AGT

Algebraic & Geometric Topology

Volume 25 (2025)

Issue 3 (pages 1265–1915)

msp

ALGEBRAIC & GEOMETRIC TOPOLOGY
msp.org/agt

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

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

Algebraic & Geometric Topology (ISSN 1472-2747 printed, 1472-2739 electronic) is published 9 times per year and continuously online, by Mathematical Sciences Publishers, 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
https://msp.org/

Algebraic & Geometric Topology **25**:3 (2025) 1265–1320
DOI: 10.2140/agt.2025.25.1265
Published: 20 June 2025

A-polynomials, Ptolemy equations and Dehn filling

JOSHUA A HOWIE
DANIEL V MATHEWS
JESSICA S PURCELL

The A-polynomial encodes hyperbolic geometric information on knots and related manifolds. Historically, it has been difficult to compute, and particularly difficult to determine A-polynomials of infinite families of knots. Here, we compute A-polynomials by starting with a triangulation of a manifold, then using symplectic properties of the Neumann–Zagier matrix encoding the gluings to change the basis of the computation. The result is a simplification of the defining equations. We apply this method to families of manifolds obtained by Dehn filling, and show that the defining equations of their A-polynomials are Ptolemy equations which, up to signs, are equations between cluster variables in the cluster algebra of the cusp torus.

57K14, 57K31, 57K32; 57K10

1 Introduction

The A-polynomial is a polynomial associated to a knot that encodes a great deal of geometric information. It is closely related to deformations of hyperbolic structures on knots, originally explored by Thurston [41]. Such deformations give rise to a one complex parameter family of representations of the knot group into $\mathrm{SL}(2,\mathbb{C})$. All representations form the *representation variety*, which was originally studied in pioneering work of Culler and Shalen [10; 11] and Culler, Gordon, Luecke and Shalen [9], and remains a very active area of research; see the survey by Shalen [39]. However representation varieties are difficult to compute, and often have complicated topology. In the 1990s, Cooper, Culler, Gillet, Long and Shalen [5] realised that a representation variety could be projected onto $\mathbb{C}^2$ using the longitude and meridian of the knot, with a simpler image. The image is given by the zero set of a polynomial in two variables, up to scaling. This is the A-polynomial.

Among its geometric properties, the A-polynomial detects many incompressible surfaces, and gives information on cusp shapes and volumes [5; 6]. It has relations to Mahler measure (Boyd [2]), and appears in quantum topology through the AJ-conjecture (Garoufalidis [18]; Garoufalidis and Lê [19]; Frohman, Gelca and Lofaro [17]). Unfortunately, A-polynomials are also difficult to compute. Unlike other knot polynomials, there are no skein relations to determine them. Originally, they were computed by finding polynomial equations from a matrix presentation of a representation, and then using resultants or Gröbner bases to eliminate variables; see, for example [6] by Cooper and Long. Unlike other knot polynomials, they are known only for a handful of infinite examples, including twist knots, some double twist knots, and small families of 2-bridge knots (Hoste and Shanahan [27]; Mathews [31]; Ham and Lee [25]; Petersen [37]; Tran [43]), some pretzel knots (Tamura and Yokota [40]; Garoufalidis and Mattman [20]), and cabled knots and iterated torus knots (Ni and Zhang [35]). Culler [7] has computed A-polynomials for all knots with up to eight crossings, most nine-crossing knots, many ten-crossing knots, and all knots that can be triangulated with up to seven ideal tetrahedra.

This paper gives a simplified method for determining A-polynomials, especially for infinite families of knots obtained by Dehn filling. Our method is to change the variables in the defining equations. Typically, defining equations for A-polynomials have high degree in the variables to eliminate, making them computationally difficult. Under a change of variables, we show that all such equations can be expressed in degree two in the variables to eliminate. For families of knots obtained by Dehn filling, even more can be said. There will be a finite, fixed number of "outside equations", and a sequence of equations determined completely by the slope of the Dehn filling. All such equations exhibit Ptolemy-like properties, with very similar behaviours to cluster algebras. We expect the method to greatly improve our ability to compute families of A-polynomials. Indeed, of all the known examples of infinite families of A-polynomials above, all except the cabled knots and iterated torus knots are obtained by Dehn filling a fixed parent manifold.

1.1 Computing the A-polynomial

Champanerkar [4] introduced a geometric way to compute the A-polynomial based on a triangulation of a knot complement. His method is to start with a collection of equations — one gluing equation for each edge of the triangulation, and two equations for the cusp — and eliminate variables. The coefficients in the gluing and cusp equations are effectively the entries in the Neumann–Zagier matrix [33]. This matrix has interesting symplectic properties: its rows form part of a standard symplectic basis for a symplectic vector space. Dimofte [12] and Dimofte and van der Veen [13] considered extending this collection of vectors into a standard basis for $\mathbb{R}^{2n}$, and then changing the basis. This yields a change of variables, and an equivalent set of equations. Eliminating variables again yields (up to technicalities) the A-polynomial; effectively this can be considered a process of symplectic reduction.

There are a few issues with Dimofte's calculations that have made them difficult to use in practice. First, the result appears in physics literature, which makes it somewhat difficult for mathematicians to read.

More importantly, to carefully perform the change of basis, in particular to nail down the correct signs in the defining equations, a priori one needs to determine the symplectic dual vectors to the vectors arising from gluing equations. These are not only nontrivial to compute, but also highly nonunique. Only after obtaining such vectors can one invert a large symplectic matrix.

In this paper, we overcome these issues. Using work of Neumann [32], we show that we may "invert without inverting". That is, we show that Dimofte's symplectic reduction can be read off of ingredients already present in the Neumann–Zagier matrix, without having to compute symplectic dual vectors. As a result, we may convert Champanerkar's (possibly complicated) equations into simpler equations that have Ptolemy-like structure.

There are other ways to compute A-polynomials. Zickert [46] and Garoufalidis, Thurston and Zickert [21] introduced one in work on extended Ptolemy varieties, inspired by Fock and Goncharov [14]. Their work also starts with a triangulation, but in the case of interest assigns six variables per tetrahedron, and relates these by what are called Ptolemy relations and identification relations. After an appropriate transformation, the corresponding variables satisfy gluing equations; see [21, Section 12]. Zickert notes a "fundamental duality" between Ptolemy coordinates and gluing equations in [46, Remark 1.13]. However, it is not clear why the duality arises. The equations we find in this paper are similar to the defining equations of Zickert, but with fewer variables. We expect that the results of this paper may provide a connection to two very different approaches to calculating A-polynomials. While we do not show that the methods of that paper and this one are equivalent, we conjecture that they are, and thus the techniques here may provide a geometric, symplectic explanation for the "fundamental duality".

1.2 Neumann–Zagier matrices and the main theorem

Let M be a hyperbolic 3-manifold with a triangulation. Then it has an associated Neumann–Zagier matrix, which we will denote by NZ. The properties of NZ are reviewed in Section 2. In short, gluing and cusp equations give a system of the form $\mathrm{NZ}\cdot Z = H + i\pi C$, where Z is a vector of variables related to tetrahedra, and H and C are both vectors of constants.

Neumann and Zagier showed that if M has one cusp, then the n rows of NZ corresponding to gluing equations have rank $n-1$. Thus a row can be removed, leaving $n-1$ linearly independent rows. Denote the matrix given by removing such a row of NZ by $\mathrm{NZ}^\flat$, and similarly denote the vector obtained from C by removing the corresponding row by $C^\flat$. We will refer to $\mathrm{NZ}^\flat$ as the *reduced Neumann–Zagier matrix*. The vector $C^\flat$ is called the *sign vector*. We will show that, after possibly relabelling the tetrahedra of a triangulation, we may assume one of the entries of $C^\flat$ corresponding to a gluing equation is nonzero. Neumann [32] has shown that there always exists an integer vector B such that $\mathrm{NZ}^\flat \cdot B = C^\flat$.

To state the main theorem, we introduce a little more notation. The last two rows of the matrix $\mathrm{NZ}^\flat$ correspond to cusp equations associated to the meridian and longitude. For ease of notation, we will denote the entries in the row associated to the meridian and longitude, respectively, by

$$(\mu_1\ \mu_1'\ \mu_2\ \mu_2'\ \dots) \quad \text{and} \quad (\lambda_1\ \lambda_1'\ \lambda_2\ \lambda_2'\ \dots).$$

Finally, suppose the edges of the tetrahedra are glued into n edges $E_1, \dots, E_n$. Label the ideal vertices of each tetrahedron 0, 1, 2, and 3, with 1, 2, 3 in anticlockwise order when viewed from 0. Then there are six edges, each labelled by a pair of integers $\alpha\beta \in \{01, 02, 03, 12, 13, 23\}$. For the j^{th} tetrahedron, let $j(\alpha\beta)$ denote the index of the edge class to which that edge is identified. That is, if the edge $\alpha\beta$ is glued to E_k, then $j(\alpha\beta) = k$.

Theorem 1.1 *Let M be a one-cusped manifold with a hyperbolic triangulation $\mathcal{T}$, with associated reduced Neumann–Zagier matrix $\mathrm{NZ}^\flat$ and sign vector $C^\flat$ as above. Also as above, denote the entries of the last two rows of $\mathrm{NZ}^\flat$ by μ_j, μ_j' in the row corresponding to the meridian, and λ_j, λ_j' in the row corresponding to the longitude. Let $B = (B_1, B_1', B_2, B_2', \dots)$ be an integer vector such that $\mathrm{NZ}^\flat \cdot B = C^\flat$.*

Define formal variables $\gamma_1, \dots, \gamma_n$, one associated with each edge of $\mathcal{T}$. For a tetrahedron Δ_j of $\mathcal{T}$, and edge $\alpha\beta \in \{01, 02, 03, 12, 13, 23\}$, define $\gamma_{j(\alpha\beta)}$ to be the variable γ_k such that the edge of Δ_j between vertices α and β is glued to the edge of $\mathcal{T}$ associated with γ_k.

*For each tetrahedron Δ_j of $\mathcal{T}$, define the **Ptolemy equation** of Δ_j by*

$$(-1)^{B_j'} \ell^{-\mu_j/2} m^{\lambda_j/2} \gamma_{j(01)}\gamma_{j(23)} + (-1)^{B_j} \ell^{-\mu_j'/2} m^{\lambda_j'/2} \gamma_{j(02)}\gamma_{j(13)} - \gamma_{j(03)}\gamma_{j(12)} = 0.$$

When we solve the system of Ptolemy equations of $\mathcal{T}$ in terms of m and ℓ, setting $\gamma_n = 1$ and eliminating the variables $\gamma_1, \dots, \gamma_{n-1}$, we obtain a factor of the $\mathrm{PSL}(2, \mathbb{C})$ A-polynomial.

In fact, we obtain the same factor as Champanerkar. The precise version of this theorem is contained in Theorem 2.58 below.

Remark 1.2 The Ptolemy equations above are always quadratic in the variables γ_j. Moreover, their form indicates intriguing algebraic structure that is not readily apparent from the gluing equations.

We find the simplicity and the algebraic structure of the equations of Theorem 1.1 to be a major feature of this paper. The defining equations of the A-polynomial are quite simple! We note that using these equations requires finding the vector B of Theorem 1.1. This is a problem in linear Diophantine equations. Because B is guaranteed to exist, it can be found by computing the Smith normal form of the matrix NZ (see, for example, Chapter II.21(c) of [34] by Newman). In practice, we were able to find B for examples with significantly less work.

Remark 1.3 The γ variables in Theorem 1.1 are precisely Dimofte's γ variables of [12], and these Ptolemy equations are essentially equivalent to those of that paper.

The word "equivalent" here conceals a projective subtlety. The gluing and cusp equations are a set of $n+2$ equations in n tetrahedron parameters and ℓ, m, but only $n+1$ of them are independent. The Ptolemy equations are however a set of n independent equations in n edge variables and ℓ, m. Nonetheless, they are homogeneous, and so $\gamma_1, \dots, \gamma_n$ can be regarded as varying on $\mathbb{CP}^{n-1}$; alternatively, one can divide through by an appropriate power of one γ_i to obtain equations in the $n-1$ variables,

$$\frac{\gamma_1}{\gamma_i}, \dots, \frac{\gamma_{i-1}}{\gamma_i}, \frac{\gamma_{i+1}}{\gamma_i}, \dots, \frac{\gamma_n}{\gamma_i},$$

which can be eliminated. Effectively, one can simply set one of the variables γ_i to 1.

A further subtlety arises because our Ptolemy equations are *not* polynomials in m and ℓ; they are rather polynomials in $m^{1/2}$ and $\ell^{1/2}$. If we set $M = m^{1/2}$ and $L = \ell^{1/2}$ then we obtain *polynomial* Ptolemy equations. Moreover, the variables L and M so defined are essentially those appearing in the $\mathrm{SL}(2,\mathbb{C})$ A-polynomial: a matrix in $\mathrm{SL}(2,\mathbb{C})$ with eigenvalues L, L^{-1} yields an element of $\mathrm{PSL}(2,\mathbb{C})$ corresponding to a hyperbolic isometry with holonomy $L^2 = \ell$. Indeed, the Ptolemy varieties of [46] are calculated from $\mathrm{SL}(2,\mathbb{C})$ representations, rather than $\mathrm{PSL}(2,\mathbb{C})$. We obtain:

Corollary 1.4 *After setting $M = \pm m^{1/2}$ and $L = \pm \ell^{1/2}$, eliminating the γ variables from the polynomial Ptolemy equations of a one-cusped hyperbolic triangulation yields a polynomial in M and L which contains, as a factor, the factor of the* $\mathrm{SL}(2,\mathbb{C})$ *A-polynomial describing hyperbolic structures.*

The precise version of this corollary is Corollary 2.59.

1.3 Ptolemy equations in Dehn filling

Our main application of Theorem 1.1 is to consider the defining equations of A-polynomials under Dehn filling.

Consider a two-component link in S^3 with component knots K_0, K_1. Consider Dehn filling K_0 along some slope p/q; K_1 then becomes a knot in a 3-manifold. A Dehn filling can be triangulated using *layered solid tori*, originally defined by Jaco and Rubinstein [30]; see also the work by Guéritaud and Schleimer [24]. Building a layered solid torus yields a sequence of triangulations of a once-punctured torus. The combinatorics of the 3-dimensional layered solid torus corresponds closely to the combinatorics of 2-dimensional triangulations of punctured tori.

Triangulations of punctured tori can be endowed with λ-lengths via work of Penner [36]. When one flips a diagonal in a triangulation, the λ-lengths are related by a Ptolemy equation. This gives the algebra formed by λ-lengths the structure of a *cluster algebra* (Fock and Goncharov [14]; Fomin, Shapiro and Thurston [15]; Gekhtman, Shapiro and Vainshtein [22]). Cluster algebras arise in diverse contexts across mathematics (see eg works by Fomin, Williams and Zelevinsky [16] and by Williams [45]).

We obtain two sets of Ptolemy equations: one for the cluster algebra of the punctured torus coming from λ-lengths, and one for the tetrahedra in the layered solid torus coming from Theorem 1.1. These are identical except for signs. Thus we can regard the algebra generated by our Ptolemy equations as a "twisted" cluster algebra, where the word "twisted" indicates some changes of sign.

Theorem 1.5 *Suppose M has two cusps, $\mathfrak{c}_0, \mathfrak{c}_1$, and is triangulated such that only two tetrahedra meet $\mathfrak{c}_1$, and generating curves $\mathfrak{m}_0, \mathfrak{l}_0$ on the cusp triangulation of $\mathfrak{c}_0$ avoid these tetrahedra. Then for any Dehn filling on the cusp $\mathfrak{c}_1$ obtained by attaching a layered solid torus, the Ptolemy equations satisfy:*

(i) *There are a finite number of fixed Ptolemy equations, independent of the Dehn filling, coming from tetrahedra outside the Dehn filling. These are obtained as in Theorem 1.1 using the reduced Neumann–Zagier matrix and B vector for the unfilled manifold.*

(ii) *The Ptolemy equations for the tetrahedra in the solid torus take the form*

$$\pm\gamma_x\gamma_y \pm \gamma_a^2 - \gamma_b^2 = 0,$$

where a, b, x, y are slopes on the torus boundary and x, y are crossing diagonals. In addition, the variable γ_y will appear for the first time in this equation, with γ_x, γ_a, and γ_b appearing in earlier equations.

A precise version of this theorem is Theorem 3.17.

Theorem 1.5 in particular implies that each of the Ptolemy equations for the solid torus can be viewed as giving a recursive definition of the new variable γ_y. These equations are explicit, depending on the slope. Since the outside Ptolemy equations are fixed, in practice this gives a recursive definition of the A-polynomial in terms of the slope of the Dehn filling. If we take a sequence of Dehn filling slopes $\{p_i/q_i\}$, then the A-polynomials of the knots $K_i = K_{p_i/q_i}$, are closely related. The Ptolemy equations defining $A_{K_{i+1}}$ are, roughly speaking, obtained from those for A_{K_i} by adding a single extra Ptolemy relation.

We illustrate this theorem by example for twist knots, which are Dehn fillings of the Whitehead link. While A-polynomials of twist knots are known (Hoste and Shanahan [27]; Mathews [31]), we still believe this example is useful in showing the simplicity of the Ptolemy equations. In a follow up paper with Thompson [28], we apply these tools to a new family of knots whose A-polynomials were previously unknown, namely twisted torus knots obtained by Dehn filling the Whitehead sister.

1.4 Structure of this paper

In Section 2, we recall work of Thurston [41] and Neumann and Zagier [33], including gluing and cusp equations, the Neumann–Zagier matrix, and its symplectic properties. We introduce a symplectic change of basis, and show this leads to Ptolemy equations that give the A-polynomial, proving Theorem 1.1.

In Section 3, we connect to Dehn fillings. We review the construction of layered solid tori, and triangulations of Dehn filled manifolds, and show how the triangulation adjusts the Neumann–Zagier matrix. Using this, we find Ptolemy equations for any layered solid torus, completing the proof of Theorem 1.5.

Section 4 works through the example of knots obtained by Dehn filling the Whitehead link.

Acknowledgements

This work was supported by the Australian Research Council, grants DP160103085 and DP210103136.

2 From gluing equations to Ptolemy equations via symplectic reduction

In this section we discuss Dimofte's symplectic reduction method and refine it to show how gluing and cusp equations are equivalent to Ptolemy equations, proving Theorem 1.1.

2.1 Triangulations, gluing and cusp equations

Let M be a 3-manifold that is the interior of a compact manifold $\overline{M}$ with all boundary components tori. Let the number of boundary tori be $n_\mathfrak{c}$, so M has $n_\mathfrak{c}$ cusps. For example, M may be a link complement $S^3 - L$, where L is a link of $n_\mathfrak{c}$ components, and $\overline{M}$ a link exterior $S^3 - N(L)$.

Suppose M has an ideal triangulation. Throughout this paper, unless stated otherwise, *triangulation* means ideal triangulation, and *tetrahedron* means ideal tetrahedron. Throughout, n denotes the number of tetrahedra in a triangulation.

Definition 2.1 An *oriented labelling* of a tetrahedron is a labelling of its four ideal vertices with the numbers $0, 1, 2, 3$, as in Figure 1, up to oriented homeomorphism preserving edges.

In an ideal tetrahedron with an oriented labelling, we call the opposite pairs of edges $(01, 23)$, $(02, 13)$, $(03, 12)$ respectively the *a-edges*, *b-edges* and *c-edges*.

In an oriented labelling, around each vertex (as viewed from outside the tetrahedron), the three incident edges are an a-, b-, and c-edge in anticlockwise order.

The number of edges in the triangulation is equal to the number n of tetrahedra, as follows: letting the numbers of edges and faces in the triangulation temporarily be E and F, $\partial\overline{M}$ is triangulated with $2E$ vertices, $3F$ edges and $4n$ triangles. As $\partial\overline{M}$ consists of tori, its Euler characteristic $2E - 3F + 4n$ is zero. Since $2F = 4n$, we have $E = n$.

Definition 2.2 A *labelled triangulation* of M is an oriented ideal triangulation of M, where

(i) the tetrahedra are labelled $\Delta_1, \dots, \Delta_n$ in some order,

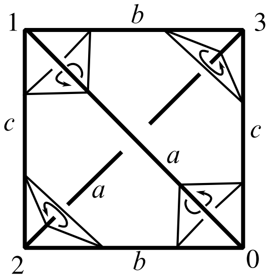


Figure 1: A tetrahedron with vertices labelled 0, 1, 2, 3 and opposite edges labelled a, b, c.

(ii) the edges are labelled $E_1, \dots, E_n$ in some order, and

(iii) each tetrahedron is given an oriented labelling.

As in the introduction, we will need to refer to the edge E_k to which an edge of tetrahedron Δ_j is glued.

Definition 2.3 For $j \in \{1, \dots, n\}$ and distinct $\mu, \nu \in \{0, 1, 2, 3\}$, the index of the edge to which the edge $(\mu\nu)$ of Δ_j is glued is denoted $j(\mu\nu)$. In other words, the edge $(\mu\nu)$ of Δ_j is identified to $E_{j(\mu\nu)}$.

Suppose now that we have a labelled triangulation of M. To each tetrahedron Δ_j we associate three variables z_j, z_j', z_j''. These variables are associated with the a-, b- and c-edges of Δ_j and satisfy the equations

$$z_j z_j' z_j'' = -1, \tag{2.4}$$

$$z_j + (z_j')^{-1} - 1 = 0. \tag{2.5}$$

If Δ_j has a hyperbolic structure then these parameters are standard tetrahedron parameters; see [42]. Each of z_j, z_j', z_j'' gives the cross ratio of the four ideal points, in some order. The arguments of z_j, z_j', z_j'' respectively give the dihedral angles of Δ_j at the a-, b- and c-edges. Note that (2.4) and (2.5) imply that none of z_j, z_j', z_j'' can be equal to 0 or 1 (ie tetrahedra are nondegenerate).

Definition 2.6 In a labelled triangulation of M, we denote by $a_{k,j}, b_{k,j}, c_{k,j}$ respectively the number of a-, b-, c-edges of Δ_j identified to E_k.

Lemma 2.7 *For each fixed j,*

$$\sum_{k=1}^{n} a_{k,j} = 2, \quad \sum_{k=1}^{n} b_{k,j} = 2 \quad \textit{and} \quad \sum_{k=1}^{n} c_{k,j} = 2. \tag{2.8}$$

Proof Each tetrahedron Δ_j has two a-edges, two b-edges and two c-edges, so for fixed j the total sum over all k must be 2. □

The nonzero terms in the first sum are $a_{j(01),j}$ and $a_{j(23),j}$. Note that $j(01)$ could equal $j(23)$; this occurs when the two a-edges of Δ_j are glued to the same edge. In that case, $a_{j(01),j}$ and $a_{j(23),j}$ are the same term, equal to 2. If the two a-edges are not glued to the same edge, then $E_{j(01)}$ and $E_{j(23)}$ are distinct, each with one a-edge of Δ_j identified to it, and $a_{j(01),j} = a_{j(23),j} = 1$. Similarly, the nonzero terms in the second sum are $b_{j(02),j}, b_{j(13),j}$ and in the third sum $c_{j(03),j}, c_{j(12),j}$.

The numbers $a_{k,j}, b_{k,j}, c_{k,j}$ can be arranged into a matrix.

Definition 2.9 The *incidence matrix*, In, of a labelled triangulation $\mathcal{T}$ is the $n \times 3n$ matrix whose kth row is $(a_{k,1}, b_{k,1}, c_{k,1}, \dots, a_{k,n}, b_{k,n}, c_{k,n})$.

Thus In has rows corresponding to the edges $E_1, \dots, E_n$, and the columns come in triples with the j^{th} triple corresponding to the tetrahedron Δ_j.

The *gluing equation* for edge E_k is then

$$\prod_{j=1}^{n} z_j^{a_{k,j}} (z_j')^{b_{k,j}} (z_j'')^{c_{k,j}} = 1. \tag{2.10}$$

When the ideal triangulation $\mathcal{T}$ is hyperbolic, the gluing equations express the fact that tetrahedra fit geometrically together around each edge.

Denote the $n_\mathfrak{c}$ boundary tori of $\overline{M}$ by $\mathbb{T}_1, \dots, \mathbb{T}_{n_\mathfrak{c}}$. A triangulation of M by tetrahedra induces a triangulation of each $\mathbb{T}_k$ by triangles. On each $\mathbb{T}_k$ we choose a pair of oriented curves $\mathfrak{m}_k, \mathfrak{l}_k$ forming a basis for $H_1(\mathbb{T}_k)$. By an isotopy if necessary, we may assume each curve is in general position with respect to the triangulation of $\mathbb{T}_k$, without backtracking. Then each curve splits into segments, where each segment lies in a single triangle and runs from one edge to a distinct edge. Each segment of $\mathfrak{m}_k$ or $\mathfrak{l}_k$ can thus be regarded as running clockwise or anticlockwise around a unique corner of a triangle; these directions are as viewed from outside the manifold. We count anticlockwise motion around a vertex as positive, and clockwise motion as negative. Each vertex (resp. face) of the triangulation of $\mathbb{T}_k$ corresponds to some edge (resp. tetrahedron) of the triangulation $\mathcal{T}$ of M; thus each corner of a triangle corresponds to a specific edge of a specific tetrahedron.

Definition 2.11 The *a-incidence number* (resp. b-, c-incidence number) of $\mathfrak{m}_k$ (resp. $\mathfrak{l}_k$) with the tetrahedron Δ_j is the number of segments of $\mathfrak{m}_k$ (resp. $\mathfrak{l}_k$) running anticlockwise (ie positively) through a corner of a triangle corresponding to an a-edge (resp. b-, c-edge) of Δ_j, minus the number of segments of $\mathfrak{m}_k$ (resp. $\mathfrak{l}_k$) running clockwise (ie negatively) through a corner of a triangle corresponding to an a-edge (resp. b-edge, c-edge) of Δ_j.

(i) Denote by $a^{\mathfrak{m}}_{k,j}, b^{\mathfrak{m}}_{k,j}, c^{\mathfrak{m}}_{k,j}$ the a-, b-, c-incidence numbers of $\mathfrak{m}_k$ with Δ_j.

(ii) Denote by $a^{\mathfrak{l}}_{k,j}, b^{\mathfrak{l}}_{k,j}, c^{\mathfrak{l}}_{k,j}$ the a-, b-, c-incidence numbers of $\mathfrak{l}_k$ with Δ_j.

To each cusp torus $\mathbb{T}_k$ we associate variables m_k, ℓ_k. The *cusp equations* at $\mathbb{T}_k$ are

$$m_k = \prod_{j=1}^{n} z_j^{a^{\mathfrak{m}}_{k,j}} (z_j')^{b^{\mathfrak{m}}_{k,j}} (z_j'')^{c^{\mathfrak{m}}_{k,j}}, \quad \ell_k = \prod_{j=1}^{n} z_j^{a^{\mathfrak{l}}_{k,j}} (z_j')^{b^{\mathfrak{l}}_{k,j}} (z_j'')^{c^{\mathfrak{l}}_{k,j}} \tag{2.12}$$

When $\mathcal{T}$ is a *hyperbolic triangulation*, meaning the ideal tetrahedra are all positively oriented and glue to give a smooth, complete hyperbolic structure on the underlying manifold, the cusp equations give m_k and ℓ_k, the holonomies of the cusp curves $\mathfrak{m}_k$ and $\mathfrak{l}_k$, in terms of tetrahedron parameters.

Any hyperbolic triangulation $\mathcal{T}$ gives tetrahedron parameters z_j, z_j', z_j'' and cusp holonomies m_k, ℓ_k satisfying the relationships (2.4)–(2.5) between the z variables, the gluing equations (2.10) and cusp equations (2.12); moreover, the tetrahedron parameters all have positive imaginary part. However, in general there may be solutions of these equations which do not correspond to a hyperbolic triangulation, for instance those with z_j with negative imaginary part (which may still give M a hyperbolic structure), or with branching around an edge (which will not). Additionally, not every hyperbolic structure on M may give a solution to the gluing and cusp equations, since the triangulation $\mathcal{T}$ may not be geometrically realisable.

2.2 The A-polynomial from gluing and cusp equations

Suppose now that $n_{\mathfrak{c}} = 1$, ie M has one cusp, and moreover, that M is the complement of a knot K in a homology 3-sphere.

In this case, there is no need for the $k = 1$ subscript in notation for the lone cusp, and we may simply write

$$\mathfrak{m} = \mathfrak{m}_1, \quad \mathfrak{l} = \mathfrak{l}_1, \quad m = m_1, \quad \ell = \ell_1,$$
$$a_j^{\mathfrak{m}} = a_{1,j}^{\mathfrak{m}}, \quad b_j^{\mathfrak{m}} = b_{1,j}^{\mathfrak{m}}, \quad c_j^{\mathfrak{m}} = c_{1,j}^{\mathfrak{m}}, \quad a_j^{\mathfrak{l}} = a_{1,j}^{\mathfrak{l}}, \quad b_j^{\mathfrak{l}} = b_{1,j}^{\mathfrak{l}}, \quad c_j^{\mathfrak{l}} = c_{1,j}^{\mathfrak{l}},$$

In this case we can take the boundary curves $(\mathfrak{m}, \mathfrak{l})$ to be a topological longitude and meridian, respectively. That is, we may take $\mathfrak{l}$ to be primitive and nullhomologous in M, and $\mathfrak{m}$ to bound a disc in a neighbourhood of K.

We orient $\mathfrak{m}$ and $\mathfrak{l}$ so that the tangent vectors $v_{\mathfrak{m}}$ and $v_{\mathfrak{l}}$ to $\mathfrak{m}$ and $\mathfrak{l}$, respectively, at the point where $\mathfrak{m}$ intersects $\mathfrak{l}$ are oriented according to the right-hand rule: $v_{\mathfrak{m}} \times v_{\mathfrak{l}}$ points in the direction of the outward normal.

The equations (2.4)–(2.5) relating the z, z', z'' variables, the gluing equations (2.10), and the cusp equations (2.12) are equations in the variables z_j, z_j', z_j'' and ℓ, m. Solve these equations for ℓ, m, eliminating the variables z_j, z_j', z_j'' to obtain a relation between ℓ and m.

Champanerkar [4] showed that the above equations can be solved in this sense to give divisors of the $\mathrm{PSL}(2,\mathbb{C})$ A-polynomial of M. Segerman showed that, if one takes a certain extended version of this variety, there exists a triangulation such that all factors of the $\mathrm{PSL}(2,\mathbb{C})$ A-polynomial are obtained [38]. See also [23] for an effective algorithm.

Theorem 2.13 (Champanerkar) *When we solve the system of equations* (2.4)–(2.5), (2.10) *and* (2.12) *in terms of m and ℓ, we obtain a factor of the* $\mathrm{PSL}(2,\mathbb{C})$ *A-polynomial.*

2.3 Logarithmic equations and Neumann–Zagier matrix

We now return to the general case where the number $n_{\mathfrak{c}}$ of cusps of M is arbitrary.

Note that equation (2.4) relating z_j, z_j', z_j'', the gluing equations (2.10), and the cusp equations (2.12) are multiplicative. By taking logarithms now we make them additive.

Equation (2.4) implies that each z_j, z'_j and z''_j is nonzero. Taking (an appropriate branch of) a logarithm we obtain

$$\log z_j + \log z'_j + \log z''_j = i\pi.$$

Define $Z_j = \log z_j$ and $Z'_j = \log z'_j$, using the branch of the logarithm with argument in $(-\pi, \pi]$, and then define Z''_j as

$$Z''_j = i\pi - Z_j - Z'_j, \tag{2.14}$$

so that indeed Z''_j is a logarithm of z''_j.

In a hyperbolic triangulation, each tetrahedron parameter has positive imaginary part. The arguments of z_j, z'_j, z''_j (ie the imaginary parts of Z_j, Z'_j, Z''_j) are the dihedral angles at the a-, b- and c-edges of Δ_j, respectively. They are the angles of a Euclidean triangle, hence they all lie in $(0, \pi)$ and they sum to π.

The gluing equation (2.10) expresses the fact that tetrahedra fit together around an edge. Taking a logarithm, we may make the somewhat finer statement that dihedral angles around the edge sum to 2π. Thus we take the logarithmic form of the gluing equations as

$$\sum_{j=1}^{n} a_{k,j} Z_j + b_{k,j} Z'_j + c_{k,j} Z''_j = 2\pi i. \tag{2.15}$$

We similarly obtain logarithmic forms of the cusp equations (2.12) as

$$\log m_k = \sum_{j=1}^{n} a^{\mathfrak{m}}_{k,j} Z_j + b^{\mathfrak{m}}_{k,j} Z'_j + c^{\mathfrak{m}}_{k,j} Z''_j, \quad \log \ell_k = \sum_{j=1}^{n} a^{\mathfrak{l}}_{k,j} Z_j + b^{\mathfrak{l}}_{k,j} Z'_j + c^{\mathfrak{l}}_{k,j} Z''_j. \tag{2.16}$$

We can then observe that any solution of (2.14) and the logarithmic gluing and cusp equations (2.15)–(2.16) yields, after exponentiation, a solution of (2.4) and the original gluing equation (2.10) and cusp equations (2.12). Moreover, any solution of (2.4), (2.10) and (2.12) has a logarithm which is a solution of (2.14) and (2.15)–(2.16).

Using (2.14) we eliminate the variables Z''_j (just as using (2.4) we can eliminate the variables z''_j). In doing so, coefficients are combined in a way that persists throughout this paper, and so we define these combinations as follows.

Definition 2.17 For a given labelled triangulation of M, we define

$$\begin{aligned}
d_{k,j} &= a_{k,j} - c_{k,j}, & d'_{k,j} &= b_{k,j} - c_{k,j}, & c_k &= \sum_{j=1}^{n} c_{k,j} & &\text{for } k = 1, 2, \dots, n, \\
\mu_{k,j} &= a^{\mathfrak{m}}_{k,j} - c^{\mathfrak{m}}_{k,j}, & \mu'_{k,j} &= b^{\mathfrak{m}}_{k,j} - c^{\mathfrak{m}}_{k,j}, & c^{\mathfrak{m}}_k &= \sum_{j=1}^{n} c^{\mathfrak{m}}_{k,j} & &\text{for } k = 1, 2, \dots, n_{\mathfrak{c}}, \\
\lambda_{k,j} &= a^{\mathfrak{l}}_{k,j} - c^{\mathfrak{l}}_{k,j}, & \lambda'_{k,j} &= b^{\mathfrak{l}}_{k,j} - c^{\mathfrak{l}}_{k,j}, & c^{\mathfrak{l}}_k &= \sum_{j=1}^{n} c^{\mathfrak{l}}_{k,j} & &\text{for } k = 1, 2, \dots, n_{\mathfrak{c}}.
\end{aligned}$$

Note that the index k in the first line steps through the n edges, while the index k in the next two lines steps through the $n_{\mathfrak{c}}$ cusps.

When $n_{\mathfrak{c}} = 1$ we can drop the k subscript on cusp terms, so we have

$$\mu_j = a_j^{\mathfrak{m}} - c_j^{\mathfrak{m}}, \quad \mu_j' = b_j^{\mathfrak{m}} - c_j^{\mathfrak{m}}, \quad c^{\mathfrak{m}} = \sum_{j=1}^{n} c_j^{\mathfrak{m}}, \quad \lambda_j = a_j^{\mathfrak{l}} - c_j^{\mathfrak{l}}, \quad \lambda_j' = b_j^{\mathfrak{l}} - c_j^{\mathfrak{l}}, \quad c^{\mathfrak{l}} = \sum_{j=1}^{n} c_j^{\mathfrak{l}}.$$

We thus rewrite the logarithmic gluing and cusp equations (2.15)–(2.16) in terms of the variables Z_j, Z_j' and ℓ_k, m_k only, as

$$\textbf{(2.18)} \qquad \sum_{j=1}^{n} d_{k,j} Z_j + d_{k,j}' Z_j' = i\pi(2 - c_k),$$

$$\textbf{(2.19)} \qquad \sum_{j=1}^{n} \mu_{k,j} Z_j + \mu_{k,j}' Z_j' = \log m_k - i\pi c_k^{\mathfrak{m}},$$

$$\textbf{(2.20)} \qquad \sum_{j=1}^{n} \lambda_{k,j} Z_j + \lambda_{k,j}' Z_j' = \log \ell_k - i\pi c_k^{\mathfrak{l}}.$$

Define the row vectors of coefficients in equations (2.18)–(2.20) as

$$\begin{aligned} R_k^G &:= (d_{k,1} \;\; d_{k,1}' \;\; \cdots \;\; d_{k,n} \;\; d_{k,n}'), \\ R_k^{\mathfrak{m}} &:= (\mu_{k,1} \;\; \mu_{k,1}' \;\; \cdots \;\; \mu_{k,n} \;\; \mu_{k,n}'), \\ R_k^{\mathfrak{l}} &:= (\lambda_{k,1} \;\; \lambda_{k,1}' \;\; \cdots \;\; \lambda_{k,n} \;\; \lambda_{k,n}'). \end{aligned}$$

So R_k^G gives the coefficients in the logarithmic gluing equation for the k^{th} edge E_k, and $R_k^{\mathfrak{m}}, R_k^{\mathfrak{l}}$ give respectively coefficients in the logarithmic cusp equations for $\mathfrak{m}_k$ and $\mathfrak{l}_k$ on the k^{th} cusp.

