q be marked points in the boundary component e. In Step 5 we will remove p and all its copies. Let r and s be the components of $e \setminus \{p, q\}$. We obtain a nonorientable surface T from S by adding two cross caps to S (retaining the orientation of the boundary components of S). The resulting surface T is shown in Figure 4.

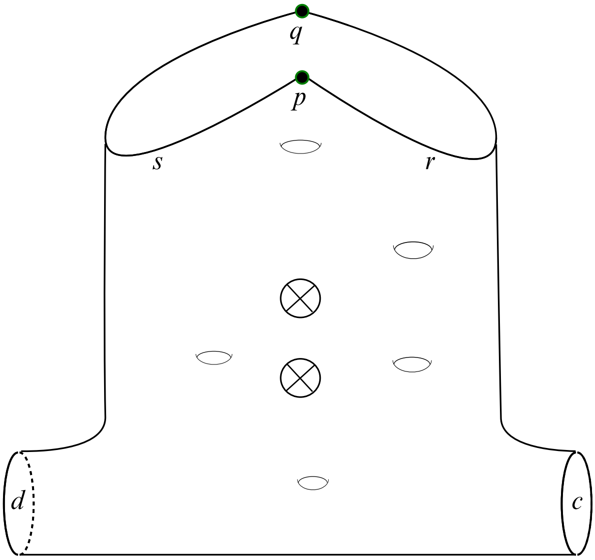


Figure 4: The surface T, which will be the building block of the construction.

Let $T_{i,j}$ be copies of the surface T, where $i, j \in \mathbb{Z}$. Let $c_{i,j}$, $d_{i,j}$ and $e_{i,j}$ be the (oriented) boundary components of $T_{i,j}$, and let $r_{i,j}$ and $s_{i,j}$ be the copies of the arcs r and s in $T_{i,j}$. Define a connected infinite surface T_∞ as the quotient

$$T_\infty := \Bigl(\bigcup_{i,j} T_{i,j}\Bigr)\Big/\!\sim$$

for all integers i and j. The gluing $\sim$ is given by orientation-reversing identifications

$$c_{i,j} \sim d_{i+1,j}, \quad r_{i,j} \sim s_{i,j+1}. \tag{7}$$

We have two natural shift maps, $\bar{\rho}_1, \bar{\rho}_2 \colon T_\infty \to T_\infty$,

$$\bar{\rho}_1 \colon T_{i,j} \mapsto T_{i+1,j}, \quad \bar{\rho}_2 \colon T_{i,j} \mapsto T_{i,j+1}.$$

Note that $\bar{\rho}_1$ and $\bar{\rho}_2$ commute. Define the surface $P_{n,k}$ as the quotient of the surface T_∞ by the covering action of the group generated by $(\bar{\rho}_1)^n$ and $(\bar{\rho}_2)^k$. Then $\bar{\rho}_1$ and $\bar{\rho}_2$ are equivariant with respect to the covering map. We denote the induced homeomorphisms of the quotient $P_{n,k}$ by ρ_1 and ρ_2. Note that later we will require that $k \geq 3$ and n is the number of punctures, given in Theorem 1.1.

Lemma 4.1 *Let*

$$g_{n,k} = (14k-2)n + 2$$

for $n \geq 1$ and $k \geq 1$. The genus of $P_{n,k}$ is $g_{n,k}$.

Proof Let $U \subset P_{n,k}$ be the subsurface

$$U = \left(\bigcup_{j=0}^{k-1} T_{0,j} \right) \Big/ {\sim'}$$

where $\sim'$ is given by (7) and by identifying $r_{i,k-1}$ and $s_{i,0}$. Then U is a compact, nonorientable surface of genus $12k$ with $2k$ boundary components. The surface $P_{n,k}$ consists of n copies of U identified along the $2k$ boundary components. Therefore the Euler characteristic of $P_{n,k}$ is

$$\chi(P_{n,k}) = n \cdot \chi(U) = n \cdot (2 - 12k - 2k) = -n(14k - 2).$$

Since $P_{n,k}$ is a nonorientable surface with empty boundary,

$$g_{n,k} = n(14k - 2) + 2. \qquad \square$$

Step 2: constructing the maps

In what is now a classical paper, Penner gives a construction of pseudo-Anosov homeomorphisms on both orientable and nonorientable surfaces [25]. Below we outline the Penner construction for nonorientable surfaces following the details of Liechti and Strenner [20, Section 2].

Inconsistent markings Let $\mathcal{N}$ be a nonorientable surface and let c be a two-sided curve in $\mathcal{N}$. There exists a neighborhood of c that is homeomorphic to an annulus. Let $\mathcal{A}_c$ be an annulus and let $\zeta_c \colon \mathcal{A}_c \to \mathcal{N}$ be the homeomorphism that maps to a neighborhood of c. The homeomorphism ζ_c is called a *marking* of c. A pair consisting of a curve c and ζ_c is called a *marked curve*. If we fix an orientation of $\mathcal{A}_c$, then we can pushforward this orientation to $\mathcal{N}$. Let (c, ζ_c) and (d, ζ_d) be two marked curves that intersect at one point p. We say that (c, ζ_c) and (d, ζ_d) are *marked inconsistently* if the pushforward of the orientation of $\mathcal{A}_c$ disagrees with the pushforward of the orientation of $\mathcal{A}_d$ in a neighborhood of p. We emphasize that we can also say that two disjoint curves are inconsistently marked.

Dehn twists We define the Dehn twist $\phi_{c,\zeta_c}(x)$ around a marked curve (c, ζ_c) as

$$\phi_{c,\zeta_c}(x) = \begin{cases} \zeta_c \circ \tau_c \circ \zeta_c^{-1}(x) & \text{if } x \in \zeta_c(\mathcal{A}_c), \\ x & \text{if } x \in \mathcal{N} - \zeta_c(\mathcal{A}_c). \end{cases}$$

Here τ_c is the left-handed Dehn twist on $\mathcal{A}_c$, ie $\tau_c(\theta, t) = (\theta + 2\pi t, t)$.

The Penner construction for nonorientable surfaces Let $\mathcal{C}$ be a set of marked essential simple closed curves in $\mathcal{N}$ such that no two curves in $\mathcal{C}$ are homotopic. A Penner construction on $\mathcal{N}$ is a composition of Dehn twists about the marked curves in $\mathcal{C}$ such that

(1) the complement of curves in $\mathcal{C}$ in $\mathcal{N}$ consists of disks with at most one puncture or marked point,

(2) the marked curves $(c_i, \zeta_i), (c_j, \zeta_j) \in \mathcal{C}$ with $i \neq j$ are marked inconsistently,

(3) a Dehn twist about each marked curve in $\mathcal{C}$ is included in the composition, and

(4) all powers of Dehn twists are positive (alternatively, all powers are negative).

Construction of $f_{n,k}$ We now construct homeomorphisms $f_{n,k}\colon P_{n,k} \to P_{n,k}$ that are defined as a composition of specific Dehn twists followed by a finite order mapping class. The key insight is that a power of this map will be a composition of Dehn twists that satisfy the criteria to be a Penner construction. Therefore $f_{n,k}$ is pseudo-Anosov. Here we are using the rotational symmetry of the $P_{n,k}$.

Let $\{\alpha_1, \dots, \alpha_8\}$ be the multicurve in $T_{0,0}$ as shown in Figure 5. Let $\{\beta_1, \dots, \beta_7\}$ be the multicurve in $T_{0,0} \cup T_{0,1} \cup T_{1,0}$ shown in Figure 5.

For any α_i, we choose a marking ζ_{α_i} to be orientation-preserving. For any β_j, let ζ_{β_j} be orientation-reversing. From here forward, we will think of α_i and β_j as (inconsistently) marked curves but we will suppress the marking maps. These choices give an inconsistent marking of $\{\alpha_1, \dots, \alpha_8\} \cup \{\beta_1, \dots, \beta_7\}$.

Let
$$\mathcal{R} = \bigcup_{i=2}^{8} \alpha_i.$$

Then $\mathcal{R}$ is a marked multicurve that is disjoint from γ. Let
$$\overline{\mathcal{R}} = \mathcal{R} \cup \rho_1(\mathcal{R}) \cup \dots \cup \rho_1^{n-1}(\mathcal{R}).$$

Let Φ_r be the composition of Dehn twists about the marked curves in $\overline{\mathcal{R}}$. Because the curves in $\overline{\mathcal{R}}$ are disjoint, the Dehn twists about the curves commute.

Let
$$\mathcal{B} = \bigcup_{j=2}^{7} \beta_j$$

in $T_{0,0} \cup T_{0,1} \cup T_{1,0}$. As above, $\mathcal{B}$ is a marked multicurve that is disjoint from γ. Let
$$\overline{\mathcal{B}} = \mathcal{B} \cup \rho_1(\mathcal{B}) \cup \dots \cup \rho_1^{n-1}(\mathcal{B}).$$

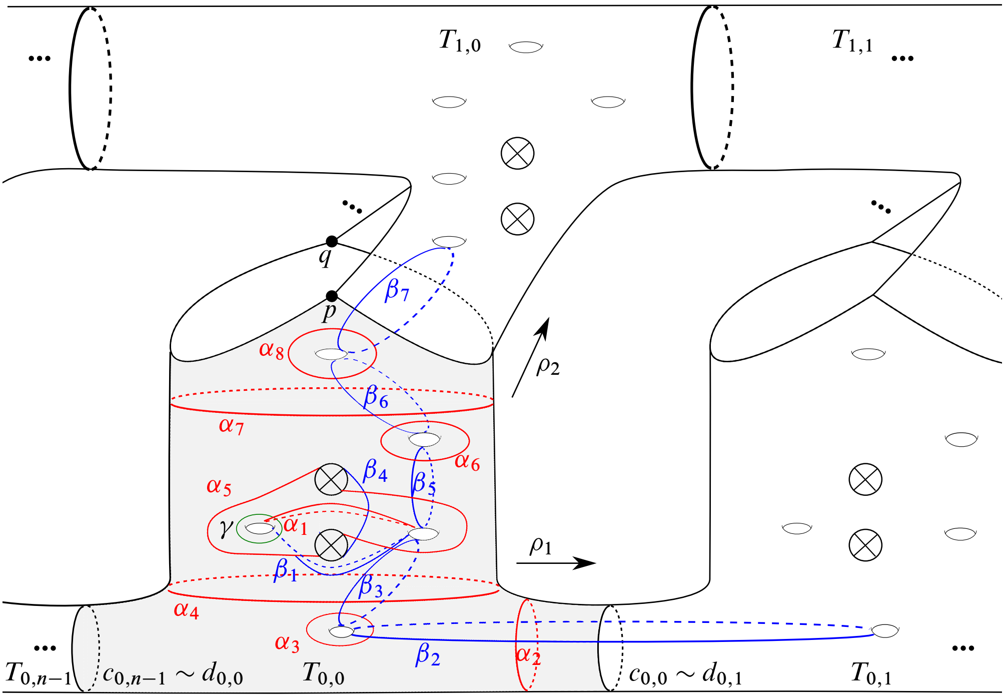


Figure 5: Part of surface $P_{n,k}$ that includes the subsurface $T_{0,0}$ and the curves α_i, β_j and γ.

Let Φ_b as the composition of Dehn twists about all of the marked curves in $\overline{\mathcal{B}}$. As with $\overline{\mathcal{R}}$, the Dehn twists about curves in $\overline{\mathcal{B}}$ commute.

Let $\alpha_1, \beta_1 \subset T_{0,0}$ be the (marked) curves in Figure 5. Let Φ be the composition of Dehn twists along all the curves $\alpha_1, \rho_1(\alpha_1), \dots, \rho_1^{n-1}(\alpha_1)$ followed by Dehn twists along all the curves $\beta_1, \rho_1(\beta_1), \dots, \rho_1^{n-1}(\beta_1)$. Define the map $f_{n,k}$ as

$$f_{n,k} := \rho_2 \circ \Phi \circ \Phi_b \circ \Phi_r.$$

Since the curves about which we twist to construct $f_{n,k}$ satisfy the conditions of Penner's construction, $f_{n,k}$ is a pseudo-Anosov homeomorphism.

Step 3: bounding the stretch factor

Following Yazdi, our next goal is to find an upper bound for the stretch factor of the pseudo-Anosov homeomorphisms $f_{n,k}$.

Train tracks Let S be a surface. A *train track* in S is graph embedded in S with that property that for every vertex v of valence three or greater, all edges adjacent to v have the same tangent vector at v. Let $\varphi\colon S \to S$ be a pseudo-Anosov homeomorphism. The map φ is equipped with a train track whose image under φ is homotopic to itself. Such a train track is an *invariant train track* associated to φ. Invariant train tracks have an associated matrix whose Perron–Frobenius eigenvalue is the stretch factor of φ.

Yazdi uses Lemma 4.2 to bound the spectral radius of the associated matrices.

Lemma 4.2 [33, Lemma 2.3] *Let A be a nonnegative integral matrix, Γ be the adjacency graph of A, and $V(\Gamma)$ the set of vertices of Γ. For each $v \in V(\Gamma)$, define v^+ to be the set of vertices $u \in V(\Gamma)$ such that there is an oriented edge from v to u. Let D and k be fixed natural numbers. Assume the following conditions hold for Γ:*

(i) *For each $v \in V(\Gamma)$, we have $\deg_{out}(v) \le D$.*

(ii) *There is a partition $V(\Gamma) = V_1 \cup \cdots \cup V_\ell$ such that, for each $v \in V_i$, we have $v^+ \subset V_{i+1}$ for any $1 \le i \le \ell$ except possibly when $i = 1$ or 3 (indices are mod ℓ).*

(iii) *For each $v \in V_1$, we have $v^+ \subset V_2 \cup V_3$.*

(iv) *For each $v \in V_3$ we have $v^+ \subset V_3 \cup V_4$, and for $u \in v^+ \cap V_3$ we have $u^+ \subset V_4$.*

(v) *For all $3 < j \le k$ and each $v \in V_j$, the set v^+ consists of a single element.*

Then the spectral radius of $A^{\ell-1}$ is at most $4D^4$.

With this result in hand, we can find an upper bound for the stretch factor of $f_{n,k}$.

Lemma 4.3 *Let $\lambda_{n,k}$ be the stretch factor of $f_{n,k}$. Then there exists a universal positive constant C' such that, for every $n \ge 1$ and $k \ge 3$, we have the upper bound*

$$\log(\lambda_{n,k}) \le C' \frac{n}{g_{n,k}}.$$

Proof We deliberately constructed our curves so that all intersections of the multicurves $\{\alpha_1, \dots, \alpha_8\}$ and $\{\beta_1, \dots, \beta_7\}$ occur in the subsurface $T_{0,0}$. The curve β_3 intersects $\rho_2(\alpha_3)$ at one point in $T_{0,1}$ and β_7 intersects $\rho_1(\alpha_8)$ at one point in $T_{1,0}$.

We define the unions of marked curves

$$\begin{aligned}
\mathcal{A} &:= \mathcal{B} \cup \mathcal{R} \cup \{\alpha_1, \beta_1\} = \bigcup_{i=1}^{8} \alpha_i \cup \bigcup_{j=1}^{7} \beta_j,\\
\bar{\mathcal{A}} &:= \mathcal{A} \cup \rho_1(\mathcal{A}) \cup \cdots \cup \rho_1^{n-1}(\mathcal{A}),\\
\hat{\mathcal{A}} &:= \bar{\mathcal{A}} \cup \rho_2(\bar{\mathcal{A}}) \cup \cdots \cup \rho_2^{k-1}(\bar{\mathcal{A}}).
\end{aligned}$$

Because $f_{n,k}$ is pseudo-Anosov, it has a corresponding invariant train track τ. Let V_τ be the space of all measured foliations that can be obtained by varying the weights on the tracks of τ. This forms a finite dimensional cone of measures, all of which can be carried by the combinatorial train track τ. Furthermore, $f_{n,k}$ acts linearly on this cone, and leaves the cone invariant, since τ is an invariant track for $f_{n,k}$. Consider now the transverse measure μ_δ for any curve δ in $\widehat{\mathcal{A}}$. This transverse measure is carried by τ, and thus μ_δ belongs in the cone of measures V_τ. Let H be the subspace spanned by $\{\mu_\delta \mid \delta \subset \widehat{\mathcal{A}}\}$. This linear subspace is also left invariant by $f_{n,k}$. Let M be the matrix representing the linear action of $f_{n,k}$ on H with respect to the basis $\{\mu_\delta \mid \delta \subset \widehat{\mathcal{A}}\}$. Let Γ be the adjacency graph for M. Work of Penner [25] tells us that the Perron–Frobenius eigenvalue of M is the stretch factor of $f_{n,k}$.

To bound the spectral radius of M, we need to show that Γ satisfies the criteria of Lemma 4.2.

(i) There exists a constant D', independent of n and k, such that, for every curve $\delta \in \widehat{\mathcal{A}}$, the geometric intersection number between δ and every curve in $\overline{\mathcal{A}}$ is at most D'. Recall that $f_{n,k} = \rho_2 \circ \Phi \circ \Phi_b \circ \Phi_r$. Let M_1, M_2, M_3 and M_4 be the matrices describing the linear action of Φ_r, Φ_b, Φ and ρ_2 on H, respectively. The matrix M can then be written as the product

$$M = M_4 M_3 M_2 M_1.$$

For a curve $\delta \in \widehat{\mathcal{A}}$, the L^1–norm of $M_i(\mu_\delta)$ is bounded above by the geometric intersection of $f_{n,k}(\delta)$ with the curves in $\overline{\mathcal{A}}$. Thus each of M_1, M_2 and M_3 will change the norm by a factor of at most $1 + D'$. Since ρ_2 will not change intersection numbers, M_4 will preserve the L^1–norm. If we let $D = (1 + D')^3$, then the outward degree of each vertex in Γ is at most D.

For the remaining conditions, we partition the vertices of Γ. Observe

$$\widehat{\mathcal{A}} = \rho_2^{-1}(\overline{\mathcal{A}}) \cup \overline{\mathcal{A}} \cup \bigcup_{i=3}^{k} \rho_2^{i-2}(\overline{\mathcal{A}}).$$

Then define V_1 as the vertices of Γ corresponding to $\rho_2^{-1}(\overline{\mathcal{A}})$, the set V_2 as the vertices of Γ corresponding to $\overline{\mathcal{A}}$, and V_i for $3 \le i \le k$ as the vertices of Γ corresponding to elements in $\rho_2^{i-2}(\overline{\mathcal{A}})$.

(ii) Suppose that $v \in V_i$, for $i \neq 1, 3$, is a vertex that corresponds to μ_δ for a curve $\delta \in \widehat{\mathcal{A}}$. Then δ is disjoint from all curves in $\overline{\mathcal{A}}$. The action of $\Phi \circ \Phi_b \circ \Phi_r$ on $\widehat{\mathcal{A}}$ will preserve the set $\rho_2^{(i-2) \bmod k}(\overline{\mathcal{A}})$ for each $i \neq 1, 3$. In particular, $\{\mu_\delta \mid \delta \subset \widehat{\mathcal{A}}\}$ will also be in $\rho_2^{(i-2) \bmod k}(\bar{\mathcal{A}})$. Then ρ_2 will rotate the curve $\Phi \circ \Phi_b \circ \Phi_r(\delta)$ into the set $\rho_2^{(i-1) \bmod k}(\bar{\mathcal{A}})$. That is, $f_{n,k} = \rho_2 \circ \Phi \circ \Phi_b \circ \Phi_r$ maps $\mu_\delta \in H$ to

$$\sum_{\zeta \in \mathcal{Z}} \mu_\zeta,$$

where $\mathcal{Z}$ is a subset of $\rho_2^{(i-1) \bmod k}(\bar{\mathcal{A}})$. Therefore $f_{n,k}$ maps v to a subset of V_{i+1}.

(iii) To verify the third condition, we first look at the vertices $v \in V_1$ such that $v^+ \not\subset V_2$. Such vertices will correspond to the curves in $\rho_2^{-1}(\bar{\mathcal{A}})$ that $\Phi \circ \Phi_b \circ \Phi_r$ maps to curves that are not in $\rho_2(\bar{\mathcal{A}})$. Because ρ_1 and ρ_2 commute, we can write the curves of $\rho_2^{-1}(\bar{\mathcal{A}})$ as

$$\rho_2^{-1}(\bar{\mathcal{A}}) = \rho_2^{-1}(\mathcal{A}) \cup \rho_1(\rho_2^{-1}(\mathcal{A})) \cup \cdots \cup \rho_1^{n-1}(\rho_2^{-1}(\mathcal{A})).$$

The elements of v^+ that are not in V_2 correspond to the images of curves in $\rho_2^{-1}(\bar{\mathcal{A}})$ under $f_{n,k}$ that are not in $\bar{\mathcal{A}}$. As in Yazdi, the only curves in $\rho_2^{-1}(\bar{\mathcal{A}})$ that intersect curves in $\bar{\mathcal{A}}$ are those in the set

$$\mathcal{X} = \{\rho_1^i(\rho_2^{-1}(\beta_7)) \mid 0 \le i \le n-1\}.$$

Therefore $\Phi \circ \Phi_b \circ \Phi_r$ maps curves in $\mathcal{X}$ to curves in $\rho_2^{-1}(\bar{\mathcal{A}}) \cup \bar{\mathcal{A}}$. Then

$$f_{n,k} = \rho_2 \circ \Phi \circ \Phi_b \circ \Phi_r$$

maps curves in $\mathcal{X}$ to curves in $\bar{\mathcal{A}} \cup \rho_2(\bar{\mathcal{A}})$. For any curve in $\mathcal{X}$, the corresponding vertex $v \in V_1$ will have $v^+ \subset V_2 \cup V_3$. Moreover, $f_{n,k}$ maps the curves $\rho_2^{-1}(\bar{\mathcal{A}}) \setminus \mathcal{X}$ to curves in $\bar{\mathcal{A}}$. Thus, for any vertex $v \in V_1$ that does not correspond to an element of $\mathcal{X}$, the set v^+ is contained in V_2.

(iv) Similarly, we look for the $v \in V_3$ such that $v^+ \not\subset V_4$. Such vertices will correspond to the curves in $\rho_2(\bar{\mathcal{A}})$ that $\Phi \circ \Phi_b \circ \Phi_r$ maps to curves that are not in $\rho_2^2(\bar{\mathcal{A}})$. As above,

$$\rho_2(\bar{\mathcal{A}}) = \rho_2(\mathcal{A}) \cup \rho_1(\rho_2(\mathcal{A})) \cup \cdots \cup \rho_1^{n-1}(\rho_2(\mathcal{A})).$$

The elements of v^+ that are not in V_4 correspond to the images of $\rho_2(\bar{\mathcal{A}})$ that intersect the curves in $\bar{\mathcal{A}}$. The only vertices of V_4 that correspond to such curves are those in the set

$$\mathcal{Y} = \{\rho_1^i(\rho_2(\alpha_8)) \mid 0 \le i \le n-1\}.$$

For any element $v \in V_3$ corresponding to a curve in $\mathcal{Y}$ and any $u \in v^+ \cap V_3$, the vertex u does not correspond to an element of $\mathcal{Y}$. Therefore $u^+ \subset V_4$.

(v) All the curves corresponding to an element of V_j, for $3 < j \le k$, are disjoint from all the curves in $\bar{\mathcal{A}}$. Thus, $f_{n,k}$ just acts by rotation.

Let $\lambda = \lambda_{n,k}$ be the stretch factor of $f_{n,k}$. By Lemma 4.2,

$$\lambda^{k-1} = \rho(M)^{k-1} = \rho(M^{k-1}) \le 4D^4.$$

Then the logarithm of λ satisfies

$$\log(\lambda^{k-1}) = (k-1)\cdot\log(\lambda) \le \log(4D^4).$$

Then, for $k \ge 2$,

$$\tfrac{1}{2}k\log(\lambda) \le (k-1)\log(\lambda) \le \log(4D^4).$$

On the other hand, we know $g_{n,k} = (14k-2)n + 2 \le 14kn$ by Lemma 4.1. Therefore,

$$\log(\lambda) \le 2\log(4D^4)\cdot\frac{1}{k} \le 2\log(4D^4)\cdot\frac{14n}{g_{n,k}}.$$

Let $C' := 28\log(4D^4)$ to complete the result. □

Step 4: the mapping torus

We have now constructed an infinite family of nonorientable surfaces $P_{n,k}$ and pseudo-Anosov homeomorphisms $f_{n,k}\colon P_{n,k} \to P_{n,k}$, but this is not enough. In Lemma 4.1, we show that $\{P_{n,k}\}$ does not include surfaces of infinitely many genera. We use the strategy of McMullen [23] and our extension of the Thurston's fibered face theory to fill in the gaps.

Next we follow the strategy of Leininger and Margalit [19] to find a surface embedded in the mapping torus of minimal genus. In our situation, this means that we will construct an embedded surface homeomorphic to $\mathcal{N}_3$.

Proposition 4.4 *Let $M_{n,k}$ be the mapping torus of $f_{n,k}$. Let $\mathcal{K}_{n,k}$ denote the fibered cone of $H^1(M_{n,k};\mathbb{R})$ corresponding to the map $f_{n,k}$. There is a relatively orientable incompressible surface $F_{n,k}$ embedded in $M_{n,k}$ that is homeomorphic to $\mathcal{N}_3$. Moreover, $F_{n,k}$ is transverse to the suspension flow direction given by $f_{n,k}$ and the Poincaré dual of $F_{n,k}$ is in the closure $\overline{\mathcal{K}}_{n,k}$.*

Proof Let $\gamma \subset T_{0,0}$ be the curve shown in Figure 6. Note that γ and $\Phi(\gamma)$ bound a nonorientable surface $\widehat{F}$ of genus 1 with boundary. For convenience, we will denote $\Phi(\gamma)$ by $\hat{\gamma}$. We are going to follow the image of γ under powers of $f_{n,k}$. Then we attach annuli to the boundary of $\widehat{F}$ to obtain $\mathcal{N}_3$. Since γ is disjoint from all curves in $\overline{\mathcal{R}}$ and $\overline{\mathcal{B}}$ (as seen in Figure 5), the maps Φ_r and Φ_b act trivially on γ. Recalling that $f_{n,k} = \rho_2 \circ \Phi \circ \Phi_b \circ \Phi_r$, we have

$$f_{n,k}(\gamma) = \rho_2 \circ \Phi \circ \Phi_b \circ \Phi_r(\gamma) = \rho_2 \circ \Phi(\gamma) = \rho_2(\hat{\gamma}).$$

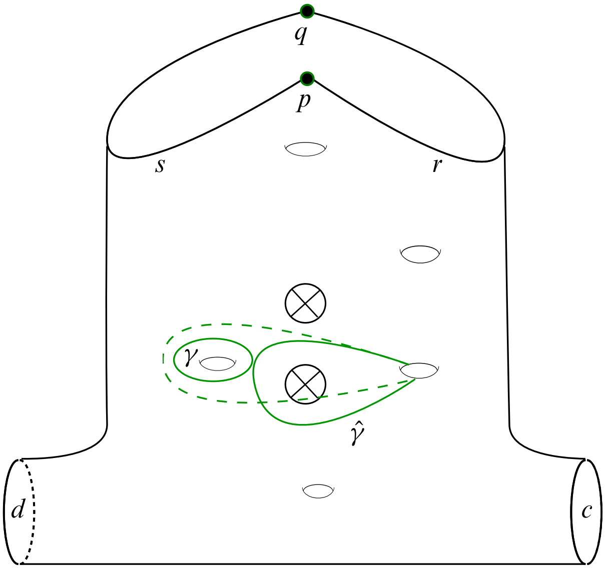


Figure 6: The curves γ and $\hat{\gamma}$ bound an a nonorientable surface of genus 1.

It follows that for all $1 \leq i \leq k$, the curve $f_{n,k}^i(\gamma)$ is $\rho_2^i(\hat{\gamma})$. For $1 \leq i \leq k$, let A_i be an annulus in $M_{n,k}$ that connects $f_{n,k}^{i-1}(\gamma)$ to $f_{n,k}^i(\gamma)$ obtained by following the suspension flow of $f_{n,k}$ around $M_{n,k}$. Let A be the union of all of the A_i, which is also an annulus. We can now construct the embedded surface $F_{n,k}$ by taking the union of A and $\widehat{F}$. The union of $\widehat{F}$ with A has empty boundary and Euler characteristic 0, so $F_{n,k}$ is homeomorphic to $\mathcal{N}_3$.

We now need to show that $F_{n,k}$ is relatively orientable. We construct a outward-pointing normal vector field by combining the outward-pointing vector fields on $\widehat{F}$ and A given by following γ along the suspension flow. Let v_1 be a vector field on $\widehat{F}$ pointing in the flow direction. Define v_2 to be a vector field on A as follows: on γ define v_2 to be the vector field pointing in to $\widehat{F}$, and flow the vector field along the suspension flow so v_2 is pointing away from $\widehat{F}$ on $\hat{\gamma}$.

Let U be a neighborhood of γ in $F_{n,k} = \widehat{F} \cup A$. Define two bump functions, c_1 and c_2, supported in U. Let c_1 be 1 on $\partial U \cap A$ and 0 on $\widehat{F}$. Let c_2 be 1 on $\partial U \cap \widehat{F}$ and 0 on A. We add the vector fields v_1 on $\widehat{F}$ and v_2 on A using these bump functions; the resulting vector field is $c_1 v_1 + c_2 v_2$. Observe that since v_1 points in the flow direction, and v_2 points into the surface, the new vector field $c_1 v_1 + c_2 v_2$ is transverse to $F_{n,k}$ in the neighborhood of γ (see Figure 7 for a picture of the resulting transverse vector field).

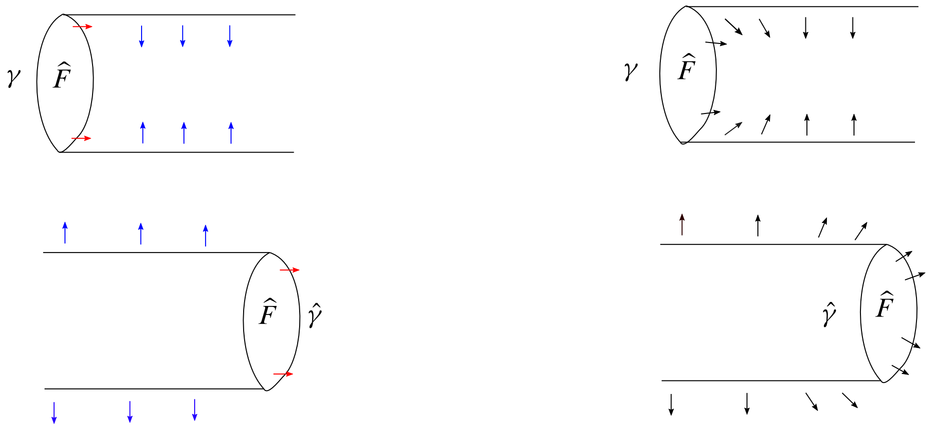


Figure 7: Left: the vector fields v_1 on $\widehat{F}$ and v_2 on A. The upper picture is a neighborhood of γ and the lower picture is a neighborhood of $\hat{\gamma}$. Right: the vector fields $c_1 v_1 + c_2 v_2$ on neighborhoods of γ and $\hat{\gamma}$.

We perform a similar construction in a small neighborhood of $\hat{\gamma}$: in this case, the fact that the vector field on $\widehat{F}$ points in the flow direction, and the vector field on A points away from the surface $\widehat{F}$ ensures that the new vector field is transverse, in a neighborhood of $\hat{\gamma}$, to the surface $F_{n,k}$.

A key fact we use in this construction is that the vector field along A that starts pointing into $\widehat{F}$ at γ comes back pointing away from the surface at $\hat{\gamma}$. This is because the homeomorphism $f_{n,kj}$ maps the inner tubular neighborhood of γ to the outer tubular neighborhood of $\hat{\gamma}$, where inner and outer tubular neighborhoods are the half tubular neighborhoods contained in $\widehat{F}$ and the complement, respectively. This fact about $f_{n,k}$ follows from its definition, ie following the four homeomorphisms whose composition is $f_{n,k}$.

The proof that $F_{n,k}$ can be isotoped to be transverse to the suspension flow is the same as the proof Yazdi uses [33], which is a restatement of that of Leininger and Margalit [19]. We include it here for completeness.

Let $N(\gamma)$ be a tubular neighborhood of γ in $\widehat{F}$. Let $\eta\colon \widehat{F} \to [0,1]$ be a smooth function supported on $N(\gamma)$ with

- $\eta^{-1}(1) = \gamma$, and
- the derivative of η vanishes on γ.

Let $\pi\colon M_{n,k} \to S^1$ be the projection map and let t_0 be such that $\widehat{F} \subset \pi^{-1}(t_0)$. Let $g\colon \widehat{F} \to M_{n,k}$ be the suspension flow of $f_{n,k}$ defined as $g(x) = (x, t_0 + k \cdot \eta(x))$.

Then the restriction of g to the interior of $\widehat{F}$ is an embedding into $M_{n,k}$ and $g(\gamma) = \hat{\gamma}$. Therefore the image of $\widehat{F}$ under g is an embedded nonorientable surface of genus 3. Moreover, $g(\widehat{F})$ is isotopic to the natural embedding of $F_{n,k}$ in $M_{n,k}$, and is transverse to the suspension flow. Therefore, the Poincaré dual of $F_{n,k}$ is in $\overline{\mathcal{K}}_{n,k}$ by Theorem 2.11.

Finally, $F_{n,k}$ is incompressible in $M_{n,k}$ because $M_{n,k}$ is hyperbolic, and $F_{n,k}$ is genus 3, the lowest possible genus for a hyperbolic nonorientable surface. □

Step 5: filling in the gaps

Recall that the family of surfaces $P_{n,k}$ that we have constructed have genera in the set $\{(14k-2)n+2\}$. We now want to construct surfaces of genera not in the set $\{(14k-2)n+2\}$ and pseudo-Anosov homeomorphisms of those surfaces that have small stretch factors. To do this we use the mapping torus $M_{n,k}$ of $P_{n,k}$ by $f_{n,k}\colon P_{n,k} \to P_{n,k}$. By Proposition 4.4, there exists a relatively incompressible surface $F_{n,k}$ in $M_{n,k}$ that is homeomorphic to $\mathcal{N}_3$. Let $P^r_{n,k}$ be the oriented sum of $P_{n,k}$ and $rF_{n,k}$, as defined in Proposition 3.3. The surfaces $P^r_{n,k}$ will be surfaces of the remaining genera.

Lemma 4.5 *The surface $P^r_{n,k}$ is of genus $g^r_{n,k} = g_{n,k} + r$. In particular, as r varies between 0 and $14n$, the genera of $\{P^r_{n,k}\}$ span the range between $g_{n,k}$ and $g_{n,k+1}$. Moreover, $P^r_{n,k}$ is isotopic to a fiber of a fibration of $M_{n,k}$ with pseudo-Anosov monodromy that fixes $2n$ of the singularities of its invariant foliation.*

Proof The Euler characteristic of an oriented sum is the sum of the Euler characteristics of the summands,

$$\chi(P^r_{n,k}) = \chi(P_{n,k}) + r\chi(F_{n,k}) = (-g_{n,k}+2) - r = -(g_{n,k}+r) + 2.$$

Since $P^r_{n,k}$ has no boundary or punctures, we have that the genus of $P^r_{n,k}$ is $g_{n,k}+r$.

By Proposition 4.4 we know that $F_{n,k}$ is incompressible and transverse to the suspension flow given by $f_{n,k}$. Therefore, by Proposition 3.2, there is a pseudo-Anosov homeomorphism $f^r_{n,k}$ of $P^r_{n,k} = P_{n,k} + rF_{n,k}$.

As in Yazdi [32, Lemma 3.5], $f_{n,k}$ fixes the $2n$ singularities of the stable foliation that are the intersection points of the axis of ρ_1 with $P_{n,k}$. By Proposition 4.4, the surface $F_{n,k}$ can be isotoped to be transverse to the suspension flow and disjoint from the orbit of the $2n$ singularities of $f_{n,k}$. Hence the monodromy $f^r_{n,k}$ still fixes the corresponding $2n$ singularities on $P^r_{n,k}$. □

We now prove the nonorientable version of the final piece of Yazdi's proof [33, Lemma 3.6].

Lemma 4.6 *Let $\lambda^r_{n,k}$ be the stretch factor of $f^r_{n,k}\colon P^r_{n,k} \to P^r_{n,k}$. Then there exists a constant $C > 0$ such that, for every $n \geq 1$, $k \geq 3$ and $0 \leq r \leq 14n$, we have the upper bound*

$$\log(\lambda^r_{n,k}) \leq C \frac{n}{g^r_{n,k}}.$$

Proof Let $\mathcal{K} = \mathcal{K}_{n,k}$ be the fibered cone in $H^1(M_{n,k};\mathbb{R})$ corresponding to $f_{n,k}$ and $h\colon \mathcal{K} \to \mathbb{R}$ the function described in Theorem 3.4. Note that $g_{n,k} \geq 42$, so

$$g^r_{n,k} = g_{n,k} + r \leq g_{n,k} + 14n < 2g_{n,k}.$$

Let ω be the Poincaré dual of $P^r_{n,k}$ and α the Poincaré dual of $P_{n,k}$. Then

$$\begin{aligned} h([\omega]) &< h([\alpha]) && \text{(convexity of } h\text{)} \\ &\leq C' \frac{n}{g_{n,k}} && \text{(Lemma 4.3)} \\ &\leq 2C' \frac{n}{g^r_{n,k}} && \text{(upper bound for } g^r_{n,k}\text{).} \end{aligned}$$ □

In the initial construction of $P_{n,k}$, there were $2n$ marked points, which were singularities of the map $f_{n,k}$. By the construction of $P^r_{n,k}$, these marked points are also singularities of $f^r_{n,k}$. Now we puncture $P^r_{n,k}$ at n of these marked points. We could think of this as removing all copies of the point p in the construction of $P_{n,k}$ in Step 1.

We can now give a proof of Theorem 1.1.

Proof of Theorem 1.1 As above, the lower bound follows easily from the lower bound in the orientable setting.

To find the upper bound, let $C' = \frac{1}{2}C$ be the value given in Lemma 4.6. Let

$$B'_2(n) = \max\{2C'n, \ell'_{1,n}, 2\ell'_{2,n}, \ldots, (40n+1)\ell'_{40n+1,n}\}.$$

By Proposition 3.2 and Lemma 4.6, $B'_2(n)$ is an upper bound for $g \cdot \ell'_{g,n}$. □

References

[1] **J W Aaber**, **N Dunfield**, *Closed surface bundles of least volume*, Algebr. Geom. Topol. 10 (2010) 2315–2342 MR Zbl

[2] **I Agol**, **C J Leininger**, **D Margalit**, *Pseudo-Anosov stretch factors and homology of mapping tori*, J. Lond. Math. Soc. 93 (2016) 664–682 MR Zbl

[3] **J Aramayona**, **C J Leininger**, **J Souto**, *Injections of mapping class groups*, Geom. Topol. 13 (2009) 2523–2541 MR Zbl

[4] **P Arnoux**, **J-C Yoccoz**, *Construction de difféomorphismes pseudo-Anosov*, C. R. Acad. Sci. Paris Sér. I Math. 292 (1981) 75–78 MR Zbl

[5] **M Bauer**, *An upper bound for the least dilatation*, Trans. Amer. Math. Soc. 330 (1992) 361–370 MR Zbl

[6] **D Calegari**, *Foliations and the geometry of* 3*–manifolds*, Oxford Univ. Press (2007) MR Zbl

[7] **J-H Cho**, **J-Y Ham**, *The minimal dilatation of a genus-two surface*, Experiment. Math. 17 (2008) 257–267 MR Zbl

[8] **A Fathi**, **F Laudenbach**, **V Poénaru**, *Thurston's work on surfaces*, Mathematical Notes 48, Princeton Univ. Press (2012) MR Zbl

[9] **D Fried**, *Flow equivalence, hyperbolic systems and a new zeta function for flows*, Comment. Math. Helv. 57 (1982) 237–259 MR Zbl

[10] **D Fried**, *Transitive Anosov flows and pseudo-Anosov maps*, Topology 22 (1983) 299–303 MR Zbl

[11] **J-Y Ham**, **W T Song**, *The minimum dilatation of pseudo-Anosov* 5*–braids*, Experiment. Math. 16 (2007) 167–179 MR Zbl

[12] **E Hironaka**, *Small dilatation mapping classes coming from the simplest hyperbolic braid*, Algebr. Geom. Topol. 10 (2010) 2041–2060 MR Zbl

[13] **E Hironaka**, **E Kin**, *A family of pseudo-Anosov braids with small dilatation*, Algebr. Geom. Topol. 6 (2006) 699–738 MR Zbl

[14] **S Hirose**, **E Kin**, *A construction of pseudo-Anosov braids with small normalized entropies*, New York J. Math. 26 (2020) 562–597 MR Zbl

[15] **N V Ivanov**, *Coefficients of expansion of pseudo-Anosov homeomorphisms*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. 167 (1988) 111–116 MR Zbl In Russian; translated in J. Soviet Math. 52 (1990) 2819–2822

[16] **E Kin**, **M Takasawa**, *Pseudo-Anosovs on closed surfaces having small entropy and the Whitehead sister link exterior*, J. Math. Soc. Japan 65 (2013) 411–446 MR Zbl

[17] **E Kin**, **M Takasawa**, *The boundary of a fibered face of the magic* 3*–manifold and the asymptotic behavior of minimal pseudo-Anosov dilatations*, Hiroshima Math. J. 46 (2016) 271–287 MR Zbl

[18] **E Lanneau**, **J-L Thiffeault**, *On the minimum dilatation of pseudo-Anosov homeromorphisms on surfaces of small genus*, Ann. Inst. Fourier (Grenoble) 61 (2011) 105–144 MR Zbl

[19] **C J Leininger**, **D Margalit**, *On the number and location of short geodesics in moduli space*, J. Topol. 6 (2013) 30–48 MR Zbl

[20] **L Liechti**, **B Strenner**, *Minimal pseudo-Anosov stretch factors on nonoriented surfaces*, Algebr. Geom. Topol. 20 (2020) 451–485 MR Zbl

[21] **M Loving**, *Least dilatation of pure surface braids*, Algebr. Geom. Topol. 19 (2019) 941–964 MR Zbl

[22] **S Matsumoto**, *Topological entropy and Thurston's norm of atoroidal surface bundles over the circle*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. 34 (1987) 763–778 MR Zbl

[23] **C T McMullen**, *Polynomial invariants for fibered* 3*–manifolds and Teichmüller geodesics for foliations*, Ann. Sci. École Norm. Sup. 33 (2000) 519–560 MR Zbl

[24] **H Minakawa**, *Examples of pseudo-Anosov homeomorphisms with small dilatations*, J. Math. Sci. Univ. Tokyo 13 (2006) 95–111 MR Zbl

[25] **R C Penner**, *A construction of pseudo-Anosov homeomorphisms*, Trans. Amer. Math. Soc. 310 (1988) 179–197 MR Zbl

[26] **R C Penner**, *Bounds on least dilatations*, Proc. Amer. Math. Soc. 113 (1991) 443–450 MR Zbl

[27] **W T Song**, **K H Ko**, **J E Los**, *Entropies of braids*, J. Knot Theory Ramifications 11 (2002) 647–666 MR Zbl

[28] **W P Thurston**, *A norm for the homology of* 3*–manifolds*, Mem. Amer. Math. Soc. 339, Amer. Math. Soc., Providence, RI (1986) MR Zbl

[29] **W P Thurston**, *Hyperbolic structures on* 3*–manifolds, II*: *Surface groups and* 3*–manifolds which fiber over the circle*, preprint (1998) arXiv math/9801045

[30] **C-Y Tsai**, *The asymptotic behavior of least pseudo-Anosov dilatations*, Geom. Topol. 13 (2009) 2253–2278 MR Zbl

[31] **A D Valdivia**, *Sequences of pseudo-Anosov mapping classes and their asymptotic behavior*, New York J. Math. 18 (2012) 609–620 MR Zbl

[32] **M Yazdi**, *Lower bound for dilatations*, J. Topol. 11 (2018) 602–614 MR Zbl

[33] **M Yazdi**, *Pseudo-Anosov maps with small stretch factors on punctured surfaces*, Algebr. Geom. Topol. 20 (2020) 2095–2128 MR Zbl

Department of Mathematics, University of Michigan
Ann Arbor, MI, United States

School of Mathematics, Georgia Institute of Technology
Atlanta, GA, United States

Department of Mathematics and Computer Science, College of the Holy Cross
Worcester, MA, United States

saykhan@umich.edu, ctpartin@gmail.com, rwinarsk@holycross.edu

https://www-personal.umich.edu/~saykhan/,
https://sites.google.com/site/rebeccawinarski/

Received: 20 July 2021 Revised: 25 October 2021

Algebraic & Geometric Topology **23**:6 (2023) 2857–2866
DOI: 10.2140/agt.2023.23.2857
Published: 7 September 2023

Infinitely many arithmetic alternating links

MARK D BAKER
ALAN W REID

We prove the existence of infinitely many alternating links in S^3 whose complements are arithmetic.

57K32; 11F06

1 Introduction

Let d be a square-free positive integer and let O_d denote the ring of integers of $\mathbb{Q}(\sqrt{-d})$. A noncompact finite-volume hyperbolic 3–manifold X is called *arithmetic* if X and the Bianchi orbifold $Q_d = \mathbb{H}^3/\mathrm{PSL}(2, O_d)$ are commensurable, that is to say they share a common finite-sheeted cover. (see Maclachlan and Reid [22, Chapters 8 and 9] for further details). If $X = S^3 \setminus L$, we call L *an arithmetic link.*

Since Thurston's original studies of hyperbolic structures on 3–manifolds [25], link complements in S^3 have played a prominent role, and indeed arithmetic links were also very much at the heart of his work. Several arithmetic link complements were constructed in [25], and, over the years, many more examples were constructed; see Aitchison, Lumsden and Rubinstein [3], Aitchison and Rubinstein [4], Baker [5; 6; 7], Baker, Goerner and Reid [9; 8], Goerner [14], Grunewald and Hirsch [16] and Hatcher [19]. Several of these arithmetic links are alternating, and although there are infinitely many arithmetic links in S^3 (for example, those links determining certain cyclic covers of the complement of the Whitehead link), whether there were infinitely many arithmetic alternating links remained open.

By relating the spectral geometry of the complement to combinatorics of an alternating diagram, Lackenby [21] showed that there are only finitely many *congruence* alternating links, and motivated by this, asks in [21], whether there are only finitely many arithmetic alternating links. More recently, the question as to whether there were infinitely many

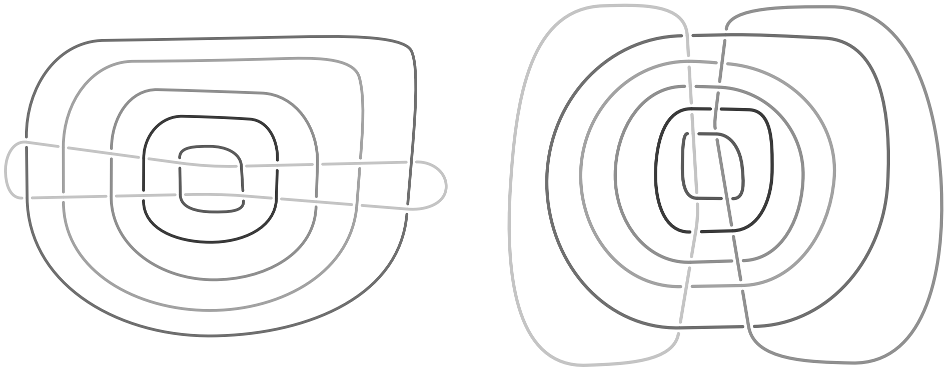


Figure 1

arithmetic alternating links was asked of the second author by D Futer in 2019. The main result of this note resolves these questions by answering Futer's question in the positive (and hence Lackenby's in the negative).

Theorem 1.1 *There are infinitely many alternating links in S^3 whose complements are arithmetic.*

Indeed, we prove something more precise. We will construct two infinite families of alternating links L_j and $\mathcal{L}_j$ whose complements are arithmetic. In more detail, the family of links L_j is built from $j+1$ concentric circles centered at the origin in the Euclidean plane, with a "horizontal" component (which we will denote by K) added intersecting each of the concentric circles in four points, and each intersection point resolved to make the diagram alternating (see Figure 1, left, where L_4 is shown). Thus L_j is an alternating link with $j+2$ components. The family of links $\mathcal{L}_j$ is constructed in a similar fashion using $j+1$ concentric circles centered at the origin in the Euclidean plane, with two additional components (which we will denote by K_1 and K_2) added intersecting each of the concentric circles in two points, and each intersection point resolved to make the diagram alternating (see Figure 1, right, where $\mathcal{L}_4$ is shown). Thus $\mathcal{L}_j$ is an alternating link with $j+3$ components.

Theorem 1.2 *L_j and $\mathcal{L}_j$ are arithmetic for all $j \geq 1$ with both $S^3 \setminus L_j \to Q_3$ and $S^3 \setminus \mathcal{L}_j \to Q_3$ of degree $60j$.*

The arithmetic nature of the link L_1 was first explicitly described by Hatcher [19, Example 5], and we recall this briefly here. As described in [19], the complement of L_1 can be obtained as the union of two regular ideal hyperbolic cubes (all of whose dihedral angles are $\pi/3$), and, as noted in [19], a regular ideal cube can be subdivided into five regular ideal hyperbolic simplices, from which Hatcher deduces that L_1 is arithmetic since the fundamental group of its complement arises as a subgroup of the group of orientation-preserving isometries of the tessellation of $\mathbb{H}^3$ by regular ideal hyperbolic simplices, which can be identified with the group $\mathrm{PGL}(2, O_3)$. Hence the link L_1 is arithmetic. In fact (see the discussion in the proof of Theorem 1.2 given in Section 2.2), the fundamental group of its complement arises as a subgroup $\mathrm{PSL}(2, O_3)$. Given the description of $S^3 \setminus L_1$ as a union of 10 regular ideal tetrahedra, its volume can be computed as $10v_0$, where v_0 is the volume of the regular ideal simplex in $\mathbb{H}^3$ (ie approximately 10.14941606...). Since the volume of Q_3 is $v_0/6$, $S^3 \setminus L_1$ is a 60–fold cover of Q_3. In [19, Example 5], Hatcher constructs a second link complement as the union of two regular ideal hyperbolic cubes, and this is homeomorphic to $S^3 \setminus \mathcal{L}_1$.

The manifolds $S^3 \setminus L_1$ and $S^3 \setminus \mathcal{L}_1$ have been reconstructed elsewhere in the literature. By volume considerations — see Adams, Hildebrand and Weeks [2] — $S^3 \setminus L_1$ (resp. $S^3 \setminus \mathcal{L}_1$) can be seen to be homeomorphic to the complement of the three-component link 8^3_4 (resp. to the complement of 8^4_1). It can be checked (eg using SnapPy [11]) that $S^3 \setminus L_1$ is also homeomorphic to a 5–fold irregular cover of the complement of the figure-eight knot (namely the so-called Roman link of Hilden, Lozano and Montesinos [20]). The complements of L_1 and $\mathcal{L}_1$ were constructed again by Aitchison and Rubinstein [4, Example 3] as well as being identified as the tetrahedral census manifolds $\mathrm{otet}10_{00006}$ and $\mathrm{otet}10_{00011}$ of Fominykh, Garoufalidis, Goerner, Tarkaev and Vesnin [13] (see also Goerner [15]).

In a different direction, neither $S^3 \setminus L_1$ nor $S^3 \setminus \mathcal{L}_1$ contains a closed embedded essential surface (see Hass and Menasco [18] for L_1 and Oertel [24] for $\mathcal{L}_1$). By comparison, in Section 3 we show that both $S^3 \setminus L_j$ and $S^3 \setminus \mathcal{L}_j$ contain a closed embedded essential surface for all $j \geq 2$.

Acknowledgements We are grateful to Dave Futer for asking the question. We are also very grateful to Will Worden for drawing the figures. Reid was supported in part by an NSF grant.

2 Proof of Theorem 1.2

Our proof will be motivated by that given in [19], but we shall certify arithmeticity in a slightly different way.

2.1 Tessellation by regular ideal cubes

Motivated by the description of $S^3 \setminus L_1$ as a union of two regular ideal cubes, we make the following definition (see [13]):

Definition 2.1 Let M be a finite-volume cusped hyperbolic 3–manifold. We call M cubical if it can be decomposed into regular ideal hyperbolic cubes.

Let $M = \mathbb{H}^3 / \Gamma$ be a cubical manifold. On lifting to the universal cover, we obtain a tessellation $\mathcal{T}(C)$ of $\mathbb{H}^3$ by regular ideal cubes, C, and so Γ is a subgroup of the group of isometries of $\mathcal{T}(C)$, which we denote by $\mathrm{Isom}(\mathcal{T}(\mathrm{C}))$ (which is a discrete group of isometries of $\mathbb{H}^3$). We will denote by $\mathrm{Isom}^+(\mathcal{T}(\mathrm{C}))$ the subgroup of $\mathrm{Isom}(\mathcal{T}(\mathrm{C}))$ of index 2 consisting of orientation-preserving isometries.

Lemma 2.2 $\mathrm{Isom}(\mathcal{T}(\mathrm{C}))$ *is an arithmetic subgroup of* $\mathrm{Isom}(\mathbb{H}^3)$ *commensurable with* $\mathrm{PSL}(2, O_3)$. *Hence any cubical manifold is arithmetic.*

A proof of Lemma 2.2 is implicit in [23], but we include a proof here for completeness. Before proving Lemma 2.2, we recall some notation. Let $\Gamma_0(2) < \mathrm{PSL}(2, O_3)$ be the image of the subgroup of $\mathrm{SL}(2, O_3)$ given by

$$\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, O_3) \;\middle|\; c \equiv 0 \bmod \langle 2 \rangle \right\}.$$

It is easy to check that $[\mathrm{PSL}(2, O_3) : \Gamma_0(2)] = 5$, that $\mathbb{H}^3 / \Gamma_0(2)$ has two cusps (corresponding to the inequivalent parabolic fixed points 0 and ∞), and that the peripheral subgroup of $\Gamma_0(2)$ fixing ∞ coincides with that of $\mathrm{PSL}(2, O_3)$, namely the image in $\mathrm{PSL}(2, O_3)$ of the subgroup

$$\left\langle \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & \omega \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} \omega & 0 \\ 0 & 1/\omega \end{pmatrix} \right\rangle, \quad \text{where } \omega^2 + \omega + 1 = 0.$$

Let ι and τ be the elements of $\mathrm{PSL}(2, \mathbb{C})$ given by the images of the elements $\left(\begin{smallmatrix} i & 0 \\ 0 & -i \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & -1/\sqrt{2} \\ \sqrt{2} & 0 \end{smallmatrix}\right)$, respectively. Note that ι and τ both have order 2, and they normalize $\Gamma_0(2)$. Hence the group $G = \langle \Gamma_0(2), \iota, \tau \rangle$ is arithmetic, containing $\Gamma_0(2)$ as a normal subgroup with quotient group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

Proof To prove Lemma 2.2, it suffices to show that $\mathrm{Isom}^+(\mathcal{T}(\mathrm{C}))$ is commensurable with $\mathrm{PSL}(2, O_3)$. To that end, we will show that the orbifolds $N_1 = \mathbb{H}^3/\mathrm{Isom}^+(\mathcal{T}(\mathrm{C}))$ and $N_2 = \mathbb{H}^3 / G$ are isometric and hence $\mathrm{Isom}^+(\mathcal{T}(\mathrm{C}))$ and G are conjugate by Mostow–Prasad rigidity. Using the remarks prior to the proof, this proves commensurability.

In the notation established above, since $\tau(0) = \infty$, the orbifold N_2 has a single cusp, and since $\iota \in G$, this is a rigid cusp of type $(2,3,6)$ (in the notation of [23]). Moreover, since the volume of Q_3 is $v_0/6$, the computation of indices given above shows that the volume of N_2 is $5v_0/24$.

Now consider the group $\mathrm{Isom}^+(\mathcal{T}(\mathrm{C}))$. This is generated by the extension to $\mathbb{H}^3$ of the orientation-preserving symmetries of a single cube C of $\mathcal{T}(C)$, along with rotations of $2\pi/6$ in the edges of C. As noted in Section 1, C can be subdivided into five regular ideal tetrahedra, and so the volume of C is $5v_0$. From this it now follows that N_1 has volume $5v_0/24$ and a rigid cusp of type $(2,3,6)$.

Finally, using Adams [1], we deduce that N_1 and N_2 are isometric, since he proved there that there is a unique orientable hyperbolic 3–orbifold of volume $5v_0/24$ and a single rigid cusp of type $(2,3,6)$. □

Remark 2.3 Part of the proof in [1] of the uniqueness of a hyperbolic 3–orbifold with a single rigid cusp of type $(2,3,6)$ was found to have a gap, but this was corrected in the recent paper [12].

Remark 2.4 As noted in [23], the group $\mathrm{Isom}(\mathcal{T}(\mathrm{C}))$, can be identified with the group generated by reflections in the faces of the tetrahedron $T[4,2,2;6,2,3] \subset \mathbb{H}^3$ in the notation of [23].

2.2 The link complements $S^3 \setminus L_j$ and $S^3 \setminus \mathcal{L}_j$ are cubical

Given Lemma 2.2, we must show that $S^3 \setminus L_j$ (for $j \geq 1$) and $S^3 \setminus \mathcal{L}_j$ (for $j \geq 1$) are cubical. We will take a slightly different perspective from Hatcher's construction of a cubical structure for $S^3 \setminus L_1$ (more in keeping with [3; 4]), which we now describe. This is what we generalize for the links L_j $(j \geq 2)$ and $\mathcal{L}_j$ $(j \geq 2)$.

Consider an alternating diagram for L_1 on some projection plane $S^2 \subset S^3$. This produces the 4–valent planar graph P_1 shown in Figure 2, left. Two-coloring the regions in checkerboard fashion and labeling these regions as $+$ and $-$ affords a decomposition of S^3 into two 3–balls, each of which is endowed with an abstract polyhedral structure. Denote these polyhedra by Π_+ and Π_-. These polyhedra are identical up to reversing all the colors and signs. Each face f_i of Π_+ is an n_i–gon (where $n_i = 2$ or 4 in this case) with a sign $\sigma_i \in \{\pm\}$, and the polyhedra Π_+ and Π_- are identified by sending f_i to the corresponding face of Π_- using a rotation of $\sigma_i 2\pi/n_i$ (with $+$ denoting clockwise). The resulting complex with vertices deleted is then homeomorphic to $S^3 \setminus L_1$ (see [3], for example).