When $n_{\mathfrak{c}} = 1$ we again drop the k subscript on cusp terms and simply write $R^{\mathfrak{m}} = R_k^{\mathfrak{m}}$ and $R^{\mathfrak{l}} = R_k^{\mathfrak{l}}$, so that $R^{\mathfrak{m}} = (\mu_1, \mu_1', \dots, \mu_n, \mu_n')$ and $R^{\mathfrak{l}} = (\lambda_1, \lambda_1', \dots, \lambda_n, \lambda_n')$.

We observe natural meanings for the new $d, d', \mu, \mu', \lambda, \lambda', c$ coefficients of Definition 2.17 by re-exponentiating. The tetrahedron parameters and the holonomies m_k, ℓ_k satisfy versions of the gluing and cusp equations without any z_j'' appearing, where the d, d' variables appear as exponents in gluing equations, $\mu, \mu', \lambda, \lambda'$ variables appear as exponents in cusp equations, and the c variables determine signs:

$$\prod_{j=1}^{n} z_j^{d_{k,j}} (z_j')^{d_{k,j}'} = (-1)^{c_k} \qquad \text{for } k = 1, \dots, n \text{ (indexing edges)}$$

$$m_k = (-1)^{c_k^{\mathfrak{m}}} \prod_{j=1}^{n} z_j^{\mu_{k,j}} (z_j')^{\mu_{k,j}'}, \quad \ell_k = (-1)^{c_k^{\mathfrak{l}}} \prod_{j=1}^{n} z_j^{\lambda_{k,j}} (z_j')^{\lambda_{k,j}'} \quad \text{for } k = 1, \dots, n_{\mathfrak{c}} \text{ (cusps).}$$

When $n_{\mathfrak{c}} = 1$, the notation for cusp equations again simplifies so we have

$$m = (-1)^{c^{\mathfrak{m}}} \prod_{j=1}^{n} z_j^{\mu_j} (z_j')^{\mu_j'} \quad \text{and} \quad \ell = (-1)^{c^{\mathfrak{l}}} \prod_{j=1}^{n} z_j^{\lambda_j} (z_j')^{\lambda_j'}.$$

The matrix with rows $R^G_1, \dots, R^G_n, R^{\mathfrak{m}}_1, R^{\mathfrak{l}}_1, \dots, R^{\mathfrak{m}}_{n_{\mathrm{c}}}, R^{\mathfrak{l}}_{n_{\mathrm{c}}}$ is called the *Neumann–Zagier matrix*, and we denote it by NZ. The first n rows correspond to the edges $E_1, \dots, E_n$, and the next rows come in pairs corresponding to the pairs $(\mathfrak{m}_k, \mathfrak{l}_k)$ of basis curves for the cusp tori $\mathbb{T}_1, \dots, \mathbb{T}_{n_{\mathrm{c}}}$. The columns come in pairs corresponding to the tetrahedra $\Delta_1, \dots, \Delta_n$. Note that the data of a labelled triangulation of Definition 2.2 give us the information to write down the matrix: the edge ordering $E_1, \dots, E_n$ orders the rows; the tetrahedron ordering $\Delta_1, \dots, \Delta_n$ orders pairs of columns; and the oriented labelling on each tetrahedron determines each pair of columns:

$$\textbf{(2.21)} \qquad \mathrm{NZ} = \begin{bmatrix} R^G_1 \\ \vdots \\ R^G_n \\ R^{\mathfrak{m}}_1 \\ R^{\mathfrak{l}}_1 \\ \vdots \\ R^{\mathfrak{m}}_{n_{\mathrm{c}}} \\ R^{\mathfrak{l}}_{n_{\mathrm{c}}} \end{bmatrix} = \begin{array}{c c} & \begin{matrix} \quad\Delta_1 & & \cdots & \Delta_n \quad \end{matrix} \\ \begin{matrix} E_1 \\ \vdots \\ E_n \\ \mathfrak{m}_1 \\ \mathfrak{l}_1 \\ \vdots \\ \mathfrak{m}_{n_{\mathrm{c}}} \\ \mathfrak{l}_{n_{\mathrm{c}}} \end{matrix} & \begin{bmatrix} d_{1,1} & d'_{1,1} & \cdots & d_{1,n} & d'_{1,n} \\ & \vdots & \ddots & \vdots & \\ d_{n,1} & d'_{n,1} & \cdots & d_{n,n} & d'_{n,n} \\ \mu_{1,1} & \mu'_{1,1} & \cdots & \mu_{1,n} & \mu'_{1,n} \\ \lambda_{1,1} & \lambda'_{1,1} & \cdots & \lambda_{1,n} & \lambda'_{1,n} \\ & \vdots & \ddots & \vdots & \\ \mu_{n_{\mathrm{c}},1} & \mu'_{n_{\mathrm{c}},1} & \cdots & \mu_{n_{\mathrm{c}},n} & \mu_{n_{\mathrm{c}},n} \\ \lambda_{n_{\mathrm{c}},1} & \lambda'_{n_{\mathrm{c}},1} & \cdots & \lambda_{n_{\mathrm{c}},n} & \lambda'_{n_{\mathrm{c}},n} \end{bmatrix} \end{array}.$$

The gluing and cusp equations can then be written as a single matrix equation, if we make the following definitions.

Definition 2.22 The Z*-vector*, z*-vector*, H*-vector* and C*-vector* are defined as

$$\begin{aligned} Z &:= (Z_1, Z'_1, \dots, Z_n, Z'_n)^T, \\ z &:= (z_1, z'_1, \dots, z_n, z'_n)^T, \\ H &:= \big(0, \dots, 0, \log m_1, \log \ell_1, \dots, \log m_{n_{\mathrm{c}}}, \log \ell_{n_{\mathrm{c}}}\big)^T, \\ C &:= \big(2 - c_1, \dots, 2 - c_n, -c^{\mathfrak{m}}_1, -c^{\mathfrak{l}}_1, \dots, -c^{\mathfrak{m}}_{n_{\mathrm{c}}}, -c^{\mathfrak{l}}_{n_{\mathrm{c}}}\big)^T. \end{aligned}$$

The vector Z contains the logarithmic tetrahedral parameters; the vector H contains the cusp holonomies, and the vector C is a vector of constants derived from the gluing data, giving sign terms in exponentiated equations.

We summarise our manipulations of the various equations in the following statement.

Lemma 2.23 *Let $\mathcal{T}$ be a labelled triangulation of M.*

(i) *The logarithmic gluing and cusp equations can be written compactly as*

$$\textbf{(2.24)} \qquad \mathrm{NZ} \cdot Z = H + i\pi C.$$

That is, logarithmic gluing and cusp equations (2.18)–(2.20) *are equivalent to* (2.24).

(ii) *After exponentiation, a solution Z of* (2.24) *gives z which, together with z_j'' defined by* (2.4), *yields a solution of the gluing equations* (2.10) *and cusp equations* (2.12).

(iii) *Conversely, any solution (z_j, z_j', z_j'') of* (2.4), *gluing equations* (2.10) *and cusp equations* (2.12) *yields z with logarithm Z satisfying* (2.24).

(iv) *Any hyperbolic triangulation yields Z and H which satisfy* (2.24). □

2.4 Symplectic and topological properties of the Neumann–Zagier matrix

The matrix NZ has nice symplectic properties, due to Neumann–Zagier [33], which we now recall.

First, we introduce notation for the standard symplectic structure on $\mathbb{R}^{2N}$, for any positive integer N. Denote by $\boldsymbol{e}_i$ (resp. $\boldsymbol{f}_i$) the vector whose only nonzero entry is a 1 in the $(2i-1)^{\text{th}}$ coordinate (resp. $2i^{\text{th}}$ coordinate). Dually, let x_i (resp. y_i) denote the coordinate function which returns the $(2i-1)^{\text{th}}$ coordinate (resp. $2i^{\text{th}}$ coordinate). We define the standard symplectic form ω as

$$\omega = dx_1 \wedge dy_1 + \cdots + dx_N \wedge dy_N = \sum_{j=1}^{N} dx_j \wedge dy_j. \tag{2.25}$$

Thus, given two vectors $V = (V_1, V_1', \dots, V_N, V_N')$ and $W = (W_1, W_1', \dots, W_N, W_N')$ in $\mathbb{R}^{2N}$,

$$\omega(V, W) = \sum_{j=1}^{N} V_j W_j' - V_j' W_j.$$

Alternatively, $\omega(V, W) = V^T J W = (JV) \cdot W$, where $\cdot$ is the standard dot product, and J is multiplication by i on $\mathbb{C}^N \cong \mathbb{R}^{2N}$, ie $J(\boldsymbol{e}_i) = \boldsymbol{f}_i$ and $J(\boldsymbol{f}_i) = -\boldsymbol{e}_i$ (hence $J^2 = -1$). As a matrix,

$$J = \begin{bmatrix} 0 & -1 & & & & & \\ 1 & 0 & & & & & \\ & & 0 & -1 & & & \\ & & 1 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & -1 \\ & & & & & 1 & 0 \end{bmatrix}.$$

The ordered basis $(\boldsymbol{e}_1, \boldsymbol{f}_1, \dots, \boldsymbol{e}_N, \boldsymbol{f}_N)$ forms a *standard symplectic basis*, satisfying

$$\omega(\boldsymbol{e}_i, \boldsymbol{f}_j) = \delta_{i,j}, \quad \omega(\boldsymbol{e}_i, \boldsymbol{e}_j) = 0, \quad \omega(\boldsymbol{f}_i, \boldsymbol{f}_j) = 0$$

for all $i, j \in \{1, \dots, N\}$. Any sequence of $2N$ vectors on which ω takes the same values on pairs is a *symplectic basis*.

Maps which preserve a symplectic form are called *symplectomorphisms*. We will need to use a few particular linear symplectomorphisms. The proof below is a routine verification.

Lemma 2.26 *In the standard symplectic vector space* $(\mathbb{R}^{2N}, \omega)$ *as above*, *the following linear transformations are symplectomorphisms*:

(i) *For* $j, k \in \{1, \dots, N\}$, $j \neq k$, *and any* $a \in \mathbb{R}$, *map* $\boldsymbol{e}_j \mapsto \boldsymbol{e}_j + a\boldsymbol{f}_k$, $\boldsymbol{e}_k \mapsto \boldsymbol{e}_k + a\boldsymbol{f}_j$, *and leave all other standard basis vectors unchanged.*

(ii) *For* $j \in \{1, \dots, N\}$ *and any* $a \in \mathbb{R}$, *map* $\boldsymbol{e}_j \mapsto \boldsymbol{e}_j + a\boldsymbol{f}_j$, *and leave all other standard basis vectors unchanged.* □

In fact, it is not difficult to show that the linear symplectomorphisms above generate the group of linear symplectomorphisms which fix all $\boldsymbol{f}_j$. If we reorder the standard basis $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n, \boldsymbol{f}_1, \dots, \boldsymbol{f}_n)$, the symplectic matrices fixing the Lagrangian subspace spanned by the $\boldsymbol{f}_j$ have matrices of the form

$$\begin{bmatrix} I & 0 \\ A & I \end{bmatrix},$$

where I is the $n \times n$ identity matrix and A is an $n \times n$ symmetric matrix. These form a group isomorphic to the group of $n \times n$ real symmetric matrices under addition.

Returning to the Neumann–Zagier matrix NZ, observe that its row vectors lie in $\mathbb{R}^{2n}$, where n (as always) is the number of tetrahedra. These vectors behave nicely with respect to ω.

Theorem 2.27 (Neumann–Zagier [33]) *With* R_k^G, $R_k^{\mathfrak{m}}$, $R_k^{\mathfrak{l}}$ *and* ω *as above*:

(i) *For all* $j, k \in \{1, \dots, n\}$, *we have* $\omega(R_j^G, R_k^G) = 0$.

(ii) *For all* $j \in \{1, \dots, n\}$ *and* $k \in \{1, \dots, n_{\mathfrak{c}}\}$, *we have* $\omega(R_j^G, R_k^{\mathfrak{m}}) = \omega(R_j^G, R_k^{\mathfrak{l}}) = 0$.

(iii) *For all* $j, k \in \{1, \dots, n_{\mathfrak{c}}\}$, *we have* $\omega(R_j^{\mathfrak{m}}, R_k^{\mathfrak{l}}) = 2\delta_{jk}$.

(iv) *The row vectors* $R_1^G, \dots, R_n^G$ *span a subspace of dimension* $n - n_{\mathfrak{c}}$.

(v) *The rank of* NZ *is* $n + n_{\mathfrak{c}}$.

In light of Theorem 2.27(iv), by relabelling edges if necessary, we can assume a labelled triangulation has the property that the first $n - n_{\mathfrak{c}}$ rows of its Neumann–Zagier matrix are linearly independent. We will make this assumption throughout.

According to Theorem 2.27, the values of ω on pairs of vectors taken from the list of $n + n_{\mathfrak{c}}$ vectors $(R_1^G, \dots, R_{n-n_{\mathfrak{c}}}^G, R_1^{\mathfrak{m}}, \frac{1}{2}R_1^{\mathfrak{l}}, \dots, R_{n_{\mathfrak{c}}}^{\mathfrak{m}}, \frac{1}{2}R_{n_{\mathfrak{c}}}^{\mathfrak{l}})$ agree with the value of ω on corresponding pairs in the list $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_{n-n_{\mathfrak{c}}}, \boldsymbol{e}_{n-n_{\mathfrak{c}}+1}, \boldsymbol{f}_{n-n_{\mathfrak{c}}+1}, \dots, \boldsymbol{e}_n, \boldsymbol{f}_n)$. For $R_1^G, \dots, R_{n-n_{\mathfrak{c}}}^G$ linearly independent, there is a linear symplectomorphism sending each vector in the first list to the corresponding vector in the second.

Accordingly, as observed by Dimofte [12] the list of $n + n_{\mathfrak{c}}$ vectors

$$\left(R_1^G, \dots, R_{n-n_{\mathfrak{c}}}^G, R_1^{\mathfrak{m}}, \tfrac{1}{2}R_1^{\mathfrak{l}}, \dots, R_{n_{\mathfrak{c}}}^{\mathfrak{m}}, \tfrac{1}{2}R_{n_{\mathfrak{c}}}^{\mathfrak{l}}\right)$$

extends to a symplectic basis for $\mathbb{R}^{2n}$,

$$\left(R_1^{\Gamma}, R_1^{G}, \dots, R_{n-n_{\mathfrak{c}}}^{\Gamma}, R_{n-n_{\mathfrak{c}}}^{G}, R_1^{\mathfrak{m}}, \tfrac{1}{2}R_1^{\mathfrak{l}}, \dots, R_{n_{\mathfrak{c}}}^{\mathfrak{m}}, \tfrac{1}{2}R_{n_{\mathfrak{c}}}^{\mathfrak{l}}\right),$$

with the addition of $n - n_{\mathfrak{c}}$ vectors, denoted $R_1^{\Gamma}, \dots, R_{n-n_{\mathfrak{c}}}^{\Gamma}$. Being a symplectic basis means that, in addition to the equations of Theorem 2.27(i)–(iii), we also have

$$\begin{aligned} &\omega(R_j^{\Gamma}, R_k^{\Gamma}) = 0 \quad \text{and} \quad \omega(R_j^{\Gamma}, R_k^{G}) = \delta_{j,k} && \text{for all } j,k \in \{1, \dots, n-n_{\mathfrak{c}}\}, \text{ and} \\ &\omega(R_j^{\Gamma}, R_k^{\mathfrak{m}}) = \omega(R_j^{\Gamma}, R_k^{\mathfrak{l}}) = 0 && \text{for all } j \in \{1, \dots, n-n_{\mathfrak{c}}\} \text{ and } k \in \{1, \dots, n_{\mathfrak{c}}\}. \end{aligned}$$

Indeed, the R_j^{Γ} may be found by solving the equations above: given $R_k^{G}, R_k^{\mathfrak{m}}, R_k^{\mathfrak{l}}$, we may solve successively for $R_1^{\Gamma}, R_2^{\Gamma}, \dots, R_{n-n_{\mathfrak{c}}}^{\Gamma}$. Being solutions of linear equations with rational coefficients, we can find each $R_j^{\Gamma} \in \mathbb{Q}^{2n}$.

Remark 2.28 The R_j^{Γ} are not unique: there are many solutions to the above equations. Distinct solutions are related precisely by the linear symplectomorphisms of $\mathbb{R}^{2n}$ fixing an $(n+n_{\mathfrak{c}})$-dimensional coisotropic subspace. Following the discussion after Lemma 2.26, such symplectomorphisms are naturally bijective with $(n-n_{\mathfrak{c}}) \times (n-n_{\mathfrak{c}})$ real symmetric matrices. Hence the space of possible $(R_1^{\Gamma}, \dots, R_{n-n_{\mathfrak{c}}}^{\Gamma})$ has dimension $\frac{1}{2}(n-n_{\mathfrak{c}})(n-n_{\mathfrak{c}}+1)$.

For $k \in \{1, \dots, n-n_{\mathfrak{c}}\}$, write

$$\left(R_k^{\Gamma} = f_{k,1} \;\; f'_{k,1} \;\; \cdots \;\; f_{k,n} \;\; f'_{k,n}\right).$$

The symplectic basis $\left(R_1^{G}, R_1^{\Gamma}, \dots, R_{n-n_{\mathfrak{c}}}^{G}, R_{n-n_{\mathfrak{c}}}^{\Gamma}, R_1^{\mathfrak{m}}, \frac{1}{2}R_1^{\mathfrak{l}}, \dots, R_{n_{\mathfrak{c}}}^{\mathfrak{m}}, \frac{1}{2}R_{n_{\mathfrak{c}}}^{\mathfrak{l}}\right)$ forms the sequence of row vectors of a symplectic matrix, which we call $\mathrm{SY} \in \mathrm{Sp}(2n, \mathbb{R})$. When $n_{\mathfrak{c}} = 1$, we have

$$\mathrm{SY} := \begin{bmatrix} R_1^{\Gamma} \\ R_1^{G} \\ \vdots \\ R_{n-1}^{\Gamma} \\ R_{n-1}^{G} \\ R^{\mathfrak{m}} \\ \frac{1}{2}R^{\mathfrak{l}} \end{bmatrix} = \begin{bmatrix} f_{1,1} & f'_{1,1} & f_{1,2} & f'_{1,2} & \cdots & f_{1,n} & f'_{1,n} \\ d_{1,1} & d'_{1,1} & d_{1,2} & d'_{1,2} & \cdots & d_{1,n} & d'_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{n-1,1} & f'_{n-1,1} & f_{n-1,2} & f'_{n-1,2} & \cdots & f_{n-1,n} & f'_{n-1,n} \\ d_{n-1,1} & d'_{n-1,1} & d_{n-1,2} & d'_{n-1,2} & \cdots & d_{n-1,n} & d'_{n-1,n} \\ \mu_1 & \mu'_1 & \mu_2 & \mu'_2 & \cdots & \mu_n & \mu'_n \\ \frac{1}{2}\lambda_1 & \frac{1}{2}\lambda'_1 & \frac{1}{2}\lambda_2 & \frac{1}{2}\lambda'_2 & \cdots & \frac{1}{2}\lambda_n & \frac{1}{2}\lambda'_n \end{bmatrix}. \tag{2.29}$$

As a symplectic matrix, SY satisfies $(\mathrm{SY})^T J (\mathrm{SY}) = J$, and for any vectors V, W,

$$\omega(V, W) = \omega(\mathrm{SY} \cdot V, \mathrm{SY} \cdot W).$$

2.5 Linear and nonlinear equations and hyperbolic structures

The symplectic matrix SY of (2.29) shares several rows in common with NZ. We will need to rearrange rows of various matrices, and so we make the following definition.

Definition 2.30 Let A be a matrix with $n + 2n_\mathfrak{c}$ rows, denoted $A_1, \dots, A_{n+2n_\mathfrak{c}}$.

(i) The submatrices A^{I}, A^{II}, A^{III} consist of the first $n - n_\mathfrak{c}$ rows, the next $n_\mathfrak{c}$ rows, and the final $2n_\mathfrak{c}$ rows. That is,

$$A^{\mathrm{I}} = \begin{bmatrix} A_1 \\ \vdots \\ A_{n-n_\mathfrak{c}} \end{bmatrix}, \quad A^{\mathrm{II}} = \begin{bmatrix} A_{n-n_\mathfrak{c}+1} \\ \vdots \\ A_n \end{bmatrix}, \quad A^{\mathrm{III}} = \begin{bmatrix} A_{n+1} \\ \vdots \\ A_{n+2n_\mathfrak{c}} \end{bmatrix}, \qquad \text{so} \quad A = \begin{bmatrix} A^{\mathrm{I}} \\ A^{\mathrm{II}} \\ A^{\mathrm{III}} \end{bmatrix}.$$

(ii) The matrix $A^\flat$ consists of the rows of A^{I} followed by the rows of A^{III}. In other words, it is the matrix of $n + n_\mathfrak{c}$ rows

$$A^\flat = \begin{bmatrix} A^{\mathrm{I}} \\ A^{\mathrm{III}} \end{bmatrix}.$$

This matrix A of Definition 2.30 includes the case of a $(n+2n_\mathfrak{c}) \times 1$ matrix, ie a $(n+2n_\mathfrak{c})$-dimensional vector.

Observe that Definition 2.30 applies to the Neumann–Zagier matrix NZ. The matrix NZ^{I} has rows $R_1^G, \dots, R_{n-n_\mathfrak{c}}^G$, which we may assume are linearly independent. By Theorem 2.27(i) and (iv), the rows of NZ^{I} form a basis of an isotropic subspace, and the rows of $\mathrm{NZ}^{\mathrm{II}}$ also lie in this subspace. The matrix $\mathrm{NZ}^{\mathrm{III}}$ has rows $R_1^\mathfrak{m}, R_1^\mathfrak{l}, \dots, R_{n_\mathfrak{c}}^\mathfrak{m}, R_{n_\mathfrak{c}}^\mathfrak{l}$. Theorem 2.27(iv) and (v) imply that the rows of $\mathrm{NZ}^\flat$ form a basis for the rowspace of NZ.

Similarly for the vector C, observe C^{I} contains the entries $(2 - c_1, \dots, 2 - c_{n-n_\mathfrak{c}})$, and C^{III} contains the entries $(-c_1^\mathfrak{m}, -c_1^\mathfrak{l}, \dots, -c_{n_\mathfrak{c}}^\mathfrak{m}, -c_{n_\mathfrak{c}}^\mathfrak{l})$. For the holonomy vector H, we have that H^{I} and H^{II} are zero vectors, while H^{III} contains cusp holonomies.

The gluing equations (2.18) can be written as

$$\begin{bmatrix} \mathrm{NZ}^{\mathrm{I}} \\ \mathrm{NZ}^{\mathrm{II}} \end{bmatrix} \cdot Z = i\pi \begin{bmatrix} C^{\mathrm{I}} \\ C^{\mathrm{II}} \end{bmatrix}. \tag{2.31}$$

The first $n - n_\mathfrak{c}$ among these equations are given by

$$\mathrm{NZ}^{\mathrm{I}} \cdot Z = i\pi C^{\mathrm{I}}. \tag{2.32}$$

We have seen that the rows of NZ^{I} span the rows of $\mathrm{NZ}^{\mathrm{II}}$, so knowing $\mathrm{NZ}^{\mathrm{I}} \cdot Z$ determines $\mathrm{NZ}^{\mathrm{II}} \cdot Z$. But it is perhaps not so clear whether $\mathrm{NZ}^{\mathrm{I}} \cdot Z = i\pi C^{\mathrm{I}}$ implies that $\mathrm{NZ}^{\mathrm{II}} \cdot Z = i\pi C^{\mathrm{II}}$. However, as we now show, in a hyperbolic situation this is in fact the case.

Lemma 2.33 *Suppose the triangulation $\mathcal{T}$ has a hyperbolic structure. Then a vector $Z \in \mathbb{C}^{2n}$ satisfies* (2.31) *if and only if it satisfies* (2.32).

Proof Hyperbolic structures (not necessarily complete) give solutions to the gluing equations $Z = (Z_1, Z_1', \dots, Z_n, Z_n') \in \mathbb{C}^{2n}$; hence the solution space of (2.31) is nonempty. Since equations (2.32) are a subset of those of (2.31), the solution space of (2.32) is also nonempty.

Since both matrices $\left[\begin{smallmatrix}\mathrm{NZ}^{\mathrm{I}}\\ \mathrm{NZ}^{\mathrm{II}}\end{smallmatrix}\right]$ and NZ^{I} have rank $n-n_{\mathfrak{c}}$, the solution spaces of both (2.31) and (2.32) have the same dimension: $2n-(n-n_{\mathfrak{c}})=n+n_{\mathfrak{c}}$. □

Thus, some of the gluing equations of (2.18), or equivalently of (2.31), are redundant. The same is true of the larger system (2.24). So $\mathrm{NZ}^{\flat}$ is a more efficient version of the Neumann–Zagier matrix, containing only necessary information for computing hyperbolic structures.

As discussed at the end of Section 2.1, the solution spaces of these equations do not in general coincide with spaces of hyperbolic structures. The solution space of (2.32) contains the space of hyperbolic structures on the triangulation $\mathcal{T}$, but is strictly larger. These equations treat Z_j and Z_j' as independent variables, but of course they are not. In a hyperbolic structure, $z_j=e^{Z_j}$ and $z_j'=e^{Z_j'}$ are related by the equations (2.5).

Indeed, the solution space of the linear equations (2.32) has dimension $n+n_{\mathfrak{c}}$, but there are a further n conditions imposed by the relations $z_j+(z_j')^{-1}-1=0$ of (2.5). As discussed in the proof of [33, Proposition 2.3], these n conditions are independent and the result is a variety of dimension $n_{\mathfrak{c}}$. However, as we just saw, this variety may contain points that do not correspond to hyperbolic tetrahedra. Moreover, it may not contain all hyperbolic structures, as not every hyperbolic structure may be able to be realised by the triangulation $\mathcal{T}$.

However, by Thurston's hyperbolic Dehn surgery theorem [42], the space of hyperbolic structures on M is also $n_{\mathfrak{c}}$-dimensional. So at a point of the variety defined by the linear equations (2.32) and the nonlinear equations (2.5) describing a hyperbolic structure, the variety locally coincides with the space of hyperbolic structures.

We summarise this section with the following statement.

Lemma 2.34 *Let $\mathcal{T}$ be a hyperbolic triangulation of M, labelled so that its Neumann–Zagier matrix* NZ *has rows $R_1^G,\dots,R_{n-n_{\mathfrak{c}}}^G$ linearly independent.*

(i) *The logarithmic gluing equations, expressed equivalently by* (2.18) *or* (2.31), *are equivalent to the smaller independent set of equations* (2.32).

(ii) *The variety V defined by the solutions of these linear equations* (2.32), *together with the nonlinear equations* (2.5), *has dimension $n_{\mathfrak{c}}$. The hyperbolic structures on $\mathcal{T}$ correspond to a subset of V. Near a point of V corresponding to a hyperbolic structure on $\mathcal{T}$, V parametrises hyperbolic structures on $\mathcal{T}$.*

(iii) *The logarithmic gluing and cusp equations for $\mathcal{T}$ are equivalent to*

$$\mathrm{NZ}^{\flat}\cdot Z=H^{\flat}+i\pi C^{\flat}. \tag{2.35}$$ □

2.6 Symplectic change of variables

Dimofte in [12] considered using the matrix SY to *change variables* in the logarithmic gluing and cusp equations.

If M is hyperbolic, by Lemma 2.34 the gluing and cusp equations are equivalent to (2.35). Observe that the rows of $\mathrm{NZ}^\flat$ are $\big($up to a factor of $\frac{1}{2}$ in the rows $R_k^{\mathfrak{l}}\big)$ a subset of the rows of SY. Indeed, obtain SY from $\mathrm{NZ}^\flat$ by multiplying $R_k^{\mathfrak{l}}$ rows by $\frac{1}{2}$, and inserting rows $R_1^\Gamma, \dots, R_{n-n_{\mathfrak{c}}}^\Gamma$.

In the equations of (2.35) $Z = (Z_1, Z_1', \dots, Z_n, Z_n')^T$ are regarded as variables, and we now change them using SY.

Definition 2.36 Given a labelled hyperbolic triangulation $\mathcal{T}$ and a choice of symplectic matrix SY, define the collection of variables

$$\Gamma = \big(\Gamma_1, G_1, \dots, \Gamma_{n-n_{\mathfrak{c}}}, G_{n-n_{\mathfrak{c}}}, M_1, \tfrac{1}{2}L_1, \dots, M_{n_{\mathfrak{c}}}, \tfrac{1}{2}L_{n_{\mathfrak{c}}}\big)^T$$

by $\Gamma = \mathrm{SY} \cdot Z$.

In other words,

$$\Gamma = \mathrm{SY}\begin{bmatrix} Z_1 \\ Z_1' \\ \vdots \\ Z_n \\ Z_n' \end{bmatrix} \iff \begin{cases} \Gamma_k = R_k^\Gamma \cdot Z & \text{for } k \in \{1, \dots, n-n_{\mathfrak{c}}\}, \\ G_k = R_k^G \cdot Z, & \text{for } k \in \{1, \dots, n-n_{\mathfrak{c}}\}, \\ M_k = R_k^{\mathfrak{m}} \cdot Z & \text{for } k \in \{1, \dots, n_{\mathfrak{c}}\}, \\ \frac{1}{2}L_k = \frac{1}{2}R_k^{\mathfrak{l}} \cdot Z & \text{for } k \in \{1, \dots, n_{\mathfrak{c}}\}. \end{cases}$$

Lemma 2.37 *Let $\mathcal{T}$ be a labelled hyperbolic triangulation, and* SY *a matrix defining the variables* Γ. *Then the logarithmic gluing and cusp equations are equivalent to*

$$G_k = i\pi(2 - c_k), \quad M_j = \log m_j - i\pi c_j^{\mathfrak{m}}, \quad L_j = \log \ell_j - i\pi c_j^{\mathfrak{l}}. \tag{2.38}$$

In the new variables, these equations are simplified. Note that the Γ_k variables do not appear in (2.38).

Proof The first $n - n_{\mathfrak{c}}$ rows of (2.35) express the gluing equations as $R_k^G \cdot Z = i\pi(2 - c_k)$, for $k \in \{1, \dots, n-n_{\mathfrak{c}}\}$. Remaining rows of (2.35) express cusp equations as $R_j^{\mathfrak{m}} \cdot Z = \log m_j - i\pi c_j^{\mathfrak{m}}$ and $R_j^{\mathfrak{l}} \cdot Z = \log \ell_j - c_j^{\mathfrak{l}}$. □

The symplectic change of variables involves writing variables Z in terms of the variables Γ. That is, we need to invert SY.

As SY is symplectic, $(\mathrm{SY})^T J (\mathrm{SY}) = J$, so its inverse is given by $\mathrm{SY}^{-1} = -J(\mathrm{SY})^T J$, or

$$\begin{bmatrix} d_{1,1}' & -f_{1,1}' & \cdots & d_{n-n_{\mathfrak{c}},1}' & -f_{n-n_{\mathfrak{c}},1}' & \frac{1}{2}\lambda_{1,1}' & -\mu_{1,1}' & \cdots & \frac{1}{2}\lambda_{n_{\mathfrak{c}},1}' & -\mu_{n_{\mathfrak{c}},1}' \\ -d_{1,1} & f_{1,1} & \cdots & -d_{n-n_{\mathfrak{c}},1} & f_{n-n_{\mathfrak{c}},1} & -\frac{1}{2}\lambda_{1,1} & \mu_{1,1} & \cdots & -\frac{1}{2}\lambda_{n_{\mathfrak{c}},1} & \mu_{n_{\mathfrak{c}},1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{1,n}' & -f_{1,n}' & \cdots & d_{n-n_{\mathfrak{c}},n}' & -f_{n-n_{\mathfrak{c}},n}' & \frac{1}{2}\lambda_{1,n}' & -\mu_{1,n}' & \cdots & \frac{1}{2}\lambda_{n_{\mathfrak{c}},n}' & -\mu_{n_{\mathfrak{c}},n}' \\ -d_{1,n} & f_{1,n} & \cdots & -d_{n-n_{\mathfrak{c}},n} & f_{n-n_{\mathfrak{c}},n} & -\frac{1}{2}\lambda_{1,n} & \mu_{1,n} & \cdots & -\frac{1}{2}\lambda_{n_{\mathfrak{c}},n} & \mu_{n_{\mathfrak{c}},n} \end{bmatrix}. \tag{2.39}$$

Thus we explicitly express the Z_j, Z'_j in terms of the variables of Γ, using $Z = (\mathrm{SY})^{-1}\Gamma$:

$$Z_j = \sum_{k=1}^{n-n_{\mathrm{c}}} (d'_{k,j}\Gamma_k - f'_{k,j} G_k) + \tfrac{1}{2}\sum_{k=1}^{n_{\mathrm{c}}} (\lambda'_{k,j} M_k - \mu'_{k,j} L_k), \tag{2.40}$$

$$Z'_j = \sum_{k=1}^{n-n_{\mathrm{c}}} (-d_{k,j}\Gamma_k + f_{k,j} G_k) + \tfrac{1}{2}\sum_{k=1}^{n_{\mathrm{c}}} (-\lambda_{k,j} M_k + \mu_{k,j} L_k). \tag{2.41}$$

2.7 Inverting without inverting

It is possible to explicitly compute a symplectic matrix SY, then invert it, express the variables Z in terms of the variables Γ by (2.40)–(2.41), and then solve to obtain the A-polynomial. However, we now show that we can perform this calculation without ever having to find SY or its inverse SY^{-1} explicitly — *provided* that we can find a certain sign term.

To see why this should be the case, note the following preliminary observation. Equations (2.40)–(2.41) express Z_j and Z'_j in terms of the Γ_k, G_k, M_i and L_i. The coefficients of the Γ_k, M_i and L_i are numbers which appear in the Neumann–Zagier matrix. The only coefficients which do not appear in NZ are the coefficients of the G_k. But the gluing equations (2.38) say $G_k = i\pi(2 - c_k)$, so upon exponentiation these terms only contribute a sign. In other words, up to sign, all the information we need to write the Z_j in terms of the variables Γ_k, G_k, L_i, M_i is already in the Neumann–Zagier matrix.

To implement this, observe that the matrix $-J(\mathrm{NZ}^\flat)^T$ shares many columns with SY^{-1}:

$$-J(\mathrm{NZ}^\flat)^T = \begin{bmatrix} d'_{1,1} & d'_{2,1} & \cdots & d'_{n-n_{\mathrm{c}},1} & \mu'_{1,1} & \lambda'_{1,1} & \cdots & \mu'_{n_{\mathrm{c}},1} & \lambda'_{n_{\mathrm{c}},1} \\ -d_{1,1} & -d_{2,1} & \cdots & -d_{n-n_{\mathrm{c}},1} & -\mu_{1,1} & -\lambda_{1,1} & \cdots & -\mu_{n_{\mathrm{c}},1} & -\lambda_{n_{\mathrm{c}},1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ d'_{1,n} & d'_{2,n} & \cdots & d'_{n-n_{\mathrm{c}},n} & \mu'_{1,n} & \lambda'_{1,n} & \cdots & \mu'_{n_{\mathrm{c}},n} & \lambda'_{n_{\mathrm{c}},n} \\ -d_{1,n} & -d_{2,n} & \cdots & -d_{n-n_{\mathrm{c}},n} & -\mu_{1,n} & -\lambda_{1,n} & \cdots & -\mu_{n_{\mathrm{c}},n} & -\lambda_{n_{\mathrm{c}},n} \end{bmatrix}. \tag{2.42}$$

In particular, for any quantities $A_1, \dots, A_{n-n_{\mathrm{c}}}, A_1^\lambda, A_1^\mu, \dots, A_{n_{\mathrm{c}}}^\lambda, A_{n_{\mathrm{c}}}^\mu$,

$$\begin{aligned} &\mathrm{SY}^{-1}\begin{bmatrix} A_1 & 0 & A_2 & 0 & \dots & A_{n-n_{\mathrm{c}}} & 0 & A_1^\lambda & A_1^\mu & \dots & A_{n_{\mathrm{c}}}^\lambda & A_{n_{\mathrm{c}}}^\mu \end{bmatrix}^T \\ &\qquad = -J(\mathrm{NZ}^\flat)^T \begin{bmatrix} A_1 & A_2 & \dots & A_{n-n_{\mathrm{c}}} & -A_1^\mu & \tfrac{1}{2}A_1^\lambda & \dots & -A_{n_{\mathrm{c}}}^\mu & \tfrac{1}{2}A_{n_{\mathrm{c}}}^\lambda \end{bmatrix}^T. \end{aligned}$$

Splitting up the Γ_k and G_k terms, using Definition 2.36 and informed by the gluing and cusp equations (2.38), we obtain

$$Z = \mathrm{SY}^{-1} \cdot \Gamma = -J(\mathrm{NZ}^\flat)^T \overline{\Gamma} + \mathrm{SY}^{-1}\overline{G}, \tag{2.43}$$

where $\overline{\Gamma}$ is the vector

$$\overline{\Gamma} = \begin{bmatrix} \Gamma_1, \dots, \Gamma_{n-n_{\mathrm{c}}}, -\tfrac{1}{2}\log \ell_1, \tfrac{1}{2}\log m_1, \dots, -\tfrac{1}{2}\log \ell_{n_{\mathrm{c}}}, \tfrac{1}{2}\log m_{n_{\mathrm{c}}} \end{bmatrix}^T$$

and $\overline{G}$ is

$$\left[0, G_1, \dots, 0, G_{n-n_{\mathfrak{c}}}, (M_1 - \log m_1), \tfrac{1}{2}(L_1 - \log \ell_1), \dots, (M_{n_{\mathfrak{c}}} - \log m_{n_{\mathfrak{c}}}), \tfrac{1}{2}(L_{n_{\mathfrak{c}}} - \log \ell_{n_{\mathfrak{c}}})\right]^T$$

The first term $-J(\mathrm{NZ}^\flat)^T\,\overline{\Gamma}$ of (2.43) only involves NZ. The final vector $\overline{G}$ consists of the precise quantities which are fixed to be constants by the gluing and completeness equations (2.38). Indeed, (2.38) says precisely that the final vector in equation (2.43) is a vector of constants essentially identical in content to $\pi i C^\flat$. We define

$$C^\# = \left[0, 2 - c_1, 0, 2 - c_2, \dots, 0, 2 - c_{n-n_{\mathfrak{c}}}, -c_1^{\mathfrak{m}}, -\tfrac{1}{2}c_1^{\mathfrak{l}}, \dots, -c_{n_{\mathfrak{c}}}^{\mathfrak{m}}, -\tfrac{1}{2}c_{n_{\mathfrak{c}}}^{\mathfrak{l}}\right]^T,$$

which is $C^\flat$, with some zeroes inserted, and some factors of one half. So the final vector in (2.43) is set to $\pi i C^\#$, and we obtain the following.

Proposition 2.44 *Given a hyperbolic triangulation, labelled so that its Neumann–Zagier matrix* NZ *has rows* $R_1^G, \dots, R_{n-n_{\mathfrak{c}}}^G$ *linearly independent, and* SY *a matrix defining the variables* Γ, *the logarithmic gluing and cusp equations are equivalent to*

(2.45)
$$Z = (-J)(\mathrm{NZ}^\flat)^T\,\overline{\Gamma} + \pi i\, \mathrm{SY}^{-1} C^\#. \qquad \square$$

Once we find a vector $B = \mathrm{SY}^{-1} C^\#$, Proposition 2.44 allows us to express the Z_j and Z_j' in terms of the variables $\Gamma_1, \dots, \Gamma_{n-1}$, and the holonomies ℓ_k, m_k of the longitudes and meridians, using only information already available in the Neumann–Zagier matrix. There is no need to find the extra vectors R_k^Γ of the symplectic basis, or the matrix SY. If in addition B is an *integer* vector, then when we exponentiate (2.45) to obtain the tetrahedron parameters $z_j = e^{Z_j}$ and $z_j' = e^{Z_j'}$, B determines a sign. Hence we refer to this term as a sign term.

The approach outlined above may sound paradoxical: we avoid calculating the symplectic matrix SY, by finding a vector $B = \mathrm{SY}^{-1} C^\#$. This seems to involve the symplectic matrix SY anyway! However, in the next section we show we can find B by solving a simpler equation, involving only the Neumann–Zagier matrix, and then *choose* SY so that $B = \mathrm{SY}^{-1} C^\#$. That is, we may use the flexibility in choosing R_k^Γ of Remark 2.28 to find appropriate SY.

2.8 The sign term

We now demonstrate the existence of an SY and an integer vector B satisfying $\mathrm{SY} \cdot B = C^\#$.

The rows of the matrix equation $\mathrm{SY} \cdot B = C^\#$ are

(2.46)
$$R_k^\Gamma \cdot B = 0 \qquad \text{for } k = 1, \dots, n - n_{\mathfrak{c}},$$

(2.47)
$$R_k^G \cdot B = 2 - c_k \qquad \text{for } k = 1, \dots, n - n_{\mathfrak{c}},$$

(2.48)
$$R_k^{\mathfrak{m}} \cdot B = -c_k^{\mathfrak{m}}, \quad R_k^{\mathfrak{l}} \cdot B = -c_k^{\mathfrak{l}} \qquad \text{for } k = 1, \dots, n_{\mathfrak{c}}.$$

Equations (2.47)–(2.48) are exactly the equations in the rows of a matrix equation with $\mathrm{NZ}^\flat$:

(2.49)
$$\mathrm{NZ}^\flat \cdot B = C^\flat.$$

This equation has been studied by Neumann; it is known to always have an integer solution.

Theorem 2.50 (Neumann [32, Theorem 2.4])

(i) *There exists an integer vector* B *satisfying* $\mathrm{NZ} \cdot B = C$.

(ii) *Given an integer vector* B_0 *such that* $\mathrm{NZ} \cdot B_0 = C$, *the set of integer solutions to* $\mathrm{NZ} \cdot B = C$ *includes*

$$B_0 + \mathrm{Span}_{\mathbb{Z}}(JR_1^G, \dots, JR_n^G) = \left\{ B_0 + \sum_{k=1}^{n} a_k JR_k^G \,\middle|\, a_1, \dots, a_n \in \mathbb{Z} \right\}.$$

Neumann's result is more precise, incorporating a parity condition on B not needed here. Additionally, we will not need part (ii) of the theorem until later, but we state it now. Note that, by taking a subset of the rows, or equations, $\mathrm{NZ} \cdot B = C$ implies $\mathrm{NZ}^\flat \cdot B = C^\flat$.

In order to solve $\mathrm{SY} \cdot B = C^\#$, it remains to satisfy the equations (2.46). As discussed above, we do this not by adjusting B, but by judicious choice of the vectors R_k^Γ, and hence the matrix SY. Recall from Section 2.4 that there is substantial freedom in choosing the vectors R_k^Γ. But first we deal with a technical condition on the triangulation, which we need for the argument. Recall $c_k = \sum_{j=1}^n c_{k,j}$ (Definition 2.17), where $c_{k,j}$ is the number of c-edges of the tetrahedron Δ_j identified to edge E_k (Definition 2.6). So c_k is just the number of c-edges of tetrahedra identified to E_k.

Lemma 2.51 *Any triangulation of* M *has a labelling such that*

(i) *its Neumann–Zagier matrix* NZ *has rows* $R_1^G, \dots, R_{n-n_\mathfrak{c}}^G$ *linearly independent*, *and*

(ii) *there exists* $k \in \{1, \dots, n - n_\mathfrak{c}\}$ *with* $c_k \neq 2$.

In other words, the conclusion of the lemma requires that some edge be incident to a number of c-edges other than 2. In fact, we will see that one can start from any labelled triangulation, and it suffices to relabel the vertices of at most one tetrahedron, and possibly reorder some edges. Moreover, we can choose any edge E_k with nonzero R_k^G, and adjust so that this particular edge is incident to $c_k \neq 2$ c-edges.

The proof of Lemma 2.51 requires that $n > n_\mathfrak{c}$. In fact, Adams and Sherman [1] proved that $n \geq 2n_\mathfrak{c}$ for any finite volume orientable hyperbolic 3-manifold with $n_\mathfrak{c}$ cusps.

Proof Take a labelled triangulation $\mathcal{T}$ of M. Choose some $k \in \{1, \dots, n\}$ such that R_k^G is nonzero. (Such k certainly exists since the R_k^G span a space of rank $n - n_\mathfrak{c} \geq 1$.) We claim that if $c_k = 2$, then $\mathcal{T}$ can be relabelled so that $c_k \neq 2$.

Let Δ_t be a tetrahedron of $\mathcal{T}$. The relabellings of Δ_t have the effect of cyclically permuting the a-, b- and c-edges, and hence cyclically permuting the triple $(a_{k,t}, b_{k,t}, c_{k,t})$; however other terms $c_{k,j}$ in the sum for c_k are unchanged. Hence, if one of $a_{k,t}$ or $b_{k,t}$ is not equal to $c_{k,t}$, then a relabelling of Δ_t will change c_k to a distinct value, not 2, as desired. Otherwise, all relabellings of Δ_t leave $c_k = 2$, and we have $a_{k,t} = b_{k,t} = c_{k,t}$, so $d_{k,t} = d'_{k,t} = 0$ (Definition 2.17).

The above argument applies to any tetrahedron Δ_t of $\mathcal{T}$. Thus, if every relabelling of any single tetrahedron leaves $c_k = 2$, then the numbers $d_{k,t} = d'_{k,t} = 0$ for all $t \in \{1, \dots, n\}$. But these are precisely the entries in the vector R_k^G forming a row of $\mathrm{NZ}^\flat$, so $R_k^G = 0$, contradicting $R_k^G \neq 0$ above. This contradiction proves the claim. Moreover, after relabelling the tetrahedron, there still exists $t \in \{1, \dots, n\}$ such that $a_{k,t}, b_{k,t}, c_{k,t}$ are not all equal, and hence R_k^G is not zero.

Thus, there exists a relabelling of a single tetrahedron that makes $c_k \neq 2$, and R_k^G remains nonzero. Call the resulting labelled triangulation $\mathcal{T}'$ and Neumann–Zagier matrix NZ'. Now by Theorem 2.27(iv), the first n row vectors of NZ' span an $(n{-}n_{\mathfrak{c}})$-dimensional space. Hence we may relabel the edges so that the edges labelled $1, \dots, n - n_{\mathfrak{c}}$ have linearly independent row vectors, and our chosen edge is among them. This relabelling satisfies the lemma. □

For a triangulation as in Lemma 2.51, the nonzero entry of $C^\flat$ provides the leverage to make a choice of vectors R_k^Γ so that they satisfy (2.46).

Lemma 2.52 *Suppose that $\mathcal{T}$ is labelled to satisfy Lemma 2.51. Let $B \in \mathbb{Z}^{2n}$ be a vector satisfying $\mathrm{NZ}^\flat \cdot B = C^\flat$. Then there exist vectors $R_1^\Gamma, \dots, R_{n-n_{\mathfrak{c}}}^\Gamma$ in $\mathbb{Q}^{2n}$ such that*

(i) $\bigl(R_1^\Gamma, R_1^G, \dots, R_{n-n_{\mathfrak{c}}}^\Gamma, R_{n-n_{\mathfrak{c}}}^G, R_1^{\mathfrak{m}}, \tfrac{1}{2}R_1^{\mathfrak{l}}, \dots, R_{n_{\mathfrak{c}}}^{\mathfrak{m}}, \tfrac{1}{2}R_{n_{\mathfrak{c}}}^{\mathfrak{l}}\bigr)$ *forms a symplectic basis*, *and*

(ii) *for all $j \in \{1, \dots, n - n_{\mathfrak{c}}\}$ we have $R_j^\Gamma \cdot B = 0$.*

Proof We start from arbitrary choices of the $R_k^\Gamma \in \mathbb{Q}^{2n}$ such that

$$\bigl(R_1^\Gamma, R_1^G, \dots, R_{n-n_{\mathfrak{c}}}^\Gamma, R_{n-n_{\mathfrak{c}}}^G, R_1^{\mathfrak{m}}, \tfrac{1}{2}R_1^{\mathfrak{l}}, \dots, R_{n_{\mathfrak{c}}}^{\mathfrak{m}}, \tfrac{1}{2}R_{n_{\mathfrak{c}}}^{\mathfrak{l}}\bigr)$$

is a symplectic basis.

Lemma 2.26 allows us to adjust the R_k^Γ, without changing any R_k^G, $R_j^{\mathfrak{m}}$ or $R_j^{\mathfrak{l}}$, so that we still have a symplectic basis. In particular, we may make the following modifications:

(i) For $j \neq k \in \{1, \dots, n - n_{\mathfrak{c}}\}$, and $a \in \mathbb{R}$, map $R_j^\Gamma \mapsto R_j^\Gamma + aR_k^G$, $R_k^\Gamma \mapsto R_k^\Gamma + aR_j^G$.

(ii) Take $j \in \{1, \dots, n - n_{\mathfrak{c}}\}$ and $a \in \mathbb{R}$, and map $R_j^\Gamma \mapsto R_j^\Gamma + aR_j^G$.

Let $R_j^\Gamma \cdot B = a_j$. We will adjust the R_j^Γ until all $a_j = 0$.

We claim there exists a $k \in \{1, \dots, n - n_{\mathfrak{c}}\}$ such that $R_k^G \cdot B \neq 0$. Indeed, as $\mathcal{T}$ satisfies Lemma 2.51, there exists a $k \in \{1, \dots, n - n_{\mathfrak{c}}\}$ such that $c_k \neq 2$. Then the k^{th} row of the equation $\mathrm{NZ}^\flat \cdot B = C^\flat$ says that $\alpha := R_k^G \cdot B = 2 - c_k$, which is nonzero as claimed.

First, modify R_k^Γ by (ii), replacing R_k^Γ with $(R_k^\Gamma)' = R_k^\Gamma - (a_k/\alpha)R_k^G$. Then

$$(R_k^\Gamma)' \cdot B = R_k^\Gamma \cdot B - \frac{a_k}{\alpha} R_k^G \cdot B = 0.$$

Thus the modification makes $a_k = 0$; the other a_j are unchanged.

Now consider $j \neq k$. If $R_j^G \cdot B \neq 0$, modify R_j^Γ by (ii) to set $a_j = 0$. Otherwise, $R_j^G \cdot B = 0$ and modify R_j^Γ and R_k^Γ by (i), replacing them with

$$(R_j^\Gamma)' = R_j^\Gamma - \frac{a_j}{\alpha} R_k^G \quad \text{and} \quad (R_k^\Gamma)' = R_k^\Gamma - \frac{a_j}{\alpha} R_j^G,$$

respectively. Then

$$(R_j^\Gamma)' \cdot B = R_j^\Gamma \cdot B - \frac{a_j}{\alpha} R_k^G \cdot B = 0 \quad \text{and} \quad (R_k^\Gamma)' \cdot B = R_k^\Gamma \cdot B - \frac{a_j}{\alpha} R_j^G \cdot B = a_k = 0.$$

Again the effect is to set $a_j = 0$ and leave the other a_i unchanged.

Modifying R_j^Γ in this way for each $j \neq k$, we obtain the desired vectors. □

We summarise the result of this section in the following proposition.

Proposition 2.53 *Let $\mathcal{T}$ be a hyperbolic triangulation labelled to satisfy Lemma 2.51. Let B be an integer vector such that* $\mathrm{NZ}^\flat \cdot B = C^\flat$ (*such a vector exists by Theorem 2.50*). *Then there exists a symplectic matrix* SY *defining variables* Γ, *such that the logarithmic gluing and cusp equations are equivalent to the equation*

$$Z = (-J)(\mathrm{NZ}^\flat)^T \overline{\Gamma} + \pi i B. \tag{2.54}$$ □

We have now realised our claim of "inverting without inverting". Proposition 2.53 allows us to convert the variables Z_i, Z_i' into the variables Γ_i, together with the cusp holonomies ℓ_i, m_i, without having to actually calculate the vectors R_i^Γ or the matrix SY! The only information we need is the Neumann–Zagier matrix NZ, and the integer vector B such that $\mathrm{NZ}^\flat \cdot B = C^\flat$.

2.9 The A-polynomial from gluing equations and from Ptolemy equations

Suppose that $n_{\mathfrak{c}} = 1$, we have a labelled triangulation $\mathcal{T}$ satisfying Lemma 2.51, and a vector $B = (B_1, B_1', \dots, B_n, B_n')^T$ such that $\mathrm{NZ}^\flat \cdot B = C^\flat$.

Proposition 2.53 converts the logarithmic gluing and cusp equations — linear equations — into the variables $\Gamma_1, \dots, \Gamma_{n-1}$, together with the cusp holonomies m, ℓ. We now convert the nonlinear equations (2.5) into these variables.

We first convert to the exponentiated variables z_j. Let $\gamma_j = e^{\Gamma_j}$. Using (2.54), and the known form of $(-J)(\mathrm{NZ}^\flat)^T$ from (2.42), we obtain

$$z_j = (-1)^{B_j} \ell^{-\mu_j'/2} m^{\lambda_j'/2} \prod_{k=1}^{n-1} \gamma_k^{d_{k,j}'}, \quad z_j' = (-1)^{B_j'} \ell^{\mu_j/2} m^{-\lambda_j/2} \prod_{k=1}^{n-1} \gamma_k^{-d_{k,j}}. \tag{2.55}$$

Then the nonlinear equation (2.5) for the tetrahedron Δ_j becomes

$$(-1)^{B_j} \ell^{-\mu_j'/2} m^{\lambda_j'/2} \prod_{k=1}^{n-1} \gamma_k^{d_{k,j}'} + (-1)^{B_j'} \ell^{-\mu_j/2} m^{\lambda_j/2} \prod_{k=1}^{n-1} \gamma_k^{d_{k,j}} - 1 = 0.$$

Since $d_{k,j} = a_{k,j} - c_{k,j}$ and $d'_{k,j} = b_{k,j} - c_{k,j}$ (Definition 2.17), we may multiply through by $\gamma^{c_{k,j}}$; then the exponents become the incidence numbers $a_{k,j}, b_{k,j}, c_{k,j}$ of the various types of edges of tetrahedra with edges of the triangulation (Definition 2.6):

$$(-1)^{B_j} \ell^{-\mu'_j/2} m^{\lambda'_j/2} \prod_{k=1}^{n-1} \gamma_k^{b_{k,j}} + (-1)^{B'_j} \ell^{-\mu_j/2} m^{\lambda_j/2} \prod_{k=1}^{n-1} \gamma_k^{a_{k,j}} - \prod_{k=1}^{n-1} \gamma_k^{c_{k,j}} = 0. \tag{2.56}$$

Each product in the above expression is simpler than it looks: it is a polynomial of total degree at most 2 in the γ_k, by Lemma 2.7! The product $\prod_{k=1}^{n-1} \gamma_k^{a_{k,j}}$ has j fixed, referring to the tetrahedron Δ_j. The product is over the various edges E_k of the triangulation; the exponent $a_{k,j}$ is the incidence number of the a-edges of Δ_j with the edge E_k. But Δ_j only has two a-edges, so at most two $a_{k,j}$ are nonzero, and the $a_{k,j}$ sum to 2 as in (2.8).

Recall the notation $j(\mu\nu)$ of Definition 2.3. For fixed j, the only nonzero $a_{k,j}$ are $a_{j(01),j}$ and $a_{j(23),j}$ (which may be the same term). Thus the product $\prod_{k=1}^{n-1} \gamma_k^{a_{k,j}}$ is equal to the product of $\gamma_{j(01)}$ and $\gamma_{j(23)}$, with the caveat that γ_n does not appear in the product. Indeed, in Definition 2.36 we only define $\Gamma_1, \dots, \Gamma_{n-1}$, so only $\gamma_1, \dots, \gamma_{n-1}$ are defined. However, it is worthwhile to introduce γ_n as a formal variable.

Definition 2.57 Let $\mathcal{T}$ be a labelled triangulation of a 3-manifold with one cusp, and let B be an integer vector such that $\mathrm{NZ}^\flat \cdot B = C^\flat$. The *Ptolemy equation* of the tetrahedron Δ_j is

$$(-1)^{B'_j} \ell^{-\mu_j/2} m^{\lambda_j/2} \gamma_{j(01)} \gamma_{j(23)} + (-1)^{B_j} \ell^{-\mu'_j/2} m^{\lambda'_j/2} \gamma_{j(02)} \gamma_{j(13)} - \gamma_{j(03)} \gamma_{j(12)} = 0.$$

The *Ptolemy equations* of $\mathcal{T}$ consist of Ptolemy equations for each tetrahedron of $\mathcal{T}$.

Equation (2.56) is the Ptolemy equation for Δ_j, with the formal variable γ_n set to 1.

Let us now put the work of this section together.

Theorem 2.58 *Let $\mathcal{T}$ be a hyperbolic triangulation of a one-cusped M, labelled to satisfy Lemma 2.51. When we solve the system of Ptolemy equations of $\mathcal{T}$ in terms of m and ℓ, setting $\gamma_n = 1$ and eliminating the variables $\gamma_1, \dots, \gamma_{n-1}$, we obtain a factor of the* $\mathrm{PSL}(2, \mathbb{C})$ *A-polynomial, which is also the polynomial of Theorem 2.13.*

(Note that the polynomial described here, arising by eliminating variables from a system of equations, is only defined up to multiplication by units, and the equality of polynomials here should be interpreted accordingly.)

Proof Theorem 2.13 tells us that solving equations (2.4)–(2.5), (2.10) and (2.12) for m and ℓ, eliminating the variables z_j, z'_j, z''_j, yields a factor of the $\mathrm{PSL}(2, \mathbb{C})$ A-polynomial. By Lemma 2.23, a solution of the logarithmic gluing and cusp equations, after exponentiation, gives a solution of (2.4), (2.10) and (2.12); and conversely any solution of (2.4), (2.10) and (2.12) has a logarithm solving the logarithmic gluing and cusp equations.

By Proposition 2.53, after introducing appropriate B and SY and variables Γ, which all exist, the logarithmic gluing and cusp equations are equivalent to (2.54). Exponentiating gives us that the equations (2.55) imply (2.4), (2.10) and (2.12). Combining these with (2.5) yields the equations (2.56), one for each tetrahedron. Therefore, any solution of the equations (2.56) for $\gamma_1, \dots, \gamma_{n-1}, m, \ell$ yields a solution of (2.4)–(2.5), (2.10) and (2.12). Conversely, any solution of (2.4)–(2.5), (2.10) and (2.12) has a logarithm satisfying the logarithmic gluing and cusp equations, hence yields solutions of (2.56).

Thus the pairs (ℓ, m) arising in solutions of (2.4)–(2.5), (2.10) and (2.12) are those arising in solutions of (2.56). The latter equations are the Ptolemy equations of $\mathcal{T}$ with γ_n set to 1. Thus, the (ℓ, m) satisfying the polynomial obtained by solving the Ptolemy equations with $\gamma_n = 1$ are also those satisfying the polynomial of Theorem 2.13. □

Corollary 2.59 *With $\mathcal{T}$ and M as above, let $A_0(\mathcal{L}, \mathcal{M})$ denote the factor of the* $\mathrm{SL}(2, \mathbb{C})$ *A-polynomial describing hyperbolic structures on $\mathcal{T}$. Letting $\mathcal{L} = \ell^{1/2}$ and $\mathcal{M} = m^{1/2}$ and solving the Ptolemy equations with $\gamma_n = 1$ as above, we obtain a polynomial in $\mathcal{M}$ and $\mathcal{L}$ which contains a factor either $A_0(\mathcal{L}, \mathcal{M})$ or $A_0(-\mathcal{L}, \mathcal{M})$.*

Proof Suppose $(\mathcal{L}, \mathcal{M})$ lies in the zero set of the factor of the $\mathrm{SL}(2, \mathbb{C})$ A-polynomial describing hyperbolic structures on $\mathcal{T}$. Then there is a representation $\pi_1(M) \to \mathrm{SL}(2, \mathbb{C})$ sending the longitude to a matrix with eigenvalues $\mathcal{L}, \mathcal{L}^{-1}$ and the meridian to a matrix with eigenvalues $\mathcal{M}, \mathcal{M}^{-1}$. Projecting to $\mathrm{PSL}(2, \mathbb{C})$ we have the holonomy of a hyperbolic structure on $\mathcal{T}$ whose cusp holonomies are given by $\mathcal{L}^2 = \ell$ and $\mathcal{M}^2 = m$, respectively. Hence (ℓ, m) and the tetrahedron parameters of the hyperbolic structure solve the gluing and cusp equations $\mathcal{T}$, and hence satisfy the polynomial of Theorem 2.58. □

3 Dehn fillings and triangulations

3.1 Layered solid tori

Suppose we have a triangulation where a cusp $\mathfrak{c}_1$ meets exactly two tetrahedra $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$ in exactly one ideal vertex per tetrahedron. (We show in Appendix A, Proposition A.1, that such a triangulation can be constructed for quite general manifolds with two or more cusps.) These two tetrahedra together give a triangulation of a manifold homeomorphic to $T^2 \times [0, \infty)$ with a single point removed from $T^2 \times \{0\}$. The boundary component $T^2 \times \{0\}$ of $\Delta_1^{\mathfrak{c}} \cup \Delta_2^{\mathfrak{c}}$ is a punctured torus, triangulated by the two ideal triangles of $\partial \Delta_1^{\mathfrak{c}}$ and $\partial \Delta_2^{\mathfrak{c}}$ that do not meet the cusp $\mathfrak{c}_1$. We will remove $\Delta_1^{\mathfrak{c}} \cup \Delta_2^{\mathfrak{c}}$ from our triangulated manifold, and obtain a space with boundary a punctured torus, triangulated by the same two ideal triangles. We will then replace $\Delta_1^{\mathfrak{c}} \cup \Delta_2^{\mathfrak{c}}$ by a solid torus with a triangulation such that the boundary is a triangulated once-punctured torus. This will give a triangulation of the Dehn filling.

A *layered solid torus* is a triangulation of a solid torus, first described by Jaco and Rubinstein [30]; see also [24]. When working with ideal triangulations, as in our situation, the boundary of a layered solid torus consists of two ideal triangles whose union is a triangulation of a punctured torus. The space of all two-triangle triangulations of punctured tori is described by the Farey graph. A layered solid torus can be built using the combinatorics of the Farey graph.

Recall first the construction of the Farey triangulation of $\mathbb{H}^2$. We view $\mathbb{H}^2$ in the disc model, with antipodal points $1/0$ and $0/1$ in $\partial\mathbb{H}^2$ lying on a horizontal line through the centre of the disc, and $1/1$ at the north pole, $-1/1$ at the south pole. Two points a/b and c/d in $\mathbb{Q}\cup\{\infty\}\subset\partial\mathbb{H}^2$ have distance measured by

$$\iota(a/b, c/d) = |ad - bc|.$$

Here $\iota(\cdot,\cdot)$ denotes geometric intersection number of slopes on a punctured torus. We draw an ideal geodesic between each pair a/b, c/d with $|ad-bc| = 1$. This gives the *Farey triangulation*. The dual graph of the Farey triangulation is an infinite trivalent tree, which we denote by $\mathcal{F}$.

Any triangulation of a once-punctured torus consists of three slopes on the boundary of the torus, with each pair of slopes having geometric intersection number 1. Denote the slopes by f, g, h. This triple determines a triangle in the Farey triangulation. Moving across an edge (f, g) of the Farey triangulation, we arrive at another triangle whose vertices include f and g; but the slope h is replaced with some other slope h'. This corresponds to changing the triangulation on the punctured torus, replacing lines of slope h with lines of slope h'.

When we wish to perform a Dehn filling by attaching a solid torus to a triangulated once-punctured torus, there are four important slopes involved. Three of the slopes are the slopes of the initial triangulation of the once-punctured solid torus. For example, these might be $0/1$, $1/0$, and $1/1$. We will typically denote the slopes by (f, g, h). These determine an initial triangle T_0 in the Farey graph. The other important slope is r, the slope of the Dehn filling.

Now consider the geodesic in $\mathbb{H}^2$ from the centre of T_0 to the slope $r\subset\partial\mathbb{H}^2$. This geodesic passes through a sequence of distinct triangles in the Farey graph, which we denote $T_0, T_1, \dots, T_{N+1}$. Each T_{j+1} is adjacent to T_j. We regard this as a walk or voyage through the triangulation; more precisely, we can regard $T_0, \dots, T_N$ as forming an oriented path in the dual tree $\mathcal{F}$ without backtracking. The slope r appears as a vertex of the final triangle T_{N+1}, but not in any earlier triangle.

We build the layered solid torus by stacking tetrahedra $\Delta_0, \Delta_1, \dots$ onto the punctured torus, replacing one set of slopes T_0 with another T_1, then another T_2, and so on. That is, two consecutive punctured tori always have two slopes in common and two that differ by a diagonal exchange. The diagonal exchange is obtained in three-dimensions by layering a tetrahedron onto a given punctured torus such that the diagonal on one side matches the diagonal to be replaced. See Figure 2.

For each edge crossed in the path from T_0 to T_N, layer on a tetrahedron, obtaining a collection of tetrahedra homotopy equivalent to $T^2\times I$. After gluing k tetrahedra $\Delta_0, \dots, \Delta_{k-1}$, the side $T^2\times\{0\}$ has the triangulation whose slopes are given by T_0, and the side $T^2\times\{1\}$ has slopes given by T_k. Two of

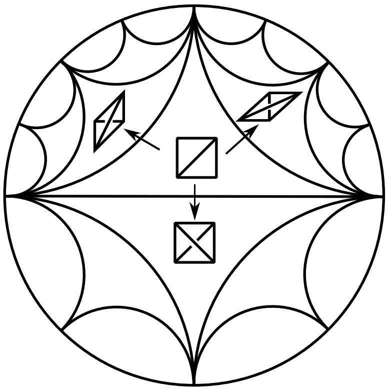


Figure 2: Constructing a layered solid torus.

the faces of Δ_{k-1} are glued to triangles of the previous layer, with slopes given by T_{k-1}, and the other two faces form a triangulation of the "top" boundary $T^2 \times \{1\}$; this triangulation has slopes given by T_k. Continue until $k = N$, obtaining a triangulated complex consisting of N tetrahedra $\Delta_0, \dots, \Delta_{N-1}$, with boundary consisting of two once-punctured tori, one triangulated by T_0 and the other by T_N.

Recall we are trying to obtain a triangulation of a solid torus for which the slope r is homotopically trivial. Note that r is a diagonal of the triangulation T_N. That is, a single diagonal exchange replaces the triangulation T_N with T_{N+1}; and T_{N+1} is a triangulation consisting of two slopes s and t in common with T_N, together with the slope r, which cuts across a slope r' of T_N. To homotopically kill the slope r, fold the two triangles of T_N across the diagonal slope r', as in Figure 3. Gluing the two triangles on one boundary component of $T^2 \times I$ in this manner gives a quotient that is homeomorphic to a solid torus, with boundary still triangulated by T_0. Inside, the slopes s and t are identified. The slope r has been folded onto itself, meaning it is now homotopically trivial. Note that N is the number of ideal tetrahedra in the layered solid torus.

There are two exceptional cases. If $N = 0$ then no tetrahedra are layered to form a layered solid torus. Instead, we fold across existing faces to homotopically "kill" the slope r that lies in one of the three Farey triangles adjacent to (f, g, h). This can be considered as attaching a degenerate layered solid torus, consisting of a single face, folded into a Möbius band.

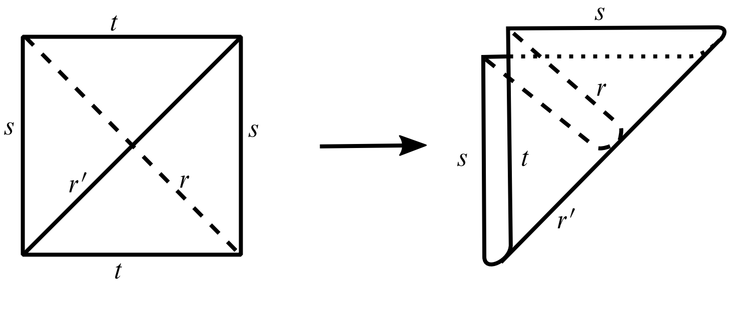


Figure 3: Folding makes the diagonal slope r homotopically trivial.

There is one other *extra-exceptional* case. In this case, the slope r is one of f, g, h. We can triangulate the Dehn filling: for example we can attach a tetrahedron covering the edge corresponding to r, performing a diagonal exchange on the once-punctured torus triangulation, then immediately fold the two new faces across the diagonal, creating an edge with valence one. This case will be ignored in the arguments below.

3.2 Notation for a voyage in the Farey triangulation

We now give notation to keep closer track of the slopes obtained at each stage of the construction of a layered solid torus.

As we have seen, each tetrahedron Δ_{k-1} replaces one set of slopes with another; the set of slopes corresponding to the triangle T_{k-1} in the Farey triangulation is replaced with the set of slopes with the triangle T_k. Thus, we associate to Δ_{k-1} an oriented edge of the dual tree $\mathcal{F}$ of the Farey triangulation, from T_{k-1} to T_k.

As $\mathcal{F}$ is an infinite trivalent tree, at each stage of a path in $\mathcal{F}$ without backtracking, after we begin and before we stop, there are two choices: turning left or right. As is standard, we denote these choices by L and R. Note that the choice of L or R is not well-defined when moving from T_0 to T_1, but thereafter the choice of L or R is well-defined. Thus, to the path $T_0, T_1, \dots, T_{N+1}$ in $\mathcal{F}$, there is a word of length N in the letters {L,R}. We call this word W. The j^{th} letter of W corresponds to the choice of L or R when moving from T_j to T_{j+1}, which also corresponds to adding tetrahedron Δ_j.

As we voyage at each stage from T_k to T_{k+1}, we pass through an edge e_k of the Farey triangulation (dual to the corresponding edge of $\mathcal{F}$), which has one endpoint to our left (port) and one to our right (starboard).[1] We leave behind an old slope, one of the slopes of T_k, namely the one not occurring in T_{k+1}. And we head towards a new slope, namely the slope of T_{k+1} which is not one of T_k.

Definition 3.1 As we pass from T_k to T_{k+1}, across the edge e_k, the slope corresponding to

(i) the endpoint of e_k to our left is denoted p_k (for port);

(ii) the endpoint of e_k to our right is denoted s_k (for starboard);

(iii) the vertex of $T_k \setminus T_{k+1}$ is denoted o_k (old);

(iv) the vertex of $T_{k+1} \setminus T_k$ is denoted h_k (heading).

Thus, the initial slopes $\{f, g, h\}$ are given by $\{o_0, s_0, p_0\}$ in some order, and the final, or Dehn filling slope is given by $r = h_N$. Adding the tetrahedron Δ_{k-1}, we pass from T_{k-1} to T_k, so the edges of Δ_{k-1} correspond to slopes $p_{k-1}, s_{k-1}, o_{k-1}, h_{k-1}$.

Lemma 3.2 (i) *If the i^{th} letter of W is an L, then $o_i = s_{i-1},\ p_i = p_{i-1},\ s_i = h_{i-1}$.*

(ii) *If the i^{th} letter of W is an R, then $o_i = p_{i-1},\ p_i = h_{i-1},\ s_i = s_{i-1}$.*

[1] As "left" and "right" are used in the context or the previous paragraph, we use the nautical terminology here.

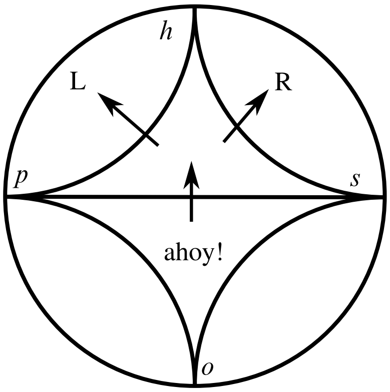


Figure 4: Labels on the slopes in the Farey graph.

Proof This is immediate upon inspecting Figure 4. If we tack left as we proceed from T_{i-1} through T_i to T_{i+1}, then we wheel around the port side; our previous heading is now to starboard, and we leave starboard behind. Similarly for turning right. □

So ye sail, me hearty, until ye arrive at ye last tetrahedron Δ_{N-1}, proceeding from triangle T_{N-1} into T_N, with associated slopes $o_{N-1}, s_{N-1}, h_{N-1}, p_{N-1}$. We have made $N-1$ choices of left or right, L or R. The boundary $T^2 \times \{1\}$ of the layered solid torus constructed to this point has triangulation with slopes given by T_N, ie with slopes $p_{N-1}, s_{N-1}, h_{N-1}$.

The final choice of L or R takes us from triangle T_N into triangle T_{N+1}, whose final heading h_N is the Dehn filling slope r. This final L or R determines how we fold up the two triangles with slopes T_N on the boundary of Δ_N. As discussed in Section 3.1, we fold the two triangular faces of the boundary torus together along an edge, so as to make a curve of slope $r = h_N$ homotopically trivial. This means folding along the edge of slope o_N. In the process, the edges of slopes p_N and s_N are identified. An example is shown in Figure 5.

If the final, N^{th} letter of W is an L, then $s_N = h_{N-1}$, $p_N = p_{N-1}$ and $o_N = s_{N-1}$; so we fold along the edge of slope s_{N-1}, identifying the edges of slopes h_{N-1} and p_{N-1} of the triangle T_N describing the slopes on the boundary torus after layering all the solid tori up to Δ_{N-1}. Similarly, if the final letter of W is an R, then $s_N = s_{N-1}$, $p_N = h_{N-1}$ and $o_N = p_{N-1}$, so we fold along the edge of slope p_{N-1}, identifying the edges of slopes s_{N-1} and h_{N-1} of T_N.

3.3 Neumann–Zagier matrix before Dehn filling

Start with the unfilled manifold, and assume there are $n_{\mathfrak{c}} \geq 2$ cusps. We consider two of these cusps $\mathfrak{c}_0, \mathfrak{c}_1$ with cusp tori $\mathbb{T}_0, \mathbb{T}_1$, respectively. Suppose the triangulation $\mathcal{T}$ has the property that $\mathbb{T}_1$ meets exactly two ideal tetrahedra Δ_1, Δ_2, each in one ideal vertex, and there exist generators $\mathfrak{m}_0, \mathfrak{l}_0$ of $H_1(\mathbb{T}_0)$ that avoid Δ_1 and Δ_2. We prove such a triangulation always exists in Proposition A.1. Cusp $\mathfrak{c}_1$ will be filled. There is a unique ideal edge e running into the cusp $\mathfrak{c}_1$; its other end is in $\mathfrak{c}_0$. The labellings on $\mathcal{T}$ are (at this stage) made arbitrarily.