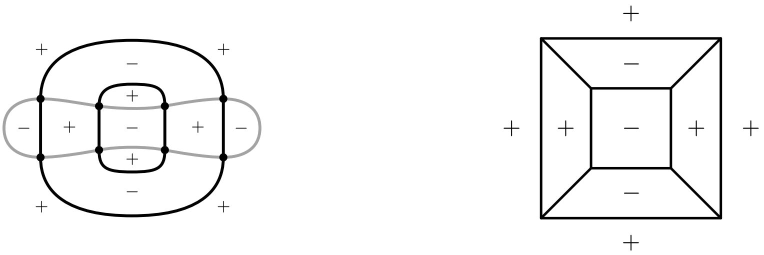


Figure 2

Note that P_1 contains four bigons, and we can collapse each of these bigons to an edge in each of the polyhedra Π_+ and Π_-, and then make the identifications described above. The resulting polyhedra obtained are cubes (see Figure 2, right), so that $S^3 \setminus L_1$ is the identification space of two cubes with vertices deleted.

This combinatorial realization can be done geometrically: namely, the identifications described above can be realized as identifications of the regular ideal cube in $\mathbb{H}^3$ with six 2–cells meeting along an edge (with dihedral angle $\pi/3$).

For the general case of L_j, we refer to Figure 3 (which shows the case of L_4) and proceed as follows.

Performing the construction above on each L_j results in a 4–valent planar graph P_j (see Figure 3, left) and polyhedra Π_+^j and Π_-^j. As above, the graphs P_j each contain exactly four bigons, and collapsing these bigons leads to the polyhedra shown in Figure 3, right. As is visible from the diagram, each of Π_+^j and Π_-^j is a union of j cubes, whose faces are identified as described above. To establish that for each $j \geq 2$ the manifold

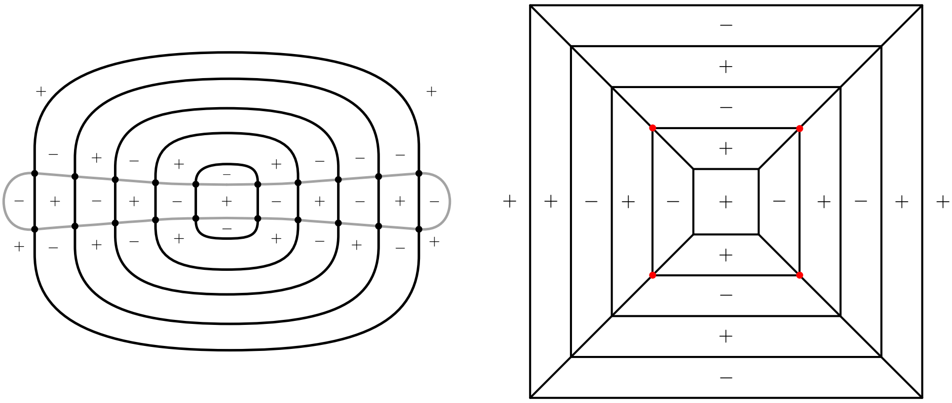


Figure 3

$S^3 \setminus L_j$ is cubical, and therefore arithmetic by Lemma 2.2, we need to ensure that the combinatorial decomposition described here can be realized geometrically.

Referring to Figure 3, right, we now view the polyhedra Π_+^j and Π_-^j as being built from copies of the regular ideal cube, so that edges of Π_+^j and Π_-^j have dihedral angle $\pi/3$ or $2\pi/3$, the latter occurring at edges where two cubes meet, eg the edges between those red vertices of Figure 3, right, and then the edges of all concentric squares except the "innermost" and "outermost" ones. From above, the polyhedra Π_+ and Π_- are identified by sending f_i to the corresponding face of Π_- using a rotation of $\sigma_i\pi/2$ (with $+$ denoting clockwise). Using this we see that edges with dihedral angle $2\pi/3$ are identified via the $\pi/2$ rotation to an edge with dihedral angle $\pi/3$. Each such edge with dihedral angle $2\pi/3$ lies in two faces of adjacent cubes and so once the identifications are completed the angle sum is 2π. Edges of the innermost and outermost squares have dihedral angles $\pi/3$. They are identified via $\pi/2$ rotations to edges also with dihedral angles $\pi/3$. Six of these edges are identified to get angle sum 2π. This proves that each $S^3 \setminus L_j$ is cubical, and hence arithmetic.

Moreover, since any arithmetic link complement commensurable with Q_3 necessarily covers Q_3 (see for example [22, Theorem 9.2.2] and note that $M(2, \mathbb{Q}(\sqrt{-3}))$ has type number one), the final part of Theorem 1.2 follows since, from above, the volume of $S^3 \setminus L_j$ is $10jv_0$, and the volume of Q_3 is $v_0/6$.

The case of $\mathcal{L}_j$ is handled in a completely similar manner using polyhedra arising as in Figure 4. We omit the details. □

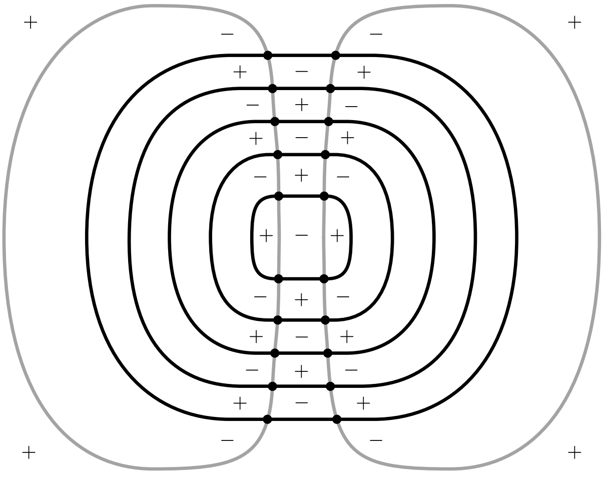


Figure 4

As was pointed out in [13, Remark 3.7], it is not always the case that a cubical manifold decomposes into regular ideal tetrahedra. However, this does hold for the manifolds $S^3 \setminus L_j$ and $S^3 \setminus \mathcal{L}_j$. The important point to note is that insertion of the diagonals on faces to create the five tetrahedra can be done so consistently (as was implicit in [19]). In particular, each of $S^3 \setminus L_j$ and $S^3 \setminus \mathcal{L}_j$ is decomposed into $10j$ regular ideal tetrahedra, and so, using this decomposition and [17], a corollary of Theorem 1.2 is:

Corollary 2.5 *$S^3 \setminus L_j$ and $S^3 \setminus \mathcal{L}_j$ are manifolds of maximal volume amongst all hyperbolic manifolds admitting a decomposition into $10j$ tetrahedra.*

3 Closed embedded essential surfaces

We first show that, for $j \geq 2$, $S^3 \setminus L_j$ contains a closed embedded essential surface. Deleting the component K of L_j results in the $(j+1)$–component unlink. The result now follows from [10, Theorem 4.1] since the $\mathrm{SL}(2,\mathbb{C})$ character variety of F_{j+1} has dimension $3(j+1)-3=3j$ and this is greater than $j+2$ for $j \geq 2$.

The case of $S^3 \setminus \mathcal{L}_j$ is handled in a similar manner. In this case, deleting the components K_1 and K_2 from $\mathcal{L}_j$ results in the $(j+1)$–component unlink and we now argue as above, applying [10, Theorem 4.1] on noting that $3(j+1)-3=3j$ is greater than $j+3$ for $j \geq 2$.

References

[1] **C C Adams**, *Noncompact hyperbolic 3–orbifolds of small volume*, from "Topology '90" (B Apanasov, W D Neumann, A W Reid, L Siebenmann, editors), Ohio State Univ. Math. Res. Inst. Publ. 1, de Gruyter, Berlin (1992) 1–15 MR Zbl

[2] **C Adams**, **M Hildebrand**, **J Weeks**, *Hyperbolic invariants of knots and links*, Trans. Amer. Math. Soc. 326 (1991) 1–56 MR Zbl

[3] **I R Aitchison**, **E Lumsden**, **J H Rubinstein**, *Cusp structures of alternating links*, Invent. Math. 109 (1992) 473–494 MR Zbl

[4] **I R Aitchison**, **J H Rubinstein**, *Combinatorial cubings, cusps, and the dodecahedral knots*, from "Topology '90" (B Apanasov, W D Neumann, A W Reid, L Siebenmann, editors), Ohio State Univ. Math. Res. Inst. Publ. 1, de Gruyter, Berlin (1992) 17–26 MR Zbl

[5] **M D Baker**, *Link complements and imaginary quadratic number fields*, PhD thesis, Massachusetts Institute of Technology (1981) MR Available at `https://www.proquest.com/docview/303188699`

[6] **M D Baker**, *Link complements and integer rings of class number greater than one*, from "Topology '90" (B Apanasov, W D Neumann, A W Reid, L Siebenmann, editors), Ohio State Univ. Math. Res. Inst. Publ. 1, de Gruyter, Berlin (1992) 55–59 MR Zbl

[7] **M D Baker**, *Link complements and the Bianchi modular groups*, Trans. Amer. Math. Soc. 353 (2001) 3229–3246 MR Zbl

[8] **M D Baker**, **M Goerner**, **A W Reid**, *All known principal congruence links*, preprint (2019) arXiv 1902.04426

[9] **M D Baker**, **M Goerner**, **A W Reid**, *All principal congruence link groups*, J. Algebra 528 (2019) 497–504 MR Zbl

[10] **D Cooper**, **D D Long**, *Derivative varieties and the pure braid group*, Amer. J. Math. 115 (1993) 137–160 MR Zbl

[11] **M Culler**, **N M Dunfield**, **M Goerner**, **J R Weeks**, *SnapPy, a computer program for studying the geometry and topology of* 3*–manifolds*, version 2.6 (2017) Available at `http://snappy.computop.org`

[12] **S T Drewitz**, **R Kellerhals**, *The non-arithmetic cusped hyperbolic* 3*–orbifold of minimal volume*, Trans. Amer. Math. Soc. 376 (2023) 3819–3866 MR Zbl

[13] **E Fominykh**, **S Garoufalidis**, **M Goerner**, **V Tarkaev**, **A Vesnin**, *A census of tetrahedral hyperbolic manifolds*, Exp. Math. 25 (2016) 466–481 MR Zbl

[14] **M Goerner**, *Visualizing regular tessellations: principal congruence links and equivariant morphisms from surfaces to* 3*–manifolds*, PhD thesis, University of California, Berkeley (2011) MR Available at `https://www.proquest.com/docview/928944884`

[15] **M Goerner**, *A census of hyperbolic platonic manifolds and augmented knotted trivalent graphs*, New York J. Math. 23 (2017) 527–553 MR Zbl

[16] **F Grunewald**, **U Hirsch**, *Link complements arising from arithmetic group actions*, Internat. J. Math. 6 (1995) 337–370 MR Zbl

[17] **U Haagerup**, **H J Munkholm**, *Simplices of maximal volume in hyperbolic n–space*, Acta Math. 147 (1981) 1–11 MR Zbl

[18] **J Hass**, **W Menasco**, *Topologically rigid non-Haken* 3*–manifolds*, J. Austral. Math. Soc. Ser. A 55 (1993) 60–71 MR Zbl

[19] **A Hatcher**, *Hyperbolic structures of arithmetic type on some link complements*, J. London Math. Soc. 27 (1983) 345–355 MR Zbl

[20] **H M Hilden**, **M T Lozano**, **J M Montesinos**, *On knots that are universal*, Topology 24 (1985) 499–504 MR Zbl

[21] **M Lackenby**, *Spectral geometry, link complements and surgery diagrams*, Geom. Dedicata 147 (2010) 191–206 MR Zbl

[22] **C Maclachlan**, **A W Reid**, *The arithmetic of hyperbolic* 3*–manifolds*, Graduate Texts in Math. 219, Springer (2003) MR Zbl

[23] **W D Neumann**, **A W Reid**, *Notes on Adams' small volume orbifolds*, from "Topology '90" (B Apanasov, W D Neumann, A W Reid, L Siebenmann, editors), Ohio State Univ. Math. Res. Inst. Publ. 1, de Gruyter, Berlin (1992) 311–314 MR Zbl

[24] **U Oertel**, *Closed incompressible surfaces in complements of star links*, Pacific J. Math. 111 (1984) 209–230 MR Zbl

[25] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `http://msri.org/publications/books/gt3m`

IRMAR, Université de Rennes 1
Rennes, France

Department of Mathematics, Rice University
Houston, TX, United States

`mark.baker@univ-rennes1.fr`, `alan.reid@rice.edu`

Received: 2 August 2021 Revised: 14 February 2022

Algebraic & Geometric Topology **23**:6 (2023) 2867–2893
DOI: 10.2140/agt.2023.23.2867
Published: 7 September 2023

Unchaining surgery, branched covers, and pencils on elliptic surfaces

TERRY FULLER

R İnanç Baykur, Kenta Hayano, and Naoyuki Monden used a technique called *unchaining* to construct a family of simply connected symplectic 4–manifolds $X'_g(i)$ for all $g \geq 3$ and $0 \leq i \leq g-1$ (Geom. Topol. 20 (2016) 2335–2395). Among this family, the manifolds $X'_g(g-2)$ are shown to be symplectic Calabi–Yau 4–manifolds. They also showed that each $X'_g(i) \# \overline{\mathbb{CP}^2}$ admits a pair of inequivalent genus g Lefschetz pencils. We show how to describe every $X'_g(i)$ as a 2–fold branched cover of a rational surface, and use this to prove that each $X'_g(i)$ is diffeomorphic to the elliptic surface $E(g-i)$. This has several notable consequences: each symplectic Calabi–Yau they construct is diffeomorphic to K3; for each $n \geq 3$ and $g \geq n$, the elliptic surface $E(n)$ admits a genus g Lefschetz pencil; and for each $n \geq 3$ and $g \geq n$, the once blown up elliptic surface $E(n) \# \overline{\mathbb{CP}^2}$ admits a pair of inequivalent genus g Lefschetz pencils.

57K40, 57K43

1 Introduction

Since the foundational work of Donaldson [6] and Gompf [10] in the 1990s, Lefschetz pencils and fibrations have been known to characterize symplectic 4–manifolds. In [4], R İnanç Baykur, Kenta Hayano, and Naoyuki Monden construct a doubly indexed family of symplectic 4–manifolds $X'_g(i)$ for all $g \geq 3$ and $0 \leq i \leq g-1$. Their examples are constructed as the total spaces of symplectic genus g Lefschetz pencils, through explicit factorizations of their monodromy. We review the specific factorizations which define $X'_g(i)$ below, but in the meantime summarize results from [4] about these manifolds:

Theorems [4] *For each $g \geq 3$ and $0 \leq i \leq g-1$, there is a genus g Lefschetz pencil on $X'_g(i)$ with the following properties*:

(a) [4, Lemma 4.7] *The manifolds $X'_g(i)$ are simply connected, with Euler characteristic $e(X'_g(i)) = 12(g-i)$ and signature $\sigma(X'_g(i)) = -8(g-i)$.*

(b) [4, Lemma 5.6] *The manifolds $X'_g(i)$ are spin if and only if $g-i$ is even.*

(c) [4, Theorem 4.8] *The manifolds $X'_g(g-1)$ are diffeomorphic to the rational elliptic surface $E(1)$.*

These statements suggest our main result:

Theorem 1 *The manifolds $X'_g(i)$ are diffeomorphic to the elliptic surface $E(g-i)$.*

This has some immediate corollaries. In [4], Baykur, Hayano, and Monden note that when $g-i$ is even, $X'_g(i)$ is irreducible (since it is spin), but the irreducibility of $X'_g(i)$ for odd $g-i$ is left open.

Corollary 2 *$X'_g(i)$ is irreducible for all $g \geq 3$ and $0 \leq i \leq g-2$.*

Additionally, in [4], the Kodaira dimensions of $X'_g(i)$ are computed only for the special cases of $g-3 \leq i \leq g-1$ and $g-i$ even. Our main theorem fills in the missing cases:

Corollary 3 *The symplectic Kodaira dimension of $X'_g(i)$ is*

$$\kappa(X'_g(i)) = \begin{cases} -\infty & \text{if } i = g-1, \\ 0 & \text{if } i = g-2, \\ 1 & \text{if } 0 \leq i \leq g-3. \end{cases} \tag{1}$$

An additional corollary concerns *symplectic Calabi–Yau* 4–manifolds. A complex Calabi–Yau surface is one with a trivial canonical class, and one can likewise define a symplectic Calabi–Yau 4–manifold to be one with a trivial symplectic canonical class. All known examples of symplectic Calabi–Yau manifolds are complex K3 surfaces or torus bundles over tori. Since any symplectic Calabi–Yau manifold must have the rational homology type of these complex surfaces (see Bauer [2] and Li [13]), it is an intriguing open question if there exist *any* symplectic Calabi–Yau 4–manifolds which are not diffeomorphic to one of these; see Friedl and Vidussi [8] and Li [14]. Baykur, Hayano, and Monden show that the manifolds $X'_g(g-2)$ are symplectic Calabi–Yau [4, Corollary 4.10], and ask if they are diffeomorphic to the standard K3 surface.

Corollary 4 *The symplectic Calabi–Yau manifolds $X'_g(g-2)$ are diffeomorphic to $K3$.*

In addition to its relevance to finding examples of symplectic Calabi–Yau manifolds, this result serves to illustrate the diversity of Lefschetz pencils on fixed 4–manifolds. The K3 surface is known to admit pencils of every genus (see Smith [15]), and it is noted in [4] that the diffeomorphism $X'_g(g-1) \cong E(1)$ implies that the same is true for the rational elliptic surface $E(1)$. The author is not aware of any other such examples.

Corollary 5 *For all $n \geq 3$, the elliptic surface $E(n)$ admits a genus g Lefschetz pencil for every $g \geq n$.*

A deeper related application concerns finding inequivalent Lefschetz pencils on a given 4–manifold with the same topological data (ie genus and number of basepoints). By using the braiding lantern substitution technique of Baykur and Hayano [3], Baykur, Hayano, and Monden prove:

Theorem [4, Corollary 6.4] *For all $g \geq 3$ and $0 \leq i \leq g-1$, the manifold $X'_g(i) \# \overline{\mathbb{CP}^2}$ admits a pair of inequivalent genus g Lefschetz pencils. In particular, the manifold $E(1) \# \overline{\mathbb{CP}^2}$ admits a pair of inequivalent genus g Lefschetz pencils for all $g \geq 3$.*

Theorem 1 strengthens this result.

Corollary 6 *For all $n \geq 3$, the once blown up elliptic surface $E(n) \# \overline{\mathbb{CP}^2}$ admits a pair of inequivalent genus g Lefschetz pencils for all $g \geq n$.*

Of course, the conclusions of corollaries 5 and 6 apply to blow ups of $E(n)$ and $E(n) \# \overline{\mathbb{CP}^2}$ at basepoints, as well.

The method of proof of Theorem 1 exploits the natural 2–fold symmetry of Baykur, Hayano, and Monden's construction. We begin by blowing up the pencil on $X'_g(i)$ to obtain an associated Lefschetz fibration $X_g(i)$, and use this symmetry to represent $X_g(i)$ as a 2–fold branched cover of a rational surface. A sequence of handle slides in the base of this cover allows one to find and blow down the required number of exceptional sections, and we arrive at a branched cover description of $X'_g(i)$. The branch surface of this cover is represented as a banded unlink diagram, of the sort studied by Mark Hughes, Seungwon Kim, and Maggie Miller in [11], with an explicitly drawn ribbon surface as (most of) the branch locus. We then use various band moves to obtain an isotopy of the branch surface, yielding a branched cover description that is recognized as one for elliptic surfaces.

In Section 2 we discuss banded unlink diagram descriptions of embedded surfaces, and review the moves on these diagrams that we will employ in the proof. The following section reviews the topology of Lefschetz pencils and fibrations. Finally, in Section 4, we define the manifolds $X'_g(i)$ and $X_g(i)$, and give the proof of Theorem 1.

2 Banded unlink diagrams

In this section we review the notion of a banded unlink diagram [11]. This describes a closed surface embedded in a closed 4–manifold X. Banded unlink diagrams can be defined using any handlebody description for X, but since in our application X will lack 1– and 3–handles, we only discuss that setting here.

Suppose X is obtained by attaching n 2–handles to a single 0–handle, and then attaching one 4–handle. The manifold X can be depicted by a Kirby diagram $\mathcal{K}$ consisting of an n–component framed link in S^3. Let X_0 denote the boundary of the 0–handle, and X_1 the union of the 0– and 2–handles. Of course, both $\partial(X_0)$ and $\partial(X_1)$ are S^3, and $\partial(X_1)$ can be described as the result of a surgery of S^3 along the components of $\mathcal{K}$.

Let L be a link in the exterior $E(\mathcal{K})$. Since L avoids the attaching region of the 2–handles, we can view L as a link in $\partial(X_0)$ and in $\partial(X_1)$. In a banded unlink diagram, we begin with an unlink in $E(\mathcal{K})$, and form a ribbon surface by attaching a disjoint collection of bands to the spanning disks of the unlink; L is the link that results from the band surgery to the unlink, and we may push the interior of the ribbon surface into X_0 to get an embedded surface. In a banded unlink diagram, we also require that L bounds a collection of disjoint disks in $\partial(X_1)$. In this way, the ribbon surface that L bounds can be capped off by these disks, giving a closed surface in X.

In [11], Hughes, Kim, and Miller give a complete set of moves for banded unlink diagrams of isotopic surfaces in a 4–manifold. As we will apply these to manifolds without 1–handles, we review only the moves that we use later: band slides, band swims, 2–handle band swims, and 2–handle band slides. These are shown in Figure 1. (The 2–handle band slides in Figure 1 can be done with any knotted attaching circle and any framing, following the usual rules of Kirby calculus; the 0–framed unknot pictured here is all that will be used later. The strands running through the attaching circle of the handle can represent other handles, bands, or unlink components.)

Two particular iterations of the swim moves will be used often, and are shown in Figures 2 and 3. In each figure, a sequence of swims is performed, moving the band

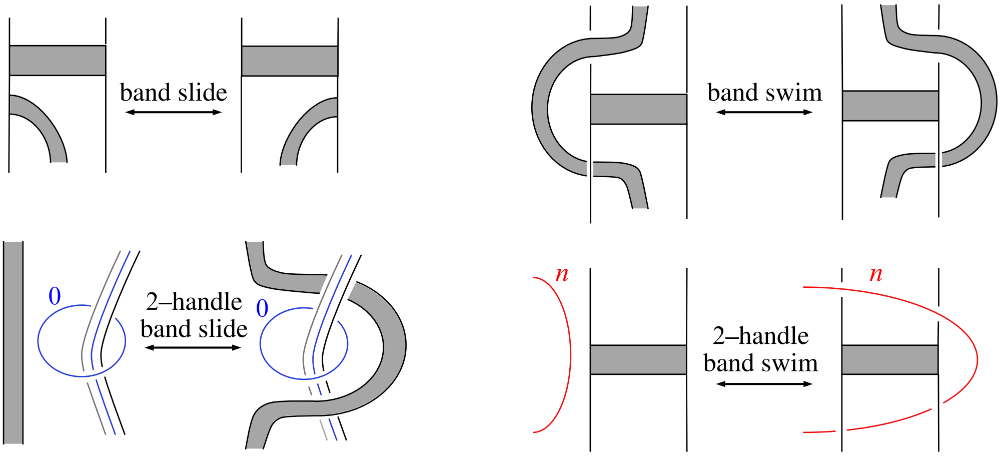


Figure 1: In the two swim moves, the band/attaching circle passes lengthwise through the interior of the horizontally drawn band.

or attaching circle from the right side of the initial diagram successively through each of the bands to its left. An intermediate step following the first swim is depicted in each figure. In later use, the initial ribbon surface from each figure will be replaced by

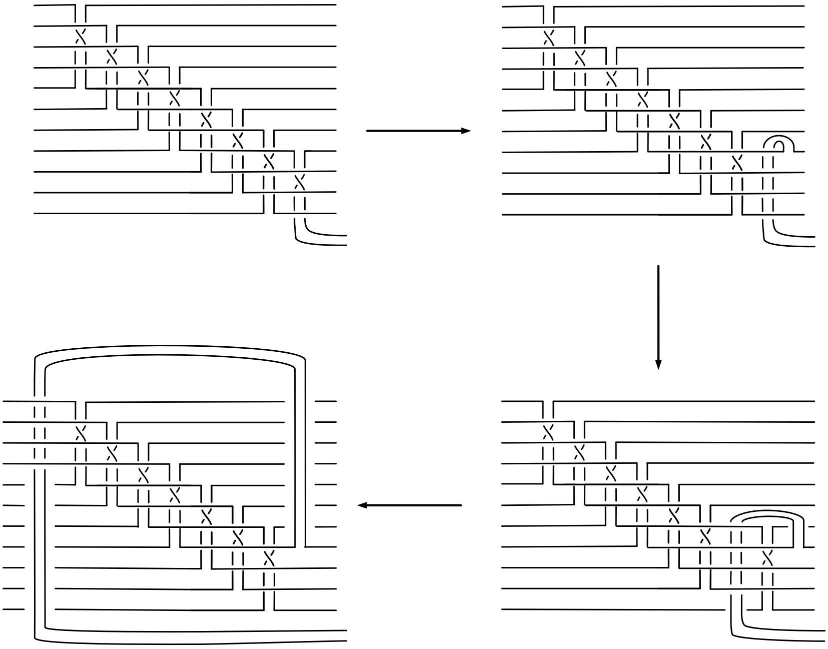


Figure 2: A band dive.

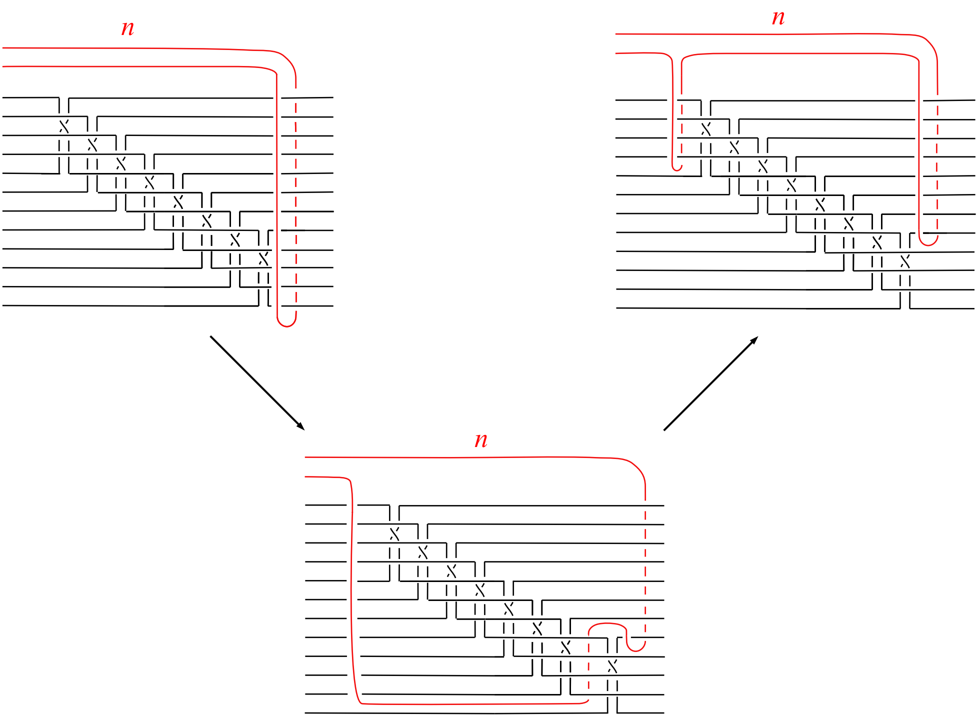


Figure 3: A 2–handle band dive.

the final one, and we will refer to these moves as *band dives* and 2*–handle band dives*, respectively.

3 Lefschetz pencils and fibrations

In this section we review the definitions of Lefschetz fibrations and Lefschetz pencils, and discuss the topology of these structures. A more comprehensive description of the topology of Lefschetz fibrations and pencils can be found in [10].

We denote a closed oriented genus g surface by Σ_g, and a compact oriented genus g surface with n boundary components by Σ_g^n. Their mapping class groups will be denoted by Γ_g and Γ_g^n, respectively. We will also denote a sphere with m marked points by $\Sigma_{0,m}$, and its mapping class group by $\Gamma_{0,m}$.

Definition Let W be a compact oriented smooth 4–manifold, and C a compact oriented smooth surface. A proper smooth map $f: W \to C$ is a *Lefschetz fibration* if

(i) the critical points of f lie in the interior of W, and

(ii) for each critical point of f in W, there are complex coordinate charts agreeing with the orientations on W and C such that locally f can be expressed as $f(z_1, z_2) = z_1^2 + z_2^2$.