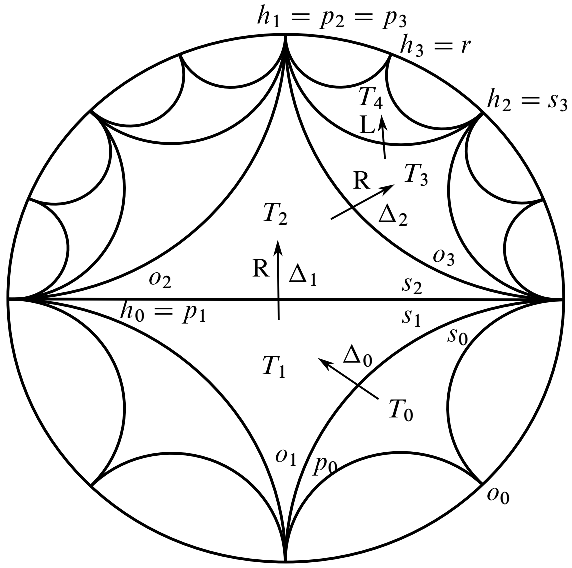


Figure 5: Example of a voyage in the Farey graph when $N = 3$. The word W is RRL. There are three tetrahedra in the layered solid torus, namely Δ_0, Δ_1, Δ_2. The slopes along the way can have several names; for example $s_0 = s_1 = s_2 = o_3$. No tetrahedron is added in the final step from T_3 to T_4.

Lemma 3.3 *Let $\mathcal{T}$, $\mathfrak{m}_0$ and $\mathfrak{l}_0$ be as above. There is a choice of curves $\mathfrak{m}_1$, $\mathfrak{l}_1$ on $\mathbb{T}_1$ generating $H_1(\mathbb{T}_1)$ such that the corresponding Neumann–Zagier matrix* NZ *has the following form*:

(i) *The row of* NZ *corresponding to edge e contains only zeroes. In the cusp triangulation of $\mathfrak{c}_0$, the unique vertex corresponding to e is surrounded by six triangles*, *corresponding to ideal vertices of Δ_1 and Δ_2 in alternating order*, *which form a hexagon $\mathfrak{h}$ around e.*

(ii) *The six vertices of $\mathfrak{h}$ correspond to the ends of three edges of $\mathcal{T}$, denoted f, g, h. After possibly relabelling Δ_1 and Δ_2, the entries of* NZ *in the corresponding rows*, *and in the columns corresponding to Δ_1, Δ_2, are as follows*:

$$\begin{array}{c|cccc|} & \multicolumn{2}{c}{\Delta_1} & \multicolumn{2}{c}{\Delta_2} \\ f & 0 & 1 & 0 & 1 \\ g & -1 & -1 & -1 & -1 \\ h & 1 & 0 & 1 & 0 \end{array}.$$

(iii) *The rows of* NZ *corresponding to $\mathfrak{m}_1$ and $\mathfrak{l}_1$ contain entries as shown below in the columns corresponding to Δ_1, Δ_2, with all other entries in those rows zero*:

$$\begin{array}{c|cccc|} & \multicolumn{2}{c}{\Delta_1} & \multicolumn{2}{c}{\Delta_2} \\ \mathfrak{m}_1 & 1 & 0 & -1 & 0 \\ \mathfrak{l}_1 & 0 & 1 & 0 & -1 \end{array}.$$

(iv) *All other rows of* NZ *contain only zeroes in the columns corresponding to Δ_1 and Δ_2.*

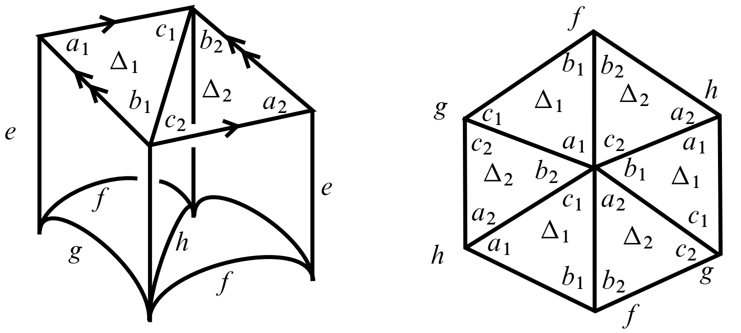


Figure 6: Left: how tetrahedra Δ_1 and Δ_2 meet the cusp $\mathfrak{c}_1$. Right: how they meet the cusp $\mathfrak{c}_0$.

Proof The proof is obtained by considering carefully the gluing. The two tetrahedra Δ_1 and Δ_2 must meet $\mathfrak{c}_1$ as shown in Figure 6, left. The three additional edge classes meeting these tetrahedra are labelled f, g, and h as in that figure. These three edges have both endpoints on $\mathfrak{c}_0$. We may determine how they meet $\mathfrak{c}_0$ by tracing a curve in $\mathfrak{c}_0$ around the edge e. This can be done by tracing a curve around the ideal vertex of the punctured torus made up of the two faces of Δ_1 and Δ_2 that do not meet $\mathfrak{c}_1$. The result is the hexagon $\mathfrak{h}$ shown on the right of Figure 6. Each of the eight ideal vertices of Δ_1 and Δ_2 have been accounted for: two on $\mathfrak{c}_1$ and six forming the hexagon $\mathfrak{h}$ on $\mathfrak{c}_0$.

Now label opposite edges of Δ_1 and Δ_2 as a-, b-, and c-edges respectively, as in Figure 6. These labels determine the 4×6 entries in the rows of the incidence matrix In, corresponding to edges e, f, g, h and tetrahedra Δ_1, Δ_2, as follows:

$$\begin{array}{c} \\ e \\ f \\ g \\ h \end{array} \begin{array}{c} \begin{array}{cc} \quad\Delta_1 & \qquad\Delta_2 \end{array} \\ \left[\begin{array}{ccccccc} 1 & 1 & 1 & & 1 & 1 & 1 \\ 0 & 1 & 0 & & 0 & 1 & 0 \\ 0 & 0 & 1 & & 0 & 0 & 1 \\ 1 & 0 & 0 & & 1 & 0 & 0 \end{array}\right] \end{array}.$$

As the entries in the e row account for all edges of tetrahedra incident with e, all other entries of In in this row are zero. Moreover, as the entries in the e, f, g, h rows account for all edges of Δ_1 and Δ_2, any other row of In has all zeroes in the columns corresponding to Δ_1 and Δ_2.

Turning to the cusp $\mathfrak{c}_1$, we can choose $\mathfrak{m}_1, \mathfrak{l}_1$ as shown in Figure 7. Then $\mathfrak{m}_1$ has a-incidence number 1 with Δ_1 and -1 with Δ_2 (Definition 2.11), and all other incidence numbers zero. In other words, $a^{\mathfrak{m}}_{1,1} = 1$

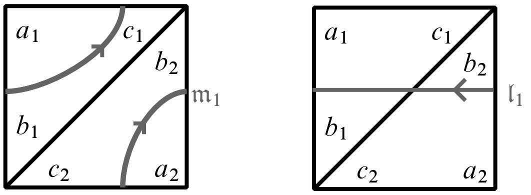


Figure 7: Choices for $\mathfrak{m}_1$ and $\mathfrak{l}_1$.

and $a^{\mathfrak{m}}_{1,2} = -1$ are the only nonzero incidence numbers $a/b/c^{\mathfrak{m}}_{1,j}$. Similarly, $\mathfrak{l}_1$ has b-incidence numbers 1 with Δ_1 and -1 with Δ_2, ie $b^{\mathfrak{l}}_{1,1} = 1$ and $b^{\mathfrak{l}}_{1,2} = -1$, and all other incidence numbers zero.

Forming the Neumann–Zagier matrix by subtracting columns of In, and subtracting incidence numbers, according to Definition 2.17, we obtain the form claimed in (i)–(iii).

It remains to show that in all rows of NZ other than the $e, f, g, h, \mathfrak{m}_1, \mathfrak{l}_1$ rows, there are zeroes in the Δ_1 and Δ_2 columns. We have seen that In contains only zeroes in the Δ_1 and Δ_2 columns in all rows other than the e, f, g, h rows. Hence NZ also has zeroes in the corresponding rows and columns. The remaining rows to consider are the $\mathfrak{m}_k$ and $\mathfrak{l}_k$ rows for $k = 0$ and $k \geq 2$. By hypothesis (or Proposition A.1(ii)), $\mathfrak{m}_0, \mathfrak{l}_0$ avoid the tetrahedra Δ_1 and Δ_2, and hence the $\mathfrak{m}_0, \mathfrak{l}_0$ rows of NZ have zero in the Δ_1, Δ_2 columns. For any $k \geq 2$, the cusp $\mathfrak{c}_k$ does not intersect Δ_1 or Δ_2, as these tetrahedra have all their ideal vertices on $\mathfrak{c}_0$ and $\mathfrak{c}_1$. Thus whatever curves are chosen for $\mathfrak{m}_k$ and $\mathfrak{l}_k$, the corresponding rows of NZ are zero in the Δ_1 and Δ_2 columns. □

Note that in the above proof, by relabelling the tetrahedra Δ_1, Δ_2 and cyclically permuting a-, b- and c-edges, the effect is to cyclically permute the f, g, h rows in the NZ entries above.

To compute the Ptolemy equations for Dehn-filled manifolds, we need a vector B as in Theorem 2.50.

Lemma 3.4 *Let $M, \mathcal{T}$, cusp curves $\mathfrak{m}_k, \mathfrak{l}_k$, tetrahedra Δ_1, Δ_2, and the matrix* NZ *be as above. Suppose $\mathcal{T}$ consists of n tetrahedra. Then there exists a vector*

$$B = (B_1, B'_1, \dots, B_n, B'_n) \in \mathbb{Z}^{2n}$$

with the following properties:

(i) $\mathrm{NZ} \cdot B = C$.

(ii) *The entries B_1, B'_1 and B_2, B'_2 corresponding to Δ_1 and Δ_2 are all zero.*

Proof By Theorem 2.50(i), there exists an integer vector $A = (A_1, A'_1, \dots, A_n, A'_n)$ such that $\mathrm{NZ} \cdot A = C$. The $\mathfrak{m}_1$ and $\mathfrak{l}_1$ rows of NZ are given by Lemma 3.3(iii), and the incidence numbers calculated in the proof show that the corresponding entries of C are $-c^{\mathfrak{m}}_1 = 0$ and $-c^{\mathfrak{l}}_1 = 0$. Thus the $\mathfrak{m}_1, \mathfrak{l}_1$ rows of $\mathrm{NZ} \cdot A = C$ give equations $A_1 - A_2 = 0$ and $A'_1 - A'_2 = 0$. Thus $A_1 = A_2$, $A'_1 = A'_2$, and the Δ_1 and Δ_2 entries of A are given by (A_1, A'_1, A_1, A'_1).

We now adjust A to obtain the desired B, using Theorem 2.50(ii). Write R^G_f and R^G_h for the row vectors in the NZ matrix corresponding to edges f and h. Lemma 3.3(ii) says that R^G_f has $(0,1,0,1)$ in the Δ_1 and Δ_2 columns, and R^G_h has $(1,0,1,0)$. Thus JR^G_f has $(-1,0,-1,0)$ in the Δ_1 and Δ_2 columns, and JR^G_h has $(0,1,0,1)$.

Now let $B = A + A_1 JR^G_f - A'_1 JR^G_h$. By Theorem 2.50(ii), $\mathrm{NZ} \cdot B = C$, and we observe that its Δ_1, Δ_2 entries are

$$(B_1, B'_1, B_2, B'_2) = (A_1, A'_1, A_1, A'_1) + A_1(-1,0,-1,0) - A'_1(0,1,0,1) = (0,0,0,0). \qquad \square$$

3.4 Neumann–Zagier matrix of a layered solid torus

Let the manifold M, triangulation $\mathcal{T}$, cusp curves, tetrahedra and Neumann–Zagier matrix NZ be as in the previous section.

To perform Dehn filling on $\mathfrak{c}_1$, we first remove tetrahedra $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$, leaving a manifold with boundary a once-punctured torus, triangulated by the boundary edges f, g, and h. Then we glue a layered solid torus to this once-punctured torus.

Because generators $\mathfrak{m}_0$, $\mathfrak{l}_0$ of $H_1(\mathbb{T}_0)$ were chosen to be disjoint from $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$ before Dehn filling, representatives of these generators avoid the hexagon $\mathfrak{h}$. When we pull out $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$, $\mathfrak{m}_0$ and $\mathfrak{l}_0$ still avoid $\mathfrak{h}$, and consequently they will form generators of $H_1(\mathbb{T}_0)$ that avoid the layered solid torus when we perform the Dehn filling.

Note that, as in Figure 6, left, the edges f, g, h are each adjacent to a unique face with an ideal vertex at $\mathfrak{c}_1$. Via these faces, each of f, g, h corresponds to one of the three edges in the cusp triangulation of $\mathfrak{c}_1$, and hence to slopes on the torus $\mathbb{T}_1$. As we add tetrahedra of the layered solid torus, each edge similarly corresponds to a slope on $\mathbb{T}_1$. We will in fact label edges by these slopes: we denote the edge corresponding to the slope s by E_s. Thus, we regard f, g, h as slopes, and these slopes form the triangle T_0 of Section 3.1 in the Farey triangulation. In the notation of Section 3.2, $\{f, g, h\} = \{o_0, s_0, p_0\}$ in some order.

As discussed in Section 3.1, the layered solid torus that we glue is determined by the slope r of the filling, and a path in the Farey triangulation from the triangle T_0 with vertices f, g, h to the slope r. This path passes through a sequence of triangles $T_0, \dots, T_{N+1}$, where T_{N+1} contains r as a vertex (and previous T_j do not). The layered solid torus contains N tetrahedra.

The j^{th} tetrahedron (Δ_{j-1} in the notation of Section 3.2) of the layered solid torus corresponds to passing from T_{j-1} to T_j. The four vertices of these triangles are the slopes $(o_{j-1}, p_{j-1}, s_{j-1}, h_{j-1})$ as discussed in Section 3.2. Each edge of the tetrahedron corresponds to one of these four slopes. By Lemma 3.2, the sequence of "old" slopes $o_0, o_1, \dots$ consists of distinct slopes. We will label each tetrahedron by its "old" slope: so rather than writing Δ_{j-1}, we will write $\Delta_{o_{j-1}}$. Then in the final step we glue the two boundary faces together along the edge of slope o_N, which identifies the edges of slopes p_N and s_N. We denote this edge by $E_{p_N = s_N}$.

We arrive at an ideal triangulation of the manifold $M(r)$ obtained by Dehn filling M along slope r on cusp $\mathfrak{c}_1$.

The tetrahedra of this triangulation are of two types: those inside and outside the layered solid torus. We split the columns of the Neumann–Zagier matrix into two blocks accordingly. The N tetrahedra of the layered solid torus are labelled by their "old" slopes, $\Delta_{o_0}, \dots, \Delta_{o_{N-1}}$.

The edges are of three types:

- those lying outside the layered solid torus;

- those lying on the boundary of the layered solid torus, ie f, g, h as above, which we call *boundary edges*; and
- (for $N \geq 1$) the edges lying in the interior of the layered solid torus, labelled by the slopes $h_0, h_1, \dots, h_{N-1}$.

Note that in the final folding, two of these edges are identified. Thus, the rows of the Neumann–Zagier matrix of the triangulated Dehn-filled manifold come in four blocks, corresponding to the three types of edges above, and the cusp rows for the remaining cusps $\mathfrak{c}_0$ and $\mathfrak{c}_k$ for $k \geq 2$.

We regard the Dehn filled manifold $M(r)$ as built up, piece by piece, as follows. Let M_0 denote the original manifold M with the two tetrahedra Δ_1, Δ_2 removed. Let M_k denote the manifold obtained from M_0 after adding the first k tetrahedra of the layered solid torus. Thus

$$M_0 \subset M_1 \subset \cdots \subset M_N.$$

Note M_k has a triangulation of its boundary torus with slopes (o_k, s_k, p_k), the vertices of the triangle T_k of the Farey triangulation.

Then $M(r)$ is obtained by folding together the two boundary faces of M_N along the edge of the boundary triangulation of slope o_N, and identifying the edges of the 3-manifold triangulation of slopes s_N and p_N.

Even though each M_k is not a cusped 3-manifold, rather having boundary components, there is still a well-defined notion of labelled triangulation and incidence matrix. Moreover, since by construction the cusp curves $\mathfrak{m}_0, \mathfrak{l}_0$ avoid the removed tetrahedra Δ_1, Δ_2, they still have well-defined incidence numbers with edges and tetrahedra. Thus there is a well-defined Neumann–Zagier matrix NZ_k for M_k, with rows for the edges and two rows for the cusp $\mathfrak{c}_0$ (but no rows for the boundary left behind from cusp $\mathfrak{c}_1$). Similarly, there is a well defined C-vector C_k for M_k (Definition 2.22).

Lemma 3.5 *The matrix* NZ_0 *of* M_0 *is obtained from the incidence matrix* NZ *of* M *by deleting the columns corresponding to the removed tetrahedra* Δ_1, Δ_2, *and deleting the rows corresponding to the removed edge* e *and cusp* $\mathfrak{c}_1$.

The vector C_0 *is obtained from the* C*-vector* C *of* M *by deleting entries corresponding to edge* e *and zeros corresponding to* $\mathfrak{m}_1$ *and* $\mathfrak{l}_1$, *and adding* 2 *to one of the entries corresponding to edges* f, g *or* h; *by labelling* Δ_1, Δ_2 *appropriately, we can specify which entry.*

Proof The deletion does not otherwise affect incidence relations, so the only effect on the Neumann–Zagier matrix is to delete entries. We similarly delete the entries from C.

In Lemma 3.3, the incidence matrix entries calculated show that one edge, g, is identified with one c-edge of Δ_1 and Δ_2, but edges f and h are not identified with any c-edges of Δ_1 or Δ_2. Thus the g entry of C_0 is 2 greater than the g entry of C.

As noted in the comment after the proof of Lemma 3.3, by labelling Δ_1, Δ_2 appropriately, we can cyclically permute the f, g, h rows, so that we add 2 to the f or h entry of C instead. □

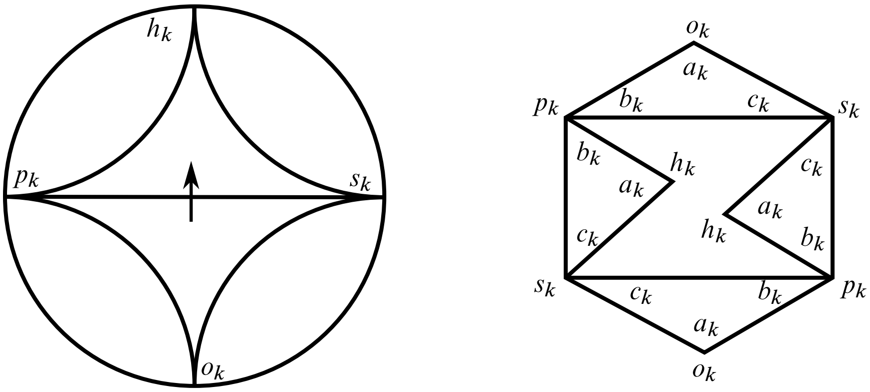


Figure 8: When attaching a nondegenerate layered solid torus, at each intermediate step a tetrahedron is attached with labels as shown on the right.

As each successive tetrahedron is glued, the effect on the cusp triangulation of $\mathfrak{c}_0$ is shown in Figure 8. The hexagon $\mathfrak{h}$ of Lemma 3.3 has been removed, leaving a hexagonal hole; this hole is partly filled in, leaving a "smaller" hexagonal hole.

Lemma 3.6 *For an appropriate labelling of the tetrahedron Δ_{k+1}, the matrix NZ_{k+1} is obtained from NZ_k as follows.*

(i) *Add a pair of columns for the tetrahedron Δ_{o_k}, and a row for the edge with slope h_k. All entries of the new row are zero outside of the Δ_{o_k} columns.*

(ii) *The only nonzero entries in the Δ_{o_k} columns are in the rows corresponding to edges of slope o_k, s_k, p_k, h_k and are as follows*:

$$\begin{array}{c} \\ E_{o_k} \\ E_{s_k} \\ E_{p_k} \\ E_{h_k} \end{array} \begin{array}{c} \Delta_{o_k} \\ \begin{bmatrix} 1 & 0 \\ -2 & -2 \\ 0 & 2 \\ 1 & 0 \end{bmatrix} \end{array}. \tag{3.7}$$

(iii) *All other entries are unchanged.*

The vector C_{k+1} is obtained from C_k by subtracting 2 from the E_{s_k} entry, and inserting an entry 2 for the row E_{h_k}.

Proof Of the six edges of Δ_{o_k}, one of them is identified to E_{o_k}, two opposite edges are identified to E_{p_k}, two opposite edges are identified to E_{s_k}, and one is the newly added edge E_{h_k}. Observe that the three slopes of a triangle in a two-triangle triangulation of a torus are in anticlockwise order if and only if they form the vertices of a triangle of the Farey triangulation in clockwise order. Since (o_k, s_k, p_k) are in anticlockwise order around the triangle T_k of the Farey triangulation, they are slopes associated to the edges of a triangle on the boundary of M_k in clockwise order. Hence we may label the edges of

Δ_{o_k} identified with E_{o_k} (hence also E_{h_k}) as a-edges, those identified with E_{p_k} as b-edges, and those identified with E_{s_k} as c-edges. This gives the entries of NZ_{k+1} and the changes to C-vectors claimed.

No other changes occur with incidence relations of edges and tetrahedra. As cusp curves avoid the layered solid torus, the cusp rows of the Neumann–Zagier matrix and the cusp entries of C_k are also unchanged. □

Finally, we examine the effect of folding up the two boundary faces of M_N, and identifying the two edges E_{p_N}, E_{s_N} into an edge $E_{p_N=s_N}$ to obtain the Dehn-filled manifold $M(r)$.

We denote the row vector of NZ_N corresponding to the edge E_s of slope s by R^G_s; and we denote the row vector of $\mathrm{NZ}(r)$ corresponding to the identified edge $E_{p_N=s_N}$ by $R^G_{p_N=s_N}$. Similarly, we denote the entry of C_N corresponding to slope s by $(C_N)_s$; and we denote the entry of $C(r)$ corresponding to the identified edge $E_{p_N=s_N}$ by $C(r)_{p_N=s_N}$.

Lemma 3.8 *The Neumann–Zagier matrix* $\mathrm{NZ}(r)$ *of* $M(r)$ *is obtained from* NZ_N *by replacing the rows corresponding to edges* E_{p_N} *and* E_{s_N} *with their sum, corresponding to the edge* $E_{p_N=s_N}$*. The* C*-vector* $C(r)$ *of* $M(r)$ *is obtained from* C_N *by replacing the entries* $(C_N)_{p_N}$, $(C_N)_{s_N}$ *corresponding to edges* E_{p_N}, E_{s_N} *with an entry* $C(r)_{p_N=s_N} = (C_N)_{p_N} + (C_N)_{s_N} - 2$, *corresponding to edge* $E_{p_N=s_N}$*.*

Thus row vectors $R^G_{p_N}$ and $R^G_{s_N}$ are replaced with $R^G_{p_N=s_N} = R^G_{p_N} + R^G_{s_N}$. Corresponding entries of C_N are also summed, but then we subtract 2 for the replacement entry.

Proof The only change in incidence relations between edges and tetrahedra after gluing is that all tetrahedra that were incident to edges E_{p_N} or E_{s_N} are now incident to the identified edge $E_{p_N=s_N}$. Thus we sum the two rows. The cusp rows are again unaffected.

Each C-vector entry corresponding to an edge E_k is of the form $2-c_k$, where $c_k=\sum_j c_{k,j}$ (Definition 2.22). When we combine the two edges, the c_k terms combine by a sum, but in place of $2+2$ we must have a single 2; hence we subtract 2. □

The effect on the cusp triangulation of $\mathfrak{c}_1$ is to close the hexagonal hole by gluing its edges together as in Figure 9.

As mentioned previously, the slopes (p_N, s_N) are equal to (p_{N-1}, h_{N-1}) if the last letter of W is an L, and equal to (h_{N-1}, s_{N-1}) if the last letter of W is an R. Either way, we observe that the slope h_{N-1} is among those being identified. Thus the last new edge in the layered solid torus appears at step $N-1$, with label h_{N-2} at that step.

Alternatively, we may write the matrix $\mathrm{NZ}(r)$ by deleting the row $E_{h_{N-1}}$ from NZ_N and adding it to the row $E_{p_{N-1}}$ or $E_{s_{N-1}}$ accordingly as the last choice is an L or R. Then the edges are regarded as having slopes $\{f, g, h\} = \{o_0, p_0, s_0\}$, together with $h_0, h_1, \dots, h_{N-2}$.

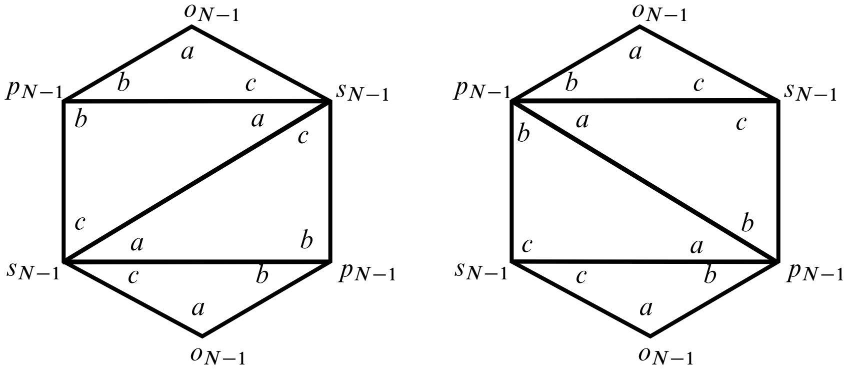


Figure 9: The last tetrahedron in the layered solid torus has its two interior triangles identified together, either by folding over the edge labelled p_{N-1} or by folding over the edge labelled s_{N-1}. The two cases are shown.

With this notation, the Neumann–Zagier matrix $\mathrm{NZ}(r)$ has pairs of columns corresponding to tetrahedra, which consist of the tetrahedra of $M \setminus (\Delta_1^{\mathfrak{c}} \cup \Delta_2^{\mathfrak{c}})$, and the tetrahedra of the layered solid torus, $\Delta_{o_0}, \dots, \Delta_{o_{N-1}}$. The rows correspond to the edges of M disjoint from $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$, and then edges $E_{o_0}, E_{s_0}, E_{p_0}$ on the boundary of the hexagon, then $E_{h_0}, E_{h_1}, \dots, E_{h_{N-2}}$ inside the layered solid torus; and cusp rows corresponding to $\mathfrak{m}_0, \mathfrak{l}_0$. The general form is shown in Figure 10.

Thus if there are n edges and tetrahedra in the triangulation, then outside the layered solid torus there are $n-N$ tetrahedra and $n-N-2$ edges.

Lemma 3.8 includes the case where $N=0$, ie where the layered solid torus is *degenerate*. In this case we go directly from M to M_0 (removing $\Delta_1^{\mathfrak{c}} \cup \Delta_2^{\mathfrak{c}}$) to $M(r)$. In this case the filling slope r is equal to h_0, so has distance 1 from two of the initial slopes f, g, h, and distance 2 from the other. These are

$$\mathrm{NZ}(r) = \begin{array}{c|cccccccccccc|}
 & \multicolumn{4}{c}{\text{tet. of } M \setminus (\Delta_1^{\mathfrak{c}} \cup \Delta_2^{\mathfrak{c}})} & \multicolumn{2}{c}{\Delta_{o_0}} & \multicolumn{2}{c}{\Delta_{o_1}} & \cdots & \multicolumn{2}{c}{\Delta_{o_{N-1}}} \\
\text{edges of } M & * & * & \cdots & * & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
\text{outside} & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\Delta_1^{\mathfrak{c}} \cup \Delta_2^{\mathfrak{c}} & * & * & \cdots & * & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
\hline
E_{o_0} & * & * & \cdots & * & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
E_{s_0} & * & * & \cdots & * & -2 & -2 & * & * & \cdots & * & * \\
E_{p_0} & * & * & \cdots & * & 0 & 2 & * & * & \cdots & * & * \\
\hline
E_{h_0} & 0 & 0 & \cdots & 0 & * & * & * & * & \cdots & * & * \\
E_{h_1} & 0 & 0 & \cdots & 0 & 0 & 0 & * & * & \cdots & * & * \\
E_{h_2} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & * & * \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
E_{h_{N-2}} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & * & * \\
\hline
\mathfrak{m}_0 & * & * & \cdots & * & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
\mathfrak{l}_0 & * & * & \cdots & * & 0 & 0 & 0 & 0 & \cdots & 0 & 0
\end{array}$$

Figure 10: Neumann–Zagier matrix of a Dehn-filled manifold.

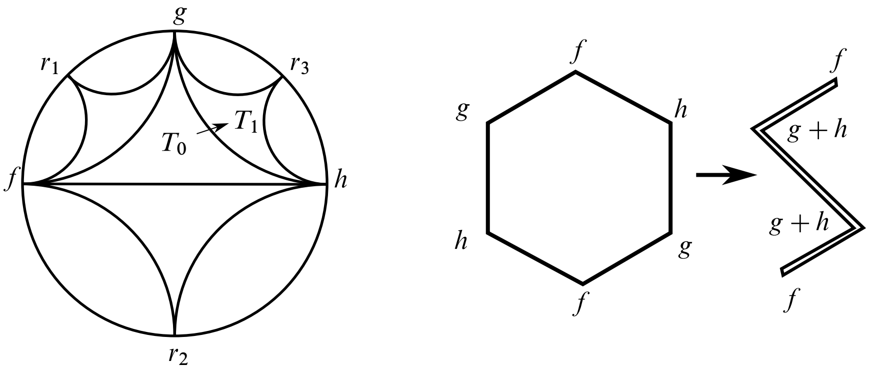


Figure 11: Left: Dehn filling along slope r_1, r_2, or r_3 attaches a degenerate layered solid torus, with no tetrahedra. Right: the effect of such a Dehn filling on the cusp triangulation of C_0 is to fold the hexagon, identifying two boundary edges together.

the slopes labelled r_1, r_2, and r_3 in Figure 11, left. No tetrahedra are added, and we skip to the final folding step, folding boundary faces of the boundary torus together along the edge of slope o_0, and identifying the edges corresponding to slopes s_0 and p_0. The effect is to combine and sum the rows of NZ_0 corresponding to E_{s_0} and E_{p_0}.

The resulting matrix $\mathrm{NZ}(r)$ is described explicitly in the following propositions; they simply describe the result of applying the previous lemmas, and their proofs are immediate from those lemmas. Figure 10 shows most of the structure described.

Proposition 3.9 *Suppose* $\mathrm{NZ}(r)$ *is the Neumann–Zagier matrix of* $M(r)$, *obtained by Dehn filling the manifold* M *of Lemma 3.3, with Neumann–Zagier matrix* NZ, *along the slope* r *on* $\mathfrak{c}_1$. *Then the rows of* $\mathrm{NZ}(r)$ *corresponding to edges outside the layered solid torus and its boundary, and the rows corresponding to* $\mathfrak{m}_0$ *and* $\mathfrak{l}_0$, *are as follows.*

(i) *Entries in columns corresponding to tetrahedra of the layered solid torus are all zero.*

(ii) *Entries in columns corresponding to tetrahedra outside the layered solid torus are unchanged from their entries in* NZ. □

In the $N = 0$ case, by Lemma 3.8 and subsequent discussion, the only edge rows of the layered solid torus are those with slopes o_0 and $s_0 = p_0$, and there are no columns corresponding to tetrahedra in the layered solid torus.

Proposition 3.10 *Suppose* $N = 0$. *Then the entries in the rows of* $\mathrm{NZ}(r)$ *corresponding to the edges of the layered solid torus are as follows.*

(i) *The row corresponding to* o_0 *has the same entries as corresponding columns of* NZ.

(ii) *The row corresponding to* $s_0 = p_0$ *is the sum of entries in* s_0 *and* p_0 *rows of* NZ. □

Proposition 3.11 *Suppose $N \geq 1$. The entries in the rows of $\mathrm{NZ}(r)$ corresponding to the edges of the layered solid torus are as follows.*

(i) *In columns corresponding to the tetrahedra outside the layered solid torus*:

  (a) *The entries in the rows corresponding to the edges with slopes $h_0, \dots, h_{N-2}$ are all zero (there are no such edges if $N = 1$).*

  (b) *The entries in the rows corresponding to the boundary edges, with slopes $\{f, g, h\} = \{o_0, p_0, s_0\}$ are the same as in the corresponding rows and columns of NZ. (The p_0 or s_0 row may be combined and summed with the h_{N-1} row in the final step, but being summed with zeroes, the entries remain the same.)*

(ii) *The entries in the pair of columns corresponding to the tetrahedron Δ_{o_j}, are as described in Lemma 3.6, except that rows corresponding to slopes p_N and s_N are summed as in Lemma 3.8. In particular, we have the following*:

  (a) *The row of slope o_0 has $(1, 0)$ in the Δ_{o_0} columns, zero in every other Δ_{o_j} column.*

  (b) *Provided $s_0 \neq s_N$, the row of slope s_0 has a sequence of pairs $(-2, -2)$, followed by $(1, 0)$ and then all zeroes. (The number of such pairs is $k + 1$, where W begins with a string of k Rs.)*

  (c) *Provided $p_0 \neq p_N$, the row of slope p_0 has a sequence of pairs $(0, 2)$, followed by $(1, 0)$ and then all zeroes. (The number of such pairs is $k + 1$, where W begins with a string of k Ls.)*

  (d) *In the two columns for Δ_{o_j}, entries in rows of slope $h_{j+1}, \dots, h_{N-2}$ are zero.* □

3.5 Building up the sign vector

We will now show how to build up a vector $B(r)$ satisfying the sign equation (2.49) for the Dehn-filled manifold $M(r)$; that is,

$$\mathrm{NZ}(r) \cdot B(r) = C(r).$$

We do this starting from the sign vector B found for the unfilled manifold M in Lemma 3.4. We build up a sequence of vectors $B_0, \dots, B_N$ associated to the manifolds $M_0, \dots, M_N$. These vectors "almost" satisfy $\mathrm{NZ}_k \cdot B_k = C_k$. From B_N we obtain the desired vector $B(r)$.

In Lemma 3.5, we showed that we can take C_0 to be obtained from C by deleting the e entry, and adding 2 to one of the entries corresponding to slopes $\{f, g, h\} = \{o_0, s_0, p_0\}$, whichever we prefer. For the following, we want the 2 to be added to the entry corresponding to slope s_0 or p_0. For definiteness, we take C_0 to be obtained by adding 2 to the s_0 entry.

Lemma 3.12 *Let B_0 be the vector obtained from B by removing the two pairs of entries corresponding to the removed tetrahedra Δ_1^c, Δ_2^c. Then $C_0 - \mathrm{NZ}_0 \cdot B_0$ consists of all zeroes, except for a 2 in the entry corresponding to the edge with slope s_0.*

Proof We have $\mathrm{NZ} \cdot B = C$. Examine the effect of changing the terms to $\mathrm{NZ}_0 \cdot B_0$ and C_0. By Lemma 3.4, the vector B has pairs of entries corresponding to Δ_1^c and Δ_2^c consisting of all zeroes. Consider the rows of NZ corresponding to edges away from $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$, together with the $\mathfrak{m}_0, \mathfrak{l}_0$ rows. These rows have all zero entries in $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$ columns, by Lemma 3.3. The corresponding rows of NZ_0 are obtained by deleting the zero entries in the $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$ columns (Lemma 3.5). Thus the corresponding entries of $\mathrm{NZ} \cdot B$ and $\mathrm{NZ}_0 \cdot B_0$ are equal. Similarly, the corresponding entries of C and C_0 are equal. So $C_0 - \mathrm{NZ}_0 \cdot B_0$ has zeroes in these entries.

By Lemma 3.5, the only remaining rows of NZ_0 are those corresponding to rows with slopes $\{f, g, h\} = \{o_0, s_0, p_0\}$.

In both $\mathrm{NZ} \cdot B$ and $\mathrm{NZ}_0 \cdot B_0$ we obtain exactly the same terms from the tetrahedra outside $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$, by Lemma 3.5 and construction of B_0. These account for all the terms in $\mathrm{NZ}_0 \cdot B_0$, but in $\mathrm{NZ} \cdot B$ there are also terms from the tetrahedra $\Delta_1^{\mathfrak{c}}$ and $\Delta_2^{\mathfrak{c}}$. However, as the corresponding entries of B are zero, these terms are zero. So $\mathrm{NZ}_0 \cdot B_0$ and $\mathrm{NZ} \cdot B$ have the same entries in these rows, and hence also C. However, as discussed above, we have chosen C_0 to differ from C by 2 in the row with slope s_0. Hence $C_0 - \mathrm{NZ}_0 \cdot B_0$ is as claimed. □

Observe from the proof that Lemma 3.12 works equally well with the slope s_0 replaced with any of $\{f, g, h\} = \{o_0, s_0, p_0\}$.

As it turns out, going from B_0 to B_1 is a little different from the general case, and so we deal with it separately.

Lemma 3.13 *Let B_1 be obtained from B_0 by adding zero entries corresponding to Δ_{o_0} Then $C_1 - \mathrm{NZ}_1 \cdot B_1$ consists of all zeroes, except for a 2 in the new entry corresponding to E_{h_0}.*

Proof By Lemma 3.6, NZ_1 is obtained from NZ_0 by adding a row for the edge with slope h_0 and a pair of columns for Δ_{o_0}, with added nonzero entries as in (3.7). Also, C_1 is obtained from C_0 by subtracting 2 from the E_{s_0} entry, and inserting an entry 2 for the row E_{h_0}.

Now each entry of $\mathrm{NZ}_0 \cdot B_0$ is equal to the corresponding entry in $\mathrm{NZ}_1 \cdot B_1$, since the terms are exactly the same, except for the terms of $\mathrm{NZ}_1 \cdot B_1$ corresponding to the added tetrahedron Δ_{o_0}, which are zero since B_1 has zero entries there. The extra entry in $\mathrm{NZ}_1 \cdot B_1$, corresponding to E_{h_0}, is also zero, since this row of NZ_1 only has nonzero entries in the terms corresponding to Δ_{o_0}, where B_1 is zero. Thus $\mathrm{NZ}_1 \cdot B_1$ is equal to $\mathrm{NZ}_0 \cdot B_0$ with a 0 appended.

Similarly, each entry of C_0 is equal to the corresponding entry of C_1, except for the entry of slope s_0, where $C_1 - C_0$ has a -2. The vector C_1 also has a 2 appended.

From Lemma 3.12, each entry of $C_0 - \mathrm{NZ}_0 \cdot B_0$ is zero, except for the s_0 entry, which is 2.

Putting these together, each entry of $C_0 - \mathrm{NZ}_0 \cdot B_0$ equals the corresponding entry of $C_1 - \mathrm{NZ}_1 \cdot B_1$, except for the entry of slope s_0, where $C_1 - \mathrm{NZ}_1 \cdot B_1$ has entry $2 - 2 = 0$. The additional entry of $C_1 - \mathrm{NZ}_1 \cdot B_1$ of slope h_0 is $2 - 0 = 2$. Thus $C_1 - \mathrm{NZ}_1 \cdot B_1$ is as claimed. □

Had we chosen C_0 to differ from C in the p_0 entry, then $C_0 - \mathrm{NZ}_0 \cdot B_0$ would have a nonzero entry for slope p_0; in this case we could take B_1 to be obtained from B_0 by adding entries $(0,1)$ and obtain the same conclusion.

We now proceed to the general case, building B_{k+1} from B_k. We use the first $N-1$ letters of the word W in the letters {L,R}.

Lemma 3.14 *Suppose $1 \le k \le N-1$. If the k^{th} letter of the word W is R (resp. L), let B_{k+1} be obtained from B_k by appending $(0,1)$ (resp. $(0,0)$) for the added tetrahedron Δ_{o_k}. Then $C_{k+1} - \mathrm{NZ}_{k+1} \cdot B_{k+1}$ consists of all zeroes except a 2 in the entry corresponding to E_{h_k}.*

Proof Proof by induction on k; Lemma 3.13 provides the base case. Assume $C_k - \mathrm{NZ}_k \cdot B_k$ has only nonzero entry 2 in the row of slope h_{k-1}, and we consider $C_{k+1} - \mathrm{NZ}_{k+1} \cdot B_{k+1}$.

Again using Lemma 3.6, C_{k+1} and C_k differ only in that C_{k+1} has a 2 in the new entry E_{h_k}, and has 2 subtracted from the E_{s_k} entry.

Suppose that the k^{th} letter of W is an R. Then by Lemma 3.2 we have $o_k = p_{k-1}$, $s_k = s_{k-1}$ and $p_k = h_{k-1}$. Thus the new entries in NZ_{k+1} are given by

$$\begin{array}{rc} & \Delta_{o_k} \\ \begin{matrix} E_{o_k}=E_{s_{k-1}} \\ E_{s_k}=E_{s_{k-1}} \\ E_{p_k}=E_{h_{k-1}} \\ E_{h_k} \end{matrix} & \begin{bmatrix} 1 & 0 \\ -2 & -2 \\ 0 & 2 \\ 1 & 0 \end{bmatrix} \end{array}.$$

So with B_{k+1} defined as stated, the entries of $\mathrm{NZ}_k \cdot B_k$ differ from the corresponding entries of $\mathrm{NZ}_{k+1} \cdot B_{k+1}$ in entries for rows of slope $p_k = h_{k-1}$ and s_k. In the row of slope $p_k = h_{k-1}$, $\mathrm{NZ}_{k+1} \cdot B_{k+1}$ is greater by 2, and in the row of slope s_k, $\mathrm{NZ}_{k+1} \cdot B_{k+1}$ is lesser by 2. The new entry in $\mathrm{NZ}_{k+1} \cdot B_{k+1}$ of slope h_k is 0.

Putting the above together, we find that $C_{k+1} - \mathrm{NZ}_{k+1} \cdot B_{k+1}$ has the same entries as $C_k - \mathrm{NZ}_k \cdot B_k$, except in the rows of slope: $p_k = h_{k-1}$, where they differ by -2; $s_k = s_{k-1}$, where they differ by $(-2) - (-2) = 0$; and h_k, where there is an extra entry of 2. Thus $C_{k+1} - \mathrm{NZ}_{k+1} \cdot B_{k+1}$ has unique nonzero entry 2 in the E_{h_k} entry as desired.

Suppose that the k^{th} letter is an L; then we have $s_i = h_{i-1}$. The argument is simpler since B_{k+1} simply appends zeroes to B_k. As we only append zeroes, there is no need to consider the new columns of NZ_{k+1} in any detail. Indeed, $\mathrm{NZ}_{k+1} \cdot B_{k+1}$ and $\mathrm{NZ}_k \cdot B_k$ have the same nonzero entries. Thus the nonzero entries in $C_{k+1} - \mathrm{NZ}_{k+1} \cdot B_{k+1}$ are those of $C_k - \mathrm{NZ}_k \cdot B_k$, with -2 added to the $s_k = h_{k-1}$ entry, and 2 inserted in the h_k entry, giving the result. □

We now consider the final step: the desired sign vector $B(r)$ is just B_N.

Lemma 3.15 *The vector B_N of Lemma 3.14 satisfies $\mathrm{NZ}(r) \cdot B_N = C(r)$.*

Proof By Lemma 3.8, $\mathrm{NZ}(r)$ is obtained from NZ_n by replacing the rows of slope p_N and s_N with their sum, corresponding to the identified edge $E_{p_N=s_N}$. The row vectors $R^G_{p_N}$ and $R^G_{s_N}$ are replaced with

$$R^G_{p_N=s_N} = R^G_{p_N} + R^G_{s_N}.$$

Similarly, $C(r)$ is obtained from C_N by replacing the corresponding entries $(C_N)_{p_N}, (C_N)_{s_N}$ with the combined entry

$$C(r)_{p_N=s_N} = (C_N)_{p_N} + (C_N)_{s_N} - 2.$$

By Lemma 3.14, $C_N - \mathrm{NZ}_N \cdot B_N$ has only nonzero entry 2 corresponding to slope h_{N-1}. Note that h_{N-1} is equal to one of the slopes p_N, s_N to be combined (accordingly as the final letter of W is an L or R).

Consider any row other than those corresponding to slopes p_N or s_N. Such a row is unaffected by the combination of rows or entries. Hence $C_N - \mathrm{NZ}_N \cdot B_N$ has zero entry in this row; and since $\mathrm{NZ}(r)$ and $C(r)$ are equal to NZ_N and C_N in these rows, $C(r) - \mathrm{NZ}(r) \cdot B_N$ has zero entry in these rows.

It remains to consider the single row obtained by combining two rows. Since these two rows include the row of slope h_{N-1}, the two corresponding entries of $C_N - \mathrm{NZ}_N \cdot B_N$ are 0 and 2 in some order. These entries are $(C_N)_{p_N} - R^G_{p_N} \cdot B_N$ and $(C_N)_{s_N} - R^G_{s_N} \cdot B_N$, so

$$(C_N)_{p_N} - R^G_{p_N} \cdot B_N + (C_N)_{s_N} - R^G_{s_N} \cdot B_N = 2.$$

Putting these together, we obtain the remaining entry of $C(r) - \mathrm{NZ}(r) \cdot B_N$ as

$$\begin{aligned} C(r)_{p_N=s_N} - R^G_{p_N=s_N} \cdot B_N &= (C_N)_{p_N} + (C_N)_{s_N} - 2 - (R^G_{p_N} + R^G_{s_N}) \cdot B_N \\ &= (C_N)_{p_N} - R^G_{p_N} \cdot B_N + (C_N)_{s_N} - R^G_{s_N} \cdot B_N - 2 = 0. \end{aligned}$$ □

We have now proved the following.

Proposition 3.16 *There exists an integer vector $B(r)$ such that $\mathrm{NZ}(r) \cdot B(r) = C(r)$. The vector $B(r)$ is given by taking a vector B for the unfilled manifold M as in Lemma 3.4, removing the two pairs of zeroes corresponding to removed tetrahedra $\Delta^{\mathfrak{c}}_1, \Delta^{\mathfrak{c}}_2$, and then appending*

(i) *a $(0,0)$ corresponding to the tetrahedron Δ_{o_0}; then*

(ii) *$N-1$ pairs $(0,1)$ or $(0,0)$, corresponding to the first $N-1$ letters of the word W. For each R we append a $(0,1)$, and for each L we append a $(0,0)$.* □

In other words, the entry of B corresponding to the tetrahedron Δ_{o_k}, for $1 \le k \le N-1$, is $(0,1)$ if the k^{th} letter of W is an R, and $(0,0)$ if the k^{th} letter of W is an L.

3.6 Ptolemy equations in a layered solid torus

We can now write down explicitly the Ptolemy equations for a Dehn filled manifold.

To do so, we will suppose M has two cusps $\mathfrak{c}_0, \mathfrak{c}_1$, and is triangulated such that exactly two tetrahedra Δ_1^1, Δ_2^1 meet $\mathfrak{c}_1$, each in a single ideal vertex. Suppose also that curves $\mathfrak{m}_0$ and $\mathfrak{l}_0$ represent generators of the first homology of $\mathfrak{c}_0$, and avoid triangles coming from Δ_1^1 and Δ_2^1 in the cusp triangulation of $\mathfrak{c}_0$. We show in Proposition A.1 that every 3-manifold of interest here admits such a triangulation, with such curves on the cusp triangulation of $\mathfrak{c}_0$.

Let $\mathrm{NZ}^\flat$ and $C^\flat$ denote the reduced Neumann–Zagier matrix and C-vector associated with this triangulation for M, where the triangulation is labelled to satisfy Lemma 2.51. Finally, suppose B is an integer vector that satisfies $\mathrm{NZ}^\flat \cdot B = C^\flat$.

Theorem 3.17 *Let M be a two-cusped manifold with cusps $\mathfrak{c}_0, \mathfrak{c}_1$, triangulated as above so that only two tetrahedra meet $\mathfrak{c}_1$, and curves $\mathfrak{m}_0, \mathfrak{l}_0$ on the cusp triangulation of $\mathfrak{c}_0$ avoid these tetrahedra. Perform Dehn filling on the cusp $\mathfrak{c}_1$ by attaching a layered solid torus with meridian slope r, consisting of tetrahedra $\Delta_{o_0}, \dots, \Delta_{o_{N-1}}$ determined by the word W in the Farey graph. Then the Ptolemy equations of the Dehn filled manifold $M(r)$ satisfy*:

(i) *There exist a finite number of* ***outside*** *equations, corresponding to tetrahedra of M and $M(r)$ lying outside the layered solid torus. These are obtained as in Definition 2.57 using the reduced Neumann–Zagier matrix $\mathrm{NZ}^\flat$ and B for the unfilled manifold M. In particular, they are independent of the Dehn filling.*

(ii) *For tetrahedra of the layered solid torus, Ptolemy equations are*

$$\begin{aligned} -\gamma_{o_k}\gamma_{h_k} + \gamma_{p_k}^2 - \gamma_{s_k}^2 &= 0 \quad \text{if } k > 0 \text{ and the } k^{th} \text{ letter of } W \text{ is an } R, \\ \gamma_{o_k}\gamma_{h_k} + \gamma_{p_k}^2 - \gamma_{s_k}^2 &= 0 \quad \text{if } k = 0 \text{ or the } k^{th} \text{ letter of } W \text{ is an } L, \end{aligned}$$

for $0 \le k \le N-1$. We also set $\gamma_{p_N} = \gamma_{s_N}$.

Proof Item (i) follows from Propositions 3.11 and 3.16: The nonzero entries of the columns of $\mathrm{NZ}(r)$ are identical to those of NZ for tetrahedra outside the layered solid torus, and entries of $B(r)$ corresponding to tetrahedra outside the layered solid torus are identical to those of B. Then (i) follows immediately from Definition 2.57.

As for (ii), the tetrahedron Δ_{o_k} has its a-edges identified to the edges E_{o_k} and E_{h_k}, both its b-edges identified to E_{p_k}, and both its c-edges identified to E_{s_k}, so the powers of γ variables are as claimed. They are disjoint from the cusp curves $\mathfrak{m}_0, \mathfrak{l}_0$, so no powers of ℓ or m appear in the Ptolemy equations. The corresponding pair of entries of B is $(0,0)$ for $k = 0$, and for $k \ge 1$, they are given by $(0,1)$ if the k^{th} letter of W is an R, and $(0,0)$ if the k^{th} letter of W is an L. At the final step the edges with slopes p_N and s_N are identified, with the effect of summing the corresponding rows of NZ matrices; this is also the effect of setting the variables $\gamma_{p_N}, \gamma_{s_N}$ equal in Ptolemy equations. Hence the Ptolemy equation of Definition 2.57 takes the claimed form. □

tetrahedron	face 012	face 013	face 023	face 123
0	3(021)	1(213)	2(130)	1(230)
1	4(102)	2(132)	0(312)	0(103)
2	2(203)	0(302)	2(102)	1(031)
3	0(021)	4(103)	4(203)	4(213)
4	1(102)	3(103)	3(203)	3(213)

Table 1: Five tetrahedra triangulation of the Whitehead link complement.

4 Example: Dehn-filling the Whitehead link

In this section, we work through the example of the Whitehead link and its Dehn fillings. The standard triangulation of the Whitehead link has four tetrahedra meeting each cusp. To apply our results, we need a triangulation with two tetrahedra meeting one of the cusps. This is obtained by a triangulation with five tetrahedra. Its gluing information is shown in Table 1, where the notation is as in Regina [3]: tetrahedra are labelled by numbers 0 through 4, with vertices labelled 0 through 3. Thus faces are determined by three labels. The notation 3(021) in row 0 under column “Face 012” means that the face of tetrahedron 0 with vertices 012 is glued to the face of tetrahedron 3 with vertices 021, with 0 glued to 0, 1 to 2, and 2 to 1. And so on. Note the software Regina [3] and SnapPy [8] can be used to confirm that the manifold produced is the Whitehead link complement.

In the triangulation, tetrahedra 3 and 4 are the only ones meeting one of the cusps, in vertices 3(3) and 4(3), respectively. We have chosen the labelling so that the Neumann–Zagier matrix satisfies the conditions of Lemma 3.3: see below. We will perform Dehn filling on the Whitehead link by replacing these two tetrahedra with a layered solid torus.

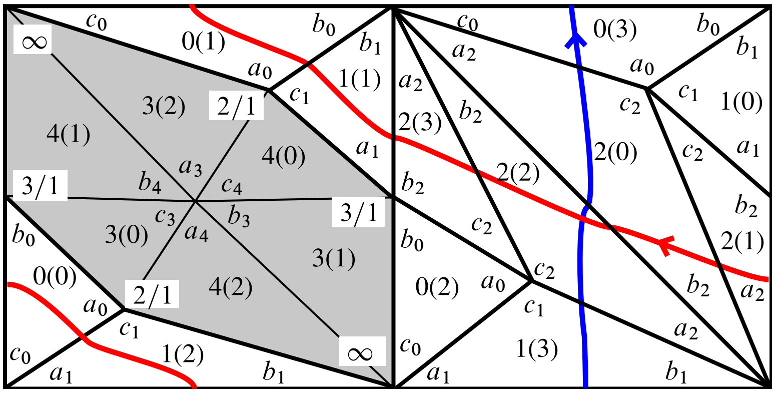


Figure 12: Cusp triangulation of the Whitehead link, with triangles corresponding to tetrahedra 3 and 4 shaded. The edge e is at the centre of the hexagon, edges with slopes $\infty = 1/0$, $3/1$, $2/1$ on the boundary of the hexagon. The additional vertex in the figure corresponds to the edge we call 0(23). Note $\mathfrak{l}$ is in red, $\mathfrak{m}$ in blue.

$$\mathrm{NZ} = \begin{array}{c|cc|cc|cc|cc|cc|}
 & \multicolumn{2}{c|}{\Delta_0} & \multicolumn{2}{c|}{\Delta_1} & \multicolumn{2}{c|}{\Delta_2} & \multicolumn{2}{c|}{\Delta_3} & \multicolumn{2}{c|}{\Delta_4} \\
E_{0(23)} & 1 & 0 & -1 & -1 & -2 & -2 & 0 & 0 & 0 & 0 \\
\hline
E_{3/1} & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\
E_{2/1} & 1 & 0 & -1 & -1 & 0 & 0 & 0 & 1 & 0 & 1 \\
E_{1/0} & -2 & -1 & 1 & 2 & 2 & 1 & -1 & -1 & -1 & -1 \\
\hline
E_e & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
\mathfrak{m}_0 & -1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mathfrak{l}_0 & -1 & -2 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mathfrak{m}_1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\
\mathfrak{l}_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1
\end{array}$$

Figure 13: The Neumann–Zagier matrix of the complement of the Whitehead link.

The cusp neighbourhood of the other cusp of the Whitehead link is shown in Figure 12. The shaded hexagon consists of triangles from tetrahedra 3 and 4. Pulling out tetrahedra 3 and 4 will leave a manifold with punctured torus boundary. The slopes of these boundary curves can be computed in terms of the usual meridian/longitude of the cusp of the Whitehead link to be $3/1$, $2/1$, and $1/0 = \infty$ (we used Regina [3] and SnapPy [8] to compare slopes under Dehn filling to identify these edges). Each slope corresponds to an edge of the punctured torus, and an edge of the triangulation, and appears twice in the hexagon of our cusp triangulation. The three slopes are labelled in Figure 12. There are two additional edges; one e only meets tetrahedra 3 and 4. The other we denote by $0(23)$ (because the edge $0(23)$ in Regina notation corresponds to this edge class). Finally, we choose generators of the fundamental group of the cusp torus to be disjoint from the hexagon in the cusp neighbourhood.

We may now read the incidence matrix of the Whitehead link complement off of the cusp triangulation, and use it to find the Neumann–Zagier matrix, which is shown in Figure 13.

The vector C is $[-1, 2, 1, -2, 0, -1, -1, 0, 0]^T$. Notice that the vector

$$B = [1, 1, 1, -1, 1, 0, 0, 0, 0, 0]^T$$

satisfies the properties of Lemma 3.4: $\mathrm{NZ} \cdot B = C$ and the last four entries of B are all zero. We now have enough information to determine the outside Ptolemy equations for *any* Dehn filling of the Whitehead link complement. By Theorem 3.17 and Definition 2.57, they are

$$\begin{aligned}
\Delta_0 :&\quad -\ell^{1/2} m^{-1/2} \gamma_{0(23)} \gamma_{2/1} - \ell^{1/2} m^{-1} \gamma_{3/1} \gamma_{1/0} - \gamma_{1/0}^2 = 0, \\
\Delta_1 :&\quad -m^{1/2} \gamma_{3/1} \gamma_{1/0} - \ell^{1/2} m^{-1/2} \gamma_{1/0}^2 - \gamma_{0(23)} \gamma_{2/1} = 0, \\
\Delta_2 :&\quad \gamma_{1/0}^2 - \gamma_{1/0} \gamma_{3/1} - \gamma_{0(23)}^2 = 0.
\end{aligned} \tag{4.1}$$

Recall that we set $\gamma_n = 1$, where n is such that the n^{th} gluing equation is redundant in the Neumann–Zagier matrix. For this example, we may always set $\gamma_{1/0} = 1$, and then use the equation from Δ_2 to write $\gamma_{3/1}$ in terms of $\gamma_{0(23)}$. Equations from Δ_0 and Δ_1 can then be used to write $\gamma_{0(23)}$ and $\gamma_{2/1}$ only in terms of ℓ and m. These may be substituted into additional Ptolemy equations that arise from Dehn filling.

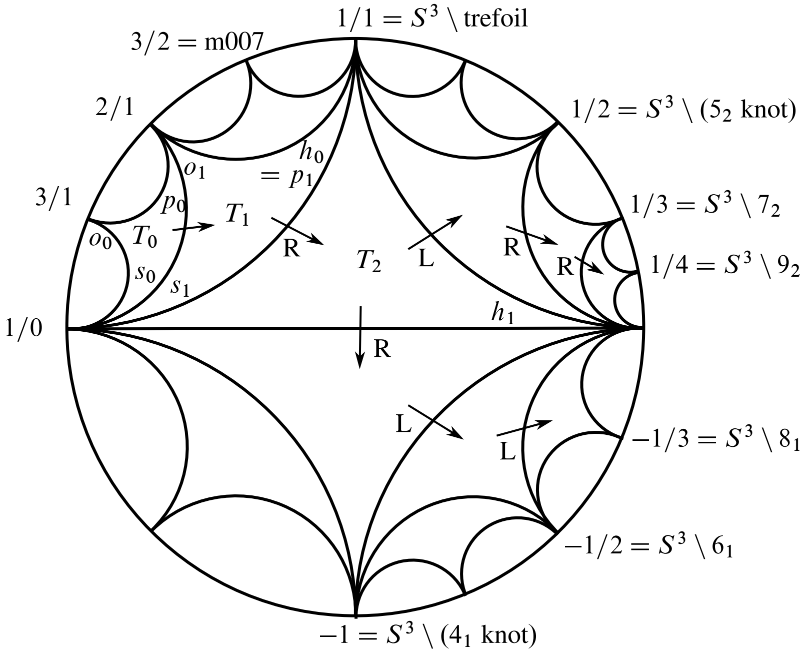

$1/1 = S^3 \setminus$ trefoil
$3/2 = \mathrm{m}007$
$2/1$
$1/2 = S^3 \setminus (5_2$ knot$)$
h_0
$= p_1$
o_1
$3/1$
p_0
T_1
$1/3 = S^3 \setminus 7_2$
o_0
T_0
R
T_2
L
R
R
$1/4 = S^3 \setminus 9_2$
s_0
s_1
$1/0$
h_1
R
L
L
$-1/3 = S^3 \setminus 8_1$
$-1/2 = S^3 \setminus 6_1$
$-1 = S^3 \setminus (4_1$ knot$)$

Figure 14: Some Dehn fillings of the Whitehead link and their location in the Farey graph.

A Dehn filling is determined by a path in the Farey graph, giving a layered solid torus. Figure 14 shows where we begin in the Farey graph, namely in the triangle T_0 with slopes $3/1, 2/1, 1/0$, and paths we take to obtain well-known Dehn fillings, in particular twist knots.

For example, if we attach a degenerate layered solid torus, folding along the edge of slope $1/0$, we will perform $1/1$ Dehn filling, which gives the trefoil knot complement. Since the trefoil is not hyperbolic, Theorem 2.58 is not guaranteed to apply, so we skip this Dehn filling. To obtain other twist knots, first cover slope $1/0$, stepping into triangle T_1 in the Farey graph, then swing R into triangle T_2. From there, the path depends on whether we wish to obtain an even twist knot or an odd one.

Consider performing $-1/1$ Dehn filling, to obtain the complement of the 4_1 knot, or figure-8 knot. This Dehn filling is obtained by attaching a layered solid torus built of two tetrahedra, $\Delta_{3/1}$ and $\Delta_{2/1}$, where our naming convention is as in Section 3.4: Tetrahedron $\Delta_{o_0} = \Delta_{3/1}$ is attached when we step from T_0 to T_1 in the Farey graph, and $\Delta_{o_1} = \Delta_{2/1}$ when we step from T_1 to T_2. Notice that this step in the Farey graph is in the direction R. Then to obtain the 4_1 knot, from T_2 we fold over the edge $E_{1/1}$, identifying $E_{0/1}$ and $E_{1/0}$.

Equations arising from the layered solid torus can be computed with reference only to Theorem 3.17, without writing down the full Neumann–Zagier matrix:

$$\begin{aligned} \Delta_{3/1}: &\quad \gamma_{3/1}\gamma_{1/1} + \gamma_{2/1}^2 - \gamma_{1/0}^2 = 0, \\ \Delta_{2/1}: &\quad -\gamma_{2/1}\gamma_{0/1} + \gamma_{1/1}^2 - \gamma_{1/0}^2 = 0. \end{aligned} \tag{4.2}$$

Observe that in the equation for $\Delta_{3/1}$, $\gamma_{3/1}$, $\gamma_{1/0}$, and $\gamma_{2/1}$ are already known in terms of m and ℓ alone. Hence direct substitution allows us to write $\gamma_{1/1}$ in terms of m and ℓ. Similarly for $\gamma_{0/1}$ in the equation from $\Delta_{2/1}$.

The equations for the figure-8 knot are finally obtained by setting the variables $\gamma_{0/1} = \gamma_{1/0}$. Then the final equation turns the system into a single equation in m and ℓ. The calculations for the figure-8 knot are carried out in Appendix B.

Now consider the 5_2 knot. This is obtained by starting with the same two tetrahedra $\Delta_{3/1}$ and $\Delta_{2/1}$ as in the case of the figure-8 knot. However, instead of folding across the edge $E_{1/1}$, we fold across the edge $E_{1/0}$, and identify $E_{1/1}$ to $E_{0/1}$; see Figure 14. Thus the Ptolemy equations look identical to those above for the figure-8 knot, except set the variables $\gamma_{1/1}$ and $\gamma_{0/1}$ to be equal. As before, substitution gives the A-polynomial. Again the calculations are in Appendix B.

For the 7_2 knot: Turn left from the triangle T_2 in the Farey graph, picking up equation

$$\Delta_{1/0}: \quad \gamma_{1/0}\gamma_{1/2} + \gamma_{1/1}^2 - \gamma_{0/1}^2 = 0,$$

and identify variables $\gamma_{1/2}$ and $\gamma_{0/1}$. Substitution allows us to write $\gamma_{1/2}$ in terms of m and ℓ, and then use this to find the A-polynomial.

For the 9_2 knot: Turn right. Pick up a new equation,

$$\Delta_{1/1}: \quad -\gamma_{1/1}\gamma_{1/3} + \gamma_{1/2}^2 - \gamma_{0/1}^2 = 0,$$

and identify variables $\gamma_{1/3}$ and $\gamma_{0/1}$.

Any twist knot with $2N+1$ crossings is obtained similarly, for $N \geq 4$. The word W in the Farey graph has the form RLRR. . .R. The Ptolemy equations include all the equations above, as well as a sequence of equations

$$-\gamma_{1/k}\gamma_{1/(k+2)} + \gamma_{1/(k+1)}^2 - \gamma_{0/1}^2 = 0 \quad \text{for } 2 \leq k \leq N-1.$$

At the end, the variables $\gamma_{0/1}$ and $\gamma_{1/N-1}$ are identified.

In all cases, a step in the Farey graph gives an equation with a single new variable; we use this equation to write the new variable in terms of m and ℓ. Then direct substitution at the final step yields the A-polynomial.

Twist knots with $2N$ crossings are obtained similarly from a word in the Farey graph of the form RRL. . .L, with corresponding adjustments to the Ptolemy equations to determine the A-polynomial.

Appendix A Nice triangulations of manifolds with torus boundaries

In this appendix, we show that every 3-manifold admits a triangulation that behaves well with Dehn filling by layered solid tori, such that the results of Section 3 apply.

Proposition A.1 *Let $\overline{M}$ be a connected, compact, orientable, irreducible, ∂-irreducible 3-manifold with boundary consisting of $m+1 \geq 2$ tori. Then, for any torus boundary component $\mathbb{T}_0$, there exists an ideal triangulation $\mathcal{T}$ of the interior M of $\overline{M}$ such that the following hold.*

(i) *If $\mathbb{T}_1, \dots, \mathbb{T}_m$ are the torus boundary components of $\overline{M}$ disjoint from $\mathbb{T}_0$, then in M, the cusp corresponding to $\mathbb{T}_j$ for any $j = 1, \dots, m$ meets exactly two ideal tetrahedra, $\Delta_{j,1}$ and $\Delta_{j,2}$. Each of these tetrahedra meets $\mathbb{T}_j$ in exactly one ideal vertex.*

(ii) *There exists a choice of generators for $H_1(\mathbb{T}_0; \mathbb{Z})$, represented by curves $\mathfrak{m}_0$ and $\mathfrak{l}_0$, such that $\mathfrak{m}_0$ and $\mathfrak{l}_0$ meet the cusp triangulation inherited from $\mathcal{T}$ in a sequence of arcs cutting off single vertices of triangles, without backtracking, and such that $\mathfrak{m}_0$ and $\mathfrak{l}_0$ are disjoint from the tetrahedra $\Delta_{j,1}$ and $\Delta_{j,2}$, for all $j = 1, \dots, m$.*

In the notation of Section 2, the number of cusps here is $n_{\mathfrak{c}} = m+1 \geq 2$.

Proof By work of Jaco and Rubinstein [29, Proposition 5.15, Theorem 5.17], $\overline{M}$ admits a triangulation by finite tetrahedra, ie with material vertices, such that the triangulation has all its vertices in $\partial\overline{M}$ and has precisely one vertex in each boundary component. Thus each component of $\partial\overline{M}$ is triangulated by exactly two material triangles.

Adjust this triangulation to a triangulation of M with ideal and material vertices, as follows. For each component of $\partial\overline{M}$, cone the boundary component to infinity. That is, attach $T^2 \times [0, \infty)$. Triangulate by coning: over the single material vertex v in $\mathbb{T}_j$, attach an edge with one vertex on the material vertex, and one at infinity. Over each edge e in $\mathbb{T}_j$, attach a 1/3-ideal triangle, with one side of the triangle on the edge e with two material vertices, and the other two sides on the half-infinite edges stretching to infinity. Finally, over each triangle T in $\mathbb{T}_j$ attach a tetrahedron with one face identified to T, with all material vertices, and all other faces identified to the 1/3-ideal triangles lying over edges of the triangulation of $\partial\overline{M}$.

Note that each cusp of M now meets exactly two tetrahedra, in exactly one ideal vertex of each tetrahedron. To complete the proof, we need to remove material vertices.

Begin by removing a small regular neighbourhood of each material vertex; each such neighbourhood is a ball B in M. Removing B truncates the tetrahedra incident to that material vertex. We will obtain the ideal triangulation by drilling tubes from the balls to the cusp $\mathbb{T}_0$, disjoint from the tetrahedra meeting the other cusps. Thus the triangulation of the distinguished cusp $\mathbb{T}_0$ will be affected, but the triangulations of the other cusps will remain in the form required for the result.

To drill a tube, we follow the procedure of Weeks [44] in Section 3 of that paper (see also [26, Figures 10 and 11] for pictures of this process). That is, truncate all ideal vertices in the triangulation of M. Truncate material vertices by removing a ball neighbourhood, giving a triangulation by truncated ideal tetrahedra of the manifold $\overline{M} - (B_0 \cup \cdots \cup B_m)$, where $B_0, \dots, B_m$ are the ball neighbourhoods of material vertices.

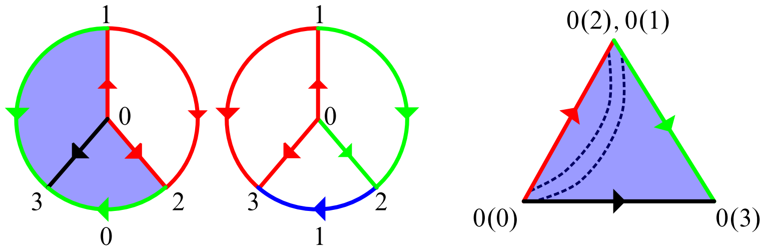


Figure 15: Gluing two tetrahedra as shown on the left yields a triangular pillowcase with a predrilled tube, as shown on the right.

There exists an edge E_0 of the truncated triangulation from $\mathbb{T}_0$ to exactly one of the B_i; call it B_0. Now inductively order the B_i and choose edges $E_1, \dots, E_m$ such that E_j has one endpoint on B_k for some $k < j$ and one endpoint on B_j. Note these edges must necessarily be disjoint from the tetrahedra meeting cusps of M disjoint from $\mathbb{T}_0$, since all edges in such a tetrahedron run from a ball to a different cusp, or from a ball back to itself. Note also that such edges $E_0, \dots, E_m$ must exist, else M is disconnected, contrary to assumption.

Starting with $i = 0$ and then repeating for each $i = 1, \dots, m$, take a triangle T_i with a side on E_i. Cut M open along the triangle T_i and insert a triangular pillow with a predrilled tube as in [44]. The gluing of the two tetrahedra to form the tube is shown in Figure 15, with face pairings given in Table 2. The two unglued faces are then attached to the two copies of T_i. This gives a triangulation of $\overline{M} - (B_{i+1} \cup \dots \cup B_m)$ by truncated tetrahedra, with the ball B_i merged into the boundary component corresponding to $\mathbb{T}_0$. Note it only adds edges, triangles, and tetrahedra, without removing any or affecting the other edges E_j.

When we have repeated the process $m + 1$ times, we have a triangulation of $\overline{M}$ by truncated ideal tetrahedra. By construction, each boundary component $\mathbb{T}_j$, $j = 1, \dots, m$, meets exactly two truncated tetrahedra $\Delta_{j,1}$ and $\Delta_{j,2}$ in exactly two ideal vertices. This gives (i).

For (ii), we trace through the gluing data in Table 2 and Figure 15 to find the cusp triangulation of the pillow with predrilled tube. These are shown in Figure 16. Note there are two connected components. One is a disk made up of vertex 3 of tetrahedron 0 and vertex 2 of tetrahedron 1. The other is an annulus, made up of the remaining truncated vertices.

The cusp triangulation of the manifold $\overline{M} - (B_0 \cup \dots \cup B_m)$ consists of two triangles per torus boundary component, along with $m + 1$ triangulated 2-spheres. When we add the first pillow, we slice open a triangle, which appears in three edges of the cusp triangulation: one on the torus $\mathbb{T}_0$, and the other two

	012	013	023	123
0	1(013)	—	—	1(012)
1	0(123)	0(012)	1(123)	1(023)

Table 2: Gluing instructions to form a triangular pillow with a predrilled tube. Notation is as in [3].

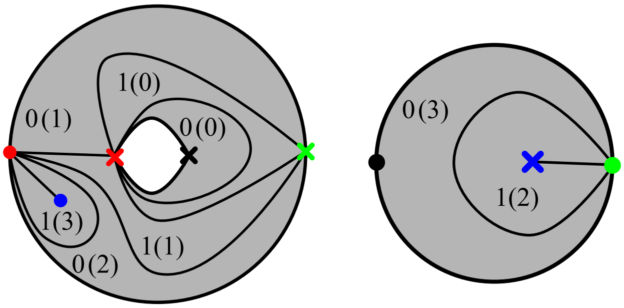


Figure 16: The cusp triangulations of the pillow. Each triangle in the cusp triangulation is labelled, with tetrahedron number (vertex).

on the boundary of the ball B_0. These edges of the cusp triangulation are sliced open, leaving a bigon on $\mathbb{T}_0$ and two bigons on B_0. When the pillow is glued in, the bigons are replaced. One, on the boundary of the ball B_0, is just filled with the disk on the right of Figure 16. One on $\mathbb{T}_0$ is filled with the annulus on the left of Figure 16. The remaining one, on the boundary of B_0, is glued to the inside of the annulus. Thus the cusp triangulation of $\mathbb{T}_0$ is changed by cutting open an edge, inserting an annulus with the triangulation on the left of Figure 16, and inserting a disk into the centre of that annulus with the (new) triangulation of the boundary of B_0.

When we repeat this process inductively for each B_i, we slice open edges of the cusp triangulation of the adjusted $\mathbb{T}_0$, and add in an annulus and disks corresponding to the triangulation of the boundary of B_i. This process only adds triangles; it does not remove or adjust existing triangles, except to separate them by inserting disks.

Now let $\mathfrak{m}_0$ and $\mathfrak{l}_0$ be any generators of $H_1(T_0;\mathbb{Z})$. We can choose representatives that are normal with respect to the triangulation of

$$\overline{M} - (B_0 \cup \cdots \cup B_m).$$

At each step, we replace an edge of the triangulation with a disk. However, note that all such disks must be contained within the centre of the first attached annulus. Now suppose $\mathfrak{m}_0$ runs through the edge that is replaced in the first stage. Then keep $\mathfrak{m}_0$ the same outside the added disk. Within the disc, let it run from one side to the other by cutting off single corners of triangles 0(2), 1(1), 1(0), and 0(1). The new curve is still a generator of homology along with $\mathfrak{l}_0$. It meets the same tetrahedra as before, and the two tetrahedra added to form the tube. It does not meet any of the vertices of the tetrahedra of the ball B_0. The curve $\mathfrak{l}_0$ can also be replaced in the same manner, by a curve cutting through the same cusp triangles, parallel to the segment of $\mathfrak{m}_0$ within these triangles. Inductively, we may replace $\mathfrak{m}_0$ and $\mathfrak{l}_0$ at each stage by curves that are identical to the previous stage, unless they meet a newly added disk, and in this case they only meet the disk in triangles corresponding to the added pillow, not in triangles corresponding to tetrahedra meeting other cusps. The result holds by induction.