We will only encounter $C = S^2$ or D^2.

Definition Let W' be a closed oriented smooth 4–manifold. Let $B \subset W'$ be a finite set of points. A smooth map $f\colon W' \setminus B \to \mathbb{CP}^1$ is a *Lefschetz pencil* if:

(i) For each critical point of f in $W' \setminus B$ there are complex coordinate charts agreeing with the orientations on W' and $\mathbb{CP}^1$ such that locally f can be expressed as $f(z_1, z_2) = z_1^2 + z_2^2$.

(ii) For each point of B there is a complex coordinate chart on W' and an identification of the base as $\mathbb{CP}^1$ such that locally f can be expressed as $f(z_1, z_2) = [z_1 : z_2]$.

The existence of a Lefschetz pencil $f\colon W' \setminus B \to \mathbb{CP}^1$ will be described by saying that there is a Lefschetz pencil *on* W'.

The points of B are called *basepoints* of the Lefschetz pencil. A Lefschetz pencil with $B = \varnothing$ is a Lefschetz fibration over $\mathbb{CP}^1 \cong S^2$. If $B \neq \varnothing$, we can blow up W' at each basepoint to get W, and the Lefschetz pencil on W' becomes a Lefschetz fibration $W \to \mathbb{CP}^1 \cong S^2$.

It is a consequence of these definitions that, for a Lefschetz fibration, a regular fiber $f^{-1}(x)$ is a closed genus g surface. For a Lefschetz pencil with $n > 0$ basepoints, $f^{-1}(x)$ is not compact, and we instead consider $f^{-1}(x) \cap (W' \setminus (U_1 \cup \cdots \cup U_n))$, where U_i is an open ball about the basepoint in each coordinate chart with property (ii) above. This fiber will be a compact genus g surface with n boundary components. We refer to *genus* g Lefschetz fibrations or pencils, accordingly.

Lefschetz pencils and fibrations are understood topologically through monodromy factorizations. Let $x_1, \dots, x_\mu$ be the critical values for f. We assume, without loss of generality, that each critical point of f lies in a separate fiber. For a pencil, we select a regular value $x_0 \in \mathbb{CP}^1$, and a disjoint collection of arcs γ_i from x_0 to x_i for each $i = 1, \dots, \mu$. (We also assume each γ_i avoids the other critical points.) We further assume the arcs $\gamma_1, \dots, \gamma_\mu$ appear in this order as we travel in a small circle about x_0. For each i, we consider a loop that begins at x_0, travels along γ_i, then counterclockwise around a small circle centered at x_i, and back to x_0 along γ_i. Using an identification of $f^{-1}(x_0)$ with Σ_g^n, the monodromy of f along this loop is known to be a right-handed Dehn twist t_{c_i} along a simple closed curve $c_i \subset \Sigma_g^n$ [12]. The curve c_i is called a *vanishing cycle*. To get a global description of a Lefschetz pencil, these local models must fit together according to the equation $t_{c_1} \dots t_{c_\mu} = t_{\delta_1} \dots t_{\delta_n}$ in Γ_g^n, where δ_j denotes a right-handed Dehn twist about a curve parallel to the j^{th} boundary component of Σ_g^n. Conversely, given any factorization in Γ_g^n of $t_{\delta_1} \dots t_{\delta_n}$ as a product of right-handed Dehn twists, one can construct a Lefschetz pencil with monodromy prescribed by the factorization.

When working with Lefschetz fibrations, one has a similar description of the local monodromy about a critical value x_i as a right-handed Dehn twist t_{c_i} about a simple closed curve $c_i \subset \Sigma_g$. To form a global Lefschetz fibration over S^2, the local monodromies must concatenate to form a relation $t_{c_1} \dots t_{c_\mu} = 1$ in Γ_g.

Any particular monodromy description of a Lefschetz pencil is far from unique, as it depends on a choice of identification of a regular fiber, as well as on a system of arcs γ_i. Modifying these choices translates into a simple set of moves on factorizations in Γ_g^n (see [10]), and two factorizations related in this way are said to be *Hurwitz equivalent*.

There is a straightforward relationship between a monodromy factorization of a Lefschetz pencil on W' and that of the Lefschetz fibration $W \to S^2$ obtained by blowing up W' at all basepoints. Under the homomorphism $\Gamma_g^n \to \Gamma_g$ obtained by capping off each boundary component of Σ_g^n with a disk, Dehn twists about the boundary parallel curves δ_j become trivial in Γ_g. A monodromy factorization $t_{c_1} \dots t_{c_\mu} = t_{\delta_1} \dots t_{\delta_n}$ in Γ_g^n for the pencil on W' then gives a monodromy factorization $t_{c_1} \dots t_{c_\mu} = 1$ in Γ_g for the fibration $W \to S^2$.

A monodromy factorization of a genus g Lefschetz fibration $f\colon W \to S^2$ also leads to a handlebody description of W in a well-understood way [10]. One begins with a handlebody description of $\Sigma_g \times D^2$ consisting of a 0–handle, 1–handles, and 2–handles. Given a factorization $t_{c_1} \dots t_{c_\mu} = 1$ in Γ_g, we form $\Sigma_g \times D^2 \cup \bigl(\bigcup_{i=1}^{\mu} H_i^2\bigr)$, where each H_i^2 is a 2–handle attached along the vanishing cycle c_i in a separate fiber $\Sigma_g \times \{\text{point}\} \subset \Sigma_g \times S^1 = \Sigma_g \times \partial D^2$. The 2–handles are attached along the S^1 factor in the order they appear in the factorization, and they have framing -1 relative to the framing on c_i induced by the product $\Sigma_g \times S^1$. Following these handle attachments, we have a handlebody describing a Lefschetz fibration over D^2 with the prescribed monodromy factorization. The boundary of $\Sigma_g \times D^2 \cup \bigl(\bigcup_{i=1}^{\mu} H_i^2\bigr)$ is Σ_g–bundle over S^1 with monodromy $t_{c_1} \dots t_{c_\mu}$; because this is isotopic to the identity, this boundary is diffeomorphic to $\Sigma_g \times S^1$. Hence we can extend the Lefschetz fibration to be one over S^2 by attaching the trivial fibration $\Sigma_g \times D^2 \to D^2$ along $\Sigma_g \times S^1$. This final attachment adds one or more 2–handles, 3–handles, and a 4–handle.

A technique for constructing new Lefschetz pencils or fibrations from old is *monodromy substitution*. Given a monodromy factorization, a monodromy substitution swaps a subword of the factorization with a different (but equal, in Γ_g^n or Γ_g) product of right-handed Dehn twists. In [4], Baykur, Hayano, and Monden employ this operation using the *odd chain relation*: suppose $c_1, c_2, \dots, c_{2h+1}$ are simple closed curves on Σ_g^n or Σ_g that form a chain; that is, c_i and c_{i+1} intersect in one point for all i, and c_i and c_j are disjoint otherwise. A regular neighborhood of $c_1 \cup \dots \cup c_{2h+1}$ is a subsurface S homeomorphic to Σ_h^2. The chain relation is $(t_{c_1} t_{c_2} \dots t_{c_{2h+1}})^{2h+2} = t_{b_1} t_{b_2}$, where b_1 and b_2 are the boundary components of S. Using this relation to replace a subword in a monodromy factorization given by the left-hand side of the chain relation with the two Dehn twists on the right is referred to as *unchaining*.

Realizing hyperelliptic Lefschetz fibrations as branched covers Let $\iota\colon \Sigma_g \to \Sigma_g$ be the hyperelliptic involution, and $\pi\colon \Sigma_g \to \Sigma_{0,2g+2}$ the branched covering that is the quotient of ι. A Lefschetz fibration on $W \to S^2$ is *hyperelliptic* if it is Hurwitz equivalent to one with a monodromy factorization where each vanishing cycle c_i satisfies $\iota(c_i) = c_i$. If all c_i are nonseparating, then W is a 2–fold branched cover of an S^2–bundle over S^2, with the Lefschetz fibration map obtained as the composition of this cover with the bundle projection [9]. This cover is crucial to the proof of Theorem 1, and we review the details.

Since all c_i are nonseparating and symmetric, the factorization $t_{c_1} \dots t_{c_\mu} = 1$ is the lift of the relation $h_{\pi(c_1)} \dots h_{\pi(c_\mu)} = 1$ in $\Gamma_{0,2g+2}$, where $h_{\pi(c_i)}$ is a right-handed disk twist about the arc $\pi(c_i)$ in $\Sigma_{0,2g+2}$. The factorization $h_{\pi(c_1)} \dots h_{\pi(c_\mu)}$ can be used to construct a ribbon surface in $S^2 \times D^2$, for which the cover branched over that surface is a Lefschetz fibration over D^2 with the required monodromy factorization. The Birman–Hilden theorem (see [5; 7]) then implies that we can always extend this cover by attaching a trivial covering of $\Sigma_g \times D^2$ over $S^2 \times D^2$, resulting in W covering an S^2–bundle over S^2 branched over a closed surface.

In practice, the base and branch set of this cover can be explicitly drawn as a banded unlink diagram. In $S^2 \times D^2$, represented as a Kirby diagram by a 0–framed unknot, we begin with $2g + 2$ disks representing {point} × D^2, drawn as meridians to the unknot, with their interiors pushed into the 0–handle. The branched cover of $S^2 \times D^2$

over these disks is $\Sigma_g \times D^2$, restricting to the hyperelliptic quotient in each fiber. A ribbon surface is then constructed by attaching left-handed half-twisted bands so that the core of each band is the arc $\pi(c_i)$ in $S^2 \times \{\text{point}\}$. By the method in [1], in the 2–fold cover of $S^2 \times D^2$ branched over this ribbon surface, each added band lifts to a 2–handle attached along c_i, with relative framing -1. Thus the lift of $S^2 \times D^2$ branched over the ribbon surface is the total space of a Lefschetz fibration over D^2, with monodromy factorization $t_{c_1} \dots t_{c_\mu}$. On the boundary, we have a Σ_g–bundle over S^1 covering an S^2–bundle over S^1, each with monodromy isotopic to the identity. To extend the branched covering over W, it is necessary to find a *fiber-isotopy* of the factorization $t_{c_1} \dots t_{c_\mu}$ to the identity (ie an isotopy through homeomorphisms which are all fiber-preserving with respect to π): using a given fiber-isotopy to the identity, we can then identify the branched covering on the boundary as $\pi \times \mathrm{id}\colon \Sigma_g \times S^1 \to S^2 \times S^1$ and extend the covering as $\pi \times \mathrm{id}\colon \Sigma_g \times D^2 \to S^2 \times D^2$. The attachment of $S^2 \times D^2$ to the base matches the boundary of disks $\{\text{point}\} \times D^2$ to the boundary of the ribbon surface, and in this way we get a closed surface as branch set. The extension attaches a 2–handle union a 4–handle to the diagram of the base, with the 2–handle attached as a meridian to the 0–framed 2–handle. When working with examples, the braid factorization $h_{\pi(c_1)} \dots h_{\pi(c_\mu)}$ plays a valuable role. The necessary fiber-isotopy to the identity can often be seen by simply observing that the braid factorization is isotopic to the identity by an isotopy that fixes the branch points at all times, in which case one obtains a fiber-isotopy of $t_{c_1} \dots t_{c_\mu}$ to the identity as its lift. We can also use the braid factorization to compute the framing of the second attached 2–handle and to see how the attaching circle links the boundary of the branch surface. To do this, we select a reference point $* \in \Sigma_{0,2g+2} \setminus B^2_{2g+2}$, where B^2_{2g+2} is a disk containing the branch points, and track a framed neighborhood of $*$ through the isotopy of $d_{h(c_1)} \dots d_{h(c_\mu)}$ to the identity.

In [9], it was shown how to modify this branched covering description of a hyperelliptic Lefschetz fibration to accomplish an unchaining monodromy substitution. Although the procedure in [9] was described only for even unchaining substitutions, the method applies equally well to the odd unchaining substitutions considered here.

4 The proof of Theorem 1

We are now ready to describe the manifolds $X'_g(i)$ constructed by Baykur, Hayano, and Monden, and prove that they are diffeomorphic to the elliptic surfaces $E(g-i)$.

4.1 The manifolds $X'_g(i)$ and $X_g(i)$

In [4], Baykur, Hayano, and Monden construct their infinite family of Lefschetz pencils by explicit monodromy factorization. Their factorizations use Dehn twists about the curves on $\Sigma_g^{2(i+1)}$ shown in Figure 4. We abbreviate the product of boundary curve twists as $\Delta = t_{\delta_{i+1}} \dots t_{\delta_2} t_{\delta_1} t_{\delta'_{i+1}} \dots t_{\delta'_2} t_{\delta'_1}$, and also let $D_g = t_{d_4} t_{d_5} \dots t_{d_{2g+1}}$ and $E_g = t_{e_{2g+1}} \dots t_{e_5} t_{e_4}$.

Theorem [4, Theorem 4.6] *For each $g \geq 3$ and $0 \leq i \leq g-1$, there are symplectic genus g Lefschetz pencils on $X'_g(i)$ with monodromy factorizations in $\Gamma_g^{2(i+1)}$,*

$$\Delta = \begin{cases} D_g E_g t_{x_{i+1}} \dots t_{x_2} t_{x_1} t_{x'_{i+1}} \dots t_{x'_2} t_{x'_1} (t_{c_1} t_{c_2} t_{c_3})^{4(g-i)}, & g \text{ odd}, \\ D_g E_g t_{x_{i+1}} \dots t_{x_2} t_{x_1} t_{x'_{i+1}} \dots t_{x'_2} t_{x'_1} (t_{c_1} t_{c_2} t_{c_3})^{4(g-1-i)+2} (t_{c_3} t_{c_2} t_{c_1})^2, & g \text{ even}. \end{cases}$$

(Here we have cyclically permuted the right-hand side from its expression in [4].)

If we cap off each boundary component of $\Sigma_g^{2(i+1)}$ with a disk, each of the curves x_j and x'_j become parallel copies of a curve x and x', respectively, on Σ_g. From the previous equation we see that the monodromy factorization of the Lefschetz fibration $X_g(i) \to S^2$ is

$$(2) \qquad 1 = \begin{cases} D_g E_g (t_x)^{i+1} (t_{x'})^{i+1} (t_{c_1} t_{c_2} t_{c_3})^{4(g-i)} & \text{if } g \text{ is odd}, \\ D_g E_g (t_x)^{i+1} (t_{x'})^{i+1} (t_{c_1} t_{c_2} t_{c_3})^{4(g-1-i)+2} (t_{c_3} t_{c_2} t_{c_1})^2 & \text{if } g \text{ is even}. \end{cases}$$

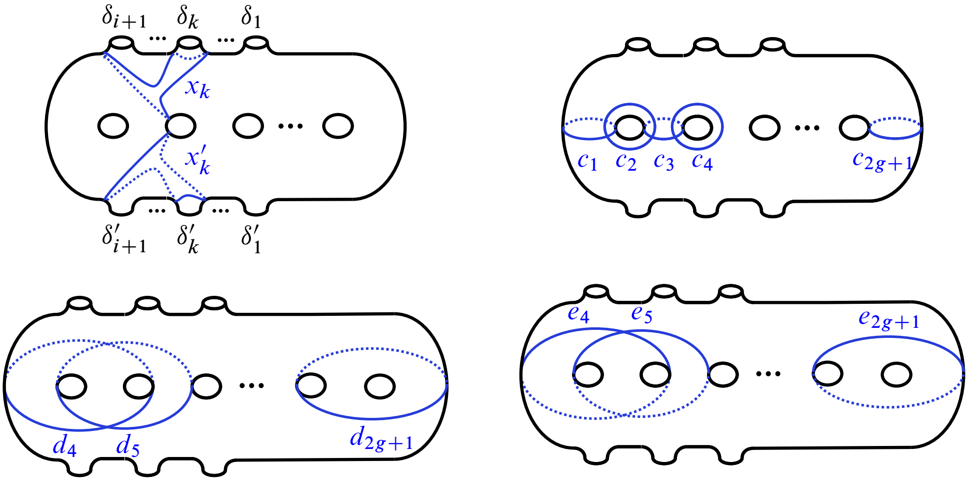


Figure 4: Curves on $\Sigma_g^{2(i+1)}$.

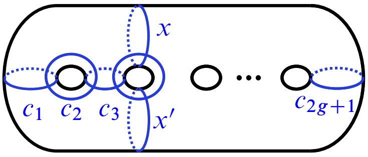


Figure 5: Curves on Σ_g.

As it will play a role later, we review Baykur, Hayano, and Monden's derivation of this monodromy factorization. They begin with the full chain relation

$$t_{\delta_1} t_{\delta_1'} = (t_{c_1} t_{c_2} \dots t_{c_{2g+1}})^{2g+2}$$

in Γ_g^2. This is well known to be the monodromy of a pencil with two basepoints on a complex surface Z_g' of general type. Through a series of lemmas, they show this is Hurwitz equivalent to the factorization

$$t_{\delta_1} t_{\delta_1'} = \begin{cases} D_g E_g (t_{c_1} t_{c_2} t_{c_3})^{4g} (t_{c_5} t_{c_6} \dots t_{c_{2g+1}})^{2g-2}, & g \text{ odd}, \\ D_g E_g (t_{c_1} t_{c_2} t_{c_3})^{4(g-1)+2} (t_{c_3} t_{c_2} t_{c_1})^2 (t_{c_5} t_{c_6} \dots t_{c_{2g+1}})^{2g-2}, & g \text{ even}. \end{cases} \tag{3}$$

They then apply unchaining monodromy substitutions to this factorization, i times to the subword $(t_{c_1} t_{c_2} t_{c_3})^4$, and once to $(t_{c_5} t_{c_6} \dots t_{c_{2g+1}})^{2g-2}$. In addition, a clever inductive use of the lantern relation shows that this relation has a lift from Γ_g^2 to $\Gamma_g^{2(i+1)}$, providing enough sections of the pencil to allow for the computation of the symplectic Kodaira dimension for some of the resulting 4–manifolds, and giving the factorization in the above theorem.

We give separate proofs that $X_g'(i) \cong E(g-i)$ for g odd and even. Each proof will have two stages: representing $X_g'(i)$ as a 2–fold branched cover, followed by modifications of the base that realize the diffeomorphism.

4.2 The proof for odd g

4.2.1 Representing $X_g'(i)$ as a branched covering Let $\mathbb{F}_n$ denote the n^{th} Hirzebruch surface. We begin by discussing how to represent $X_g'(i)$ for odd g as the 2–fold branched cover of the rational surface $\mathbb{F}_{i+1}$, branched over an embedded surface. The base of the covering and the branch surface will be represented as a banded unlink diagram.

Recalling the derivation of the factorization in the theorem of Section 4.1, we discuss first the Lefschetz fibration $Z_g \to S^2$ that comes from blowing up the Lefschetz pencil defined by (3). This Lefschetz fibration on Z_g has monodromy given by the relation

$$D_g E_g (t_{c_1} t_{c_2} t_{c_3})^{4i} (t_{c_5} t_{c_6} \dots t_{c_{2g+1}})^{2g-2} (t_{c_1} t_{c_2} t_{c_3})^{4(g-i)} = 1 \tag{4}$$

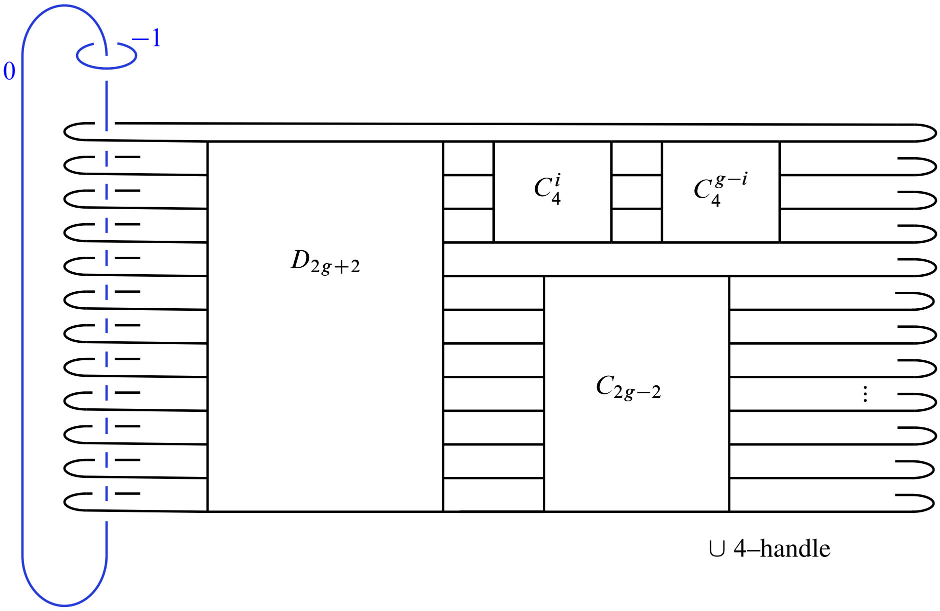


Figure 6: The 2–fold branched cover is Z_g.

in Γ_g. This is a hyperelliptic Lefschetz fibration, and from the discussion in Section 3, we see that Z_g can be described as the 2–fold cover of $\mathbb{F}_1$ branched over the surface described in Figure 6. The visible part of the branch surface is the ribbon surface consisting of $2g+2$ horizontal disks together with the collection of bands C_4, C_{2g-2}, and D_{2g+2} defined in Figures 7 and 8. (The exponents for C_4 denote repeated copies.) The branched cover of the 0–handle union the 0–framed 2–handle branched over the ribbon surface is a Lefschetz fibration over D^2 with monodromy given by (4). It can be checked directly using the Alexander method (see [7]) that the projection of (4)

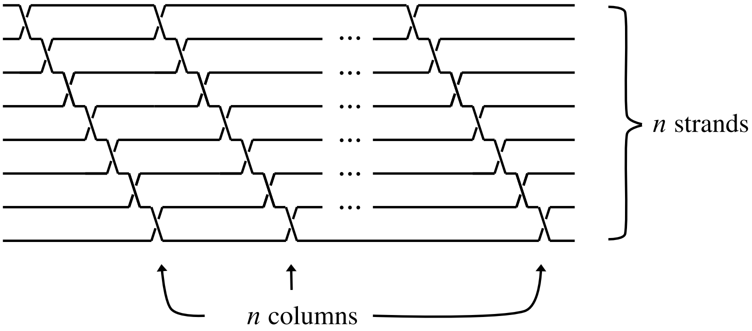


Figure 7: The braid C_n.

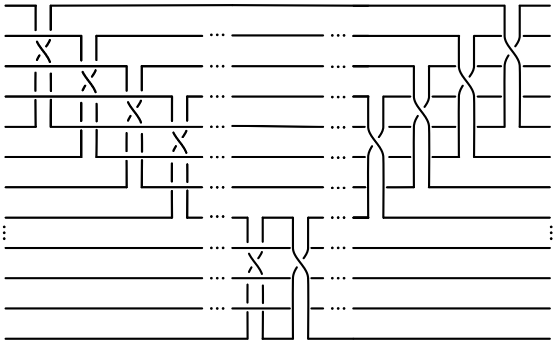


Figure 8: The braid D_n, with n strands.

to a homeomorphism of $\Sigma_{0,2g+2}$ equals a right-handed Dehn twist about a circle which encloses all marked branch points. This is isotopic to the identity by an isotopy that fixes all branch points throughout, providing a fiber-isotopy to the identity, as required. This isotopy also fixes a reference point $* \in \Sigma_{0,2g+2} \setminus B^2_{2g+2}$, and rotates a framed neighborhood of $*$ once in a left-handed direction. Thus if we attach the second 2–handle as shown in Figure 6, along a meridian with framing -1, we match $2g+2$ disks to the boundary of the ribbon surface, and we see Z_g as the cover of the surface given as a banded unlink diagram, as claimed.

We now consider unchaining substitutions on (4), i times on the subword $(t_{c_1} t_{c_2} t_{c_3})^{4i}$ and once on $(t_{c_5} t_{c_6} \dots t_{c_{2g+1}})^{2g-2}$. Doing so yields Baykur, Hayano, and Monden's relation (2) that defines the Lefschetz fibration $X_g(i)$. As described in [9], we can realize this substitution pictorially by "blowing up" the chain boxes in Figure 6, that is, by replacing them with -1–framed 2–handles, as shown in Figure 9. Each of the newly introduced 2–handles will lift to two 2–handles with relative product framing -1, attached along the pair of vanishing cycles x and x'. This figure still represents a banded unlink diagram, with $2g+2$ disks in the 4–handle, attached to the boundary of the ribbon surface. Thus $X_g(i)$ is the 2–fold cover of $\mathbb{F}_1 \# (i+1)\overline{\mathbb{CP}^2}$ branched over the surface shown in Figure 9.

We next execute a series of moves to the base of the branched covering. We begin by isotoping the newly added 2–handles by swinging them around the back of the ribbon surface so that they appear on the left, as in Figure 10.

We next slide the upper left -1–framed 2–handle over the lower one, producing Figure 11. Next the -2–framed 2–handle is slid over the parallel 0–framed one, giving

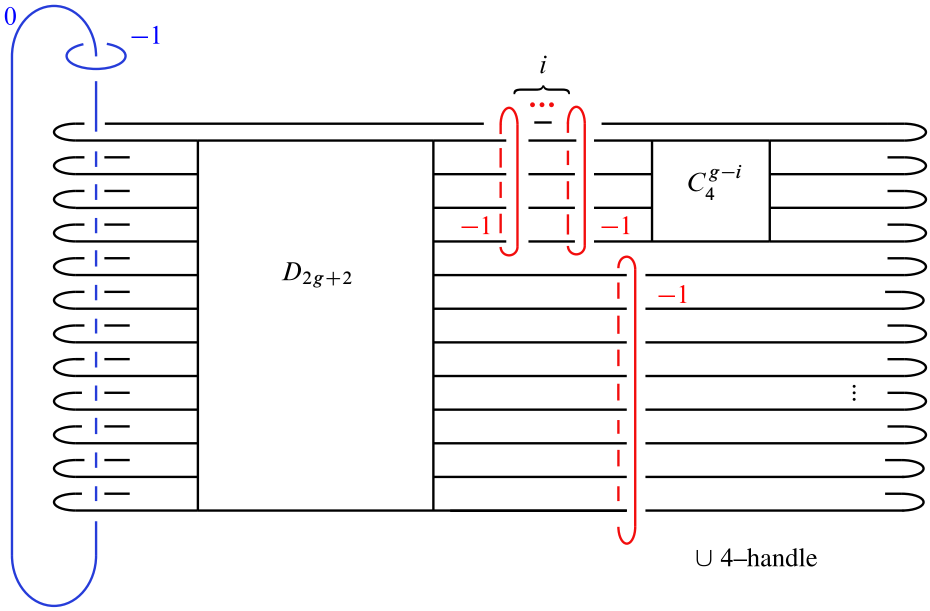


Figure 9: The 2–fold branched cover is $X_g(i)$.

Figure 12, and then slid over the -1–framed 2–handle that links it as a meridian. The result is Figure 13.

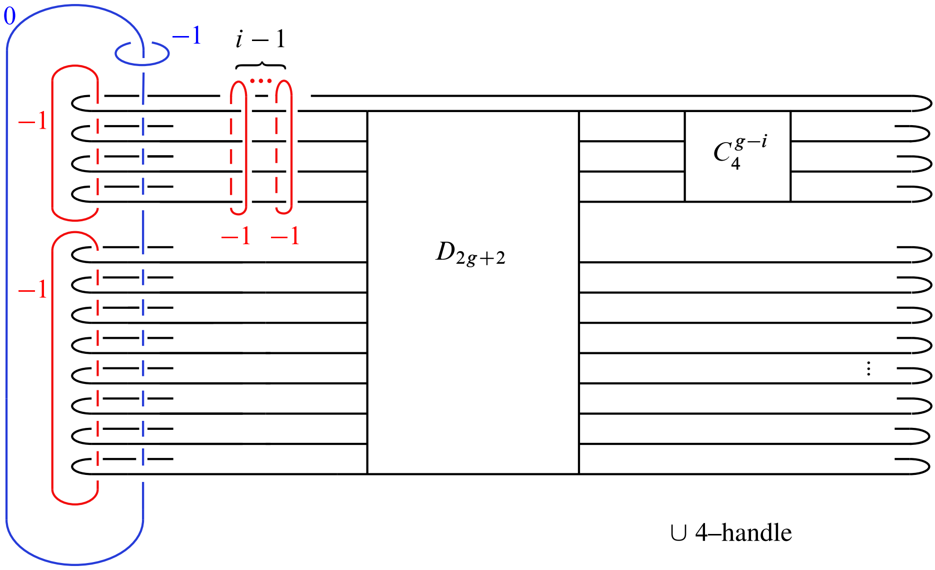


Figure 10

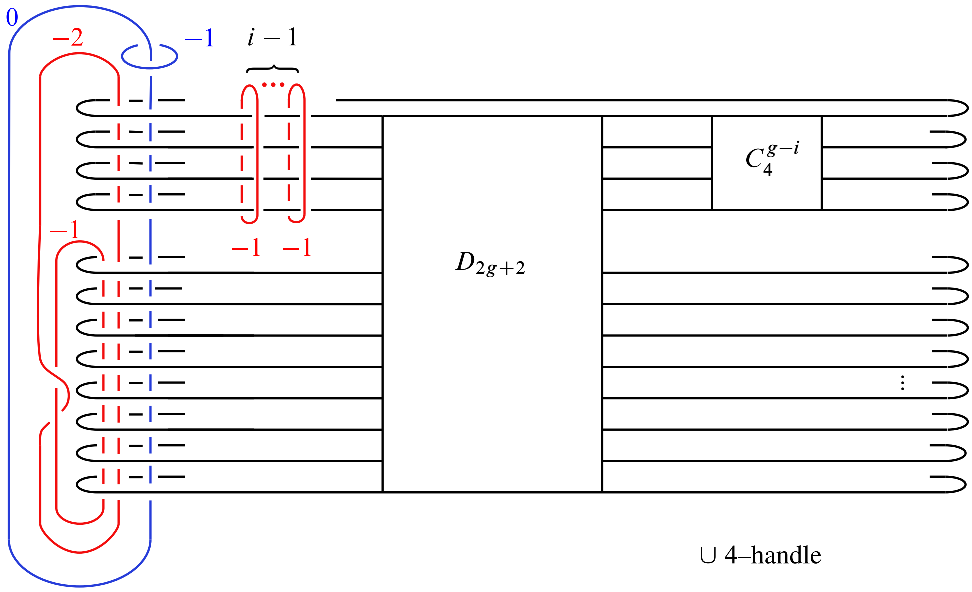


Figure 11

We repeat this series of slides for each of the remaining −1–framed 2–handles at the top of the picture, resulting in Figure 14.