Complete the proof by replacing truncated tetrahedra by ideal tetrahedra. □

Appendix B Calculations for some twist knots

In Section 4, we found Ptolemy equations for Dehn fillings of the Whitehead link. In this short appendix, we explain how to use them and direct substitution to find an A-polynomial. This will not immediately look like the standard A-polynomial, because we have chosen a nonstandard longitude and because our equations have extra factors and square roots. After conjugation and a change of basis, we obtain the usual A-polynomials.

To compute the polynomials, we use the equations corresponding to the tetrahedra Δ_0, Δ_1, and Δ_2 of the Whitehead link that lie outside the cusp we will fill, as in (4.1), as well as the equation $\gamma_{1/0} = 1$. Via direct substitution, Δ_2 gives an equation for $\gamma_{3/1}$ in terms of $\gamma_{0(23)}$, which can then be substituted into Δ_1 to give an equation for $\gamma_{2/1}$ in terms of ℓ, m, and $\gamma_{0(23)}$, which can then be substituted into Δ_0 to obtain an equation of $\gamma_{0(23)}$ in terms of ℓ and m. Substituting this into the equations for $\gamma_{2/1}$ and $\gamma_{3/1}$, we obtain

$$\begin{gathered}\gamma_{0(23)}^2 = \frac{m\ell^{1/2} + \ell - \ell^{1/2} - m}{\ell^{1/2}m - \ell^{1/2}},\\ \gamma_{2/1} = \frac{1}{\gamma_{0(23)}} \frac{m^2 - \ell}{m^{1/2}\ell^{1/2}(1-m)}, \quad \gamma_{3/1} = \frac{\ell - m}{\ell^{1/2}(1-m)}, \quad \gamma_{1/0} = 1.\end{gathered} \tag{B.1}$$

Note we have left $\gamma_{0(23)}$ in the equation for $\gamma_{2/1}$ for now, since it is a square root with possible positive or negative sign.

We obtain two more Ptolemy equations from (4.2); the first gives us $\gamma_{1/1}$ in terms of m and ℓ:

$$\gamma_{1/1} = \frac{\ell^{1/2} - m^2}{(-1 + \ell^{1/2})m}. \tag{B.2}$$

We can then use the second to solve for $\gamma_{0/1}$ in terms of m and ℓ (and $\gamma_{0(23)}$):

$$\gamma_{0/1} = -\gamma_{0(23)} \frac{\ell^{1/2}(-1+m)^2(1+m)}{(-1+\ell^{1/2})^2 m^{3/2}}. \tag{B.3}$$

B.1 Figure-8 knot

An A-polynomial for the Figure-8 knot is now obtained by setting $\gamma_{0/1} = \gamma_{1/0} = 1$. To remove (some of) the square roots coming from the $\gamma_{0(23)}$ term, square both sides of (B.3), obtaining

$$1 = \frac{\ell^{1/2}(-1+m)^3(1+m)^2(\ell^{1/2}+m)}{(-1+\ell^{1/2})^3 m^3}.$$

Multiplying through the denominator and moving all terms to the left-hand side, we obtain the following PSL A-polynomial:

$$(\ell^{1/2} - m^2)(\ell^{1/2} + m - \ell^{1/2}m - 2\ell^{1/2}m^2 - \ell^{1/2}m^3 + \ell m^3 + \ell^{1/2}m^4).$$

This will not give the usual PSL A-polynomial for the figure-8 knot, because our choice of longitude ℓ differs from the standard longitude. In fact, checking against SnapPy [8], the red curve shown in Figure 12 is isotopic to the "shortest" curve intersecting the meridian once, under the Euclidean metric inherited from the hyperbolic structure. Thus the standard longitude differs from that shown by subtracting two meridians. Propositions 5.11 and 5.12 of [28] then give the required change of basis for any Dehn filling of the Whitehead link. For the figure-8 knot, the required change of basis is

$$(\ell, m) \mapsto (\ell m^{-2}, m),$$

and after clearing the denominator, the PSL A-polynomial becomes

$$(\ell^{1/2} - m^3)(m^2 + \ell^{1/2}(1 - m - 2m^2 - m^3 + m^4) + \ell m^2).$$

Following Corollary 2.59, we note that the second factor gives the usual SL A-polynomial when we take $L = -\ell^{1/2}$ and $M = m^{1/2}$; compare to [7]:

$$(-L - M^6)(M^4 - L(1 - M^2 - 2M^4 - M^6 + M^8) + L^2 M^4).$$

B.2 The 5_2 knot

An A-polynomial for the 5_2 knot is obtained by setting $\gamma_{0/1} = \gamma_{1/1}$. Set (B.2) equal to (B.3), square both sides and subtract, to obtain the following PSL A-polynomial for the 5_2 knot:

$$\ell + \ell^{1/2}m - 2\ell m + \ell^{3/2}m - \ell^{1/2}m^2 - 2\ell m^2 + 2\ell^{1/2}m^4 + \ell m^4 - m^5 + 2\ell^{1/2}m^5 - \ell m^5 - \ell^{1/2}m^6.$$

Again we change the basis via $(\ell, m) \mapsto (\ell m^{-2}, m)$, and clear the denominator:

$$\ell + \ell^{3/2} - 2m\ell + m^2(\ell^{1/2} - 2\ell) - m^3\ell^{1/2} + m^4\ell + m^5(2\ell^{1/2} - \ell) + 2m^6\ell^{1/2} + m^7(-1 - \ell^{1/2}).$$

To obtain the SL A-polynomial, following Corollary 2.59, we set $L = \pm\ell^{1/2}$ and $M = \pm m^{1/2}$. Again, $L = -\ell^{1/2}$ does the trick. To obtain a formula matching that of Culler [7], we then need to map L to L^{-1}, which corresponds to considering the mirror image of the 5_2 knot. After clearing denominators and multiplying through by -1, the result is

$$1 - L(1 - 2M^2 - 2M^4 + M^8 - M^{10}) - L^2M^4(-1 + M^2 - 2M^6 - 2M^8 + M^{10}) + L^3M^{14}.$$

References

[1] **C Adams**, **W Sherman**, *Minimum ideal triangulations of hyperbolic* 3*-manifolds*, Discrete Comput. Geom. 6 (1991) 135–153 MR Zbl

[2] **D W Boyd**, *Mahler's measure and invariants of hyperbolic manifolds*, from "Number theory for the millennium, I", A K Peters, Natick, MA (2002) 127–143 MR Zbl

[3] **B A Burton**, **R Budney**, **W Pettersson**, et al., *Regina*: *software for low-dimensional topology* (1999–2019) Available at `https://regina-normal.github.io`

[4] **A A Champanerkar**, *A-polynomial and Bloch invariants of hyperbolic* 3*-manifolds*, PhD thesis, Columbia University (2003) Available at `https://www.proquest.com/docview/305332823`

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

[6] **D Cooper**, **D D Long**, *Remarks on the A-polynomial of a knot*, J. Knot Theory Ramifications 5 (1996) 609–628 MR Zbl

[7] **M Culler**, *A-polynomials*, electronic resource (2013) Available at `http://homepages.math.uic.edu/~culler/Apolynomials/`

[8] **M Culler**, **N M Dunfield**, **M Goerner**, **J R Weeks**, *SnapPy, a computer program for studying the geometry and topology of* 3*-manifolds* (2016) Available at `http://snappy.computop.org`

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

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

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

[12] **T Dimofte**, *Quantum Riemann surfaces in Chern–Simons theory*, Adv. Theor. Math. Phys. 17 (2013) 479–599 MR Zbl

[13] **T Dimofte**, **R van der Veen**, *A spectral perspective on Neumann–Zagier*, preprint (2014) arXiv 1403.5215

[14] **V Fock**, **A Goncharov**, *Moduli spaces of local systems and higher Teichmüller theory*, Publ. Math. Inst. Hautes Études Sci. 103 (2006) 1–211 MR Zbl

[15] **S Fomin**, **M Shapiro**, **D Thurston**, *Cluster algebras and triangulated surfaces, I: Cluster complexes*, Acta Math. 201 (2008) 83–146 MR Zbl

[16] **S Fomin**, **L Williams**, **A Zelevinsky**, *Introduction to cluster algebras*: *chapters* 1–3, preprint (2016) arXiv 1608.05735

[17] **C Frohman**, **R Gelca**, **W Lofaro**, *The A-polynomial from the noncommutative viewpoint*, Trans. Amer. Math. Soc. 354 (2002) 735–747 MR Zbl

[18] **S Garoufalidis**, *On the characteristic and deformation varieties of a knot*, from "Proceedings of the Casson Fest", Geom. Topol. Monogr. 7, Geom. Topol. Publ., Coventry (2004) 291–309 MR Zbl

[19] **S Garoufalidis**, **T T Q Lê**, *The colored Jones function is q-holonomic*, Geom. Topol. 9 (2005) 1253–1293 MR Zbl

[20] **S Garoufalidis**, **T W Mattman**, *The A-polynomial of the $(-2, 3, 3+2n)$ pretzel knots*, New York J. Math. 17 (2011) 269–279 MR Zbl

[21] **S Garoufalidis**, **D P Thurston**, **C K Zickert**, *The complex volume of $\mathrm{SL}(n,\mathbb{C})$-representations of* 3*-manifolds*, Duke Math. J. 164 (2015) 2099–2160 MR Zbl

[22] **M Gekhtman**, **M Shapiro**, **A Vainshtein**, *Cluster algebras and Weil–Petersson forms*, Duke Math. J. 127 (2005) 291–311 MR Zbl

[23] **M Goerner**, **C K Zickert**, *Triangulation independent Ptolemy varieties*, Math. Z. 289 (2018) 663–693 MR Zbl

[24] **F Guéritaud**, **S Schleimer**, *Canonical triangulations of Dehn fillings*, Geom. Topol. 14 (2010) 193–242 MR Zbl

[25] **J-Y Ham**, **J Lee**, *An explicit formula for the A-polynomial of the knot with Conway's notation $C(2n, 3)$*, J. Knot Theory Ramifications 25 (2016) art. id. 1650057 MR Zbl

[26] **K Hikami**, **R Inoue**, *Braids, complex volume and cluster algebras*, Algebr. Geom. Topol. 15 (2015) 2175–2194 MR Zbl

[27] **J Hoste**, **P D Shanahan**, *A formula for the A-polynomial of twist knots*, J. Knot Theory Ramifications 13 (2004) 193–209 MR Zbl

[28] **J A Howie**, **D V Mathews**, **J S Purcell**, **E K Thompson**, *A-polynomials of fillings of the Whitehead sister*, Int. J. Math. 34 (2023) art. id. 2350085 MR Zbl

[29] **W Jaco**, **J H Rubinstein**, *0-efficient triangulations of 3-manifolds*, J. Differential Geom. 65 (2003) 61–168 MR Zbl

[30] **W Jaco**, **J H Rubinstein**, *Layered-triangulations of 3-manifolds*, preprint (2006) arXiv math/0603601

[31] **D V Mathews**, *An explicit formula for the A-polynomial of twist knots*, J. Knot Theory Ramifications 23 (2014) art. id. 1450044 MR Zbl Correction in 23 (2014) art. id. 1492001

[32] **W D Neumann**, *Combinatorics of triangulations and the Chern–Simons invariant for hyperbolic 3-manifolds*, from "Topology '90", Ohio State Univ. Math. Res. Inst. Publ. 1, de Gruyter, Berlin (1992) 243–271 MR Zbl

[33] **W D Neumann**, **D Zagier**, *Volumes of hyperbolic three-manifolds*, Topology 24 (1985) 307–332 MR Zbl

[34] **M Newman**, *Integral matrices*, Pure Appl. Math. 45, Academic Press, New York (1972) MR Zbl

[35] **Y Ni**, **X Zhang**, *Detection of knots and a cabling formula for A-polynomials*, Algebr. Geom. Topol. 17 (2017) 65–109 MR Zbl

[36] **R C Penner**, *The decorated Teichmüller space of punctured surfaces*, Comm. Math. Phys. 113 (1987) 299–339 MR Zbl

[37] **K L Petersen**, *A-polynomials of a family of two-bridge knots*, New York J. Math. 21 (2015) 847–881 MR Zbl

[38] **H Segerman**, *A generalisation of the deformation variety*, Algebr. Geom. Topol. 12 (2012) 2179–2244 MR Zbl

[39] **P B Shalen**, *Representations of 3-manifold groups*, from "Handbook of geometric topology", North-Holland, Amsterdam (2002) 955–1044 MR Zbl

[40] **N Tamura**, **Y Yokota**, *A formula for the A-polynomials of $(-2, 3, 1+2n)$-pretzel knots*, Tokyo J. Math. 27 (2004) 263–273 MR Zbl

[41] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `https://url.msp.org/gt3m`

[42] **W P Thurston**, *Three-dimensional geometry and topology, I*, Princeton Math. Ser. 35, Princeton Univ. Press (1997) MR Zbl

[43] **A T Tran**, *The A-polynomial 2-tuple of twisted Whitehead links*, Int. J. Math. 29 (2018) art. id. 1850013 MR Zbl

[44] **J Weeks**, *Computation of hyperbolic structures in knot theory*, from "Handbook of knot theory", Elsevier, Amsterdam (2005) 461–480 MR Zbl

[45] **L K Williams**, *Cluster algebras: an introduction*, Bull. Amer. Math. Soc. 51 (2014) 1–26 MR Zbl

[46] **C K Zickert**, *Ptolemy coordinates, Dehn invariant and the A-polynomial*, Math. Z. 283 (2016) 515–537 MR Zbl

School of Mathematics, Monash University
Clayton VIC, Australia

`josh.howie@monash.edu`, `daniel.mathews@monash.edu`, `jessica.purcell@monash.edu`

Received: 10 August 2021 Revised: 6 February 2024

Algebraic & Geometric Topology **25**:3 (2025) 1321–1375
DOI: 10.2140/agt.2025.25.1321
Published: 20 June 2025

The Alexandrov theorem for $2+1$ flat radiant spacetimes

LÉO MAXIME BRUNSWIC

Fillastre showed that one can realize the universal covering of any locally Euclidean surface Σ with conical singularities of angle bigger than 2π as the boundary of a convex Fuchsian polyhedron in 3-dimensional Minkowski space in a unique manner, up to the action of $\mathrm{SO}(1,2)\ltimes\mathbb{R}^3$, the affine isometry group of Minkowski space. The proof used a so-called deformation method, which is nonconstructive. We adapt a variational method previously used by Volkov, Bobenko, Izmestiev, and Fillastre on similar problems to provide an effective proof of Fillastre's theorem. In passing, we extend Fillastre's theorem as follows. Without assumptions on the conical angles θ_i of Σ and for any choice of nonnegative $(\kappa_i)_{i\in[\![1,s]\!]}$ such that $\kappa_i<\theta_i$ and $\kappa_i\le 2\pi$, there exists a unique couple (M,P) where M belongs to a class of singular locally Minkowski manifolds we define with s singular lines of respective conical angle κ_i, and P is a convex polyhedron in M whose boundary ∂P is a Cauchy surface isometric to Σ, the i^{th} conical singularity of ∂P lying on the i^{th} singular line of M. Our result unifies Fillastre's theorem and instances of Penner–Epstein convex hull constructions, corresponding respectively to $\kappa_i=2\pi$ and $\kappa_i=0$ for all i.

51M05, 52B10, 52B70, 53C50, 57K35; 53C42, 57M60

1 Introduction

1.1 The Alexandrov theorem

Let C be a cube in the 3-dimensional Euclidean space $\mathbb{E}^3$ and consider $\Sigma:=\partial C$ its boundary, as represented in Figure 1. On the one hand, Σ is a surface homeomorphic to the 2-dimensional sphere $\mathbb{S}^2$; on the other hand, Σ is naturally endowed with a locally Euclidean metric with six conical singularities, each of angle $\frac{3}{2}\pi$.

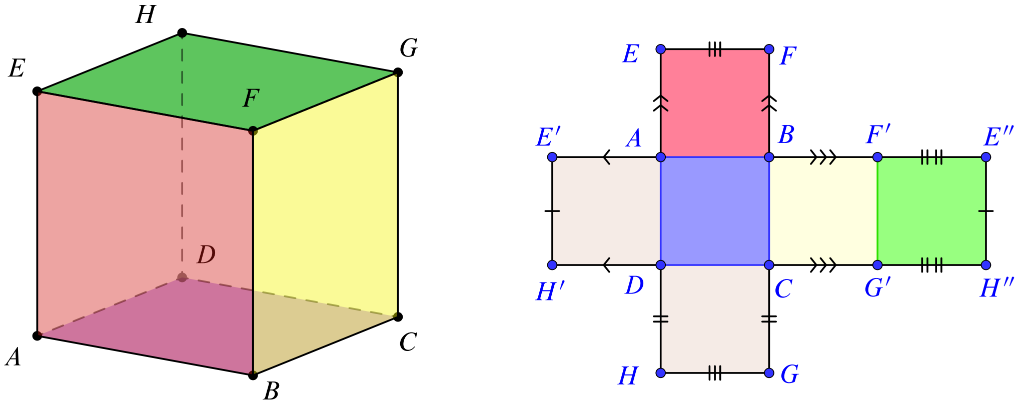


Figure 1

More generally, the boundary of any compact convex polyhedron in $\mathbb{E}^3$ is homeomorphic to the 2-dimensional sphere. It is naturally endowed with a locally Euclidean metric with conical singularities of angles less than 2π.

A classical theorem of Alexandrov [2] shows that this construction is actually bijective:

Theorem [2] *Let Σ be a locally Euclidean surface with conical singularities of angles less than 2π and homeomorphic to the sphere $\mathbb{S}^2$. There exists a compact convex polyhedron P in $\mathbb{E}^3$ such that ∂P is isometric to Σ. Furthermore, two such polyhedra are congruent.*

Using a so-called deformation method, Alexandrov proved generalizations to convex polyhedrons in $\mathbb{H}^3$ and $\mathbb{S}^3$; this method is, however, not effective since it does not provide an efficient way to construct the convex polyhedra these theorems predict.

1.2 Generalizations to space forms and main result

In the 2000s, Izmestiev and Bobenko gave a new proof of the Alexandrov theorem by a variational, therefore effective, method. See Kane, Price, and Demaine [23] for a complexity analysis of the resulting algorithm. Rivin, Hodgson, Schlenker, and Fillastre proved generalizations to Lorentzian space forms (Minkowski, de Sitter, and anti-de Sitter), in which case conical singularities of the locally Euclidean surface have angles greater than 2π. The Alexandrov problem can then be stated in a more general context that has been recently studied systematically by Fillastre and Izmestiev.

Problem *Let Σ be a closed surface of genus g endowed with a singular metric of constant curvature $K \in \{-1, 0, 1\}$ and cone angles all bigger that 2π (case $\varepsilon = -$) or all less than 2π (case $\varepsilon = +$). Denote by X_K^ε the model space of constant curvature K. It is Riemannian if $\varepsilon = +$ and Lorentzian if $\varepsilon = -$.*

Is there a convex polyhedral Fuchsian realization of Σ in X_K^ε? Furthermore, is this polyhedron unique up to congruence?

g	K	ε	DM	VM
0	0	$+$	[2]	[6]
0	-1	$+$	[3]	
0	1	$+$	[3]	
0	1	$-$	[21]	
1	-1	$+$		[18]
1	1	$-$		[19]
≥ 2	-1	$+$	[15]	
≥ 2	-1	$-$	[17]	
≥ 2	0	$-$	[17]	[B]
≥ 2	1	$-$	[31]	

Table 1: See Alexandrov [2; 3], Bobenko and Izmestiev [6], Fillastre [15; 17], Fillastre and Izmestiev [18; 19], Hodgson and Rivin [21], and Schlenker [31].

The signature of the X_K^ε and the Gauss–Bonnet formula impose constraints on (g, K, ε). Table 1 is based upon work of Fillastre [16] and sums up all possible situations, together with references to proofs by deformation (DM) and/or variational (VM) methods; [B] refers to the present work.

Proving Fillastre's theorem — the case where $(g, K, \varepsilon) = (\geq 2, 0, -)$ and X_K^ε is Minkowski space $\mathbb{E}^{1,2}$ — by a variational method is the primary motivation of the present work. Here "convex polyhedral Fuchsian realization" means that we build a triple (ρ, ι, P), where ρ is a representation of $\pi_1(\Sigma) \to \mathrm{Isom}(\mathbb{E}^{1,2})$, ι is a ρ-equivariant embedding $\iota\colon \widetilde{\Sigma} \to \mathbb{E}^{1,2}$ of the universal covering of Σ, and P is a convex globally ρ-invariant polyhedron, with the additional hypothesis that ρ fixes a point and acts cocompactly on the hyperboloid model of the hyperbolic plane $\mathbb{H}^2 = \{(t, x, y) \mid t^2 - x^2 - y^2 = 1, t > 0\} \subset \mathbb{E}^{1,2}$.

To this end, we adapt the variational method successfully used by Bobenko, Fillastre, and Izmestiev [6; 18; 19]; we derive Alexandrov–Fillastre and obtain a generalization to a class of singular locally Minkowski 3-manifolds: radiant singular flat spacetimes, which we shall describe thereafter.

Theorem *Let Σ be a closed locally Euclidean surface of genus g with s marked conical singularities*[1] *of angles $(\theta_i)_{i \in [\![1,s]\!]}$. For all*

$$\kappa \in \left(\prod_{i=1}^{s} [0, \min(\theta_i, 2\pi)] \right) \setminus \{(\theta_i)_{i \in [\![1,s]\!]}\},$$

there exists a radiant singular flat spacetime M homeomorphic to $\Sigma \times \mathbb{R}$ with exactly s singular lines of angles $\kappa_1, \dots, \kappa_s$ and a convex polyhedron $P \subset M$ whose boundary is isometric to Σ. The boundary of P is a Cauchy surface of M.

Furthermore, if for all $i \in [\![1,s]\!]$, $\kappa_i < \theta_i$, then (M, P) is unique up to equivalence.

Finally, if for some $i \in [\![1,s]\!]$, $\theta_i \leq \pi$ and $\kappa \in \mathbb{R}_+^s$ is such that $\kappa_i > \theta_i$, then there is no such couple (M, P).

[1] We allow marked conical singularities with angle 2π, which are hence not singular but marked nonetheless.

Remark By taking all the $\theta_i > 2\pi$ and $\kappa_i = 2\pi$ we obtain a manifold M whose universal covering is isomorphic to a subdomain of Minkowski space X_0^- (via a theorem of Mess [25]). Fillastre's theorem thus follows.

1.3 Radiant spacetimes

Before giving the outline of the variational method, we quickly describe radiant spacetimes. A more thorough description is given in the appendix, together with technical results. We denote by $\mathbb{E}^{1,2}$ the 3-dimensional Minkowski space (the oriented affine space $\mathbb{R}^3$ together with the quadratic form[2] $\boldsymbol{g} := \mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2$ written in some fixed choice of Cartesian coordinates t, x, y) and by $\mathrm{Isom}_0(\mathbb{E}^{1,2})$ the identity component of the Lie group of affine isometries of $\mathbb{E}^{1,2}$, namely $\mathrm{SO}_0(1,2) \ltimes \mathbb{R}^3$. We denote by $O := (0,0,0) \in \mathbb{E}^{1,2}$ the origin of $\mathbb{E}^{1,2}$. A vector $u \neq 0$ is spacelike (resp. timelike, lightlike, causal) if $\boldsymbol{g}(u) < 0$ (resp. $\boldsymbol{g}(u) > 0$, $\boldsymbol{g}(u) = 0$, $\boldsymbol{g}(u) \geq 0$). A causal vector is future (resp. past) if its t coordinate is positive (resp. negative). Minkowski space is naturally endowed with two order relations: the causal order $\leq$ and the chronological order $\lll$ (the associated strict relation is denoted by $\ll$). Given $p, q \in \mathbb{E}^{1,2}$ then $p < q$ (resp. $p \ll q$) if $q - p$ is future causal (resp. future timelike). The group $\mathrm{Isom}_0(\mathbb{E}^{1,2})$ preserves the orientation of $\mathbb{E}^{1,2}$ as well as the causal and the chronological orders. We define the causal future of p, denoted by $J^+(p) := \{q \in M \mid p \leq q\}$, as well as the chronological future of p, denoted by $I^+(p) := \{q \in M \mid p \ll q\}$. The causal past, as well as the chronological past, are defined accordingly. A plane in $\mathbb{E}^{1,2}$ is spacelike (resp. timelike, lightlike) if the induced quadratic form is positive definite (resp. definite, degenerated), and a normal to such a plane is a timelike vector (resp. spacelike vector, lightlike vector). By convention, all spacelike and lightlike planes are oriented by a future normal vector.

Radiant spacetimes are obtained via gluings of cones in $J^+(O)$ of triangular basis, ie

$$C = \{ru \mid r \in \mathbb{R}_+^*, u \in T\},$$

with T some affine spacelike triangle in $J^+(O)$. We will not consider any such gluing with boundary.

Such gluings have a natural $(\mathrm{SO}_0(1,2), I^+(O))$-structure in the sense of Ehresmann [12], Thurston [33], or Goldman [20] on the complement of the edges of the cones (the 1-facet of the simplicial complex). These "singular" edges are one of two types:

- Timelike edges are locally modeled on so-called massive particles (the plane orthogonal to the given edge is a Euclidean conical singularity of some angle $\kappa > 0$).
- Lightlike edges are locally modeled on so-called extreme BTZ-like singularities (see the appendix and Barbot, Bonsante and, Schlenker [4] for more details). The convention is that such an edge bears a cone angle $\kappa = 0$.

[2] Beware we chose a sign convention for $\boldsymbol{g}$ different from most of the literature to favor positive values of $\boldsymbol{g}$ on the relevant domains and avoid defining two different quadratic forms.

For brevity sake, we will write $\mathcal{F}$ instead of $I^+(O)$ and $\mathcal{F}$-manifold instead of $(\mathrm{SO}_0(1,2),\mathcal{F})$-manifold. Cones in $J^+(O)$ have a natural $\mathrm{SO}_0(1,2)$-invariant 1-dimensional foliation formed by the rays from the origin of the form $\mathcal{R}_u := \{ru \mid r > 0\}$ with u in $J^+(O)$; therefore each radiant spacetime comes with such a foliation. The statement "the surface Σ is a Cauchy surface of the radiant spacetime M" is understood in our context as "the surface Σ is spacelike and intersects all rays of the natural foliation".

Equivalence in our context has to be understood in the following way: two couples (M,P) and (M',P') are equivalent if there exists an isomorphism $M \to M'$ of singular $\mathbb{E}^{1,2}$-manifolds (a homeomorphism sending regular domain to regular domain and which is an $\mathbb{E}^{1,2}$-morphism on the regular domain) which induces a bijection $P \to P'$.

1.4 The variational method

Now that the terminology is clarified, the variational method proceeds as follows:

(1) Consider a closed locally Euclidean surface Σ of genus g with $s \in \mathbb{N}^*$ marked conical singularities $\theta_1,\dots,\theta_s \in \mathbb{R}_+^*$ and define S the set of marked points.

(2) Choose an arbitrary couple $(\tau,\mathcal{T})$ with $\tau\colon S \to \mathbb{R}_+$ and $\mathcal{T}$ a triangulation of Σ whose set of vertices is S.

(3) For each triangle T of $\mathcal{T}$, choose a direct affine isometric embedding

$$\iota\colon T \to J^+(O) := \{t > 0,\ \boldsymbol{g} \ge 0\} \subset \mathbb{E}^{1,2}$$

in such a way that for each vertex s of T we have $\boldsymbol{g} \circ \iota(s) = \tau(s)$.

(4) To each triangle T is then associated the cone of rays from $O := (0,0,0)$ through T in $\mathbb{E}^{1,2}$; glue these cones together following the same combinatorics as $\mathcal{T}$. The gluing is a 3-manifold M endowed with a flat Lorentzian metric on the complement of the rays through the vertices of $\mathcal{T}$. Furthermore we have a natural embedding $\iota\colon \Sigma \to M$ in such a way that $\iota(\Sigma)$ is the boundary of the polyhedron $P := J^+(\iota(\Sigma))$ of M.

(5) Study the domain of $\tau \in (\mathbb{R}_+)^S$ such that the polyhedron P is convex; ι is then called convex, and show that for a given τ there is at most one triangulation $\mathcal{T}$ (up to equivalence) for which the embedding ι is convex; a τ is then admissible if it has such a triangulation.

(6) Choose some target Lorentzian angles $\bar{\kappa}$ and define an Einstein–Hilbert functional on the space of admissible $\tau \in (\mathbb{R}_+)^S$ in such a way that each of its critical points induces a manifold M with Lorentzian cone angle $\bar{\kappa}$ around the rays through the vertices of $\mathcal{T}$.

(7) Finally, study this functional and show it admits a unique critical point.

1.5 The special case $\kappa = 0$

Penner gives another viewpoint on our result [27; 28], constructing a cellulation of the decorated Teichmüller space of a closed surface Σ with s marked points $S = \{\sigma_1,\dots,\sigma_s\}$ viewed as the space of marked finite-volume complete hyperbolic surfaces with s cusps homeomorphic to $\Sigma \setminus S$ together with a choice of a positive number on each cusp. Consider such a surface Σ^*. The universal covering of Σ^* naturally identifies with the usual hyperbolic plane $\mathbb{H}^2 := \{(t,x,y) \in \mathbb{E}^{1,2} \mid \boldsymbol{g}(t,x,y) = 1, t > 0\}$ in $\mathbb{E}^{1,2}$, and the positive number λ_σ on each cusp σ corresponds to a point on the lightlike rays corresponding to the cusp:

- There exists a unique horocycle $\mathcal{H}_{\sigma,\lambda_\sigma}$ of length λ_σ around σ.
- Consider a ray $\mathcal{R}$ fixed by a parabolic holonomy of Σ^* and a point $p \in \mathcal{R}$. The intersection of the future light cone of p (the set $\{q \in \mathbb{E}^{1,2} \mid \boldsymbol{g}(q-p) = 0, t(q-p) > 0\}$) with $\mathbb{H}^2$ is a horocycle around $\mathcal{R}$, and every horocycle is obtained in this manner.

Penner then considers the surface obtained as the boundary of the convex hull of these points.[3] He shows the surface obtained is locally Euclidean, its quotient by the holonomy of Σ^* is a locally Euclidean surface $\Sigma_{\mathbb{E}^2}$ with s conical singularities. Furthermore, the convex hull is a polyhedron, the faces of which induce a cellulation on $\Sigma_{\mathbb{E}^2}$ with marked points S. He notes that this cellulation is simply the Delaunay cellulation of $(\Sigma_{\mathbb{E}^2}, S)$. It is not hard to see that

(1) this construction actually defines a natural bijection from the decorated Teichmüller space of (Σ, S) to the deformation space of locally Euclidean metrics on Σ with arbitrary conical singularities on S,

(2) the quotient by the holonomy of Σ^* of the union of $I^+(O)$ with the rays fixed by parabolic holonomy of Σ^* is a radiant spacetime with s conical singularities of angle 0.

Penner construction can thus be seen as the special case of our theorem where $\kappa = 0$ and (Σ, S) runs through all locally Euclidean surfaces with s conical singularities at S of arbitrary angles.

Acknowledgments

This work was initiated as part of a PhD project at Laboratoire de Mathématiques d'Avignon, Université d'Avignon et des Pays de Vaucluse, under Thierry Barbot's supervision, then continued as part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement ERC advanced grant 740021–ARTHUS, PI: Thomas Buchert). The author thanks Thierry Barbot and Thomas Buchert for their continuous support, encouragement, and valuable remarks, François Fillastre and Marc Troyanov for their many corrections and comments on earlier versions of the manuscript, Graham Smith for his remarks that led to significant simplifications, the reviewers for their useful comments, as well as Masoud Hasani, Jean-Marc Schlenker, Frédéric Paulin, Gye-Seon Lee, Suhyoung Choi, Francis Bonahon, Anna Wienhard, Ivan Izmestiev, Erwann Delay, Miguel Sánchez, Abdelghani Zeghib, Philippe Delanoë, Daniel Monclair, Vincent Pecastaing, Roman Prosanov, Rabah Souam, Andrea Seppi, Tengren Zhang, Jeffrey Danciger, Nicolas Tholozan, Qiyu Chen, and Clément Guérin for valuable discussions.

[3] A joint work of Penner and Epstein [13] generalizes this construction.

2 Convex τ-suspension and polyhedral embedding

In the present section, we shall define and study τ-suspension of a singular locally Euclidean surface (Σ, S). A cellulation of Σ is a homeomorphism between Σ and a gluing of affine convex dimension-n polyhedra along $(n-1)$-facets. We identify k-facets with their image in Σ. All cellulations considered in this section have totally geodesic facets.

Definition 2.1 Let (Σ, S) be a compact Euclidean surface with conical singularities with a finite subset S of marked points such that $\mathrm{Sing}(\Sigma) \subset S$, and let $\mathcal{C}$ be a cellulation of (Σ, S). $\mathcal{C}$ is adapted if the set of vertices of $\mathcal{C}$ is exactly S.

Definition 2.2 Let (Σ, S) be a compact Euclidean surface with conical singularities with a finite subset S of marked points such that $\mathrm{Sing}(\Sigma) \subset S$. Let M be a singular $\mathbb{E}^{1,2}$-manifold. An embedding $\iota\colon \Sigma \to M$ is polyhedral if there exists a geodesic adapted cellulation $\mathcal{C}$ of (Σ, S) such that on each cell C, the restriction of ι to $\mathrm{Int}(C)$ is an isometric affine map into the regular locus of L.

The notion of an isometric affine map is well defined in this context. Indeed, both $\mathbb{E}^2$ and $\mathbb{E}^{1,2}$ are affine spaces endowed with a semi-Riemannian metric; the regular loci of Σ and M are endowed with an $\mathbb{E}^2$-structure and an $\mathbb{E}^{1,2}$-structure, respectively.

The quadratic form on $\mathbb{E}^{1,2}$ is a $\mathrm{SO}_0(1,2)$-invariant function defined on the underlying vector space $\overrightarrow{\mathbb{E}^{1,2}}$:

$$\boldsymbol{g}\colon \overrightarrow{\mathbb{E}^{1,2}} \to \mathbb{R}, \quad (t,x,y) \mapsto t^2 - x^2 - y^2.$$

We extend the definition of $\boldsymbol{g}$ to $\mathbb{E}^{1,2}$ via the identification $\mathbb{E}^{1,2} \to \overrightarrow{\mathbb{E}^{1,2}}$, $x \mapsto x - O$. The map $\boldsymbol{g}$ is positive on the future of the origin in $\mathbb{E}^{1,2}$, namely $J^+(O) := \{(t,x,y) \in \mathbb{R}^3 \mid t^2 - x^2 - y^2 \geq 0 \text{ and } t > 0\}$; furthermore, it induces a Cauchy time function on $I^+(O)$, ie an increasing map $(I^+(O), \leq) \to (\mathbb{R}_+^*, \leq)$ whose restriction to any nonextendible future causal curve of $I^+(O)$ is surjective (see the appendix for more details on the structure of singular $\mathcal{F}$-manifolds). Since $\boldsymbol{g}$ is $\mathrm{SO}_0(1,2)$-invariant, it induces a well-defined nondecreasing function on every radiant singular flat spacetime.

In a radiant singular flat spacetime, the surface $\boldsymbol{g} = 1$ is a hyperbolic surface with conical singularities and cusps, which is complete and has finite volume. One can prove that the association $M \mapsto \{\boldsymbol{g} = 1\}$ induces a bijection from the deformation space of marked radiant singular flat spacetimes to the deformation space of marked finite-volume complete hyperbolic surfaces with conical singularities and cusps; see Theorem 6 in the appendix.

2.1 Affine embedding of triangles into $\mathbb{E}^{1,2}$

The goal of this section is mainly to introduce terminology that will be used throughout the paper and to prove a parametrization of polyhedral embeddings into radiant singular flat spacetimes of a singular locally Euclidean surface by the class of *distance-like* function we introduce. This last point is the object of Theorem 1.

Lemma 2.3 *Let $T=[ABC]$ be a nondegenerated Euclidean triangle and let $\tau\colon\{A,B,C\}\to\mathbb{R}$. There exists a unique couple $(\tau_0,\omega)\in\mathbb{R}\times\mathbb{E}^2$ such that the map*

$$\tilde\tau\colon\mathbb{E}^2\to\mathbb{R},\quad x\mapsto\tau_0-d(x,\omega)^2$$

extends τ.

Furthermore, if $\tau\geq 0$ then $\tau_0>0$ and $\tilde\tau>0$ on the triangle $[ABC]$, except possibly at A, B, or C.

Proof Identify $\mathbb{E}^2$ to $\mathbb{R}^2$ via Cartesian coordinates (x,y); without loss of generality, we can assume $A=(0,0)$, and we write $B=(x_B,y_B)$ and $C=(x_C,y_C)$. Finding $\tilde\tau$ is equivalent to solving the following system in $\omega=(x_\omega,y_\omega)$ and τ_0:

$$\begin{cases}\tau_A=\tau_0-x_\omega^2-y_\omega^2,\\ \tau_B=\tau_0-(x_\omega-x_B)^2-(y_\omega-y_B)^2,\\ \tau_C=\tau_0-(x_\omega-x_C)^2-(y_\omega-y_C)^2,\end{cases}\iff\begin{cases}x_\omega^2+y_\omega^2+\tau_A=\tau_0,\\ \tau_B-\tau_A+x_B^2+y_B^2=2x_\omega x_B+2y_\omega y_B,\\ \tau_C-\tau_A+x_C^2+y_C^2=2x_\omega x_C+2y_\omega y_C.\end{cases}$$

Since A, B and C are in general position, the second and third lines form a nonsingular linear system of unknown (x_ω,y_ω). The first line is already solved. Existence and uniqueness of $\tilde\tau$ follows.

Assume $\tau\geq 0$, since A, B and C are distinct, ω is distinct from one of them, say $P\in\{A,B,C\}$. Then $0\leq\tau_P=\tau_0-d(P,\omega)^2<\tau_0$. Furthermore, $\tilde\tau$ is strictly concave, so its minimum on $[ABC]$ is reached in the set of extremal points, eg $\{A,B,C\}$ and nowhere else. □

Lemma 2.4 *Let $A,B,A',B'\in J^+(O)$, $A\neq B$ and $A'\neq B'$ be such that $\boldsymbol{g}(A)=\boldsymbol{g}(A')$, $\boldsymbol{g}(B)=\boldsymbol{g}(B')$ and $\boldsymbol{g}(B-A)=\boldsymbol{g}(B'-A')$. Then there exists a unique isometry $\gamma\in\mathrm{SO}_0(1,2)$ such that $\gamma A=A'$ and $\gamma B=B'$. Furthermore, if C is on a given side of the oriented plane (OAB), then γC is on the same side of $(OA'B')$.*

Proof The group $\mathrm{SO}_0(1,2)$ acts transitively on each of the sets $(\boldsymbol{g}|_{J^+(O)\setminus\{O\}})^{-1}(\tau_0)$ for $\tau_0\geq 0$. There thus exists some $\gamma_0\in\mathrm{SO}_0(1,2)$ such that $\gamma_0A=A'$. The stabilizer of A' under the action of $\mathrm{SO}_0(1,2)$ is a 1-parameter subgroup (either parabolic or elliptic depending on whether (OA') is lightlike or timelike); under its action, the orbit of γ_0B is

$$\{x\in J^+(O)\mid\boldsymbol{g}(x-A')=\boldsymbol{g}(\gamma_0B-A')\text{ and }\boldsymbol{g}(x)=\boldsymbol{g}(\gamma_0B)\}.$$

The stabilizer of A' acts freely on this set, so there exists a unique γ with the wanted properties. Finally, $\mathrm{SO}_0(1,2)$ preserves orientation, and the result follows. □

Proposition 2.5 *Suppose that $T=[ABC]$ is an oriented nondegenerated Euclidean triangle and let $\tau\colon\{A,B,C\}\to\mathbb{R}_+$. There exists a direct isometric affine embedding $\iota\colon T\to J^+(O)$ such that $\tau=\boldsymbol{g}\circ\iota|_{\{A,B,C\}}$, where $\iota(T)$ is endowed with the orientation induced by a future-pointing normal vector.*

Furthermore,

- *such an embedding is unique up to the action of $\mathrm{SO}_0(1,2)$,*
- *$\boldsymbol{g}\circ\iota=\tilde\tau$, where $\tilde\tau$ is given by Lemma 2.3.*

Proof Endow $\mathbb{E}^{1,2}$ with Cartesian coordinates (t,x,y), write $O=(0,0,0)$ the origin, and identify $\mathbb{E}^2$ with $\{t=0\}\subset\mathbb{E}^{1,2}$. Take $(\tau_0,\omega)\in\mathbb{R}\times\mathbb{E}^2$ and $\tilde{\tau}$ given by Lemma 2.3, and define

$$\iota\colon T\to\mathbb{E}^{1,2},\qquad x\mapsto x+\vec{u}\quad\text{with } \vec{u}=\begin{pmatrix}\sqrt{\tau_0}\\ -\overrightarrow{O\omega}\end{pmatrix}.$$

Write $\omega=(x_\omega,y_\omega)$. For $(x,y)\in T$, we have

$$\boldsymbol{g}\circ\iota(x,y)=\sqrt{\tau_0}^2-(x-x_\omega)^2-(y-y_\omega)^2=\tilde{\tau}(x,y).$$

Since $\tau\geq 0$, by Lemma 2.3 $\tilde{\tau}\geq 0$; hence $\boldsymbol{g}\circ\iota|_T\geq 0$. Moreover, $\sqrt{\tau_0}>0$, thus $\iota(T)\subset J^+(O)$. The existence statement follows, as well as the second additional point.

If ι and ι' are two such embeddings, by Lemma 2.4 there exists a unique isometry sending $\iota(A)$ on $\iota'(A)$ and $\iota(B)$ on $\iota'(B)$. There thus exist exactly two points $P_1,P_2\in J^+(O)$ such that $\boldsymbol{g}(P_i)=\tau(C)$, $\mathrm{d}(A,C)^2=\boldsymbol{g}(\iota(P_i)-\iota(A))$ and $\mathrm{d}(B,C)^2=\boldsymbol{g}(\iota(P_i)-\iota(B))$ for $i\in\{1,2\}$. Since these two points are each other's images by the reflection across the plane $(O\iota(A)\iota(B))$ which is orientation-reversing and preserves $\leq$, exactly one induces the right orientation. □

Definition 2.6 (f-triangulation) Let (Σ,S) be a singular locally Euclidean surface and let $f\colon\Sigma\to\mathbb{R}$. A triangulation $\mathcal{T}$ is an f-triangulation if $\mathcal{T}$ is a geodesic triangulation of Σ whose set of vertices contains S and such that for all triangles $T\in\mathcal{T}$, there exists $\omega\in\mathbb{E}^2$ and $\tau_0\in\mathbb{R}$ such that

$$\text{for all } x\in T,\quad f(x)=\tau_0-\mathrm{d}(\mathcal{D}(x),\omega)^2,$$

where $\mathcal{D}\colon T\to\mathbb{E}^2$ is a developing map of T.

Definition 2.7 (distance-like function) Let (Σ,S) be a singular locally Euclidean surface. A function $f\colon\Sigma\to\mathbb{R}$ is distance-like if it admits an f-triangulation.

Remark Let (Σ,S) be a singular locally Euclidean surface, and let M be a radiant spacetime. For any polyhedral embedding $\iota\colon\Sigma\to M$, the map $\boldsymbol{g}\circ\iota\colon\Sigma\to\mathbb{R}_+$ is distance-like.

Proposition 2.8 *Let (Σ,S) be singular locally Euclidean surface. Let $\mathcal{T}$ be an adapted triangulation of (Σ,S).*

For all $\tau\colon S\to\mathbb{R}$, there exists a unique distance-like extension $\tilde{\tau}$ such that $\mathcal{T}$ is a $\tilde{\tau}$-triangulation.

Proof Apply Lemma 2.3 to each triangle of $\mathcal{T}$. □

Definition 2.9 Let (Σ,S) be a singular locally Euclidean surface. Let $\mathcal{T}$ be an adapted triangulation of (Σ,S) and let $\tau\colon S\to\mathbb{R}_+$. We denote by $\tilde{\tau}_{\tau,\mathcal{T}}$ the extension of τ given by Proposition 2.8.

Definition 2.10 (equivalent triangulations) Let (Σ,S) be singular locally Euclidean surface. Let $\tau\colon S\to\mathbb{R}_+$. Two adapted triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$ of (Σ,S) are τ-equivalent if

$$\tilde{\tau}_{\tau,\mathcal{T}_1}=\tilde{\tau}_{\tau,\mathcal{T}_2}.$$

Definition 2.11 (τ-suspension) Let (Σ, S) be a singular locally Euclidean surface and $f : \Sigma \to \mathbb{R}_+$ be distance-like.

Choose an f-triangulation $\mathcal{T}$ not necessarily adapted to (Σ, S). For each $T \in \mathcal{T}$, denote by $\iota_T : T \to J^+(O)$ the affine embedding of T given by Proposition 2.5 and define $C_T := \{t \cdot \iota_T(x) \mid t \in \mathbb{R}_+^*, x \in T\}$. For each edge e of $\mathcal{T}$ bounding $T_1, T_2 \in \mathcal{T}$, let γ_e be the isometry given by Lemma 2.4 sending the face of C_{T_2} associated to e to the face of C_{T_1} associated to e.

Define $M(f)$ as the radiant spacetime obtained by gluing the family $(C_T)_{T \in \mathcal{T}}$ via the isometries $(\gamma_e)_{e \in \mathrm{edges}(\mathcal{T})}$

Proposition 2.12 *Let (Σ, S) be a singular locally Euclidean surface and $f : \Sigma \to \mathbb{R}_+$ be distance-like. The spacetime $M(f)$ does not depend on the choice of the f-triangulation $\mathcal{T}$.*

Proof Consider two geodesic f-triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$. There exists a geodesic f-triangulation of (Σ, S) such that any 2-facet of $\mathcal{T}_1$ or $\mathcal{T}_2$ is a union of adjacent 2-facets of $\mathcal{T}$. It thus suffices to show that on a given triangle $T \subset \Sigma$ on which $\tilde{\tau}$ is $\mathscr{C}^1$, any decomposition of T into smaller triangles $(T_i)_{i \in [\![1,n]\!]}$ induces a gluing isomorphic to C_T. We may assume T is obtained by inductively gluing T_{k+1} to $\bigcup_{i=1}^k T_i$ for $k \in [\![1, n-1]\!]$. We give ourselves an embedding $\iota_0 : T \to J^+(O)$ given by Proposition 2.5. Start from T_1 with an embedding $\iota : T_1 \to J^+(O)$, using Lemma 2.4. Without loss of generality, we may assume that $\iota|_{0T_1} = \iota$, then glue the C_{T_k} for $k \in [\![2, n]\!]$ naturally extending $\iota : \bigcup_{i=1}^k T_i \to J^+(O)$. By Lemma 2.4, at each step, there is only one way to glue a cone $C_{T_{k+1}}$ to $\bigcup_{i=1}^k C_{T_i}$ so that $\tilde{\tau} = \boldsymbol{g} \circ \iota$. Hence at each step there is at most one extension of ι to $\bigcup_{i=1}^k T_i$; the embedding ι thus coincides with the restriction of ι_0 at each step, and thus on the whole T. Finally, C_T is isomorphic to the gluing of the $(C_{T_i})_{i \in [\![1,n]\!]}$. □

Definition 2.13 (equivalent polyhedral embedding) Let (Σ, S) be a singular locally Euclidean surface and let (M_1, ι_1) and (M_2, ι_2) be two radiant spacetimes together with a polyhedral embedding of (Σ, S).

We say that (M_1, ι_1) is equivalent to (M_2, ι_2) if there exists an isomorphism $\varphi : M_1 \to M_2$ such that $\iota_2 = \varphi \circ \iota_1$.

Theorem 1 *Denoting by $\sim$ the equivalence relation among polyhedral embeddings, the function*

$$\{(M, \iota) \mid M \text{ radiant, } \iota \text{ polyhedral embedding}\}/\!\sim \;\to \{\tilde{\tau} \mid \tilde{\tau} : \Sigma \to \mathbb{R}_+ \text{ distance-like}\}, \quad (\iota, M) \mapsto \boldsymbol{g} \circ \iota$$

is bijective with inverse $\tilde{\tau} \mapsto M(\tilde{\tau})$.

Remark The proof depends on a description of radiant spacetimes as suspensions of singular hyperbolic surfaces; we give it in the appendix.

Proof Denote by Φ the function above. For any $\tilde{\tau}$ distance-like on (Σ, S), by Proposition 2.5 the construction of $M(\tilde{\tau})$ ensures $\Phi(M(\tilde{\tau})) = \tilde{\tau}$. Hence Φ is surjective. Let (M_1, ι_1) be the polyhedral embedding of (Σ, S), let $\tilde{\tau} := \Phi(M_1, \iota_1)$, and let $M_2 = M(\tilde{\tau})$ with its polyhedral embedding $\iota_2 : \Sigma \to M_2$.

By Theorem 6, for $i \in \{1,2\}$, M_i is isomorphic to $\operatorname{susp}(\Sigma_i)$ with Σ_i the space of rays of the natural causal foliation of M_i endowed with its $\mathbb{H}^2_{\geq 0}$-structure. Define the natural projections $\pi_i \colon M_i \to \Sigma_i$. Denote by $\mathcal{R} \colon \mathcal{F} \to \mathbb{H}^2$ the map that associates to any $x \in \mathcal{F}$ the intersection point of the ray through x with $\mathbb{H}^2 \subset \mathcal{F}$.

For $i \in \{1,2\}$, the map $\pi_i \circ \iota_i \colon \Sigma \to \Sigma_i$ is a homeomorphism. The map $h := \pi_2 \circ \iota_2 \circ (\pi_1 \circ \iota_1)^{-1}$ is then a homeomorphism. We shall prove g is an a.e. $\mathbb{H}^2$-morphism from Σ_1 to Σ_2 and hence that $\operatorname{susp}(h) \colon M_1 \to M_2$ is an isomorphism.

Choose a geodesic triangulation $\mathcal{T}$ of Σ adapted to $\tilde{\tau}$. Its image by $\pi_i \circ \iota_i$ is a geodesic triangulation of Σ_i. Note that h sends a cell of Σ_1 to a cell of Σ_2. Thus in order to prove that h is an $\mathbb{H}^2$-morphism, it suffices to prove that its restrictions to each cell of Σ_1 are isometries.

Let $T \in \mathcal{T}$, $x \in T \setminus S$, and, for $i \in \{1,2\}$, choose a chart $(\mathcal{U}_i, \mathcal{V}_i, \varphi_i)$ of M_i around $\iota_i(x)$ such that $\mathcal{V}_i$ is a cone of $\mathcal{F}$. Let $T_\Sigma \subset T \setminus S$ be a triangle of Σ containing x. For $i \in \{1,2\}$, write $T_{M_i} := \iota_i(T_\Sigma)$, $T_{\Sigma_i} := \pi_i \circ \iota_i(T_\Sigma)$, $T^{(i)}_{\mathcal{F}} := \varphi_i(T_{M_i})$, and $T^{(i)}_{\mathbb{H}^2} := \mathcal{R} \circ \varphi_i(T_{M_i})$. By construction of the $\mathbb{H}^2$-structure on Σ_i, φ_i induces a chart $\bar{\varphi}_i \colon T_{\Sigma_i} \to T^{(i)}_{\mathbb{H}^2}$. By Lemma 2.4 there exists a unique $\phi \in \mathrm{SO}_0(1,2)$ such that $\varphi_2 \circ \iota_2 = \phi \circ \varphi_1 \circ \iota_1$. Since $\mathcal{R}$ commutes with the action of $\mathrm{SO}_0(1,2)$, we then have $\mathcal{R} \circ \varphi_2 \circ \iota_2 = \phi \circ \mathcal{R} \circ \varphi_1 \circ \iota_1$. The following commutative diagram sums up the situation:

$$\begin{tikzcd}
\Sigma \arrow[rrrrr, equal] \arrow[ddd, "\wr"', "\pi_1\circ\iota_1"] & & & & & \Sigma \arrow[ddd, "\pi_2\circ\iota_2"', "\wr"] \\
& T_\Sigma \arrow[ul, hook] \arrow[rrr, equal] \arrow[d, "\wr"', "\iota_1"] & & & T_\Sigma \arrow[ur, hook'] \arrow[d, "\wr"', "\iota_2"] & \\
& T_{M_1} \arrow[r, "\varphi_1", "\sim"'] \arrow[d, "\wr"', "\pi_1"] & T^{(1)}_{\mathcal{F}} \arrow[r, dashed, "\exists!\phi\in\mathrm{SO}_0(1,2)"] \arrow[d, "\wr"', "\mathcal{R}"] & T^{(2)}_{\mathcal{F}} \arrow[d, "\wr"', "\mathcal{R}"] & T_{M_2} \arrow[l, "\varphi_2"', "\sim"] \arrow[d, "\wr"', "\pi_2"] & \\
\Sigma_1 & T_{\Sigma_1} \arrow[l, "\supset"'] \arrow[r, "\bar{\varphi}_1", "\sim"'] & T^{(1)}_{\mathbb{H}^2} \arrow[r, dashed, "\phi"] & T^{(2)}_{\mathbb{H}^2} & T_{\Sigma_2} \arrow[l, "\bar{\varphi}_2"', "\sim"] \arrow[r, "\subset"] & \Sigma_2
\end{tikzcd}$$

Therefore the (co)restriction of h from T_{Σ_1} to T_{Σ_2} is an isometry. It follows that h is an isometry from a triangle of $\pi_1 \circ \iota_1(\mathcal{T})$ to a triangle of $\pi_2 \circ \iota_2(\mathcal{T})$. □

2.2 Convex embeddings

We start by clarifying the notion of a convex embedding in Definition 2.14, and translate the notion in terms of a *Q-convex* distance-like function. Proposition 2.23 is the main result of this subsection. It provides a parametrization of convex polyhedral embeddings by a domain of $\mathbb{R}^S_+$. Throughout the section, (Σ, S) is a marked locally Euclidean surface with conical singularities included in the set of marked points S.

Definition 2.14 (convex polyhedral embedding) Let M be a radiant spacetime with $\iota \colon \Sigma \to M$ a polyhedral embedding.

The embedding ι is convex if $J^+(\iota(\Sigma))$ is convex in the sense that for any spacelike geodesic $c\colon [a,b] \to M$, if $\{c(a), c(b)\} \subset J^+(\iota(\Sigma))$ then $c([a,b]) \subset J^+(\iota(\Sigma))$.

Definition 2.15 (Q-convexity on $\mathbb{R}$) Let $I \subset \mathbb{R}$ be an interval. A function $f\colon I \to \mathbb{R}$ is Q-convex (resp. Q-concave) if f is continuous, piecewise $\mathscr{C}^1$ and if for all $t_0 \in I$,

$$\lim_{t_0^-} f' \leq \lim_{t_0^+} f' \quad \bigl(\text{resp. } \lim_{t_0^-} f' \geq \lim_{t_0^+} f'\bigr).$$

Definition 2.16 (Q-convexity on an $\mathbb{E}^2_{>0}$-surface) A function $\tilde{\tau}\colon \Sigma \to \mathbb{R}$ is Q-convex (resp. Q-concave) if for all geodesics $c\colon I \to \Sigma \setminus S$, the restriction of $\tilde{\tau}$ to c is Q-convex (resp. Q-concave).

Lemma 2.17 *Let $f, g\colon [a,b] \to \mathbb{R}$ be two continuous functions piecewise of the form $x \mapsto -x^2 + \alpha x + \beta$ with f of class $\mathscr{C}^1$.*

- *If g is Q-convex with $f(a) \geq g(a)$ and $f(b) \geq g(b)$ then $g \leq f$.*
- *If g is Q-concave with $f(a) \leq g(a)$ and $f(b) \leq g(b)$ then $f \leq g$.*

Furthermore, if the Q-convexity (resp. Q-concavity) is strict, the inequalities are strict on $]a,b[$.

Proof First, $g - f$ is piecewise affine; since f is $\mathscr{C}^1$, the Q-convexity of $g - f$ (and hence its convexity) is the same as the Q-convexity of f. In the first (resp. second) case, since $g - f$ is nonpositive (resp. nonnegative) at a and b, it is thus nonpositive (resp. nonnegative) on $[a,b]$. The strict case is obtained the same way. □

Lemma 2.18 *Let M be a radiant singular flat spacetime and let $\Sigma \subset \mathcal{M}$ be a Cauchy surface. Denote by $\mathcal{R}\colon M \to \Sigma$ the function that associates to $x \in M$ the unique intersection point with Σ of the ray through x of the natural foliation of M; denote by $M_{>0}$ the complement in M of the singular lightlike lines.*

Then

$$J^+_M(\Sigma) = \overline{\{x \in M_{>0} \mid \boldsymbol{g}(x) \geq \boldsymbol{g}(\mathcal{R}(x))\}}.$$

Proof Since Σ is a Cauchy surface of M, $J^+_M(\Sigma) \cap J^-_M(\Sigma) = \Sigma$ and $J^+_M(\Sigma) \cup J^-_M(\Sigma) = M$. Since $\boldsymbol{g}$ is increasing toward the future along the timelike rays of the natural foliation of M,

$$\{x \in M_{>0} \mid \pm\boldsymbol{g}(x) \geq \pm\boldsymbol{g}(\mathcal{R}(x))\} \subset J^\pm_M(\Sigma).$$

Furthermore, since M is globally hyperbolic and Σ compact, $J^\pm_M(\Sigma)$ are closed. Hence

$$\overline{\{x \in M_{>0} \mid \pm\boldsymbol{g}(x) \geq \pm\boldsymbol{g}(\mathcal{R}(x))\}} \subset J^\pm_M(\Sigma).$$

Since $M_{>0}$ is dense in M, we have

$$\bigcup_{\epsilon \in \{+,-\}} \overline{\{x \in M_{>0} \mid \epsilon\boldsymbol{g}(x) \geq \epsilon\boldsymbol{g}(\mathcal{R}(x))\}} = M.$$

Furthermore

$$\Sigma \subset \bigcap_{\epsilon\in\{+,-\}} \overline{\{x \in M_{>0} \mid \epsilon \boldsymbol{g}(x) \geq \epsilon \boldsymbol{g}(\mathcal{R}(x))\}} \subset J^+_M(\Sigma) \cap J^-_M(\Sigma) = \Sigma,$$

and it follows that

$$\overline{\{x \in M_{>0} \mid \pm \boldsymbol{g}(x) \geq \pm \boldsymbol{g}(\mathcal{R}(x))\}} = J^\pm_M(\Sigma). \qquad \square$$

Proposition 2.19 *Let $\tilde{\tau}\colon \Sigma \to \mathbb{R}_+$ be distance-like, and $M := M(\tilde{\tau})$ with its associated polyhedral embedding $\iota\colon \Sigma \to M$.*

The embedding ι is convex if and only if $\tilde{\tau}$ is Q-convex.

Proof We identify Σ with $\iota(\Sigma)$ and denote by $\mathcal{R}\colon M \to \Sigma$ the map that associates to any $x \in M$ the intersection point of the ray (of the natural foliation) through x with Σ. Consider a spacelike geodesic $c\colon [a,b] \to M$ such that $c(a), c(b) \in J^+(\Sigma)$. A direct computation in a chart gives that both $\boldsymbol{g} \circ c$ and $\boldsymbol{g} \circ \mathcal{R} \circ c$ are continuous piecewise of the form $s \mapsto -s^2 + \alpha s + \beta$ and that $\boldsymbol{g} \circ c$ is $\mathscr{C}^1$. Furthermore, the derivatives of $\boldsymbol{g} \circ \mathcal{R} \circ c$ and $\tilde{\tau} \circ \mathcal{R} \circ c$ may be discontinuous at $s \in [a,b]$ only when the ray through $c(s)$ encounters an edge of Σ. At such an s, these two functions $\boldsymbol{g} \circ \mathcal{R} \circ c$ have the same Q-convexity.

- Assume that $\tilde{\tau}$ is Q-convex and consider a spacelike geodesic $c\colon [a,b] \to M$ such that $c(a), c(b) \in J^+(\Sigma)$. By Lemma 2.17, $\boldsymbol{g} \circ c - \boldsymbol{g} \circ \mathcal{R} \circ c$ is nonnegative and by Lemma 2.18 we thus have $c([a,b]) \subset J^+(\Sigma)$. Finally, $J^+(\Sigma)$ is convex, and hence ι is convex.

- Assume that $\tilde{\tau}$ is not Q-convex. There thus exists an edge e in Σ around which $\tilde{\tau}$ is strictly Q-concave. Consider two points x and y in Σ, each on a different side of said edge. We can choose x and y close enough so that they lie in a chart of M around $\iota(e)$. Then consider the geodesic $c\colon [a,b] \to M$ in this chart from x to y. It follows from Lemma 2.17 that $\boldsymbol{g} \circ c < \boldsymbol{g} \circ \mathcal{R} \circ c$ on $]a,b[$. Thus by Lemma 2.18 $c(]a,b[)$ is not in $J^+(\Sigma)$ and hence $J^+(\Sigma)$ is not convex; neither is ι. □

Proposition 2.20 *Let $\tau \in \mathbb{R}_+^S$. Up to equivalence there is at most one adapted triangulation $\mathcal{T}$ such that the distance-like extension $\tilde{\tau}_{\tau,\mathcal{T}}\colon \Sigma \to \mathbb{R}_+$ is Q-convex.*

Proof Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two adapted triangulations (Σ, S) such that both $f_1 := \tilde{\tau}_{\tau,\mathcal{T}_1}$ and $f_2 := \tilde{\tau}_{\tau,\mathcal{T}_2}$ are Q-convex. For all edges e of $\mathcal{T}_1$, the function $f|_{1e}$ is continuous quadratic while the function $f|_{2e}$ is piecewise quadratic and Q-convex; also, they are equal on the vertices of e. By Lemma 2.17 it thus follows that $f_2 \leq f_1$ on e. For any triangle T of $\mathcal{T}_1$, $f_1 \geq f_2$ on ∂T, and applying again Lemma 2.17 along any segment $[a,b]$ of T with $a, b \in \partial T$, we deduce that $f_1 \geq f_2$ on T. Therefore $f_1 \geq f_2$ on the whole Σ. We show in the same way that $f_1 \leq f_2$, and hence $f_1 = f_2$. The triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$ are then equivalent. □

Corollary 2.21 *Let $\tau \in \mathbb{R}_+^S$. There is at most one Q-convex distance-like extension $\tilde{\tau}$ of τ to the whole Σ.*

Definition 2.22 (admissible times) Define $\mathcal{P}$ to be the set of $\tau \in \mathbb{R}_+^S$ such that there exists an adapted triangulation $\mathcal{T}$ of (Σ, S) inducing a Q-convex distance-like extension $\tilde{\tau}_{\tau,\mathcal{T}}$. Elements of $\mathcal{P}$ are called admissible times.

For $\tau \in \mathcal{P}$, we denote by $\mathcal{T}_\tau$ the unique adapted triangulation of Σ (up to equivalence) such that $\tilde{\tau}_{\tau,\mathcal{T}_\tau}$ is Q-convex. We define as well $\tilde{\tau}_\tau := \tilde{\tau}_{\tau,\mathcal{T}_\tau}$ and $M(\tau) := M(\tilde{\tau}_\tau)$.

As a corollary of Proposition 2.20 and Theorem 1, we obtain the following:

Proposition 2.23 *With $\sim$ the equivalence relation between polyhedral embeddings, the function*

$$\{(M,\iota) \mid M \text{ radiant}, \iota\colon \Sigma \to M \text{ polyhedral convex embedding}\}/\sim \;\to \mathcal{P}, \quad (\iota, M) \mapsto (\boldsymbol{g} \circ \iota)|_S$$

is bijective.

3 The domain of admissible times

For this whole section, we give ourselves a marked locally Euclidean surface with conical singularities (Σ, S). While Proposition 2.23 parametrizes polyhedral embeddings by the domain $\mathcal{P} \subset \mathbb{R}^S$, for now, little is known about it, and before studying the image of $\tau \mapsto M(\tau)$ we shall provide a thorough description. More precisely, we prove the following:

Theorem 2 *Let $\mathbf{1}_S$ the indicator function of S, H the linear hyperplane of $\mathbb{R}^S$ orthogonal to $\mathbf{1}_S$, and π the orthogonal projection onto H. Define $\overline{\mathcal{P}} = \pi(\mathcal{P}) \subset H$. Then we have the following properties:*

(a) *$\overline{\mathcal{P}}$ is a convex compact polyhedron.*

(b) $\mathcal{P} = (\overline{\mathcal{P}} + \mathbb{R} \cdot \mathbf{1}_S) \cap \mathbb{R}_+^S$.

(c) *The interior of $\overline{\mathcal{P}}$ contains $0 \in \mathbb{R}^S$.*

(d) *With $\mathscr{T} := \{\mathcal{T}_\tau \mid \tau \in \mathcal{P}\}$, each $\overline{\mathcal{P}}_{\mathcal{T}} := \{\pi(\tau) \mid \mathcal{T}_\tau = \mathcal{T}\} \subset \overline{\mathcal{P}}$ is a convex polyhedron of H for $\mathcal{T} \in E$. Furthermore, the family $(\overline{\mathcal{P}}_{\mathcal{T}})_{\mathcal{T} \in \mathscr{T}}$ is a finite cellulation of $\overline{\mathcal{P}}$.*

(e) *The support planes Π of $\mathcal{P}$ whose intersection with $\mathcal{P}$ has nonempty interior relative to Π are either of the form "$\tau_\sigma = 0$" for some $\sigma \in S$ or "$Q^*(\tau) = 0$" for some unflippable immersed hinge Q around an edge of a triangulation $\mathcal{T}_\tau$ for some $\tau \in \mathcal{P}$ (see Definitions 3.1, 3.5, 3.8 and 3.15).*

The starting point is to study "local" criteria for Q-convexity. By local, we mean at each edge of a given triangulation; the following definitions make this notion precise:

Definition 3.1 (hinge) A hinge is a quadrilateral $[ABCD] \subset \mathbb{E}^2$ together with a diagonal $[AC]$ such that $[AC] \subset [ABCD]$.

Beware that the quadrilateral of a hinge need not be convex. If convex with vertices in general positions, a quadrilateral may define two hinges: one for each interior diagonal. Otherwise only one hinge may be defined.

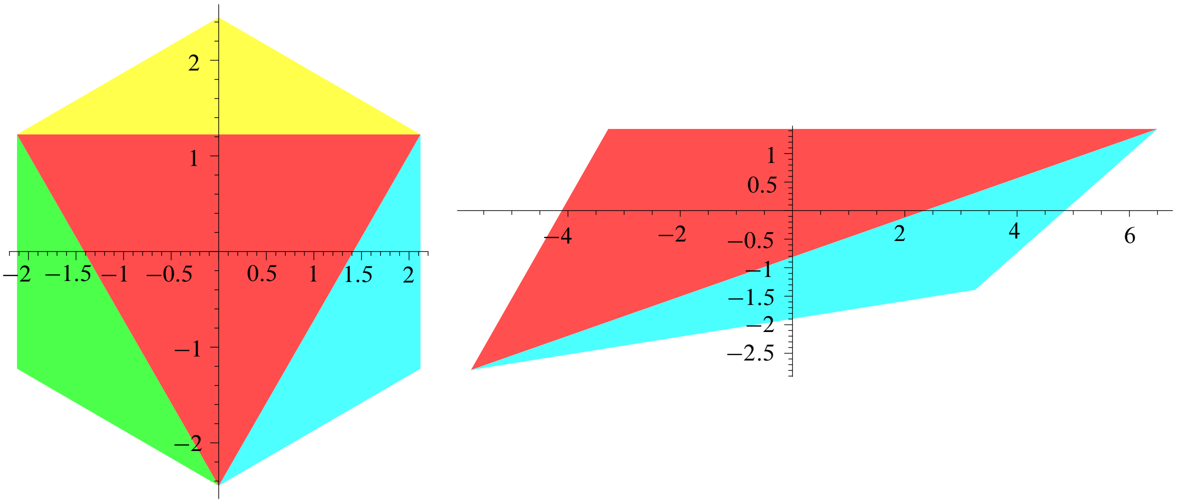


Figure 2: Projection of the domain of admissible τ. On the left, the domain $\overline{\mathcal{P}}$ of the surface is obtained by gluing two copies of an equilateral triangle's edges to edges. The central cell (red) corresponds to the Delaunay triangulation of the surface. In contrast, each of the other cells corresponds to the triangulations obtained after flipping an edge of the Delaunay triangulation. On the right, the domain $\overline{\mathcal{P}}$ of the surface is obtained by two copies of the triangle of vertices $(0,0)$, $(1,1)$ and $(0,3)$. The upper triangle corresponds to the Delaunay triangulation, while the lower one corresponds to the triangulation obtained after the only flip possible from the Delaunay triangulation. The domains are represented in an orthonormal basis of the plane H. The pictures were generated using SageMath [29].

Definition 3.2 (flippable hinge and hinge flipping) Let $Q=([ABCD],[AC])$ be a hinge. If $[ABCD]$ is convex and the four points A, B, C and D are in general position, then Q is flippable, and its flipping is the hinge $Q'=([ABCD],[DB])$. If $[ABCD]$ is not convex or A, B, C and D are not in general position, then Q is unflippable.

Definition 3.3 (weighted hinge) A weighted hinge is the datum of a hinge, $Q=([ABCD],[AC])$, and a function $\tau\colon\{A,B,C,D\}\to\mathbb{R}$.

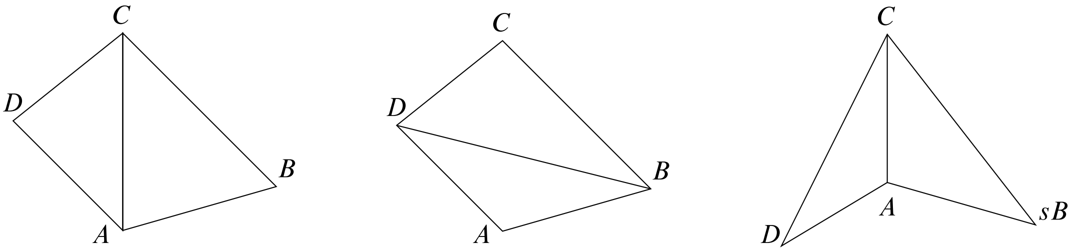


Figure 3: Different hinges. Left: a hinge $([ABCD],[AC])$. Center: its flipping $([ABCD],[DB])$. Right: a nonconvex hinge.

Definition 3.4 (τ-legal/τ-critical hinge) Let (Q,τ) be a weighted hinge. Denote by $\tilde{\tau}_{\tau,Q}\colon Q\to\mathbb{R}$ the distance-like function induced by the triangulation $\mathcal{T}=([ABC],[ADC])$. A hinge Q is τ-legal (resp. τ-critical, τ-illegal) if $\tilde{\tau}_{\tau,Q}$ is Q-convex (resp. $\mathscr{C}^1$, strictly Q-concave).

Each edge e of a given triangulation $\mathcal{T}$ provides a hinge; indeed e bounds two triangles $T_1,T_2\in\mathcal{T}$, and the gluing of these two triangles along e is a hinge. Beware that two such triangles might actually be the same in $\mathcal{T}$ (a triangle glued to itself), but we take two copies to construct the hinge. More generally, we will need to consider immersed hinges.

Definition 3.5 An immersed hinge is a couple (Q,η) with Q a hinge in $\mathbb{E}^2$ and $\eta\colon Q\to\Sigma$ an isometric immersion. An immersed hinge (Q,η) is embedded if the restriction $\eta|_{\mathrm{Int}(Q)}$ to the interior of Q is an embedding.

The hinge associated with an edge is embedded if and only if the triangles bounded by e are different in $\mathcal{T}$.

After an analysis of criteria ensuring τ-legality of a given hinge, we notice the set of τ for which a given hinge is τ-legal is the set of solutions of an affine inequality, and hence a convex set. Then, we turn to the whole surface and try to construct triangulations for which all hinges are τ-legal for a given τ.

Definition 3.6 (τ-Delaunay triangulation) Let $\mathcal{T}$ be an adapted triangulation of Σ.

The triangulation $\mathcal{T}$ is τ-Delaunay if the following equivalent properties are satisfied:

(i) $\tilde{\tau}_{\tau,\mathcal{T}}$ is Q-convex.

(ii) Every hinge of $\mathcal{T}$ is τ-legal.

For a given triangulation $\mathcal{T}$, the set of $\tau\in\mathbb{R}_+^S$ such that $\mathcal{T}$ is τ-Delaunay is the set solutions of a system of affine inequalities, and hence a convex set; hence the first part of Theorem 2(d). However, $\mathcal{P}$ is a possibly infinite union of such domains; therefore Theorem 2(a) and the second part of (d) are not direct corollaries. We thus reverse the problem and construct a τ-Delaunay triangulation with τ given a priori.

The definition of τ-Delaunay triangulation is coherent with the usual definition of Delaunay triangulation. Indeed, an adapted triangulation of (Σ,S) is a subtriangulation of the Delaunay cellulation if and only if it is 0-Delaunay. The Delaunay cellulation can either be constructed as the dual of the Voronoi cellulation (see [24] for a thorough exposition) or via a flipping algorithm starting from a given adapted triangulation. The flipping algorithm is based upon the following remark (Lemma 3.9): for a given τ, if a hinge is τ-illegal, then its flipping (if it exists) is τ-legal. The algorithm then proceeds by flipping τ-illegal hinges one by one in the hope that after finitely many iterations there will not be any τ-illegal hinges left. Proposition 3.17 ensures the algorithm behaves mostly as expected: it stops after finitely many iterations on a triangulation without any flippable τ-illegal hinges. To complete the analysis of the flipping algorithm, we show the resulting triangulation is τ-Delaunay if and only if there exists such a triangulation.

We end the section applying the results obtained on the flipping algorithm to prove Theorem 2.

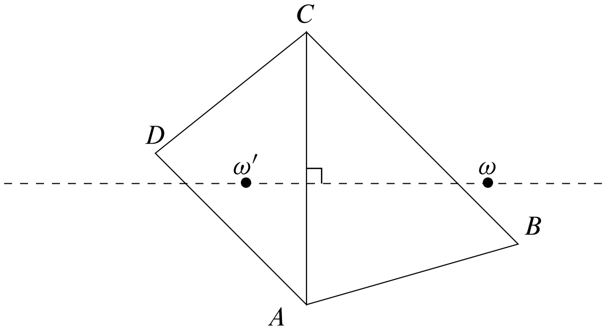


Figure 4

3.1 Q-convexity on hinges

Before going any further, we notice that the group $\mathrm{Isom}(\mathbb{E}^2)$ acts naturally on weighted hinges and preserves legality.