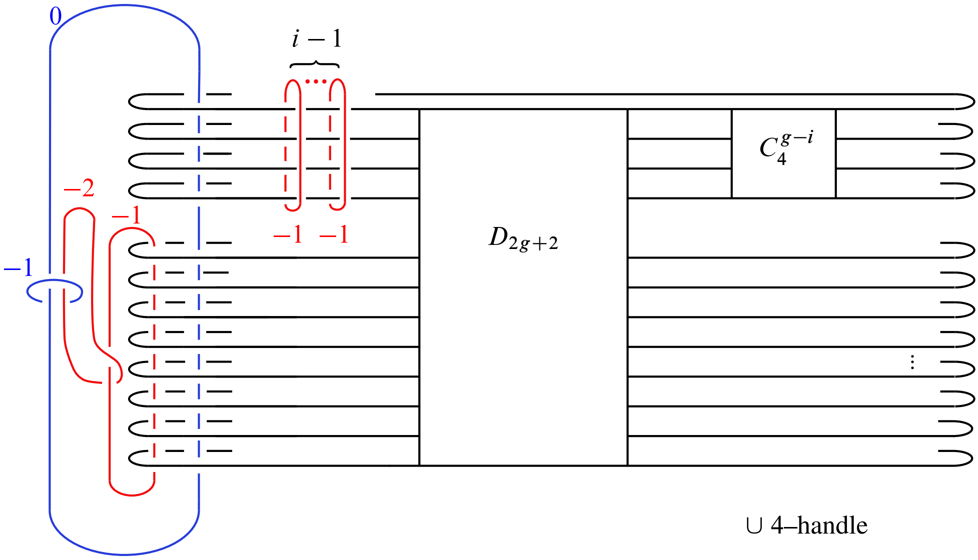


Figure 12

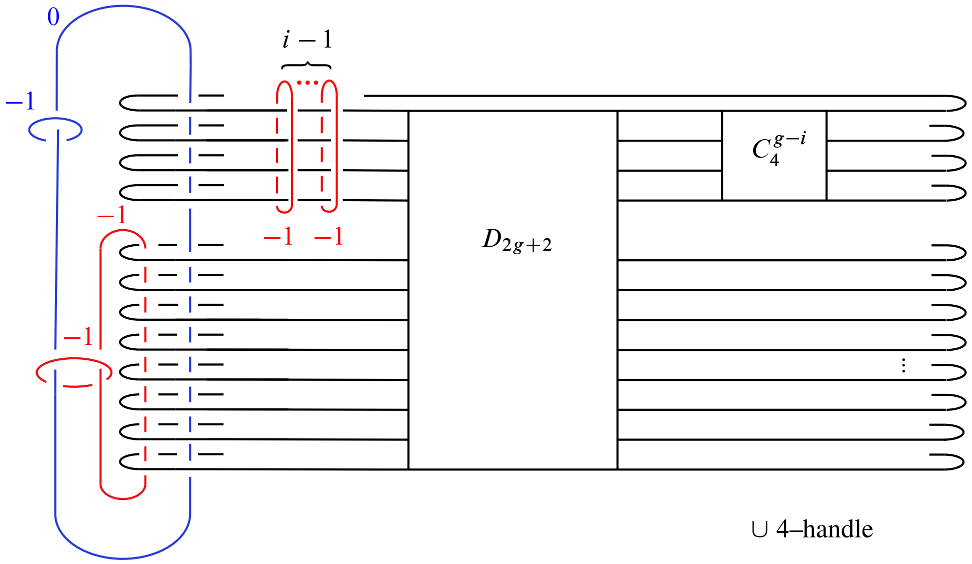


Figure 13

Next we slide the lower −1–framed 2–handle over the blue 0–framed 2–handle, then slide the result over the (blue) −1–framed 2–handle, giving Figure 15. Finally, we blow

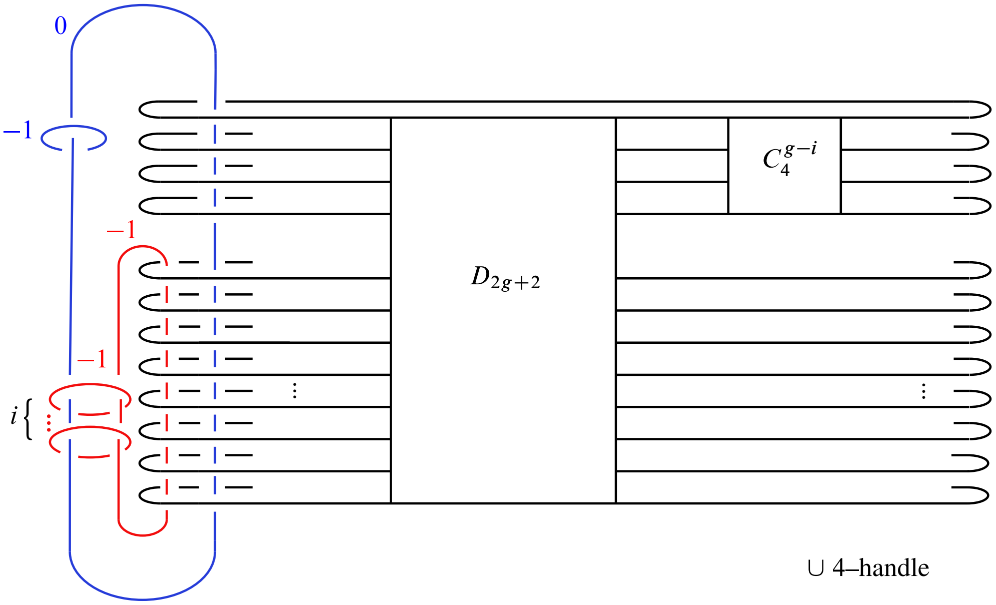


Figure 14

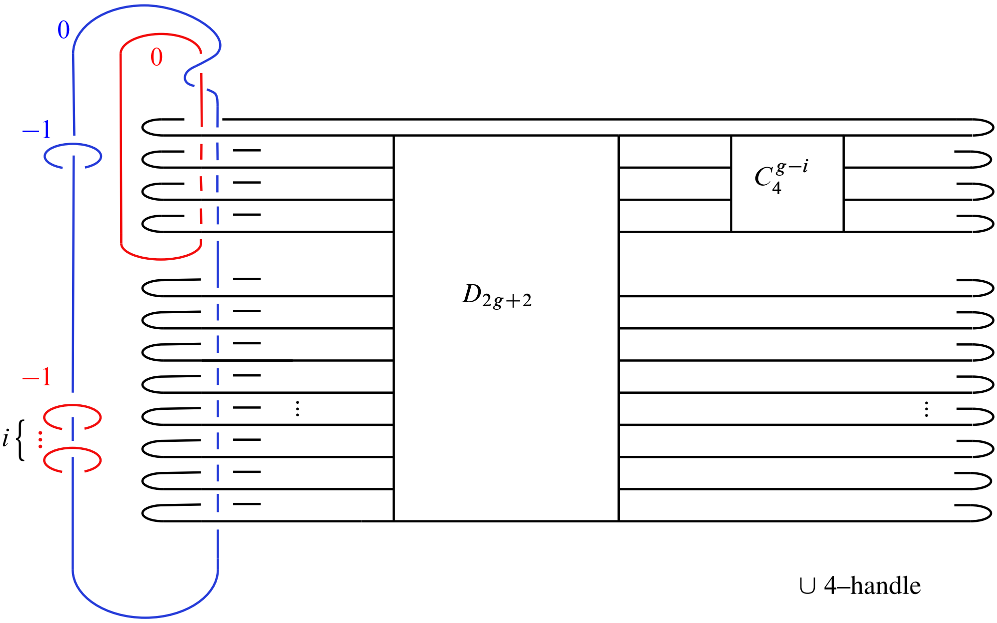


Figure 15

down each of the −1–framed 2–handles that link the 0–framed 2–handle, to arrive at Figure 16.

We pause here for an important observation: in this last step, each of the 2–handles that we are blowing down are attached along meridians to the 0–framed 2–handle.

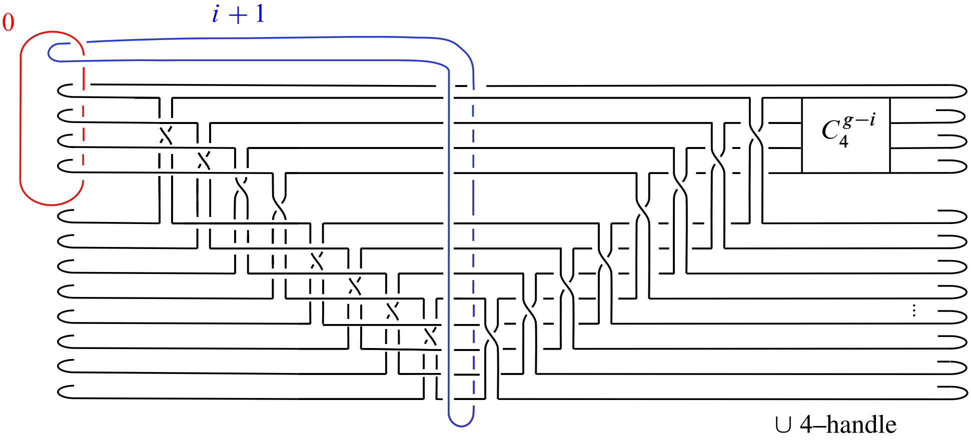


Figure 16: The 2–fold branched cover is $X'_g(i)$.

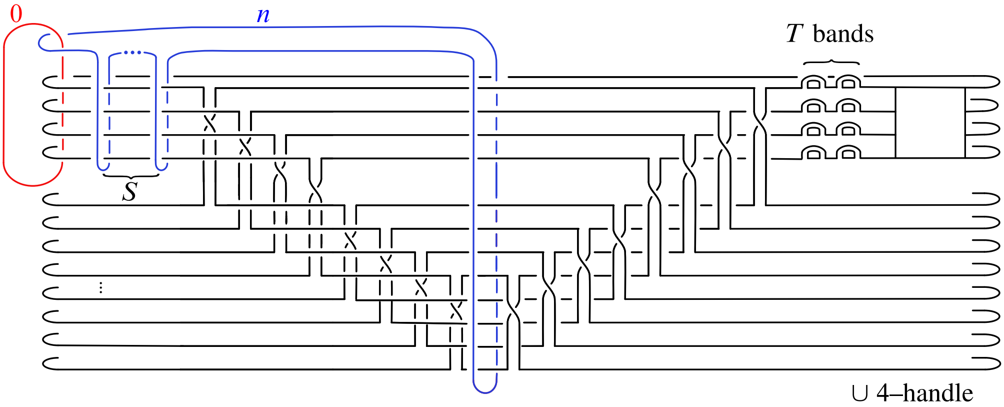


Figure 17: The ribbon surface $F(R,S,T)$ has R horizontal disks.

Retracing the diffeomorphism that goes between Figures 16 and 9, we see that the spheres given by these handles will each lift to two sections of the Lefschetz fibration on $X_g(i)$, of square -1. Because we have blown down $2(i+1)$ sections of the fibration $X_g(i)$ with square -1, it follows that the 2–fold branched cover of $\mathbb{F}_{i+1}$ branched over the embedded surface described in Figure 16 is $X'_g(i)$.

We next show that description of $X'_g(i)$ as the branched cover in Figure 16 can be used to show that it is diffeomorphic to $E(g-i)$. This relies on a key lemma.

4.2.2 A key lemma

To set up the statement, let $F(R,S,T)$ denote any ribbon surface in the 4–manifold $\mathbb{F}_n$ of the form shown in Figure 17. The box can represent

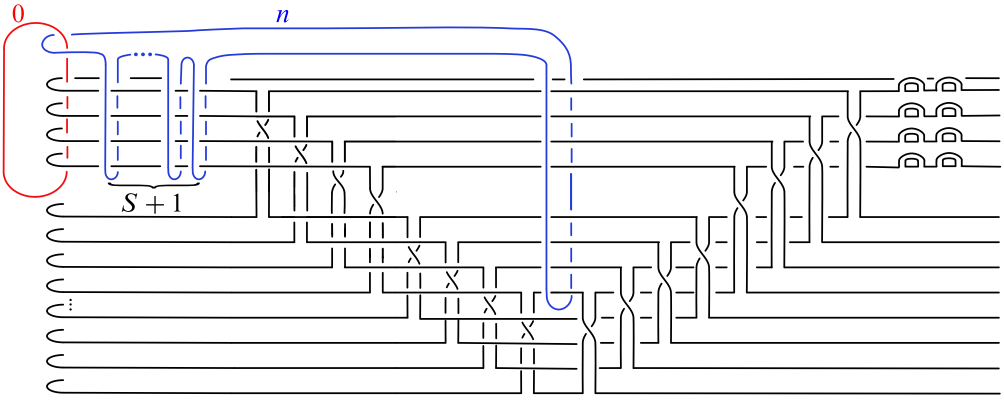


Figure 18

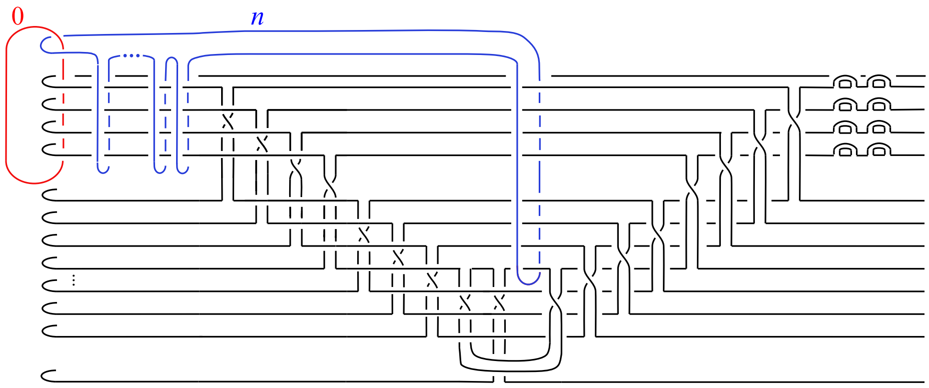


Figure 19

any collection of bands, with the condition that any bands located there are attached to the top four horizontal disks, and avoid the disks below.

The notation records that

- the ribbon surface has R horizontal disks,
- the n–framed attaching circle links the horizontal disks S times positively in the indicated region, and
- there are T trivial bands attached to the top four horizontal disks.

In applications of Lemma 7, T will be divisible by four, and the trivial bands will be distributed evenly among the top four horizontal disks.

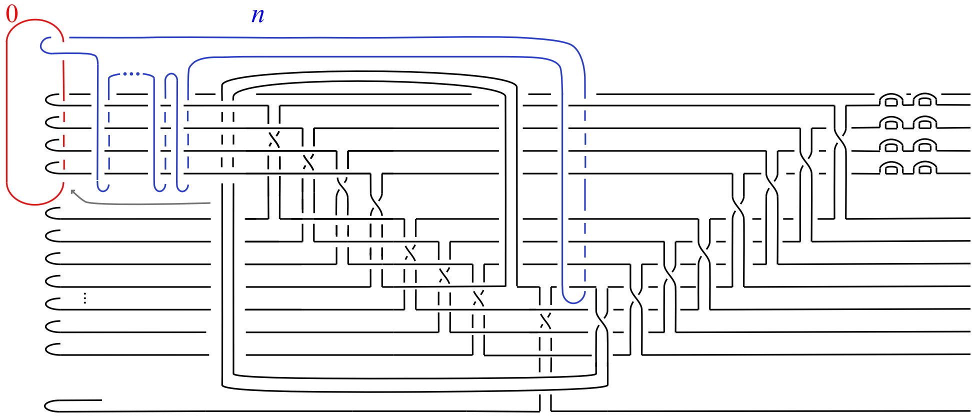


Figure 20

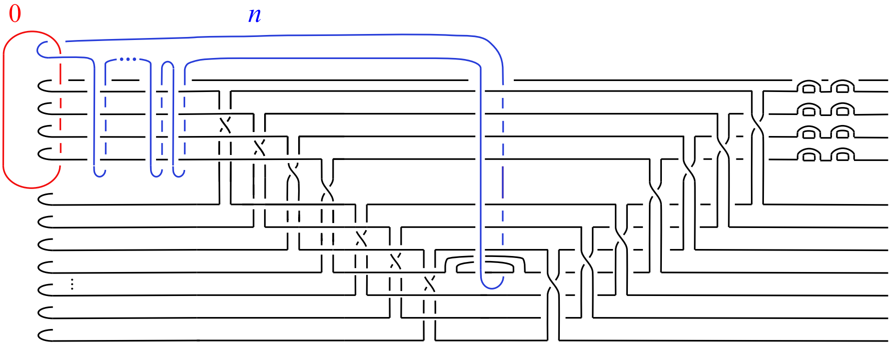


Figure 21

Lemma 7 *For $R \geq 8$, the ribbon surface $F(R,S,T)$ is isotopic to the ribbon surface $F(R-4,S+1,T+4)$.*

Proof Beginning with $F(R,S,T)$ as shown in Figure 17, we obtain Figure 18 by a 2–handle band dive of the n–framed 2–handle. This increases the linking in the upper left of the picture to $S+1$. A band slide results in Figure 19, and a band dive of that same band gives Figure 20.

At this point, we may cancel the bottom horizontal disk with the remaining attached band. In addition, we do a 2–handle band slide over the 0–framed 2–handle, using a band indicated by the gray arrow; the slide disengages the band from the top four horizontal disks, and it can be isotoped to the trivial band shown in Figure 21.

The transition from Figure 18 to Figure 21 resulted in the cancellation of the bottom horizontal disk, and added a trivial band in the process. We can repeat these steps three

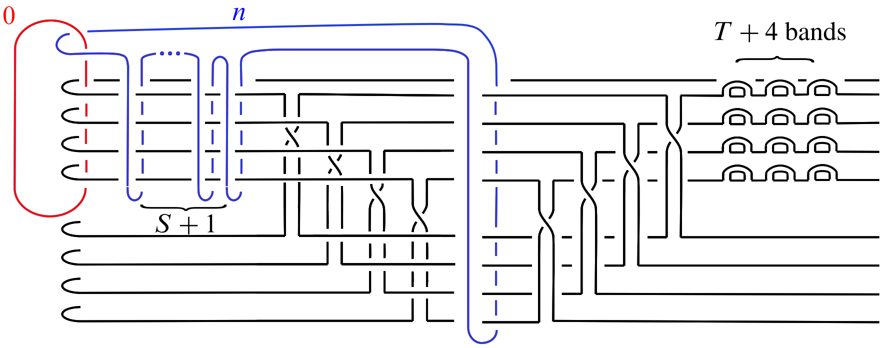


Figure 22

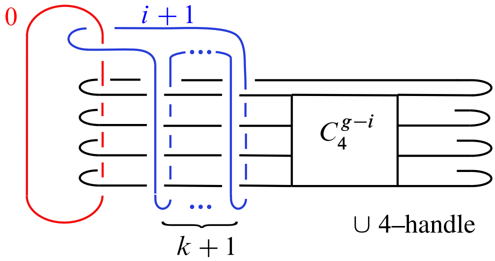


Figure 23

times to remove the bottom three horizontal disks, as shown in Figure 22. In this figure, we have moved the trivial bands from their position in Figure 21, by sliding them over the long bands to their right, so that they are now attached to the top four disks. In total we have removed the four bottom horizontal disks, and added four trivial bands; thus the values of R and T change to $R-4$ and $T+4$, respectively. □

4.2.3 An isotopy of the branch surface Let $g=2k+1$. Returning to the proof of Theorem 1, Figure 16 shows that $X_g'(i)$ is diffeomorphic to the 2–fold branched cover of $\mathbb{F}_{i+1}$ branched over a surface of the form $F(2g+2,0,0)=F(4k+4,0,0)$. Then k iterations of Lemma 7 give that $X_g'(i)$ is diffeomorphic to the 2–fold branched cover branched over a surface of the form $F(4,k,4k)$. Recall that the full surface in Figure 16 includes $2g+2$ unseen disks attached to the boundary of the ribbon surface, with their interiors in the 4–handle. Using $4k=2g-2$ of these disks to cancel the trivial bands, we have that $X_g'(i)$ is diffeomorphic to the cover of the manifold in Figure 23. (Note that four disks remain in the 4–handle.) We arrive at Figure 24 by sliding the $(i+1)$–framed 2–handle over the 0–framed one $k+1$ times. The new framing is $(i+1)-2(k+1)=i-2k-1=-(g-i)$, as shown.

The proof for odd g is completed by recognizing that the branched cover of $\mathbb{F}_{g-i}$ over the surface in Figure 24 is $E(g-i)$. This is immediate from the discussion

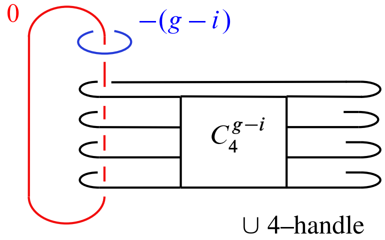


Figure 24

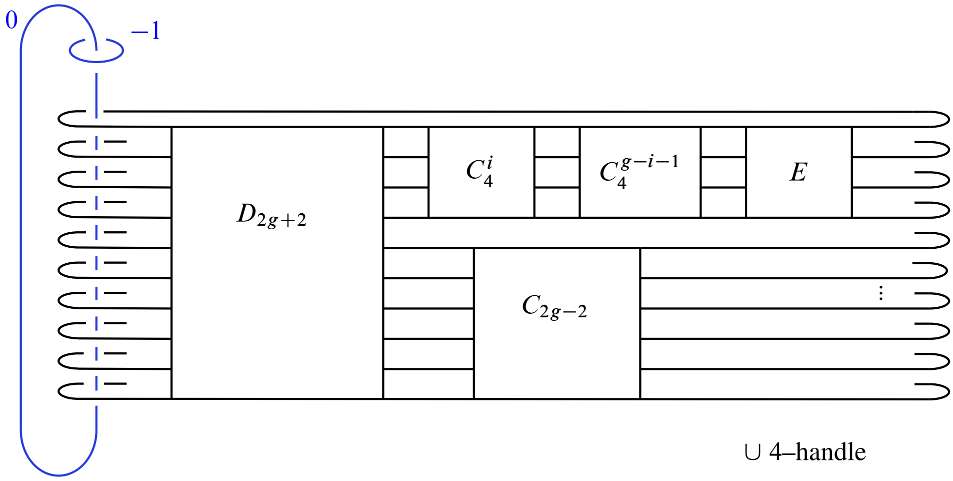


Figure 25

in Section 3. The lift of the branched cover of the 0–handle union the 0–framed 2–handle branched over the ribbon surface is a genus 1 Lefschetz fibration over D^2 with monodromy $(t_{c_1}t_{c_2}t_{c_3})^{4(g-i)}$. The braid $(d_{\pi(c_1)}d_{\pi(c_2)}d_{\pi(c_3)})^{4(g-i)}$ is equal to $g-i$ full right-handed Dehn twists about a circle enclosing all branch points. This isotopy of this to the identity fixes a reference point in $\Sigma_{0,4} \setminus B_4^2$ while rotating a framed neighborhood $g-i$ times in a left-handed direction. Thus adding a 2–handle with the indicated location and framing shows that the branched cover of $\mathbb{F}_{g-i}$ over the rest of the surface extends to a total space which is a genus 1 Lefschetz fibration over S^2, whose monodromy matches a well-known factorization of $E(g-i)$.

4.3 The proof for even g

The proof for even g is essentially the same as for odd g. However, because $2g+2$ is no longer divisible by four, we must include two additional iterations of the basic moves used in the proof of Lemma 7. Also, because the different form of the monodromy of $X'_g(i)$ makes for a different ribbon branch surface, the final step of recognizing the total space of the cover as an elliptic surface is somewhat different.

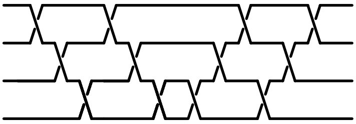


Figure 26: The ribbon surface E.

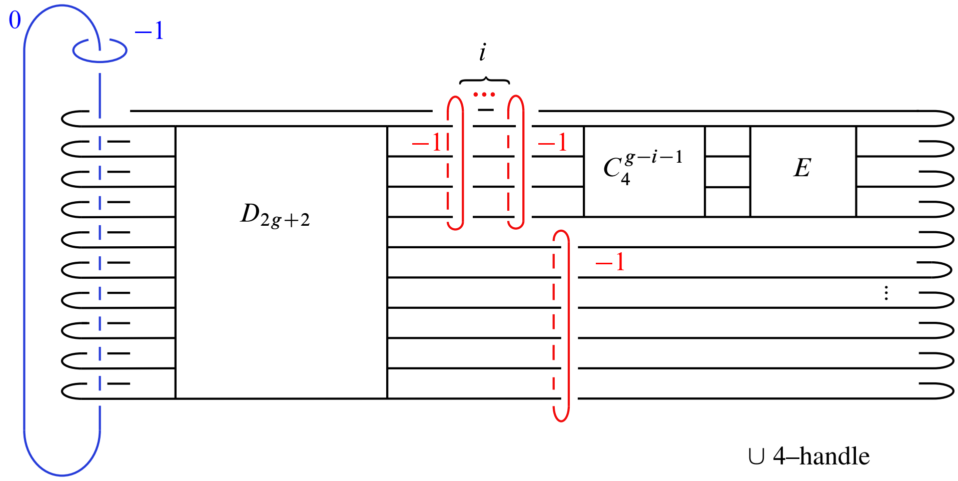


Figure 27

As a starting point for even g, we begin with the Lefschetz fibration on Z_g, which from (3) has a monodromy factorization given by the relation

$$D_g E_g (t_{c_1} t_{c_2} t_{c_3})^{4i} (t_{c_5} t_{c_6} \dots t_{c_{2g+1}})^{2g-2} (t_{c_1} t_{c_2} t_{c_3})^{4(g-i-1)} (t_{c_1} t_{c_2} t_{c_3})^2 (t_{c_3} t_{c_2} t_{c_1})^2 = 1. \tag{5}$$

As before, this hyperelliptic Lefschetz fibration can be described as the 2–fold cover of $\mathbb{F}_1$ branched over the surface described in Figures 25 and 26. Once again, it can be checked that the projection of (5) to a homeomorphism of $\Sigma_{0,2g+2}$ equals a right-handed Dehn twist about a circle that encloses all marked branch points. The unseen part of the branch surface is $2g+2$ disks attached to the boundary of the ribbon surface,

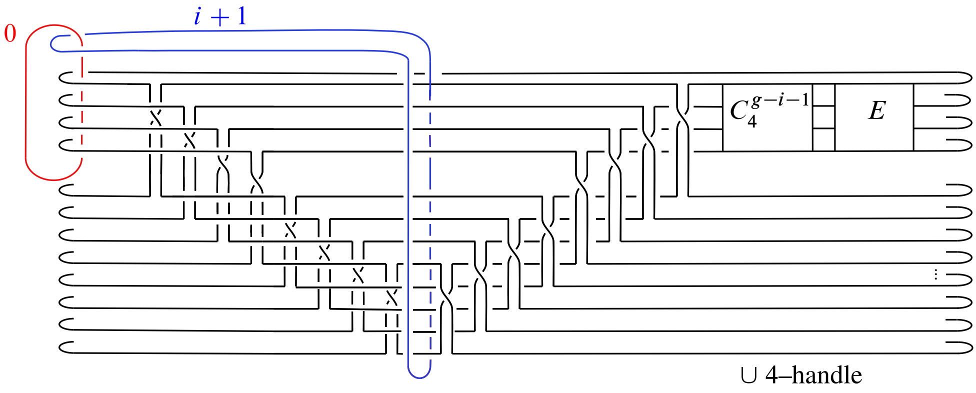


Figure 28

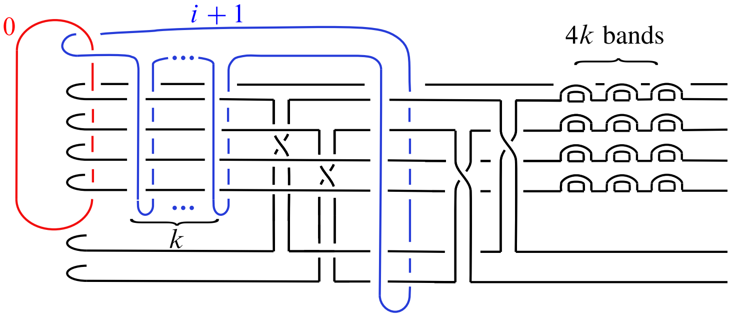


Figure 29

with interiors in the 4–handle, exactly as in the odd g case. Thus Figure 25 depicts a banded unlink diagram, as before.

Performing unchaining monodromy substitutions gives that $X_g(i)$ is the 2–fold cover of $\mathbb{F}_1 \# (i+1)\overline{\mathbb{CP}^2}$, branched over the surface in Figure 27. Mimicking the 2–handle slides from the odd case yields $2(i+1)$ sections which are blown down to give $X'_g(i)$ as the 2–fold cover of $\mathbb{F}_{i+1}$ branched over the surface in Figure 28.

Let $g = 2k+2$. The ribbon surface in Figure 28 is of the form

$$F(2g+2,0,0) = F(4k+6,0,0).$$

Then k iterations of Lemma 7 give that $X'_g(i)$ is diffeomorphic to the cover of $\mathbb{F}_{i+1}$ branched over $F(6,k,4k)$, shown in Figure 29.

We next cancel the bottom two horizontal disks as follows. A 2–handle band dive gives Figure 30. We can then twice more use the sequence of moves in the proof of Lemma 7: a band slide, followed by a band dive, followed by a 2–handle band slide. (See the transition from Figure 18 to Figure 21.) This adds two more trivial bands to

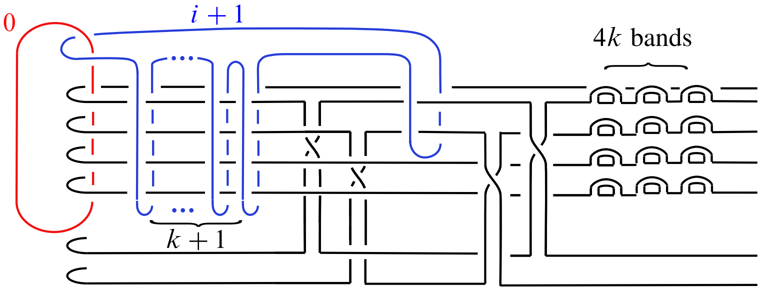


Figure 30

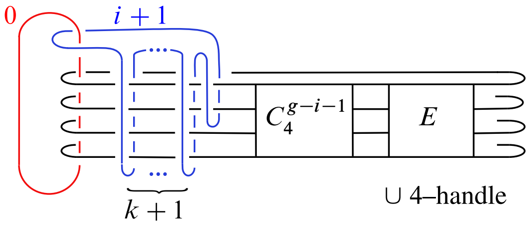


Figure 31

the picture, but we cancel all $4k+2=2g-2$ of them using disks from the 4–handle. This results in Figure 31.

We next slide the $(i+1)$–framed 2–handle $k+1$ times over the 0–framed handle. The new framing is $(i+1)-2(k+1)=i-2k-1=-(g-i-1)$. This is Figure 32.

It remains to see that the branched cover described by Figure 32, right, is $E(g-i)$. The lift of the branched cover of the 0–handle union the 0–framed 2–handle branched over the ribbon surface is a genus 1 Lefschetz fibration over D^2 with monodromy

$$(t_{c_1}t_{c_2}t_{c_3})^{4(g-i-1)}(t_{c_1}t_{c_2}t_{c_3})^2(t_{c_3}t_{c_2}t_{c_1})^2.$$

The location and framing of the other attaching circle is explained by tracking a framed neighborhood of a reference point $* \in \Sigma_{0,2g+2} \setminus B^2_{2g+2}$ through an isotopy from the braid

$$(d_{\pi(c_1)}d_{\pi(c_2)}d_{\pi(c_3)})^{4(g-i-1)}(d_{\pi(c_1)}d_{\pi(c_2)}d_{\pi(c_3)})^2(d_{\pi(c_3)}d_{\pi(c_2)}d_{\pi(c_1)})^2$$

to the identity. This isotopy first undoes $g-i-1$ right-handed Dehn twists, which fixes $*$ while rotating its neighborhood $g-i-1$ times oppositely, followed by an isotopy that pushes $*$ around a circle passing through the middle two marked points without twisting its neighborhood. Thus the branched cover of $\mathbb{F}_{g-i-1}$ extended over the rest of the surface gives a total space which is a genus 1 Lefschetz fibration over S^2. Finally, we note that the monodromy factorization of this fibration is easily seen to be equivalent to other well-known factorizations for elliptic fibrations on $E(g-i)$.

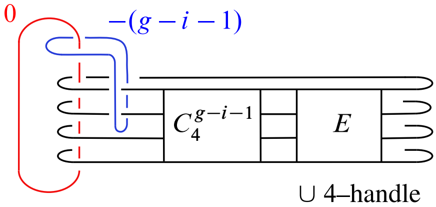


Figure 32

References

[1] **S Akbulut**, **R Kirby**, *Branched covers of surfaces in* 4*–manifolds*, Math. Ann. 252 (1979/80) 111–131 MR Zbl

[2] **S Bauer**, *Almost complex* 4*–manifolds with vanishing first Chern class*, J. Differential Geom. 79 (2008) 25–32 MR Zbl

[3] **R I Baykur**, **K Hayano**, *Multisections of Lefschetz fibrations and topology of symplectic* 4*–manifolds*, Geom. Topol. 20 (2016) 2335–2395 MR Zbl

[4] **R I Baykur**, **K Hayano**, **N Monden**, *Unchaining surgery and topology of symplectic* 4*–manifolds*, Math. Z. 303 (2023) art. id. 77 MR Zbl

[5] **J S Birman**, **H M Hilden**, *On isotopies of homeomorphisms of Riemann surfaces*, Ann. of Math. 97 (1973) 424–439 MR Zbl

[6] **S K Donaldson**, *Lefschetz pencils on symplectic manifolds*, J. Differential Geom. 53 (1999) 205–236 MR Zbl

[7] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR Zbl

[8] **S Friedl**, **S Vidussi**, *On the topology of symplectic Calabi–Yau* 4*–manifolds*, J. Topol. 6 (2013) 945–954 MR Zbl

[9] **T Fuller**, *Hyperelliptic Lefschetz fibrations and branched covering spaces*, Pacific J. Math. 196 (2000) 369–393 MR Zbl

[10] **R E Gompf**, **A I Stipsicz**, 4*–Manifolds and Kirby calculus*, Graduate Studies in Math. 20, Amer. Math. Soc., Providence, RI (1999) MR Zbl

[11] **M C Hughes**, **S Kim**, **M Miller**, *Isotopies of surfaces in* 4*–manifolds via banded unlink diagrams*, Geom. Topol. 24 (2020) 1519–1569 MR Zbl

[12] **A Kas**, *On the handlebody decomposition associated to a Lefschetz fibration*, Pacific J. Math. 89 (1980) 89–104 MR Zbl

[13] **T-J Li**, *Quaternionic bundles and Betti numbers of symplectic* 4*–manifolds with Kodaira dimension zero*, Int. Math. Res. Not. 2006 (2006) art. id. 37385 MR Zbl

[14] **T-J Li**, *Symplectic* 4*–manifolds with Kodaira dimension zero*, J. Differential Geom. 74 (2006) 321–352 MR Zbl

[15] **I Smith**, *Lefschetz pencils and divisors in moduli space*, Geom. Topol. 5 (2001) 579–608 MR Zbl

Department of Mathematics, California State University, Northridge
Northridge, CA, United States

terry.fuller@csun.edu

Received: 23 August 2021 Revised: 30 November 2021

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

Algebraic & Geometric Topology 23:6 (2023) 2895–2924
DOI: 10.2140/agt.2023.23.2895
Published: 7 September 2023

Bifiltrations and persistence paths for 2–Morse functions

RYAN BUDNEY
TOMASZ KACZYNSKI

We study the homotopy type of bifiltrations of compact manifolds induced as the preimage of filtrations of $\mathbb{R}^2$ for generic smooth functions $f\colon M \to \mathbb{R}^2$. The primary goal of the paper is to allow for a simple description of the multigraded persistent homology associated to such filtrations. Our main result is a description of the evolution of the bifiltration of f in terms of cellular attachments. Analogs of the Morse–Conley equation and Morse inequalities along so-called persistence paths are derived, and a scheme for computing pathwise barcodes is proposed.

57R35; 55M99, 55N31

1 Introduction

In the past two decades, the Morse theory of smooth functions on manifolds, and singularity theory, its extension to functions with multidimensional values, have driven a lot of attention in the applied mathematics and theoretical computer science communities due to their applications in imaging, visualisation and most recently, topological data analysis (TDA). While those theories have been extensively developed for nearly a century, new and potential applications bring different perspectives.

Morse theory is a tool that allows one to use real-valued functions on a manifold to give a combinatorial description of that manifold, in the language of handle decompositions or CW–complexes. A topological model for M is built following changes in sublevel sets $M_{g\le y} = g^{-1}((-\infty, y])$ of a Morse function (ie smooth and generic) $g\colon M \to \mathbb{R}$. The central theorem — see Milnor [22] — about the filtration of M by sublevel sets is that:

(1) The homotopy type of $M_{g\le y}$ does not change for $y \in [a, b]$ provided there are no critical values of g in the interval $[a, b]$.

(2) If there is precisely one critical value in (a,b) then the $M_{g\leq b}$ is obtained from $M_{g\leq a}$ by a handle attachment, which up to a homotopy equivalence, is a cell attachment.

In imaging and TDA, the interest shifts to the function itself. The domain of the image is typically a well-understood space such as $\mathbb{R}^n$ or a triangulated sphere S^n. That is a typical setting in works on the shape similarity by size function methods such as in Biasotti, Cerri, Frosini, Giorgi and Landi [4]. When it comes to the study of functions with multidimensional values, there are new challenges and more differences between the classical singularity theory setting and the applied context.

Given the success of Morse theory, the study of generic smooth mappings from manifolds to surfaces $f\colon M \to \Sigma$ is a natural next step. The most basic elements of the theory involves the description of the stratification of the manifold M by the singularity types, together with the local properties of the mapping around singular points. This was first worked out by Whitney [29] — see also Guillemin and Pollack [16] — when M is a surface, and fully generalized in the subsequent decades; see Saeki [26] and Wan [28]. Perhaps the main difference between the study of functions taking values in $\mathbb{R}$ vs in a surface is that the set of fibres $\{f^{-1}(a) \mid a \in \Sigma\}$ lack a linear order on them, so a poset relation has to be taken into account. In contrast, the real numbers have the relatively canonical poset $\{(\infty, a] \mid a \in \mathbb{R}\}$ of half-infinite intervals.

To be specific, let us state the posets studied in this paper. Consider $f\colon M \to \mathbb{R}^2$, where M is an m–manifold of dimension $m \geq 2$, and the plane $\mathbb{R}^2$ is endowed with the poset relation

$$(a,b) \preceq (a',b') \iff a \leq a' \text{ and } b \leq b'.$$

Any such function gives rise to a *bifiltration* of M, which is defined as the family $M_f = \{M_{(a,b)}\}_{(a,b)\in\mathbb{R}^2}$ of subsets of M given by

$$M_{(a,b)} = \{p \in M \mid f(p) \preceq (a,b)\}.$$

Equivalently, the sets $M_{(a,b)}$ are the preimages of the quadrants

$$C_{(a,b)} = (-\infty, a] \times (-\infty, b]$$

under f. They are nested with respect to inclusions; that is, $M_{(a,b)} \subseteq M_{(a',b')}$ for every $(a,b) \preceq (a',b')$.

Persistence consists of analyzing homological changes occurring along the bifiltration as the point (a,b) varies. Note that the boundary $\partial C_{(a,b)}$ of the quadrant $C_{(a,b)}$ is not a submanifold of $\mathbb{R}^2$: it can be viewed as a manifold with a corner. The problem of bifiltration has been addressed in the presented setting by Smale in 1975 [27] and further investigated by Wan [28]. As it is pointed out by Smale, the study is historically motivated by the *Pareto optimal problem* of simultaneously maximizing several functions. Our work is an extension of the work done in [27; 28], with the same topic viewed from a different perspective.

There has been progress in computing persistent homology for multifiltrations which include functions $g\colon M \to \mathbb{R}^k$ as a special case for any $1 < k < \dim M$. We refer the reader to Carlsson and Zomorodian [8] and Cavazza, Cerri, Di Fabio, Ethier, Ferri, Frosini, Kaczynski and Landi [9; 10]. However, most of the dimension-independent work on computing persistent homology, often in a discrete setting, is "geometry blind" in the sense that it does not give much insight to the particular types of singularities one may encounter. Providing that insight is the main motivation for this paper. In particular, in Allili, Kaczynski, Landi and Masoni [1], a Forman-type discrete analogy of multidimensional Morse functions is investigated. In the conclusion of that paper, it is pointed that an appropriate application-driven extension of the Morse theory to multifiltrations for smooth functions is not much investigated yet, and it would help in understanding the discrete analogy. The present work is a step in that direction. A study of discrete Forman type multidimensional Morse functions is currently under way by Landi and Scaramuccia, for instance, in [18]. A study of smooth multifiltrations on manifolds with similar geometric motivation as ours and complementary goals is currently under way by Bubenik and Catanzaro [6] and Assif and Baryshnikov [2].

We begin Section 2 with the definition of a 2–Morse function $f\colon M \to \mathbb{R}^2$, following Gay and Kirby [14] and Wan [28]. This allows us to define the (oriented) signature invariant; see Definition 2.2. We follow this definition with a few simple examples where one can explicitly compute the homotopy types of the filtration $M_{(a,b)}$ for all $(a,b) \in \mathbb{R}^2$. The main result of the paper is a characterisation of the *singular points* of the bifiltration. In short, these are the locations where the homotopy type of the bifiltration changes; see Definition 2.1.

Theorem 1.1 *If $f\colon M \to \mathbb{R}^2$ is a 2–Morse function then the bifiltration M_f has singular points consisting entirely of corner and tail singular points.*

In short, this theorem gives us a stratification of the plane $\mathbb{R}^2$ such that the homotopy type of $M_{(a,b)}$ is constant in the connected codimension zero strata. We follow that up by a description of how the homotopy type of $M_{(a,b)}$ changes as (a,b) crosses a codimension one stratum.

Our Lemma 2.8 is the analogue of (1), in that it tells us that generically the homotopy type of the manifolds $M_{(a,b)} = f^{-1}(C_{(a,b)})$ is locally constant. The nature of the proof of Lemma 2.8 is significant to the rest of the paper, describing a rather flexible technique of vector field flows, allowing us to construct conjugate isotopies (ie fibre-preserving isotopies, also known as isotopies that are horizontal diffeomorphisms with respect to the map f) in both M and the plane $\mathbb{R}^2$. This allows us the freedom to frame our remaining arguments in the language of how the homotopy type of $f^{-1}(C_t)$ changes when C_t is an arbitrary "smoothly varying" 2–manifold in the plane. There are however some points in the plane where the homotopy type of $M_{(a,b)}$ does change; this is described in Theorem 2.9. The main feature of Theorem 2.9 is that the homotopy type of $M_{(a,b)}$ changes via handle (or cell) attachments. In the proof we see one of the handle attachments comes directly from a classical Morse theory argument. The second type of handle attachment uses global features of the singular point set of f, and is perhaps best thought of as a Bott-style handle attachment. We give a brief account of Bott's variant of Morse theory. Proposition 2.10 summarizes elements of the proof of Theorem 2.9, describing the dimension of the cell attachments in terms of the oriented signature invariant. One last feature of Section 2 is the observation that at "cubic" points $(a,b) \in \mathbb{R}^2$, while the homotopy type of $M_{(a,b)}$ generally does not change, the fibrewise homotopy type (with respect to the map f) does change. Roughly speaking, these cubic points correspond to pairs of cancelling handles (or cells).

In Section 3, we turn our attention to bifiltered persistent homology. We briefly review descriptive techniques, and relations to one-dimensional persistence. In particular, the foliation method that has been introduced in [4] for size functions, used by Cagliari, Di Fabio and Ferri [7] and [10] for multidimensional persistent homology and later named as fibred barcode in the context of persistence modules by [10] and Lesnick and Wright [20]. As it is visible in examples of Section 2, in the presence of the poset relation, there are multiple ways of building the topology of M by crossing different arcs of the critical value set while respecting the poset relation. That leads us to Definition 3.1 of persistence paths. It is a new concept which is somewhat analogous to the mentioned foliation method of [10]. It also can be viewed as an analogy of a flow induced by the generalized gradient in [28], in that our persistence paths apply to functions f which may have cycles in the sense of [28, Definition 6.4 and 6.5]. We prove an analogy of the Morse–Conley equation — see Rybakowski and Zehnder [25] —

in Theorem 3.2, and derive from it Corollary 3.4 on strong Morse inequalities for persistence paths. This gives us a flexible family of Morse inequalities associated to f, extending the work of Wan [28]. We conclude Section 3 by introducing pathwise barcodes in Definition 3.5 and describing a scheme for computing the barcodes based on a small representable subfamily $\mathrm{Rep}(f)$ of all persistence paths. While Carlsson and Zomorodian [8] outline an argument that there is no complete and discrete invariant of multigraded persistent homology, the primary result of this paper strikes a more optimistic note in the case of multifiltrations induced by smooth functions, implying that our filtrations are tame; see Corollary 2.11.

In Section 4 we discuss some possible future research directions.

Acknowledgements The authors would like to thank Hyam Rubinstein, Marian Mrozek, and Michael Lesnick for helpful suggestions.

Research of Kaczynski was supported by a Discovery Grant from NSERC of Canada.

2 Fold and cubical singularities

In the classical Morse theory of smooth real-valued functions and, respectively, singularity theory of functions with values in a 2–dimensional manifold Σ, a critical point or singular point is a point $p \in M$ at which $Df(p)$ is not of maximal rank. The corresponding point $c = f(p)$ in the target space is called a critical, respectively, singular value of f. The terminology found in the literature is not consistent: sometimes the terms critical and singular are interchanged.

In computational topology, we deal with nonsmoothness and degeneracy, so a topological definition is more appropriate. It is also helpful in describing handle attachments. In addition, in the presence of the poset structure of bifiltrations, as we shall see soon, there is a substantial difference between singularity in the differential sense and criticality in the topological sense. We shall adopt the following definition.

Definition 2.1 A *homotopy regular value* of f *with respect to the bifiltration* of M is a point $(a,b) \in \mathbb{R}^2$ such that, in some neighbourhood $U_{(a,b)}$ of (a,b), for all $(a',b'),(a'',b'') \in U_{(a,b)}$ with

$$(a',b') \preceq (a'',b''),$$

the inclusion $M_{(a',b')} \hookrightarrow M_{(a'',b'')}$ is a homotopy equivalence. If this condition fails, (a,b) is called a *homotopy critical value*.

A weakening of this definition suited to persistent homology is the notion of *homological regular* and *critical values* defined by replacing homotopy equivalence by isomorphism induced in homology. This coincides with the definition of given in [9, Definition 3.4].

When $f\colon M \to \mathbb{R}$ is a Morse function, the sets of critical points and values in the differential and topological sense coincide. For $\mathbb{R}^2$–valued functions, even the generic ones, they are substantially different. We shall adopt the terms of *singular* points and values for those given by differential definition and *critical* to those given by Definition 2.1. Given $f\colon M \to \mathbb{R}^2$. we consider the sets

$$\begin{aligned} &\mathit{Singp} = \{p \in M \mid \operatorname{rank} Df(p) < 2\}, && \mathit{Singv} = f(\mathit{Singp}), \\ &\mathit{Critv} = \{(a,b) \in \mathbb{R}^2 \mid (a,b) \text{ is homotopy critical}\}, && \mathit{Critp} = f^{-1}(\mathit{Critv}). \end{aligned}$$

We shall soon see that the arcs of *Singv* along which both coordinates (a,b) increase are homotopy regular, so they are not subsets of *Critv*. The topologically significant arcs are those whose normal vectors have both coordinates of the same sign. Conversely, *Critv* contains horizontal or vertical half-lines passing through the vertex of $C_{(a,b)}$ and "kissing" points on the singularity *Singv* but not contained in it. In Proposition 2.10 we give a classification of different types of criticality.

As we just noticed, Definition 2.1 also applies to points $(a,b) \in \mathbb{R}^2$ which are not necessarily the values of f, that is, are not in the image $f(M)$. For that reason we will refer to them as points rather than values and whether we speak about points in M or in $\mathbb{R}^2$ should be made clear from the context.

Following [11], the set *Critv* will be referred as to the *extended Pareto grid.* We begin with a definition from Gay and Kirby [15; 14], and earlier Wan [28].

Definition 2.2 A 2–*Morse function* (also called Morse 2–function) is a smooth function

$$f\colon M \to \Sigma,$$

where M is an m–manifold, $m \geq 2$ and Σ is a 2–manifold satisfying a local condition. For any point $p \in M$ there are neighbourhoods $U_p \subset M$ of p in M, $V_{f(p)} \subset \Sigma$ of $f(p)$ in Σ, U_0' of 0 in $\mathbb{R}^m$, and V_0' of 0 in $\mathbb{R}^2$ together with diffeomorphisms $\phi\colon U_p \to U_0' \subset \mathbb{R}^m$ and $\psi\colon V_{f(p)} \to V_0'$ with $\phi(p) = \psi(f(p)) = 0$ making the diagram

$$\begin{tikzcd} U_p \arrow[r, "f|_{U_p}"] \arrow[d, "\phi"] & V_{f(p)} \arrow[d, "\psi"] \\ U_0' \arrow[r] & V_0' \end{tikzcd}$$

commute, where the bottom horizontal arrow must be one of the following three:

- $(x_1, x_2, \cdots, x_m) \mapsto (x_1, x_2)$; for this, p is a *regular point.*
- $(x_1, x_2, \cdots, x_m) \mapsto (x_1, \pm x_2^2 + \cdots + \pm x_m^2)$; for this, p is a *fold point.*
- $(x_1, x_2, \cdots, x_m) \mapsto (x_1, x_2^3 + x_1 x_2 + \pm x_3^2 + \cdots + \pm x_m^2)$; for this p is a *cubic point.*

Just like with Morse functions, there are elementary transversality conditions equivalent to Definition 2.2 [28, Section 1]. This allows the conclusion that, for any smooth function $f\colon M \to \Sigma$ where Σ is a 2–manifold, via a small perturbation of f we may convert f into a 2–Morse function, ie 2–Morse functions form an open and dense subset of the space of smooth functions $M \to \Sigma$.

The curves of the fold singularities come equipped with transverse-oriented indices. This is analogous to the index of a critical point of a Morse function, but made slightly more complex by the codomain of our function being $\mathbb{R}^2$.

The index has the form of a triple (v, i, j) where v is a vector transverse to the singular value set, and i is the dimension of the eigenspace that is folded into the v direction, while j is the dimension of the eigenspace that is folded into the $-v$ direction. Thus $i + j = m - 1$. Due to this convention we need the equivalence relation $(v, i, j) \sim (-v, j, i)$. Further notice that due to the nature of the cubic singularity there are two fold-type singularities that merge, with one fold being of type (v, i, j) and the other fold being of the type $(v, i+1, j-1)$. In our diagrams we will typically draw the v vectors, and only plot the pair (i, j). In general i is an integer in the set $\{0, 1, 2, \ldots, m-1\}$; see Figure 1. Our oriented index makes sense only on the fold points. We give a more precise definition in the next paragraph.

An elementary observation that may help the reader acclimatize to 2–Morse functions is the observation that if $S \subset \Sigma$ is a smoothly embedded copy of $\mathbb{R}$, with $f\colon M \to \Sigma$ a 2–Morse function, then provided S is transverse to the critical values of f, ie disjoint from the cubic points and without "kissing" tangencies to the fold points, then the restriction of f as a map $f^{-1}(\Sigma) \to \Sigma \simeq \mathbb{R}$ is a Morse function. We use this in Section 3 to define persistence paths. It is also used in the proof of Theorem 2.9.

Definition 2.3 Given a Morse 2–function $f\colon M \to \mathbb{R}^2$, and a point $p \in M$ in the fold singular points,

$$Hf_p\colon T_pM \otimes T_pM \to \mathbb{R}^2$$

is a bilinear function taking values in a 1–dimensional subspace of $\mathbb{R}^2$, complementary to the image of Df_p. Choosing $v \in \mathbb{R}^2$ spanning this subspace, we can treat Hf_p

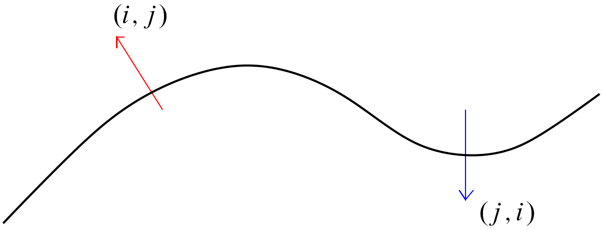


Figure 1: Depiction of the symmetry of the index of fold points.

as real-valued bilinear function, ie by considering $Hf_p \cdot v\colon T_pM \otimes T_pM \to \mathbb{R}$. As this is a symmetric bilinear function, Sylvester's law of inertia gives us a well-defined signature invariant, (i,j), that can be thought of as the dimensions of the maximal subspaces where the form is positive or negative definite, respectively.

Notice that at a cubic singular point the Hessian is degenerate, ie $i+j=m-2<m-1$, with the nullspace together with the image of Df_p spanning the cusp's plane of curvature.

Before we begin the examples, it is important to be aware that a Morse function $f\colon M \to \mathbb{R}$ gives rise to a cell decomposition of M [22]. These cell decompositions are computable in terms of flow lines of vector fields conditioned by the derivative of f. The cellular descriptions of M in their most natural state are homotopy-theoretic in nature, ie these techniques give homotopy equivalences between M and CW–complexes, not homeomorphisms. That said, CW–complexes are far from ideal tools to describe manifolds. The adaptation of CW–complexes to smooth manifolds are called *handle decompositions*, developed by Smale in his proof of the h–cobordism theorem. A k–cell for an m–manifold M is a map $D^k \to M$ that satisfies various properties, such as being an embedding on the interior. A k–handle for an m–manifold is a smooth embedding $D^k \times D^{m-k} \to M$, ie handles are not only fully embedded, but they contain the data of both the cell and a tubular neighbourhood of the cell. This allows handle decompositions to not just describe the homotopy type of M, but also its smooth structure. A subtlety of handle attachments is that a k–handle is attached only on part of its boundary, ie $(\partial D^k) \times D^{m-k}$, thus there is a risk that we are entering the class of manifolds with cubical corners. The exposition of Kosinski [17] gives careful consideration to this problem, keeping the constructions purely in the language of manifolds with boundary. A Morse function $f\colon M \to \mathbb{R}$ gives a handle decomposition of M; moreover this handle decomposition describes the smooth structure on M.

Starting from illustrative examples, we investigate the relation between bifiltration and the classical singularity theory.

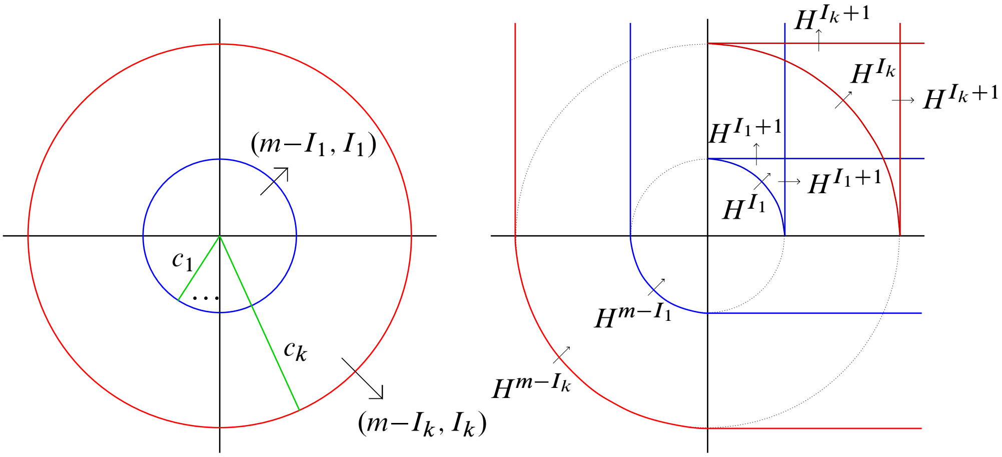


Figure 2: Singular values for f with oriented index in Example 2.4 (left) and extended Pareto grid (right). The dimension of the manifold $S^1 \times M$ is $m+1$. Only the first and last values, c_1 and c_k, are displayed. The notation H^k indicates k–handle attachments.

Example 2.4 If $g\colon M \to \mathbb{R}$ is a Morse function, then

$$f\colon S^1 \times M \to \mathbb{R}^2$$

given by $f(z,p) = z \cdot g(p)$ is a 2–Morse function on the $(m+1)$–dimensional manifold $S^1 \times M$ with only fold singularity types.

If the singular values of g consist of positive real numbers $0 < c_1 < c_2 < \cdots < c_k$ then the singular values of f consist of the circles of radius $c_1, c_2, \dots, c_k$ centred at the origin.

If the singular value c_i (of g) has index I_i, then the circle at f of radius c_i is also a fold-type singular value set of index $(\hat{r}, m-I_i, I_i)$, where $\hat{r}$ is the unit outward-pointing radial vector.

The persistence diagram for the preimages $M_{(a,b)}$ is a union of the *descending part* of the singular values of f together with some vertical and horizontal lines at the endpoints.

In Figure 2, the diagram on the left depicts the singular values of the function f. These are the circles of rotation of the singular values of g. Say the red circle corresponds to a singular point of index I_i. An alternative way of saying this is that the homotopy type of the space $g^{-1}((-\infty, t])$ as t transitions through the point c_i changes by an I_i–cell attachment.

In the figure on the right, we describe how the preimages $M_{(a,b)}$ change as the points $(a,b) \in \mathbb{R}^2$ vary. The I_i–handle attachments are labelled by H^{I_i}. Only a portion of

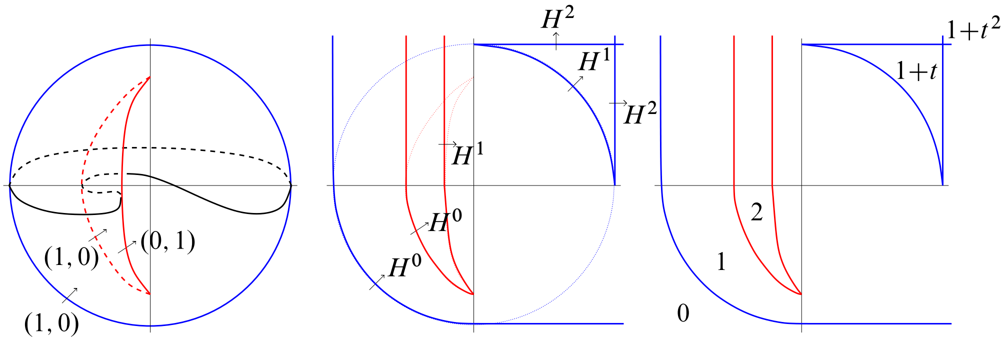


Figure 3: Cupped sphere projection.

the circle from the left diagram remains in the right, since at those (dotted) points the homotopy type of the filtration does not change.

Let us take the blue circle for example, on the left. This is the singular value c_1 of index I_1. On the right, this singular circle gives us two singular arcs. The lower blue arc is properly embedded in $\mathbb{R}^2$, and transitioning through it corresponds to a $m - I_1$ handle attachment. This should be thought of as a dual handle to c_1 (of g). The other singular value is a "fishtail", divided into three properly embedded arcs. The round arc corresponds to a handle attachment of index I_1, while the two straight lines correspond to handle attachments of index $I_1 + 1$. The handles of index $I_1 + 1$ should be thought of as cancelling handles to the index I_1 handle. Thus attachment of all three handles of index I_1, $I_1 + 1$ and $I_1 + 1$ has the same effect on the homotopy type as a single attachment of a handle of index $I_1 + 1$.

Example 2.5 Given a round sphere $S^2 \subset \mathbb{R}^3$, the orthogonal projection map $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ when restricted to S^2 has singular values the unit circle, corresponding to an equatorial circle in S^2 of singular points. Imagine the sphere being made of rubber. We grab a small section of the sphere (away from the equator) and fold it over itself, creating a cupped sphere. This introduces an eye singularity in the projection map, as depicted in Figure 3.

The pure fold singularity, the equator, is in blue. The red singularity is an "eye" type singularity, with precisely two cubic (cusp) points. This is depicted on the left. In the central figure we describe the handle attachments of the bifiltration. In the figure on the right we describe the Poincaré polynomials of the bifiltration, ie the bifiltration is regular at the white points, with transitions only at the red and blue points.

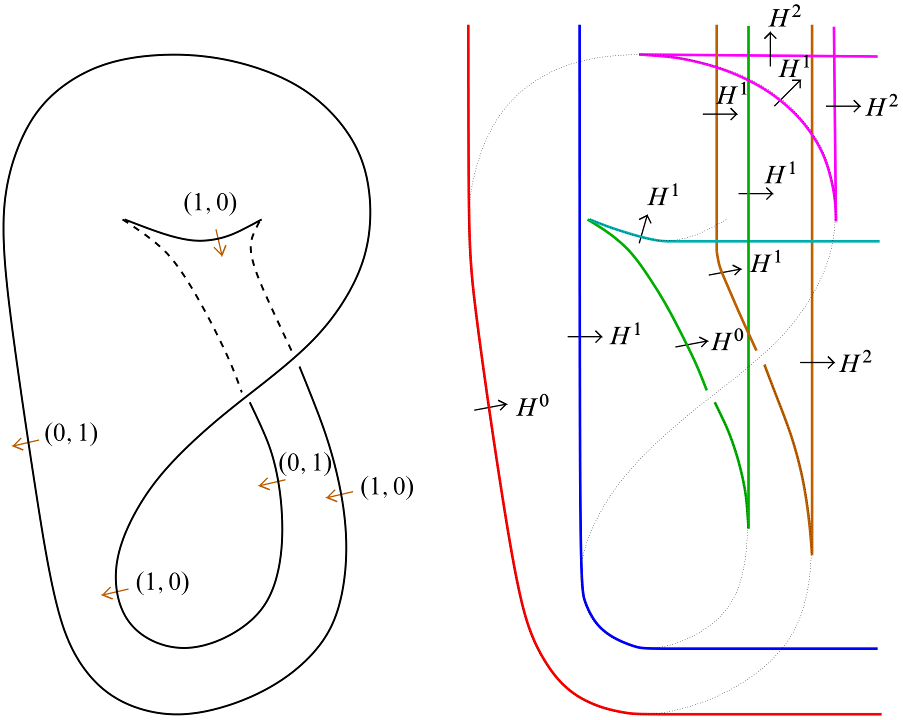


Figure 4: Klein bottle projection.

We give a fairly general example with cubic singularities.

Example 2.6 Cerf theory tells us that a 1–parameter family of real-valued functions on a manifold is not (generically) Morse at all parameter times. There will be finitely many times where the Morse singularities devolve into cubic singularities. Thus take a generic 1–parameter family of functions on M, $F\colon S^1 \to C(M,\mathbb{R})$, and form the function

$$f\colon S^1 \times M \to \mathbb{R}^2$$

given by $f(v,p) = F(v)(p)\cdot v$. The function f is 2–Morse. The bifiltration $M_{(a,b)}$ will be described in Theorem 2.9.

Notice Example 2.6 is a direct generalization of Example 2.4; ie Example 2.4 can be derived by setting F to be the constant function.

Example 2.7 A rather colourful example comes from orthogonal projection $\mathbb{R}^4 \to \mathbb{R}^2$ precomposed with one of the standard embeddings of the Klein bottle $K^2 \to \mathbb{R}^4$.

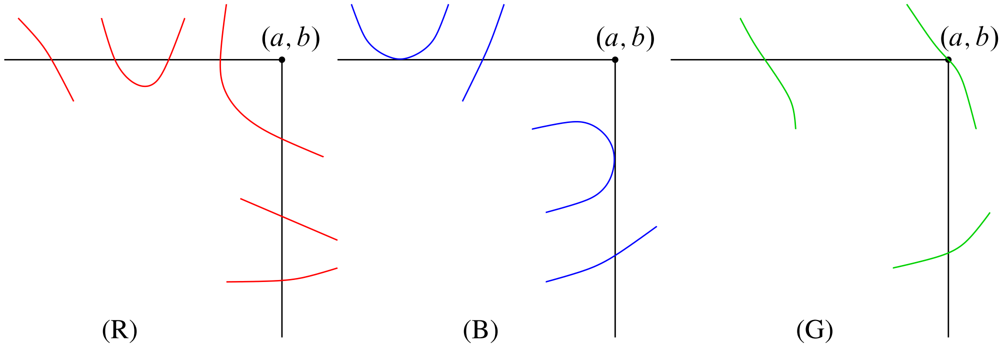


Figure 5: Three intersection types with dimension 1 stratum of the singular values. Type (R) consists of transverse intersections with the 0– and 1–dimensional strata of the singular value set for f. Type (B) consists of both transverse intersections and simple "kissing" nontransverse intersections with the 1–strata of the singular value set. Type (G) consists of transverse intersections together with a corner-type intersection. Thus (R) is generic, ie codimension 0 in the filtration, while types (B) and (G) are of higher codimension.

This example appears in [28]. The singularity theory for mappings of 2–manifolds into the plane, of which this is a good demonstration, was originally discovered by Whitney [29].

We have seen in the previous examples that the singular points of the filtration consist of a subset of the singular points of the mapping f, together with some regular points of the original mapping — these were a collection of vertical and horizontal rays. We divide singular points of the filtration into two classes, *corner singular points* and *tail singular points*:

- We say a point $(a,b) \in \mathbb{R}^2$ is a corner singular point if for all suitably small neighbourhoods U of (a,b) in $\mathbb{R}^2$ there are points $(a',b') \in U$ such that $U \cap C_{(a',b')}$ intersects the singular set of f in both the horizontal and vertical boundary edges of $C_{(a',b')}$. If we write the coordinates of $\mathbb{R}^2$ as (x,y) then this happens when locally writing the singular values of f as the graph of a function $y(x)$, then the function $y(x)$ would be decreasing at $x = a$. A corner singular point is demonstrated in Figure 5(G).

- For (a,b) to be a tail singular point, we require that $C_{(a,b)}$ intersects the singular values of f tangentially, on either the interior of the horizontal or vertical boundary curve. In a neighbourhood of the tangential intersection we require the singular set to be on one side of the cube, ie either contained in the cube or in the closure of its exterior.

Thus it is a “kissing” tangency. The two tail singular point types are demonstrated in Figure 5(B).

Notice that Figure 5(R) describes a regular point (a,b) of the filtration M_f. We should note that while it is true a point can be both corner singular and tail singular at the same time, this is a codimension two condition, thus it is relatively rare. On the other hand, tail and corner singular points are codimension one conditions.

The proof of Theorem 2.9 has several special cases, but there is one elemental argument that is common to all cases. We put this in the next lemma.

Lemma 2.8 *If $f\colon M \to \mathbb{R}^2$ is a 2–Morse function, provided the point $(a,b) \in \mathbb{R}^2$ is regular for f, and the boundary of $C_{(a,b)}$ intersects the singular values of f transversely without double-points then (a,b) is not only a regular point for the filtration M_f, but the filtration is* ***locally trivial*** *near (a,b).*

Proof Precisely, there is a neighbourhood U of $(a,b) \in \mathbb{R}^2$ such that for any $(a',b') \in U$ there is a diffeomorphism $\tilde{\phi}\colon M \to M$ covering a diffeomorphism $\phi\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $\tilde{\phi}(f^{-1}(C_{(a,b)})) = f^{-1}(C_{(a',b')})$ and $\phi(C_{(a,b)}) = C_{(a',b')}$. When we say $\tilde{\phi}$ “covers” ϕ we mean the following diagram commutes:

$$\begin{tikzcd}
M \arrow[r, "\tilde{\phi}"] \arrow[d, "f"] & M \arrow[d, "f"] \\
\mathbb{R}^2 \arrow[r, "\phi"] & \mathbb{R}^2
\end{tikzcd}$$

The map $\tilde{\phi}$ is sometimes called a fibre-preserving diffeomorphism of f, or a *horizontal* diffeomorphism. A consequence of our proof will be that $\tilde{\phi}$ and ϕ are close to the identity diffeomorphism, where “close” is controlled by the size of the neighbourhood U. That such a neighbourhood exists can be deduced from the transversality stability theorem [16].

The idea of the proof is to find a vector field in the plane whose flow maps $C_{(a,b)}$ to $C_{(a',b')}$ provided (a',b') is near enough to (a,b). We construct the vector field in a manner that allows us to lift it to a vector field on M; thus the flow of this vector field will send $f^{-1}(C_{(a,b)})$ to $f^{-1}(C_{(a',b')})$. Given that the derivative of f is not an epimorphism at singular points of f, we have to take some care defining the vector field. At fold points of f the derivative of f is only onto the tangent space of the singular value set. Thus our neighbourhood U will be constrained by the sole demand

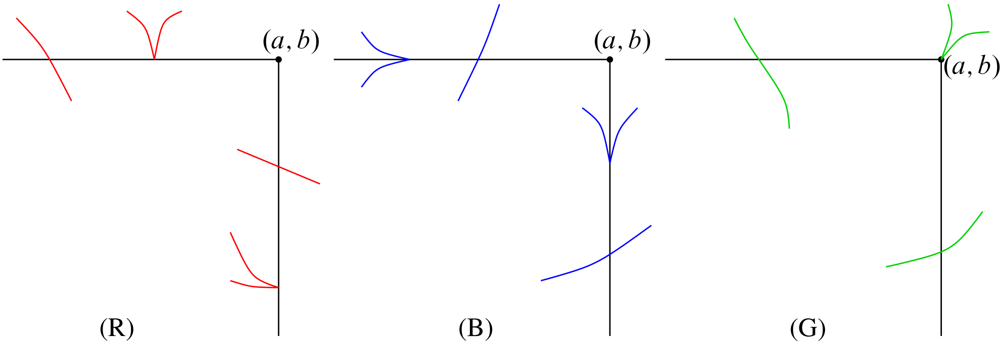


Figure 6: Three intersection types with dimension 0 stratum of the singular values. Type (R) consists of generic outward and inward intersections with the 0–dimensional stratum of the singular value set for f. Type (B) consists of nongeneric intersections with the 0–dimensional stratum. Type (G) consists of a generic corner-type intersection with the 0–dimensional stratum. In the filtration parameters, (G) is of codimension two. Type (B) is of codimension one if it exists, but for a generic 2–Morse function these singularity types are avoidable; one can convert them into type (R) by applying a small isotopy to f.

that the singular value set needs to be transverse to $\partial C_{(a',b')}$ for all $(a',b') \in U$. An example illustration of a valid U is depicted in the green region illustrated is the set of points $UC = \{p \in \partial C_{(a',b')} \mid (a',b') \in U\}$.

For the sake of argument, let's assume $b' = b$, ie we break the proof into two steps, step 1 with $b' = b$ and step 2 with $a' = a$. We further assume $a' > a$ as the $a' < a$ case is analogous. Let $\mathit{Singv}(f)$ denote the singular values of f, ie $\mathit{Singv}(f) \subset \mathbb{R}^2$. Consider the curves of $\mathit{Singv}(f) \cap UC$. On the path-components of $\mathit{Singv}(f) \cap UC$ that live in the vertical portion of UC, we define the vector field to be the unique vector field that is tangent to $\mathit{Singv}(f)$, and whose x–component is, in particular, positive. On the path components of $\mathit{Singv}(f) \cap UC$ that are in the horizontal portion of UC we define the vector field to be zero. In the horizontal portion of UC we extend the vector field to be zero. In the vertical portion of UC we interpolate between the definition on $\mathit{Singv}(f) \cap UC$ and the vector field $(1,0)$, using a tubular neighbourhood of $\mathit{Singv}(f)$ in UC. Doing this we can ensure the vector field in the vertical portion of UC always has unit x–component. We extend the vector field to all of $\mathbb{R}^2$, choosing any extension that keeps the length of the vector field bounded, ie so that its flow is complete. This gives us a flow on $\mathbb{R}^2$ that sends $C_{(a,b)}$ to $C_{(a',b)}$. Our vector field

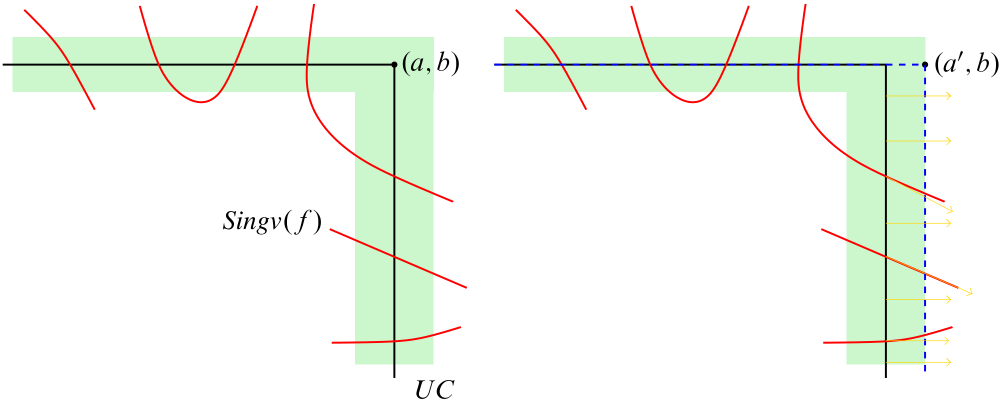


Figure 7: Neighbourhood of $\partial C_{(a,b)}$ and the tangency types for *Singv*.

lifts to M since the derivative of f is onto the tangent spaces of *Singv*(f), and for regular points, the derivative has rank two. By the existence and uniqueness theorem for solutions to ODEs, the flow of an f–lifted vector field is conjugated (by f) to the flow of the original vector field on $\mathbb{R}^2$. Thus the flow on M is fibre-preserving and sends $f^{-1}(C_{(a,b)})$ to $f^{-1}(C_{(a',b)})$. □

Lemma 2.8 has several natural generalizations. For example, let C and C' be compact 2–dimensional submanifolds of $\mathbb{R}^2$. Then provided there is an isotopy between C and C' such that ∂C is transverse to the singular value set of f through the entire isotopy (technically one needs to include intersections pairs of *Singv*(f) curves as 0–strata in *Singv*(f) for this statement to be true), then $f^{-1}(C)$ and $f^{-1}(C')$ are fibrewise diffeomorphic. The condition that the isotopy is transverse to the singular value set through all parameter times guarantees that the boundary of C does not pass over a cubic point, or ever become tangent to the singular value set. These are the events that can trigger changes in the fibrewise homotopy type of the preimage.

Similarly, provided (a,b) is a regular value of f we can *round the corner*, turning $C_{(a,b)}$ into a smooth manifold $C'_{(a,b)}$ such that $f^{-1}(C_{(a,b)})$ and $f^{-1}(C'_{(a,b)})$ are homotopy equivalent.

We choose to let C be a compact submanifold of $\mathbb{R}^2$ for the following arguments; ie rather than working with quadrants $C_{(a,b)}$ we choose to work with compact smooth manifolds, as it exposes the essential features of the argument.

The next theorem states that if the boundary of C (or quadrant $C_{(a,b)}$) passes over a cubic point in the isotopy, the fibrewise homotopy type changes but the homotopy type

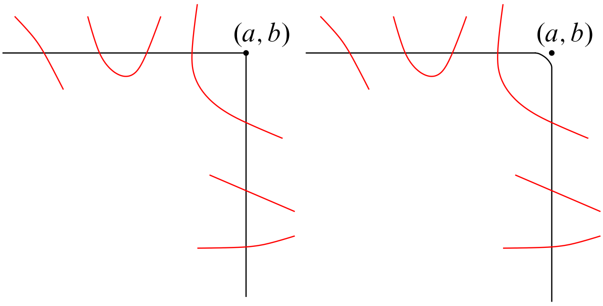


Figure 8: Rounding $C_{(a,b)}$ to produce $C'_{(a,b)}$.

does not. Moreover, if the boundary of C passes across the singular value set — at a tangency or corner, ie Figure 5(B) and (G) — then the homotopy type changes via a cell attachment. We also give enough details that allow the computation of the attaching maps.

Theorem 2.9 *If $f\colon M \to \mathbb{R}^2$ is a 2–Morse function then the bifiltration M_f has singular points consisting entirely of corner and tail singular points. Further, provided the two height functions $\pi_i\colon \mathbb{R}^2 \to \mathbb{R}$ given by $\pi_1(x,y) = x$ and $\pi_2(x,y) = y$ restrict to Morse function on the fold singular values of f, with distinct critical heights, then the transitions to the homotopy type of $M_{(a,b)}$ when (a,b) is either a corner or tail singular point are given by individual cell attachments.*

Proof Rather than using the restrictive language of quadrants, let C be a compact submanifold of $\mathbb{R}^2$ and we investigate the change in homotopy type of $f^{-1}(C)$ through an isotopy of C. We have two cases to consider.

Case 1 is a regular tangency — analogous to a type-2 Reidemeister move of the planar diagram, in that it creates two points of intersection between the boundary of C and the singular value set. Roughly speaking, there are two types of regular tangency moves. This move can be described via a "bigon modification" where one appends a bigon to the manifold C, attaching along one of the edges. The second edge of the bigon belongs to $Singv(f)$. In the "nonengulfing" move, $Singv(f)$ points out of C after the bigon is appended, while in the engulfing version, $Singv(f)$ points into C as one departs the bigon.

In the nonengulfing version of case 1, the move corresponds to a cell attachment of index i provided the index of $Singv(f)$ is of the form (v,i,j) where v is in the direction of ∂C as it sweeps over $Singv(f)$.

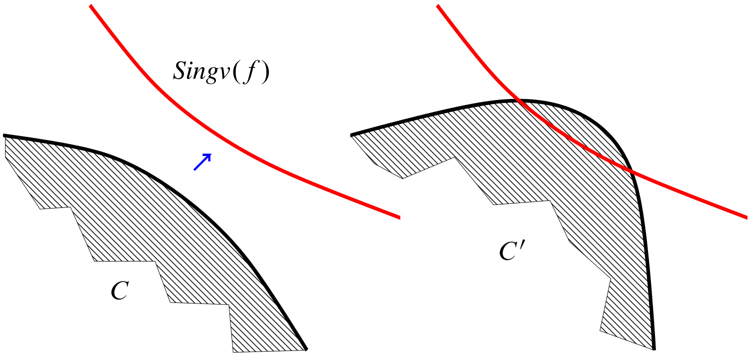


Figure 9: Case 1, nonengulfing.

Let C' denote the submanifold of $\mathbb{R}^2$ after the isotopy of C has been applied, ie as in the right hand side of Figure 8. Using an argument analogous to Lemma 2.8 we see that $f^{-1}(C')$ has the same homotopy type as $f^{-1}(C) \cup f^{-1}(B)$, where B is the blue arc in Figure 10.

The restriction of f to $f^{-1}(B)$, and after identifying B with an interval in $\mathbb{R}$, is a 1–Morse function; thus $f^{-1}(B)$ has the homotopy type of $f^{-1}(B \cap C)$ attach an i–cell, by the Morse lemma. More specifically, this is proven in [22, Theorem 3.2].

For the engulfing version of case 1, the cell attachment is of index $i+1$ provided the index of $Singv(f)$ is (v, i, j), and the attaching map is analogous to the previous case, but it should be thought of as an unbased version of a Whitehead product of the attaching map in the previous case, with the red interval disjoint from C in the diagram in Figure 11. Specifically, the characteristic map will be a product $D^i \times I$, where the D^i maps transversely to the red interval, and I can be identified with the red interval.

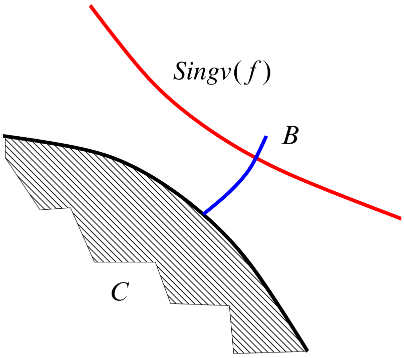


Figure 10: Case 1, nonengulfing.

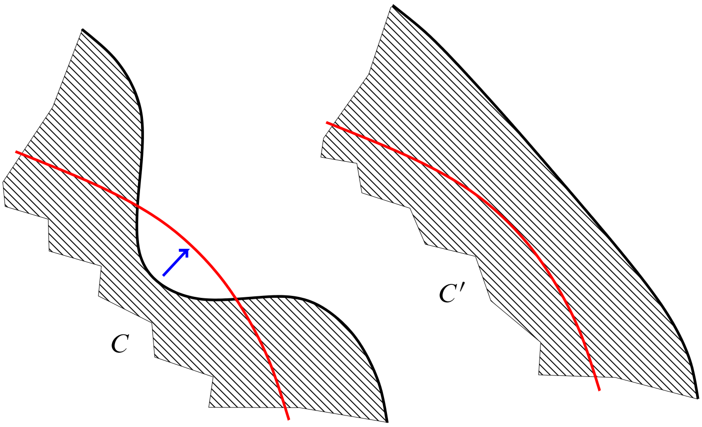


Figure 11: Case 1, engulfing.

Figure 11 indicates the rationale. Specifically, $f^{-1}(C')$ is $f^{-1}(C)$ union a relative Bott-type handle. This handle should be thought of as $I \times D^i \times D^{m-i-1}$, where (v, i, j) is the index of $Singv(f)$. This is because $\pi_v \circ f$ is a Bott-style Morse function on $f^{-1}(B)$; see Figure 12. The function $\pi_v \colon \mathbb{R}^2 \to \mathbb{R} \cdot v$ is orthogonal projection onto the line spanned by v. The "box" B is diffeomorphic to a product $B \simeq I \times I$ where the first interval factor corresponds to the red arc of $Singv(f)$ disjoint from C in Figure 12, while the second interval I is in the transverse direction (ie can be taken to be parallel to v). Thus $f^{-1}(B)$ is an interval cross an i–handle, being attached to $f^{-1}(C)$ along $(I \times \partial D^i) \cup ((\partial I) \times D^i)$, ie $\partial(I \times D^i)$. This could be thought of as an unbased version of a Whitehead product.

For details on Bott-style Morse functions, and how they give disc-bundle adjunctions for manifolds, see the paper of Bott [5, below (3.6)]. For a gentler introduction, see [3].

Case 2 is the case where the boundary of C passes over a cubic point. We will see that the homotopy type of $f^{-1}(C)$ does not change in this instance. Like case 1 there is are "engulfing" and "nonengulfing" subcases. We restrict to the nonengulfing case, as the engulfing case is similar. The main idea of the proof is that this transition corresponds to a 1–parameter family of cancelling i and $(i+1)$–handle attachments; thus we are attaching a ball along a hemisphere, which results in no change in the homotopy type.

A small variant of Lemma 2.8 occurs when the boundary of C transitions over a double-point in the singular set as in Figure 14. While the fibre homotopy type of $f^{-1}(C)$ changes during this kind of transition, the homotopy type of $f^{-1}(C)$ does not. The proof is exactly as in Lemma 2.8, in that we define the vector fields first on

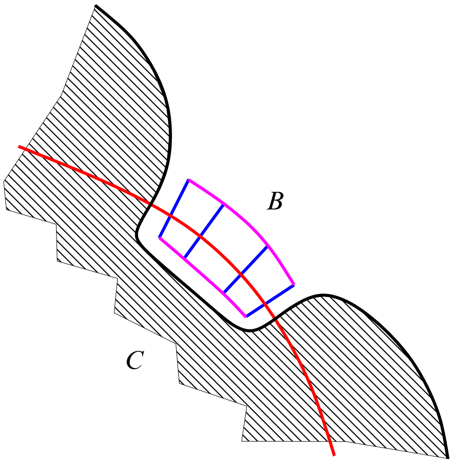


Figure 12: Case 1, engulfing.

the red curves, and then lift to M. The problem with this argument is that there is no consistent way to define the vector fields on the union of the two red curves. But this is okay, as we can lift the definition on the individual red curves (as their preimages are disjoint critical manifolds in M), and define the vector field on M directly, ie the flow of the vector field on M cannot be made to be equivariant with respect to a flow on $\mathbb{R}^2$. □

We should point out that case 2 has one special case that does result in a homotopy type change. This is depicted in Figure 6(B). These types of 2–Morse functions are not

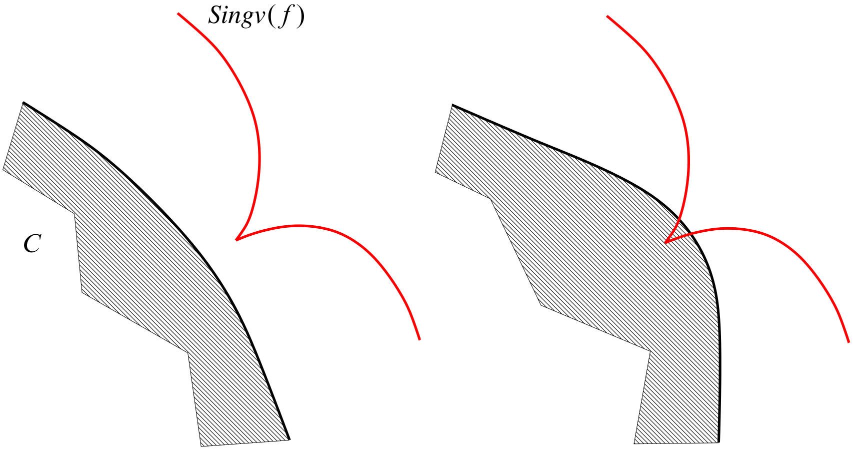


Figure 13: Case 2.

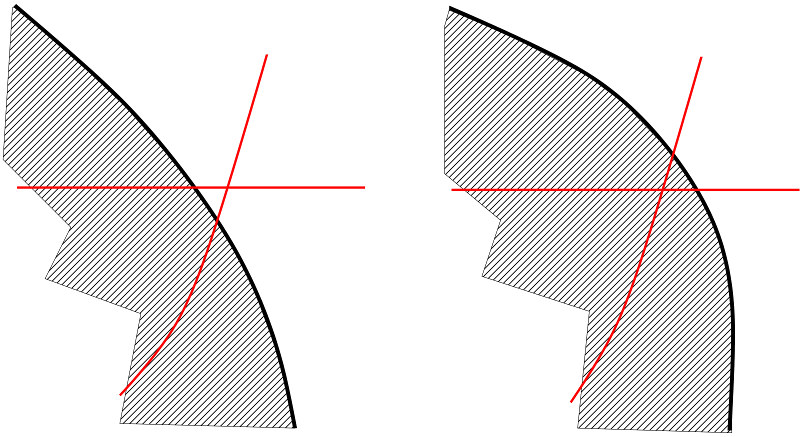


Figure 14: Case 3, over a double-point.

generic. A small isotopy of f allows one to ensure the tangent vectors at the cusps are neither vertical or horizontal, ie this at least codimension 1 in the space of smooth functions $M \to \mathbb{R}^2$, thus this situation is avoidable.

Theorem 2.9 allows us to draw the singular point set of the filtration M_f from the singular values of f, allowing automatic deduction of Examples 2.4, 2.5 and 2.7. Specifically, the singular points of the filtration are the Pareto curves of the defining map f together with the relevant vertical and horizontal rays (extending to $+\infty$) at the corresponding vertical and horizontal tangent points.

An analogous result to Theorem 2.9 appears in [2].

Proposition 2.10 *We summarize the cell attachments at the singular values of the filtration M_f.*

- *For corner singular points, the index of the cell attachment for M_f is the same as the index for f.*
- [28] *For tail singular points, if the singular values of f near the kissing tangency are exterior to the cube, then the cell attachment has the same index as f. This corresponds to Wan terminating a Pareto arc with a positive sign.*
- [28] *For a tail singular point, if the singular values of f near the kissing tangency are in the interior of the cube then the index of the cell attachment is one greater than that of the corresponding singular value for f. This corresponds to Wan terminating a Pareto arc with a negative sign.*

As a consequence of Theorem 2.9, the bifiltration of M associated to a 2–Morse function divides the plane into finitely many regions according to the homotopy type of the preimage $f^{-1}(C_{(a,b)})$. The notion of tameness [21] requires a further *no monodromy* condition, which is the requirement that there is a canonical isomorphism between $f^{-1}(C_{(a,b)})$ and a fixed representative for the region, and this isomorphism has to be natural in the sense that there is a canonical homomorphism between regions (if one exists). Our division of the plane is into contractible subspaces. By a cubical subdivision of the regions (akin to the argument that open subsets of the plane are triangulable), and taking a maximal tree in the dual 1–skeleton, one can construct a canonical zigzag of maps between any two points in a common region. The argument that there is no monodromy amounts to observing that the only avoidable handle attachments in a path from one region to another are cancelling pairs.

Corollary 2.11 [21] *Assuming the same conditions of Theorem 2.9, the bifiltration of $f\colon M \to \mathbb{R}^2$ is tame.*

We should note Wan [28] gives a filtration of the manifold M when $f\colon M \to \mathbb{R}^2$ is 2–Morse, provided the 2–Morse function satisfies the *no cycle* condition; see [28, Proposition 6.3]. Central to Wan's construction is the usage of flowlines of "generalized gradient" vector fields — roughly these are vector fields where both coordinates are increasing (away from the Pareto points). When one has a cycle, one can loop endlessly between Pareto points, but when there are no cycles, the process of connecting Pareto points via paths of generalized gradients exhausts the manifold M and linearly orders the critical intervals of Pareto sets. In our work there are a multitude of filtrations whether or not f has cycles. All the examples provided so far in this paper — and all examples in Wan's work [28] — satisfy the no cycle condition.

The simplest example of a 2–Morse function with a cycle in Wan's sense is a function of the form $f\colon S^1 \times D^2 \to \mathbb{R}^2$ having two critical arcs of index $(1,1)$, with the critical arcs being properly embedded in $S^1 \times D^2$. There are generalized gradient flows on the endpoints connecting the arcs in a cyclic ordering. While this function is only defined on a manifold diffeomorphic to $S^1 \times D^2$, with a little work one can embed this 2–Morse function into a closed 3–manifold, but one needs to add additional critical values. There is a rather simple cyclic example if one allows the use of 2–Morse functions of the sort $f\colon S^3 \to S^2$. We obtain this map as the composite of the 2–sheeted branched cover $S^3 \to S^3$ over the Hopf link together with the Hopf fibration $S^3 \to S^2$, provided the Hopf fibration projection of the Hopf link is a 2–crossing diagram in S^2.

3 Persistence paths and pathwise barcodes

In a 1–dimensional persistent homology, barcodes represent collections of parameter intervals at which homology generators are born and killed. In multifiltered persistent homology, in particular, in our 2–dimensional case, there is no simple barcode analogy, and, as Carlsson and Zomorodian pointed out in [8], there is no complete discrete invariant. Many authors have studied *rank invariants* in a module theory setting [8; 19]. A somewhat more elementary notion of *persistent Betti number* (*PBN*) *functions* is presented in Cerri, Di Fabio, Ferri, Frosini and Landi [10, Definition 2.2]. These are collections of functions $\{\beta_{f,q}\colon \Delta^+ \to N \cup \infty\}_{q\in\mathbb{Z}}$,

$$\beta_{f,q}((a,b),(a',b')) = \operatorname{rank} H_q(i^{((a,b),(a',b'))}),$$

where

$$\Delta^+ = \{((a,b),(a',b')) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid (a,b) \preceq (a',b')\},$$

and $i^{((a,b),(a',b'))}\colon M_{(a,b)} \hookrightarrow M_{(a',b')}$ is the inclusion of sublevel sets.

For computational purposes, the authors of [10] use a reduction to one-dimensional persistence diagrams via so called *foliation method*. It consists of applying the one-dimensional rank invariant along the lines defined by positive coordinate vectors in chosen finite grids. That method is restated as a *fibred barcode* in the context of persistence modules by Lesnick and Wright [20, Section 1.5].

As we observed in Section 2 on our 2–Morse function examples, although there are uncountably many singular points, the changes in topology can be finitely characterised. They either occur when we cross an arc of the singularity *Singv* in the poset-increasing direction, or when we cross a horizontal or vertical half-line passing through the vertex (a,b) of $C_{(a,b)}$ and "kissing" the singularity. We will refer to both types of components of *Critv* as to *Pareto critical value arcs* or, for short, *Pareto arcs*. Note that in [28], the term *Pareto set* refers to a subset of *Singp* $\subset M$ and *critical intervals* to its components, while our Pareto arcs are the corresponding subsets of the extended Pareto grid *Critv* $\subset \mathbb{R}^2$. There are finitely many homotopically distinct paths, with $M_{(a,b)}$ starting with an empty set and ending with the whole manifold. Each one can give a different sequence of handle attachments creating new generators of homology or cancelling previous ones, all giving $H_*(M)$ at the end of the day.

This observation leads to the notion of persistence paths which is a substitute for either Cerri's foliation method [10] or Lesnick and Wright fibred barcode [20]. It can be also be viewed as a discrete analogy of Wan's generalized gradient (whose choice is also

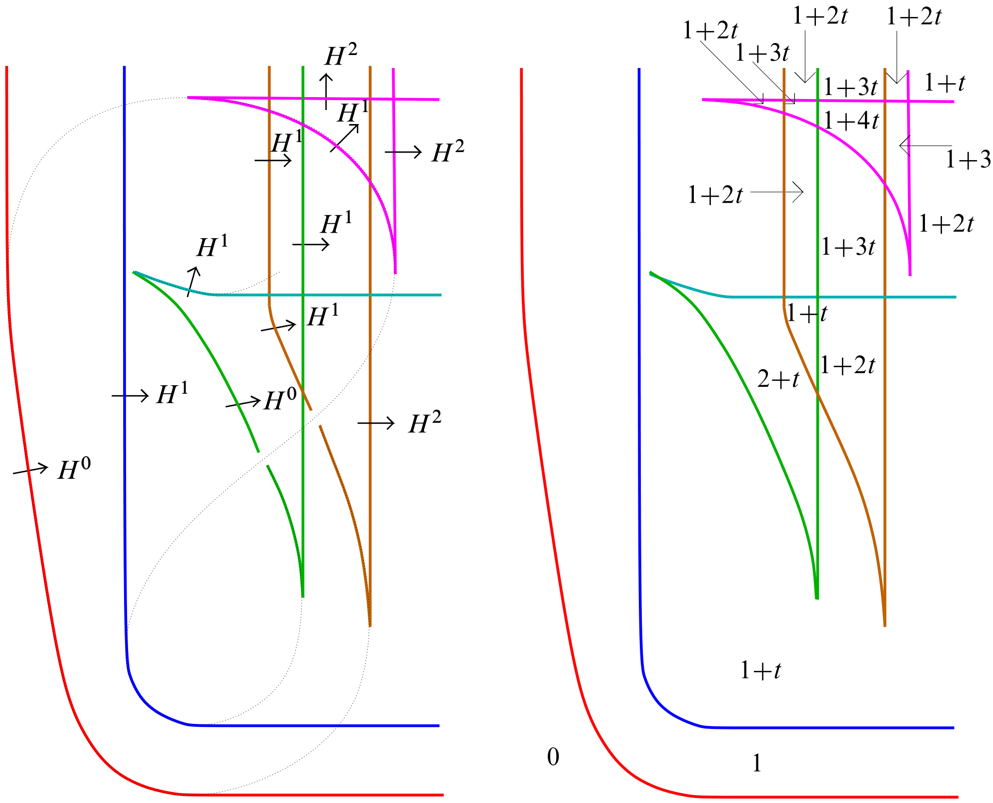


Figure 15: Klein bottle projection, with Poincaré polynomials.

not unique) in [28]. Before we proceed, let us introduce some terminology. As far as rank invariants or persistent Betti numbers are of concern, a convenient way to record the homological information carried in sublevel sets is the *Poincaré polynomial*

$$P(t,M)=\sum_{k=0}^{n}\beta_k t^k,$$

where $\beta_k=\operatorname{rank} H_k(M)$ and n is the dimension of M.

In Figure 3 and Figure 15, left, we see the Poincaré polynomials $P(t,M_{(a,b)})$ for points (a,b) located in regions bounded by Pareto arcs. We are also interested in increments $\Delta P(t,H^j)$ arising as we cross a Pareto arc increasingly in (a,b). A j–handle can either create a j generator (new component, creating a hole or a cavity) or kill a $(j-1)$ generator (merging components, filling a hole or a cavity). In the first case, we get $\Delta P(t,H^j)=t^j$, and in the second case we get $\Delta P(t,H^j)=-t^{j-1}$. Thus the index of a handle can be read out from ΔP. If it is t^k, we have a creating k–handle and if it is $-t^k$, we have a cancelling $(k+1)$–handle.

The term Pareto arc includes half-lines defined by quadrants $C_{(a,b)}$. Crossing their vertex (a,b) may create "multiple handles" where ΔP is not just one term. For example in the vertex of fish tail visible in Figure 16, $\Delta P = -t + t^2$. A point at which a single handle is attached will be called *generic*.

We choose a generic point (a,b) on each Pareto arc and let $H_{(a,b)}$ be the corresponding handle. At this time, the choice is arbitrary but we may want to chose endpoints of an arc, when we study metric sensitive barcodes.

We let $T = [0,1]$ and $R = [r_1, R_1] \times [r_2, R_2]$ be a fixed rectangle in $\mathbb{R}^2$ containing $f(M)$ in its interior.

Definition 3.1 Let $\{(a_i, b_i)\}_{i=0,1,\dots m} \subset R$ be a sequence of generic points on Pareto arcs such that $M_{(a,b)} = \varnothing$ for all (a,b) downward-left of (a_0, b_0), (a_{i+1}, b_{i+1}) can be reached from (a_i, b_i) going upward-right through the region enclosed by the two arcs, and $M_{(a_m,b_m)} = M$. A *persistence path* is a continuous function $\rho\colon I \to R$ with $\rho(0) = (r_1, r_2)$ and $\rho(1) = (R_1, R_2)$ which is nondecreasing in both coordinates, and joins the points of the sequence.

It can be seen that one can find sequences on the arcs so to get piecewise linear persistence (PL) paths with line segments between two consecutive points. This is useful in showing that we get a discrete characterization. For simplicity of notation, we let $H_i = H_{(a_i,b_i)}$ and $M_i = M_{(a_i,b_i)}$. We have a linear filtration

$$M_0 \subset M_1 \subset \cdots \subset M_m = M.$$

Figure 16, left, shows two persistence paths for the cupped sphere presented in Example 2.5. The path displayed in dark green avoids the pocket, the one in orange passes through it.

Note that [28] needs a no-cycle condition to apply the generalized gradient. Our persistence paths can be defined even in the presence of Wan's cycles, because they are more restrictive than Wan's admissible curves [28, Definition 6.3]: A persistence path leaves a Pareto arc at the same point as it enters it. Along the path, there may be no cycles, because it is increasing with respect to the poset relation.

We now shift our attention to a multidimensional analogy of the Morse inequalities. Our results may be useful for the continuation of the work on the discrete multidimensional Morse–Forman theory initiated in [1].

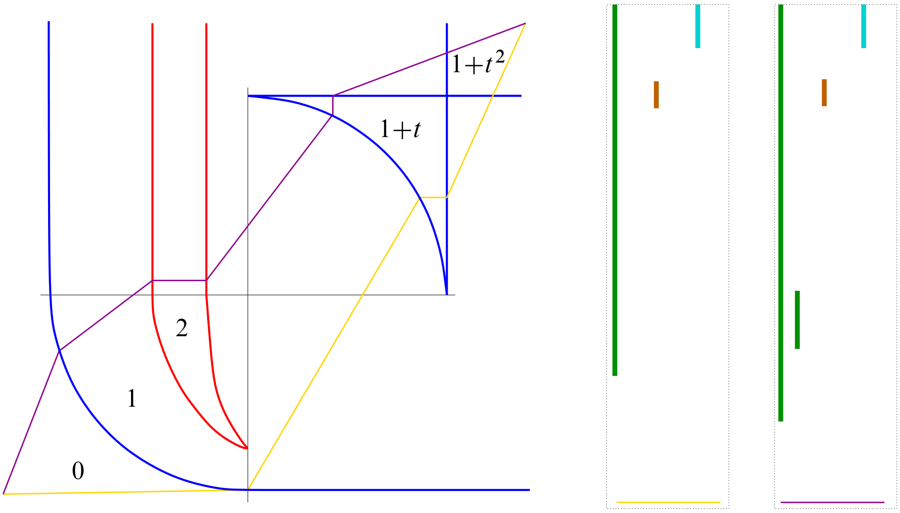


Figure 16: Left: two persistence paths for the function in Example 2.5. Right: their corresponding barcodes in rectangles marked with the same colour as the corresponding path; β_0 barcodes displayed by green lines, β_1 by brown lines, and β_2 by cyan lines.

The following result is an analogy of the Morse equation in the Conley index theory [23; 25].

Theorem 3.2 (Morse–Conley equation for persistence paths) *Let ρ be a persistence path for $((a_i, b_i))_{i=0,1,\dots m}$ and let c_j be the number of j–handles associated to its points. Then there exists a polynomial Q with nonnegative integer coefficients such that*

$$\sum_{j=0}^{n} c_j t^j = P(t, M) + (1+t)Q(t). \tag{1}$$

Proof A direct consequence of the definition of ΔP and that of persistence path is

$$\sum_{i=0}^{m} \Delta P(t, H_{(a_i,b_i)}) = \sum_{k=0}^{n} \beta_k t^k = P(t, M). \tag{2}$$

If all handles of ρ create new generators, then, in light of the preceding discussion, the left-hand side of (2) is exactly the left-hand side of (1). Thus (1) holds with $Q(t) = 0$. If a j–handle kills a $(j-1)$ generator, then the sum on the left-hand side of (2) misses two terms, t^{j-1} and t^j, contributing the sum on the left-hand side of (1). By adding all these missing terms to both sides of (2), we get (1) with Q built by terms $t^{j-1} + t^j = (1+t)t^{j-1}$. □

By taking $t = -1$ in (1), we get the following corollary.

Corollary 3.3 (Euler characteristics) *For any persistence path ρ,*

$$\sum_{j=0}^{n} (-1)^j c_j = \chi(M), \tag{3}$$

where $\chi(M) = \sum_{k=0}^{n} (-1)^k \beta_k$ is the Euler–Poincaré characteristic of M.

Equation (3) is a part of the set of classical Morse inequalities. Since two polynomials are equal if and only if all their coefficients are equal, (1) also gives *weak Morse inequalities*,

$$c_j \geq \beta_j \quad \text{for all } j = 1, 2, \dots n. \tag{4}$$

We conclude this section by deriving a classical result of Morse theory on *strong Morse inequalities*. The reader is referred to the book by Milnor [22] for the classical formulation. For the sake of completeness, we present a neat and short proof of an unknown source we have been told about by Marian Mrozek.

Corollary 3.4 (strong Morse inequalities) *For any persistence path ρ and any $k \geq 0$,*

$$c_k - c_{k-1} + c_{k-2} + \cdots \pm c_0 \geq \beta_k - \beta_{k-1} + \beta_{k-2} + \cdots \pm \beta_0. \tag{5}$$

Proof Knowing that $c_j = \beta_j = 0$ for all $j > n$, we can treat (1) as a power series equation

$$\sum_{j=0}^{\infty} c_j t^j = \sum_{j=0}^{\infty} \beta_j t^j + (1+t)Q(t). \tag{6}$$

Multiplying both sides of (6) by $\sum_{i=0}^{\infty} (-1)^i t^i$, the power series inverse of $(1+t)$, we get

$$\sum_{k=0}^{\infty} \left(\sum_{i=0}^{k} (-1)^i c_{k-i} \right) t^k = \sum_{k=0}^{\infty} \left(\sum_{i=0}^{k} (-1)^i \beta_{k-i} \right) t^k + Q(t).$$

Since two power series are equal if and only if all their coefficients are equal and the coefficients of $Q(t)$ are nonnegative, we get (5). □

Our Morse inequalities should be compared with the work of Wan [28]. Perhaps the main difference between our work and his is that we convert functions $f\colon M \to \mathbb{R}^2$ into families of filtrations of the manifold M. Wan uses essentially all of the Pareto arcs to define his filtration, which is often larger than our filtrations. Moreover he requires special "acyclic" Morse 2–functions to even define a filtration of M, while any Morse 2–function works for us.

We now turn our attention to the computability of persistent homology via persistence paths. We associate *pathwise barcodes* to any persistence path ρ as follows. First, we want to normalise lengths of persistence paths so to have them all of length 1. Given a point $(a, b) \in \rho(I)$, let $s(a, b)$ be the euclidean distance from (r_1, r_2) to (a, b) along the path ρ divided by the total length of ρ.

Definition 3.5 The *ρ–barcode* in homology of dimension k is a function on representatives of the H_k generators, whose values are subintervals of [0, 1]. When an H_k generator is born by a handle attachment at the point (a_i, b_i) and it is killed at the point (a_j, b_j) with $i < j < m$, the corresponding barcode interval is $[s(a_i, b_i), s(a_j, b_j)]$. The *lifetime* of that generator is $s(a_j, b_j) - s(a_i, b_i)$. If a generator persists until the point (R_1, R_2) of the chosen rectangle, it will also persist if the values of (R_1, R_2) increase. Thus it is reasonable to declare that its lifetime is infinite and the corresponding barcode interval is $[s(a_i, b_i), \infty)$.

Figure 16, right, shows barcodes of the two persistence paths displayed on the right. It is visible that the lifetime of the second generator of H_0 created when crossing the pocket is short and it may be null, if we choose the path in dark green that avoids the pocket. Similarly, the lifetime of the H_1 generator is short.

We shall now briefly discuss prospects for numerical implementations of pathwise barcodes. We should emphasize that the aim of our paper is to only provide a theoretical background for computation.

Following predecessors [8; 10] who set up computing methods for multifiltrations, we should consider the family of all piecewise linear persistence paths ρ built on points (a_i, b_i) in a given finite grid. However, that is a huge family and this choice is likely to lead to computational complexity issues. The size of the family of such paths is most likely similar to that of *Young diagrams* [13]. Moreover, the number of nodes to join by paths, decisive for the size of the family, increases quadratically with grid subdivisions.

For pathwise barcodes, we postulate that it should be sufficient to consider a finite representable family $\mathrm{Rep}(f)$ of persistence paths built of specific points on Pareto curves: centre points, nearly lower-right and upper-left endpoints of Pareto curves, as well as their intersections with horizontal and vertical lines passing through or touching other endpoints. We claim that $\mathrm{Rep}(f)$ is a small and exhaustive representation. Moreover, the size of $\mathrm{Rep}(f)$ does not increase with grid subdivisions.

By *exhaustive representation*, we mean here that any additional paths give rise to *equivalent barcodes*. That, in turn, means that their barcodes have the same number of intervals for each homology dimension, they may vary by length but appear in the same sequence according to birth and death dates.

We are conscious of the fact that, proceeding this way, we are missing the postulate that the persistence should be computed blindly from data, without knowing the exact manifold M and exact function f. But it may also be interesting to consider the case when we have M and f given by formulas that enable computing singularities.

4 Extensions

When applying pathwise barcodes to functions which do not satisfy Wan's no-cycle property [28, Definitions 6.4 and 6.5], it would be interesting to see what is the information carried by the barcodes of those persistence paths of $\mathrm{Rep}(f)$ which cross and go about the cycles of f.

The filtration of $\mathbb{R}^2$ by quadrants $C_{(a,b)}$ has a complementary filtration by quadrant exteriors

$$E_{(a,b)} = \{(x,y) \in \mathbb{R}^2 \mid x \geq a \text{ or } y \geq b\}.$$

Provided the boundary of $C_{(a,b)}$ is transverse to the singular points of $f\colon M \to \mathbb{R}^2$, one has that $f^{-1}(C_{(a,b)})$ is a manifold with corners. This allows us to use a Poincaré duality isomorphism

$$H_k(f^{-1}(C_{(a,b)})) \simeq H^{m-k}(M, f^{-1}(E_{(a,b)})).$$

Given that quadrant exteriors are the union of three quadrants, this gives a fairly detailed relationship between the persistent homologies of filtrations corresponding to the four quadrant families:

$$C_{f_1 \leq a, f_2 \leq b} = C_{(a,b)}, \quad C_{f_1 \geq a, f_2 \leq b}, \quad C_{f_1 \leq a, f_2 \geq b}, \quad C_{f_1 \geq a, f_2 \geq b}.$$

This technique could be thought to be a strong parallel to the theory of trisections of 4–manifolds [14; 15] as developed by Gay and Kirby. It also gives a formal setup analogous to *extended persistence* of Morse functions, considered in [12].

Another direction one could take to extrapolate this research would be using smooth functions $M \to \mathbb{R}^k$ for $k > 2$. This topic is of a great interest to the topological data analysis community. The computational methods of multiparameter persistent homology such as those in [1; 8; 9; 19] are dimension-independent but, on the other hand, they do not have the same insight into the geometry of the encountered singularities as the one we present here for the $\mathbb{R}^2$–valued functions. There are a variety of useful "Morse theory" type tools to describe the singularities of functions of this kind. The analogous theory of multisections of manifolds is developed by Rubinstein and Tillman [24].

Yet another direction to undertake is the practical implementation of our suggested method for computing pathwise barcodes on the basis of a representable family Rep(f).

References

[1] **M Allili**, **T Kaczynski**, **C Landi**, **F Masoni**, *Acyclic partial matchings for multidimensional persistence: algorithm and combinatorial interpretation*, J. Math. Imaging Vision 61 (2019) 174–192 MR Zbl

[2] **M Assif P K**, **Y Baryshnikov**, *Biparametric persistence for smooth filtrations*, preprint (2021) arXiv

[3] **A Banyaga**, **D Hurtubise**, *Lectures on Morse homology*, Kluwer Texts in the Mathematical Sciences 29, Kluwer, Dordrecht (2004) MR Zbl

[4] **S Biasotti**, **A Cerri**, **P Frosini**, **D Giorgi**, **C Landi**, *Multidimensional size functions for shape comparison*, J. Math. Imaging Vision 32 (2008) 161–179 MR

[5] **R Bott**, *Lectures on Morse theory, old and new*, Bull. Amer. Math. Soc. 7 (1982) 331–358 MR Zbl

[6] **P Bubenik**, **M J Catanzaro**, *Multiparameter persistent homology via generalized Morse theory*, preprint (2021) arXiv

[7] **F Cagliari**, **B Di Fabio**, **M Ferri**, *One-dimensional reduction of multidimensional persistent homology*, Proc. Amer. Math. Soc. 138 (2010) 3003–3017 MR Zbl

[8] **G Carlsson**, **A Zomorodian**, *The theory of multidimensional persistence*, Discrete Comput. Geom. 42 (2009) 71–93 MR Zbl

[9] **N Cavazza**, **M Ethier**, **P Frosini**, **T Kaczynski**, **C Landi**, *Comparison of persistent homologies for vector functions: from continuous to discrete and back*, Comput. Math. Appl. 66 (2013) 560–573 MR Zbl

[10] **A Cerri**, **B Di Fabio**, **M Ferri**, **P Frosini**, **C Landi**, *Betti numbers in multidimensional persistent homology are stable functions*, Math. Methods Appl. Sci. 36 (2013) 1543–1557 MR Zbl

[11] **A Cerri**, **M Ethier**, **P Frosini**, *On the geometrical properties of the coherent matching distance in 2D persistent homology*, J. Appl. Comput. Topol. 3 (2019) 381–422 MR Zbl

[12] **H Edelsbrunner**, **J L Harer**, *Computational topology: An introduction*, Amer. Math. Soc., Providence, RI (2010) MR Zbl

[13] **W Fulton**, *Young tableaux: With applications to representation theory and geometry*, London Mathematical Society Student Texts 35, Cambridge Univ. Press (1997) MR Zbl

[14] **D T Gay**, **R Kirby**, *Indefinite Morse 2–functions: broken fibrations and generalizations*, Geom. Topol. 19 (2015) 2465–2534 MR Zbl

[15] **D Gay**, **R Kirby**, *Trisecting 4–manifolds*, Geom. Topol. 20 (2016) 3097–3132 MR Zbl

[16] **V Guillemin**, **A Pollack**, *Differential topology*, Prentice-Hall, Englewood Cliffs, NJ (1974) MR Zbl

[17] **A A Kosinski**, *Differential manifolds*, Pure and Applied Mathematics 138, Academic, Boston, MA (1993) MR Zbl

[18] **C Landi**, **S Scaramuccia**, *Relative-perfectness of discrete gradient vector fields and multi-parameter persistent homology*, J. Comb. Optim. 44 (2022) 2347–2374 MR Zbl

[19] **M Lesnick**, *Lecture notes for Math 840: Multiparameter persistence*, lecture notes, SUNY Albany (2019) Available at `https://www.albany.edu/~ML644186/AMAT_840_Spring_2019/Math840_Notes.pdf`

[20] **M Lesnick**, **M Wright**, *Interactive visualization of 2–D persistence modules*, preprint (2015) arXiv

[21] **E Miller**, *Homological algebra of modules over posets*, preprint (2020) arXiv

[22] **J Milnor**, *Morse theory*, Annals of Mathematics Studies 51, Princeton Univ. Press (1963) MR Zbl

[23] **M Mrozek**, *The Morse equation in Conley's index theory for homeomorphisms*, Topology Appl. 38 (1991) 45–60 MR Zbl

[24] **J H Rubinstein**, **S Tillmann**, *Multisections of piecewise linear manifolds*, Indiana Univ. Math. J. 69 (2020) 2209–2239 MR Zbl

[25] **K P Rybakowski**, **E Zehnder**, *A Morse equation in Conley's index theory for semiflows on metric spaces*, Ergodic Theory Dynam. Systems 5 (1985) 123–143 MR Zbl

[26] **O Saeki**, *Topology of singular fibers of differentiable maps*, Lecture Notes in Math. 1854, Springer (2004) MR Zbl

[27] **S Smale**, *Global analysis and economics: Pareto optimum and a generalization of Morse theory*, Synthese 31 (1975) 345–358 MR Zbl

[28] **Y H Wan**, *Morse theory for two functions*, Topology 14 (1975) 217–228 MR Zbl

[29] **H Whitney**, *On singularities of mappings of euclidean spaces, I: Mappings of the plane into the plane*, Ann. of Math. 62 (1955) 374–410 MR Zbl

Mathematics and Statistics, University of Victoria
Victoria, BC, Canada

Département de Mathématiques, Université de Sherbrooke
Sherbrooke, QC, Canada

`rybu@uvic.ca`, `tomasz.kaczynski@usherbrooke.ca`

Received: 25 October 2021 Revised: 9 March 2022

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 23 Issue 6 (pages 2415–2924) 2023