In this subsection, we give ourselves a hinge $Q = ([ABCD], [AC])$ and some weights τ. For simplicity's sake, we choose a Cartesian coordinate system (x, y) of $\mathbb{E}^2$, set $A = O$ as the origin of this coordinate system, and put C on the vertical axis above A. Denote by ω and τ_0 (resp. ω' and τ_0') the parameters given by Lemma 2.3 on $[ABC]$ (resp. $[ADC]$) for the weights τ; define

$$\tau_{ABC} \colon \mathbb{E}^2 \to \mathbb{R}, \quad x \mapsto \tau_0 - \mathrm{d}(x,\omega)^2, \quad \tau_{ADC} \colon \mathbb{E}^2 \to \mathbb{R}, \quad x \mapsto \tau_0' - \mathrm{d}(x,\omega')^2.$$

Figure 4 sums up the situation.

Remark Note that $\mathrm{d}(\omega, C)^2 - \mathrm{d}(\omega, A)^2 = \mathrm{d}(\omega', C)^2 - \mathrm{d}(\omega', A)^2$ and hence $\overrightarrow{\omega\omega'} \perp \overrightarrow{AC}$. More generally, from the proof of Lemma 2.3, one sees that the orthogonal projection of ω on the line (AC) only depends on A, C, τ_A, and τ_C.

Proposition 3.7 (Q-convexity criteria) *Under this subsection's hypotheses, the following are equivalent:*

(i) $\tilde{\tau}_{\tau,Q}$ *is Q-convex.*

(ii) $\tilde{\tau}_{\tau,Q}$ *is Q-convex along some segment crossing* $[AC]$.

(iii) $\tau_{ABC} \le \tau_{ACD}$ *on* $[ACD]$ *and* $\tau_{ABC} \ge \tau_{ACD}$ *on* $[ABC]$.

(iv) $\tau_{ABC}(D) \le \tau_{ACD}(D)$ *or* $\tau_{ABC}(B) \ge \tau_{ACD}(B)$.

(v) $x_\omega \ge x_{\omega'}$.

(vi) $$\left(\frac{y_B}{|x_B|} + \frac{y_D}{|x_D|}\right)\tau_C + \left(\frac{AC - y_B}{|x_B|} + \frac{AC - y_D}{|x_D|}\right)\tau_A \le \frac{AC}{|x_D|}\tau_D + \frac{AC}{|x_B|}\tau_B + K$$

with

$$K = \frac{AC}{|x_B|}(AB^2 - AC y_B) + \frac{AC}{|x_D|}(AD^2 - AC y_D).$$

(vii) *Denoting by* $\vec{u}\wedge\vec{v}$ *the determinant* $|\vec{u}\vec{v}|$,

$$(\overrightarrow{AB}\wedge\overrightarrow{AD})\tau_C+(\overrightarrow{CD}\wedge\overrightarrow{CB})\tau_A-(\overrightarrow{CA}\wedge\overrightarrow{CB})\tau_D-(\overrightarrow{AC}\wedge\overrightarrow{AD})\tau_B-K\le 0$$

with

$$K=\overrightarrow{AC}\wedge\overrightarrow{AD}(\overrightarrow{AB}\cdot\overrightarrow{CB})+\overrightarrow{CA}\wedge\overrightarrow{CB}(\overrightarrow{AD}\cdot\overrightarrow{CD}).$$

Proof • **(i) $\Longrightarrow$ (ii)** This follows by definition.

• **(ii) $\Longrightarrow$ (i)** Since the line $(\omega\omega')$ is perpendicular to (AC) it follows that $\partial\tau_{ABC}/\partial y=\partial\tau_{ACD}/\partial y$. Then $\overrightarrow{\operatorname{grad}}\,\tau_{[ABC]}-\overrightarrow{\operatorname{grad}}\,\tau_{[ACD]}$ is horizontal and the sign of $\langle\overrightarrow{\operatorname{grad}}\,\tau_{[ABC]}-\overrightarrow{\operatorname{grad}}\,\tau_{[ACD]}\mid\vec{u}\rangle$ does not depend on $\vec{u}$ as long as $\vec{u}$ is directed toward increasing x.

• **(i) $\Longrightarrow$ (v) and (v) $\Longrightarrow$ (ii)** We have that (v) is equivalent to $\partial\tau_{ABC}/\partial x\ge\partial\tau_{ACD}/\partial x$, which is equivalent to Q-convexity along the direction perpendicular to $[AC]$.

• **(i) $\Longrightarrow$ (iii)** Let $P\in[ABC]$ and choose some $P'\in[ACD]$ such that $[P'P]$ crosses $[AC]$. The function $\tau_{[ACD]}$ is $\mathscr{C}^1$ while $\tilde{\tau}_{\tau,Q}$ is Q-convex along $[P'P]$. The same argument as in the proof of Lemma 2.17 gives the first inequality. The second is proven the same way.

• **(iii) $\Longrightarrow$ (iv)** This is trivial.

• **(iv) $\Longrightarrow$ (ii)** Consider any segment $[PB]$ with $P\in[ACD]$. Along such a segment, $\tilde{\tau}_{\tau,Q}$ is either Q-convex or strictly Q-concave. The inequality $\tau_{ABC}(B)\ge\tau_{ACD}(B)$ implies it is the former. The same argument shows $\tau_{ABC}(D)\le\tau_{ACD}(D)\Longrightarrow$ (ii).

• **(v) $\Longleftrightarrow$ (vi)** Solve explicitly the system in the proof of Lemma 2.3 for both sides in (v).

• **(vii) $\Longleftrightarrow$ (vi)** These are geometric rewritings of each other, which can be checked by rewriting terms in coordinates. □

The previous proposition shows that Q-convexity is an affine constraint on τ for a given hinge. Since we will have to consider multiple hinges for multiple triangulations, we introduce the following:

Definition 3.8 (affine form of a hinge) Letting $Q=([ABCD],[AC])$ be a hinge, define the affine form associated to Q by

$$Q^*\colon\mathbb{R}_+^{\{A,B,C,D\}}\to\mathbb{R},\quad \tau\mapsto\lambda_C\tau_C+\lambda_A\tau_A-\lambda_D\tau_D-\lambda_B\tau_B-K,$$

where

$$\lambda_C=\overrightarrow{AB}\wedge\overrightarrow{AD},\quad\lambda_A=\overrightarrow{CD}\wedge\overrightarrow{CB},\quad\lambda_D=\overrightarrow{CA}\wedge\overrightarrow{CB},\quad\lambda_B=\overrightarrow{AC}\wedge\overrightarrow{AD},$$
$$K=\overrightarrow{AC}\wedge\overrightarrow{AD}(\overrightarrow{AB}\cdot\overrightarrow{CB})+\overrightarrow{CA}\wedge\overrightarrow{CB}(\overrightarrow{AD}\cdot\overrightarrow{CD}).$$

Remark The affine form Q^* is defined in such a way that $\tilde{\tau}_{\tau,Q}$ is Q-convex if and only if $Q^*(\tau)\le 0$.

Remark If (Q,η) is an immersed hinge of (Σ,S) with η sending vertices into S and with $Q=([ABCD],[AC])$, we can then define a corresponding affine form $\mathbb{R}_+^S\to\mathbb{R}$

$$\mathbb{R}_+^S\to\mathbb{R},\quad \tau\mapsto Q^*(\tau\circ\eta|_{\{A,B,C,D\}}).$$

If there is no ambiguity, we shall also denote it by Q^*.

Remark A hinge Q is τ-critical if and only if $Q^*(\tau)=0$.

Lemma 3.9 *Let $Q=([ABCD],[AC])$ be a flippable hinge and let Q' be its flipped hinge. As functions $\mathbb{R}^{\{A,B,C,D\}}\to\mathbb{R}$ we have*

$$Q'^*=-Q^*.$$

Proof This can, of course, be checked directly in coordinates, but we provide a more geometric proof. Following the notation of Definition 3.8 we write

$$Q^*\colon\mathbb{R}_+^{\{A,B,C,D\}}\to\mathbb{R},\quad \tau\mapsto\lambda_C\tau_C+\lambda_A\tau_A-\lambda_D\tau_D-\lambda_B\tau_B-K,$$
$$Q'^*\colon\mathbb{R}_+^{\{A,B,C,D\}}\to\mathbb{R},\quad \tau\mapsto\lambda'_C\tau_C+\lambda'_A\tau_A-\lambda'_D\tau_D-\lambda'_B\tau_B-K',$$

where

$$\lambda_C=\overrightarrow{AB}\wedge\overrightarrow{AD},\qquad \lambda_A=\overrightarrow{CD}\wedge\overrightarrow{CB},\qquad \lambda_D=\overrightarrow{CA}\wedge\overrightarrow{CB},\qquad \lambda_B=\overrightarrow{AC}\wedge\overrightarrow{AD},$$
$$\lambda'_D=-\overrightarrow{BC}\wedge\overrightarrow{BA},\quad \lambda'_B=-\overrightarrow{DA}\wedge\overrightarrow{DC},\quad \lambda'_A=-\overrightarrow{DB}\wedge\overrightarrow{DC},\quad \lambda'_C=-\overrightarrow{BD}\wedge\overrightarrow{BA}.$$

We check that

$$\lambda'_A=-\overrightarrow{DB}\wedge\overrightarrow{DC}=-(\overrightarrow{DC}+\overrightarrow{CB})\wedge\overrightarrow{DC}=-\overrightarrow{CB}\wedge\overrightarrow{DC}=-\overrightarrow{CD}\wedge\overrightarrow{CB}=-\lambda_A,$$

and we check the same way that $\lambda'_B=-\lambda_B$, $\lambda'_C=-\lambda_C$, and $\lambda'_D=-\lambda_D$.

A quick way to prove that $K'=-K$ is to notice that

$$K=(AB\cdot CB\cdot CD\cdot DA)\sin(\widehat{BAD}+\widehat{DCB}),\quad K'=(AB\cdot CB\cdot CD\cdot DA)\sin(\widehat{CBA}+\widehat{ADC}),$$

and that $\widehat{BAD}+\widehat{DCB}+\widehat{CBA}+\widehat{ADC}=0 \bmod 2\pi$. □

Corollary 3.10 *Let (Q,τ) be a weighted flippable hinge. Then Q is τ-critical if and only if its flipping Q' is τ-critical.*

Corollary 3.11 *Let (Q,τ) be a weighted flippable hinge and Q' the flipping of Q. If Q is not τ-critical, then the following are equivalent*:

(i) *Q is τ-legal.*

(ii) *Q' is τ-illegal.*

Lemma 3.12 *For any hinge Q, the indicator function $\mathbf{1}_S$ is in the kernel of the linear part of Q^*, eg*

$$\text{for all } \tau\in\mathbb{R}^S \text{ and } \lambda\in\mathbb{R},\quad Q^*(\tau+\lambda\mathbf{1}_S)=Q^*(\tau).$$

Proof Using the notation of Definition 3.8, we have

$$\lambda_A+\lambda_C-\lambda_B-\lambda_D=\overrightarrow{CD}\wedge\overrightarrow{CB}+\overrightarrow{AB}\wedge\overrightarrow{AD}-\overrightarrow{AC}\wedge\overrightarrow{AD}-\overrightarrow{CA}\wedge\overrightarrow{CB}=\overrightarrow{AD}\wedge\overrightarrow{CB}+\overrightarrow{CB}\wedge\overrightarrow{AD}=0. \qquad \square$$

Corollary 3.13 *For all $\tau\in\mathcal{P}$ and all $\lambda\in\mathbb{R}$,*

$$\tau+\lambda\mathbf{1}_S\in\mathcal{P} \iff \tau+\lambda\mathbf{1}_S\geq 0.$$

Corollary 3.14 *With the notation of Theorem 2,*

$$\mathcal{P}=(\overline{\mathcal{P}}+\mathbb{R}\cdot\mathbf{1}_S)\cap\mathbb{R}_+^S.$$

3.2 The flipping algorithm

Let $\mathcal{T}$ be an adapted triangulation of (Σ,S). Consider (Q,η) an immersed hinge given by an edge of $\mathcal{T}$. We would like to flip (Q,η), ie construct a new triangulation of (Σ,S) with $\eta(Q)$ replaced by $\eta(Q')$ with Q' the flip of Q. There are three cases:

- η is not an embedding. Then the diagonal one wants to replace is also a side of the hinge. Hence one cannot simply replace it without modifying the triangulation $\mathcal{T}$ elsewhere.
- η is embedded but Q is not flippable.
- η is embedded and Q is flippable. Then the flipped hinge Q' is well defined, $\eta\colon Q'\to\Sigma$ is well defined, $\eta(Q')=\eta(Q)$ so that we only modify $\mathcal{T}$ locally, and the new triangulation $\mathcal{T}'$ is composed of nondegenerated triangles.

This remark motivates the following definitions:

Definition 3.15 (flippable immersed hinge) An immersed hinge (Q,η) is flippable if it is embedded and Q is flippable; it is unflippable otherwise.

Definition 3.16 (flipping algorithm) Let $\mathcal{T}_0$ be any adapted triangulation of (Σ,S) and let $\tau\colon S\to\mathbb{R}_+$. The flipping algorithm proceeds as follows:

(1) Set $i=0$.

(2) Let L_i be the set of τ-illegal flippable embedded hinges (Q,η) induced by the edges of the current triangulation $\mathcal{T}_i$.

(3) If L_i is nonempty,

(a) choose some immersed hinge (Q,η) in L_i,

(b) replace the hinge (Q,η) by its flipping (Q',η) in $\mathcal{T}_i$ to obtain a new triangulation $\mathcal{T}_{i+1}$,

(c) increment i and go to step (2).

(4) If L_i is empty, the algorithm stops and returns $\mathcal{T}_i$.

The goal of the section is to prove the following:

Proposition 3.17 *Let $\tau\colon S \to \mathbb{R}_+$. For any starting triangulation $\mathcal{T}_0$, the flipping algorithm for τ starting at $\mathcal{T}_0$ stops on some triangulation $\mathcal{T}_\tau$ after finitely many iterations and every flippable immersed hinge in $\mathcal{T}_\tau$ is τ-legal. Furthermore,*

- *$\tau \in \mathcal{P}$ if and only if $\mathcal{T}_\tau$ is τ-Delaunay,*
- $\max_\Sigma \tilde{\tau}_{\tau,\mathcal{T}_\tau} \le \max_S \tau + \max_\Sigma \tilde{\tau}_{0,\mathcal{T}_0}$.

Remark The notation $\mathcal{T}_\tau$ of this last proposition is consistent with the one introduced in Definition 2.22.

Two lemmas are key to the proof; the first is Lemma 3.18, which states that $\tilde{\tau}_{\tau,\mathcal{T}_i}$ is decreasing along the iterations of the algorithm; the second is Lemma 3.22, which implies that immersed unflippable hinges are always τ-legal for $\tau \in \mathcal{P}$, even those that are not associated to an edge. Lemma 3.22 will again be useful in the following section.

Lemma 3.18 *Let $\tau\colon S \to \mathbb{R}_+$ and let $\mathcal{T}_0$ be an adapted triangulation. Let $(\mathcal{T}_i)_{i\in I}$ be the sequence of triangulation given by the flipping algorithm with weights τ and starting at $\mathcal{T}_0$, where $I = [\![0,n]\!]$ or $\mathbb{N}$.*

Then the associated sequence of distance-like functions $(\tilde{\tau}_{\tau,\mathcal{T}_i})_{i\in I}$ is decreasing:

- *for all $i,j \in I$ with $i \le j$ we have $\tilde{\tau}_{\tau,\mathcal{T}_i} \ge \tilde{\tau}_{\tau,\mathcal{T}_j}$,*
- *for all $i,j \in I$ with $i < j$ there exists $x \in \Sigma$ such that*
$$\tilde{\tau}_{\tau,\mathcal{T}_i}(x) > \tilde{\tau}_{\tau,\mathcal{T}_j}(x).$$

Proof Let $i \in I$ be such that $i+1 \in I$. The triangulation $\mathcal{T}_{i+1}$ is obtained from $\mathcal{T}_i$ by flipping an embedded hinge, say (Q,η), of $\mathcal{T}_i$ with $Q = ([ABCD],[AC])$. Then:

- For all $x \in \Sigma \setminus \eta(\mathrm{Int}(Q))$, $\tilde{\tau}_{\tau,\mathcal{T}_i}(x) = \tilde{\tau}_{\tau,\mathcal{T}_{i+1}}(x)$. Indeed, for $x \notin \eta(Q)$, the triangle containing x is the same in $\mathcal{T}_i$ and $\mathcal{T}_{i+1}$.
- For all $x \in \eta(\mathrm{Int}(Q))$, $\tilde{\tau}_{\tau,\mathcal{T}_i}(x) > \tilde{\tau}_{\tau,\mathcal{T}_{i+1}}(x)$. Indeed, $\tilde{\tau}_{\tau,Q}$ and $\tilde{\tau}_{\tau,Q'}$ are equal on $[AB]$, $[BC]$, $[CD]$, and $[DA]$; by hypothesis $\tilde{\tau}_{\tau,Q}$ is strictly Q-concave and, from Corollary 3.11, $\tilde{\tau}_{\tau,Q'}$ is strictly Q-convex. Applying Lemma 2.17 on segments going from side to side of $[ABCD]$ we obtain
$$\text{for all } x \in \mathrm{Int}(Q), \quad \tilde{\tau}_{\tau,Q} > \tilde{\tau}_{\tau,Q'}.$$ □

Corollary 3.19 *No triangulation appears twice in the sequence $(\mathcal{T}_i)_{i\in I}$ given by the flipping algorithm.*

Lemma 3.20 *Let $\tilde{\tau}$ be a nonnegative distance-like function on (Σ,S). If $\tilde{\tau}$ is $\mathscr{C}^1$ on some geodesic of length ℓ then*
$$\max \tilde{\tau} \ge \tfrac{1}{4}\ell^2.$$

Proof Let $c\colon [a,b] \to \Sigma$ be an arc length parametrization of such a geodesic and let $f := \tilde{\tau} \circ c$. We have

$$f\colon [a,b] \to \mathbb{R}, \quad x \mapsto -x^2 + \alpha x + \beta,$$

for some $\alpha, \beta \in \mathbb{R}$. Furthermore $\tilde{\tau} \geq 0$, and so $f(a) \geq 0$ and $g(b) \geq 0$.

Define $u\colon [a,b] \to \mathbb{R}$ to be the unique affine function such that $u(a) = f(a)$ and $u(b) = b$. We thus have for all $x \in [a,b]$, $f(x) = u(x) - (x-a)(x-b)$. On the one hand, $f(a)$ and $f(b)$ are nonnegative, so u is nonnegative. On the other hand,

$$\max_{x \in [a,b]} (-(x-a)(x-b)) = \tfrac{1}{4}(b-a)^2 = \tfrac{1}{4}\ell^2. \qquad \square$$

Lemma 3.21 *For $C \in \mathbb{R}_+^*$, let E_C be the set of adapted triangulations $\mathcal{T}$ of (Σ, S) such that*

$$\text{there exists } \tau \in \mathbb{R}_+^S \text{ with } \max \tilde{\tau}_{\tau,\mathcal{T}} \leq C.$$

Then E_C is finite.

Proof Let $\mathcal{T}$ be an adapted triangulation such that there exists $\tau \in \mathbb{R}_+^S$ with $\max \tilde{\tau}_{\tau,\mathcal{T}} \leq C$. Choose such a τ. Let e be the longest edge of $\mathcal{T}$. From Lemma 3.20 with $L = \text{length}(e)$

$$\tfrac{1}{4}L^2 \leq \max_e \tilde{\tau}_{\tau,\mathcal{T}} \leq C,$$

and thus $L \leq 2\sqrt{C}$. Therefore the triangulation $\mathcal{T}$ only has edges of length less than $2\sqrt{C}$.

Consider a finite covering $\hat{\Sigma}$ of Σ branched above S such that all cone angles of $\hat{\Sigma}$ are bigger than 2π. Note that $\hat{\Sigma}$ is locally CAT(0), so its universal (unbranched) covering $\tilde{\hat{\Sigma}}$ is CAT(0) by [1, Theorem 3.3.1], and hence for any two points in $\tilde{\hat{\Sigma}}$ above S there exists at most one geodesic; see [1, Section 2.2]. Furthermore, any geodesic of length at most $2\sqrt{C}$ in Σ from a point A of S to a point B of S lifts to a geodesic in $\tilde{\hat{\Sigma}}$ of the same length starting from a fixed $\hat{A}$ to some unfixed lift $\hat{B}$ of B in the ball of radius $2\sqrt{L}$ around $\hat{A}$. There are finitely many such $\hat{B} \in \tilde{\hat{\Sigma}}$, thus finitely such geodesics in $\tilde{\hat{\Sigma}}$. There are thus only finitely many geodesics of Σ from S to S of length bounded by $2\sqrt{C}$; hence there are only finitely many triangulations with edges of length at most $2\sqrt{C}$. □

Lemma 3.22 *Let Q be an unflippable hinge with $Q = ([ABCD], [AC])$. If there exists some distance-like Q-convex function f on $[ABCD]$ extending $\tau\colon \{A,B,C,D\} \to \mathbb{R}$, then Q is τ-legal.*

Remark Beware that f-triangulations of $[ABCD]$ may be very different from the one induced by the hinge, ie $([ABC], [ACD])$.

Proof Without loss of generality, we may assume that C is in the convex hull of $[ABD]$. Define $g := \tilde{\tau}_{\tau,Q}$ and h the distance-like extension of $\tau|_{\{A,B,D\}}$ on $[ABD]$ given by Lemma 2.3. Both functions f and g are defined on $[ABCD] \subset [ABD]$ and h is defined on $[ABD]$. Furthermore, g is either Q-convex or Q-concave.

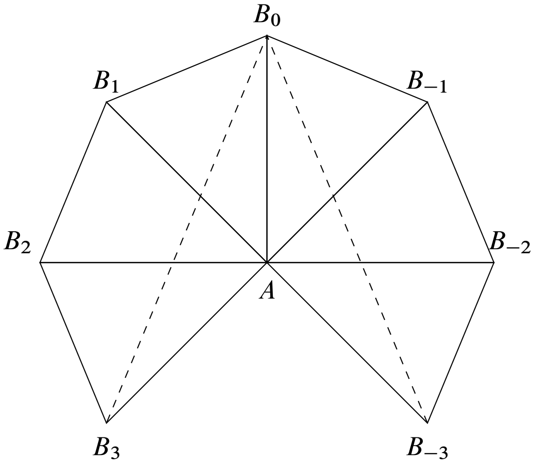


Figure 5

Applying Lemma 2.17 on sides of the hinge Q and then on any edge within Q and with extremities on the sides, we see that $f \le h$ and that either $h \ge g$ or $h \le g$, depending on whether g is Q-convex or Q-concave.

Since $g-h$ is affine on each triangle $[ACB]$ and $[ACD]$ and null at A, B, and D, we see that $g-h$ is nonpositive if and only if $g(C)-h(C)\le 0$. However, $g(C)=f(C)\le h(C)$, so $g\le h$, and hence $g=\tilde{\tau}_{\tau,Q}$ is Q-convex. □

Lemma 3.23 *Let $\tau\colon S\to\mathbb{R}_+$ and let (Q,η) be an immersed hinge with $Q=([AB_{-1}B_0B_1],[AB_0])$ such that $[AB_0B_1]$ is obtained from $[AB_{-1}B_0]$ via a rotation and $\eta([AB_0])=\eta([AB_1])=\eta([AB_{-1}])$.*

Then there exists an immersed hinge $(\widehat{Q},\hat{\eta})$ with $\widehat{Q}$ unflippable such that (Q,η) is τ-legal if and only if $(\widehat{Q},\hat{\eta})$ is τ-legal.

Proof Let $\theta=\widehat{B_0AB_1}$ and $n=\lceil\pi/\theta\rceil-1$. If $\theta\ge\frac{1}{2}\pi$ then take $\widehat{Q}=Q$ and $\hat{\eta}=\eta$.

Otherwise, construct the polygon $[AB_{-n}\cdots B_0\cdots B_n]$ such that for each $k\in[\![1-n,n]\!]$, the triangle $[AB_kB_{k+1}]$ is obtained from $[AB_0B_1]$ via the rotation of center A and angle $\alpha_k=k\theta$. Define $\widehat{Q}:=[AB_nB_0B_{-n}]$ and

$$\hat{\eta}\colon \widehat{Q}\to\Sigma,\quad x\in[AB_{2k}B_{2k+1}]\mapsto\eta(\rho_{-2\alpha_k}(x)),\quad x\in[AB_{2k-1}B_{2k}]\mapsto\eta(\rho_{-2\alpha_k}(x)),$$

where ρ_β denotes the rotation of center A and angle β; see Figure 5.

We have for all $k\in[\![-n,n]\!]$, $\hat{\eta}(B_k)=\eta(B_0)$. The weights $\tau\in\mathbb{R}^S$ thus induce weights $\hat{\tau}:=\tau\circ\hat{\eta}$ such that $\hat{\tau}(B_k)=\tau(B_0)=\hat{\tau}(B_1)=\hat{\tau}(B_{-1})$. For $I,J,K\in\{A,B_{-n},\dots,B_n\}$, denote by $\omega_{[IJK]}$ the center of $\tilde{\tau}_{\tau,[IJK]}$ and by $\omega_{[IJ]}$ the orthogonal projection of $\omega_{[IJK]}$ on the line (IJ). From the remark before Proposition 3.7, the orthogonal projection of $\omega_{[IJK]}$ on the line (IJ) only depends on $\hat{\tau}(I)$, $\hat{\tau}(J)$, and $[IJ]$; in other words $\omega_{[IJ]}$ does not depend on K.

Since $\hat{\tau}(B_n)=\hat{\tau}(B_0)=\hat{\tau}(B_{-1})$, we have that $\omega_{[B_0B_n]}$ (resp. $\omega_{[B_0B_{-n}]}$) is the middle of $[B_0B_n]$ (resp. of $[B_0B_{-n}]$). Since the lengths $(AB_k)_{k\in[\![-n,n]\!]}$ are equal, the perpendicular bisectors of $[B_0B_{-n}]$ and $[B_0B_n]$ intersect at A. Therefore $\omega_{AB_{-n}B_0}$ is on the right of $\omega_{AB_nB_0}$ on the perpendicular to (AB_0) at $\omega_{[AB_0]}$ if and only if $\omega_{[AB_0]}$ is on the ray $[AB_0)$. Hence, by Proposition 3.7(v), the hinge $\widehat{Q}$ is $\hat{\tau}$-legal if and only if $\omega_{[AB_0]}$ is on the ray $[AB_0)$.

The same argument shows Q is $\hat{\tau}$-legal if and only if $\omega_{[AB_0]}$ is on the ray $[AB_0)$. Finally, (Q,η) is τ-legal if and only if $(\widehat{Q},\hat{\eta})$ is τ-legal. □

Proof of Proposition 3.17 By Lemma 3.18, the sequence of distance-like functions given by the flipping algorithm is bounded above by the first of the sequence $\tilde{\tau}_{\tau,\mathcal{T}_0}$. Since $\tilde{\tau}_{\tau,\mathcal{T}_0}-\tilde{\tau}_{0,\mathcal{T}_0}$ is affine on each triangle of $\mathcal{T}_0$ it is bounded by its value on S, and thus by $\max_S \tau$. Hence, for all i, $\tilde{\tau}_{\tau,\mathcal{T}_i}\le \max_S \tau+\max\tilde{\tau}_{0,\mathcal{T}_0}$. By Lemma 3.21, the flipping algorithm runs through a finite set of triangulations. Finally, by Corollary 3.19, the algorithm reaches a given triangulation at most once and thus stops after finitely many steps, say $n\in\mathbb{N}^*$. The algorithm stops when the set of flippable τ-illegal hinges is empty, so $\mathcal{T}_n$ has no flippable τ-illegal hinges.

If the final triangulation $\mathcal{T}_n$ is τ-Delaunay then by definition $\tau\in\mathcal{P}$. Assume $\tau\in\mathcal{P}$ and consider (Q,η) some unflippable hinge of $\mathcal{T}_n$. Either η is an embedding, in which case the weighted hinge $(Q,\tau\circ\eta)$ satisfies the hypotheses of Lemma 3.22 and the immersed hinge (Q,η) is then τ-legal, or η is not an embedding, in which case (Q,η) satisfies the hypotheses of Lemma 3.23, so the immersed hinge $(\widehat{Q},\hat{\eta})$ provided by Lemma 3.23 satisfies the hypotheses of Lemma 3.22, thus being τ-legal, and so (Q,η) is τ-legal as well. Finally, $\mathcal{T}_n$ is τ-Delaunay. □

3.3 Description of the domain of admissible times

We may interpret Lemma 3.22 together with Lemma 3.23 in the following way: if $\tau\in\mathcal{P}$, then all unflippable immersed hinges of (Σ,S) with vertices in S are τ-legal. Furthermore, Proposition 3.17 shows the converse: the flipping algorithm stops on a triangulation $\mathcal{T}$, whose flippable hinges are all τ-legal if all unflippable hinges of (Σ,S) are τ-legal, in particular those of $\mathcal{T}$ are τ-legal, and hence $\mathcal{T}$ is τ-Delaunay. We thus proved the following:

Proposition 3.24 *Let* UFlip *be the set of the unflippable immersed hinges of* (Σ,S) *with vertices in* S*. Then*

$$\mathcal{P}=\bigcap_{(Q,\eta)\in\mathrm{UFlip}}(Q^*)^{-1}(\mathbb{R}_-).$$

In particular $\mathcal{P}$ *is a convex domain of* $\mathbb{R}_+^S$*.*

Remark Lemma 3.29 implies that UFlip is nonempty. We take the convention that the intersection is $\mathbb{R}_+^S$ if $\mathrm{UFlip}=\varnothing$.

Proposition 3.25 *For $\tau \in \mathcal{P}$, if $\tilde{\tau}$ is the unique Q-convex distance-like extension of τ to (Σ, S) then*

$$\tilde{\tau} = \min_{\mathcal{T}'} \tilde{\tau}_{\tau,\mathcal{T}'},$$

where $\mathcal{T}'$ runs through all adapted triangulations of (Σ, S).

Proof Take any adapted triangulation $\mathcal{T}$ of (Σ, S) and consider T a triangle of $\mathcal{T}$. On T, $\tilde{\tau}_{\tau,\mathcal{T}}$ is $\mathscr{C}^1$ while $\tilde{\tau}$ is Q-convex. By Lemma 2.17, $\tilde{\tau} \leq \tilde{\tau}_{\tau,\mathcal{T}}$ on T. The triangle T is arbitrary; thus $\tilde{\tau} \leq \tilde{\tau}_{\tau,\mathcal{T}}$ on Σ. □

Proposition 3.26 *The indicator function $\mathbf{1}_S$ of S is in the interior of $\mathcal{P}$.*

Proof To begin with, by [24, Theorem 4.4], each cell of the Delaunay cellulation $\mathcal{C}$ of (Σ, S) is isometric to a polygon inscribed into a circle of $\mathbb{E}^2$ whose center is a vertex of the Voronoi cellulation. For any given cell C of the Delaunay cellulation, with R_C the radius and $\omega \in \mathbb{E}^2$ the center of the circumscribed circle of the image of a development $\mathcal{D}\colon C \to \mathbb{E}^2$, the function

$$f : C \to \mathbb{R}_+, \quad x \mapsto R_C^2 - \mathrm{d}(\mathcal{D}(x), \omega)^2,$$

is distance-like $\mathscr{C}^1$ on C and $f(p) = 0$ for any vertex p of C; hence, for any adapted subtriangulation $\mathcal{T}$ of $\mathcal{C}$, for all $x \in C$, $\tilde{\tau}_{0,\mathcal{T}}(x) = f(\mathcal{D}(x))$.

Let e be an edge of the Delaunay cellulation, let C and C' be the two cells on each side of e, and denote by $\widetilde{C}$ and $\widetilde{C}'$ lifts in a covering branched above S such that $\widetilde{C} \neq \widetilde{C}'$ and such that $\widetilde{C} \cap \widetilde{C}' = \tilde{e}$ with $\tilde{e}$ a lift of e. Choose a development $\mathcal{D}$ of $\widetilde{C} \cup \widetilde{C}'$. By abuse of notation let ω and ω' denote the centers of the images of $\widetilde{C}$ and $\widetilde{C}'$, respectively.

Denote by $Q^*_{e,\mathcal{T}}$ the affine form associated with the hinge of axis e for any subtriangulation $\mathcal{T}$ of $\mathcal{C}$.

Claim $$Q^*_{e,\mathcal{T}}(0) \neq 0.$$

This is equivalent to $\omega \neq \omega'$. Assume for the sake of contradiction that $\omega = \omega'$. Then vertices of $\mathcal{D}(\widetilde{C}) \cup \mathcal{D}(\widetilde{C}')$ are cocyclic; hence C and C' are in the same Delaunay cell, ie C is glued to itself via e.

Without loss of generality, we may assume that ω is on the side (inclusively) of $\mathcal{D}(\widetilde{C}')$; hence e is strictly longer than every other edge of C. We deduce that C cannot be glued to itself via e, a contradiction.

Claim $$Q^*_{e,\mathcal{T}}(0) \leq 0.$$

We may assume that the hinge at e is developed as in Figure 4. We take the notation of the proposition. Notice that assuming condition (v) is not satisfied, either B is in the interior of the circumscribed circle of ACD or D is in the interior of the circumscribed circle of ABC. This violates a characterization of the Delaunay cellulation.

Define

$$\mathcal{U} := \mathbb{R}^S_+ \cap \bigcap_{\mathcal{T} \in \boldsymbol{D}} \bigcap_{e} {Q^*_{e,\mathcal{T}}}^{-1}(\mathbb{R}^*_-),$$

where $\boldsymbol{D}$ is the set of adapted subtriangulations of the Delaunay cellulation, and e runs through the edges of the Delaunay cellulation. The intersection is finite since there are only finitely many such subtriangulations and edges. $\mathcal{U}$ is thus an open subset of $\mathbb{R}_+^S$ which contains $\mathbb{R}_+\mathbf{1}_S$.

We now show $\mathcal{U} \subset \mathcal{P}$. Apply the flipping algorithm for some $\tau \in \mathcal{U}$ and start from some $\mathcal{T}_0 \in \boldsymbol{D}$ of the Delaunay cellulation. Let $\mathcal{T}_0, \dots, \mathcal{T}_n$ be the sequence of triangulations given by the flipping algorithm. By induction we have $\mathcal{T}_0 \in \boldsymbol{D}$, and assuming $\mathcal{T}_k \in \boldsymbol{D}$ for some $k < n$, the conditions $Q^*_{e,\mathcal{T}_k}(\tau) < 0$ ensure that the edges e are τ-legal and thus not flipped. Hence $\mathcal{T}_{k+1} \in \boldsymbol{D}$. From Proposition 3.17, the triangulation $\mathcal{T}_n$ is such that all flippable hinges are τ-legal. On the one hand, the edges of $\mathcal{C}$ are τ-legal since $\tau \in \mathcal{U}$. On the other hand, all hinges inside a cell of the Delaunay cellulation are flippable. Finally, all the edges of $\mathcal{T}_n$ are τ-legal and $\mathcal{U}$ is a subset of $\mathcal{P}$. □

In order to obtain a finite cellulation of $\mathcal{P}$ as well as characterize its boundary, we prove its transverse compactness. By transverse compactness of $\mathcal{P}$ we mean that the projection of $\mathcal{P}$ into the hyperplane $\left\{\tau \in \mathbb{R}^S \mid \sum_{s\in S} \tau(s) = 0\right\}$ is compact. Note that, for instance, if $\mathcal{P}$ were equal to the whole $\mathbb{R}_+^S$ then it wouldn't be transversely compact in this sense. The proof that $\mathcal{P}$ is transversely compact relies upon the construction of affine constraints of the form $\tau_A - \tau_C \leq \varepsilon(\tau_A + \tau_B + \tau_C + \tau_D) + K$ with $\varepsilon > 0$ arbitrarily small, and A and C arbitrary in S. Such constraints are provided by type-(x, L) hinges; see Definition 3.27, via Lemma 3.28. Lemma 3.29 focuses on the construction of such immersed hinges.

Definition 3.27 (type-(x, L) hinge) Let $x, L > 0$. A hinge $([ABCD], [AC])$ of $\mathbb{E}^2$ is of type (x, L) if it is nonconvex with $C \in [ABD]$ and

$$\mathrm{d}(B, \Delta) \leq x, \quad \mathrm{d}(D, \Delta) \leq x, \quad AB > L, \quad AD > L,$$

where Δ is the line (AC).

Lemma 3.28 *Let $l > 0$ and $x > 0$. For a hinge Q, write*

$$Q^* \colon \tau \mapsto \alpha(Q)\tau_A + \beta(Q)\tau_B + \gamma(Q)\tau_C + \delta(Q)\tau_D + K(Q)$$

for the affine form associated to Q.

Then, for all sequences $(Q_n)_{n\in\mathbb{N}}$ of hinges such that for all $n \in \mathbb{N}$, Q_n is of type (x, n) and axis length l, we have

$$\lim_{n\to+\infty} \frac{\alpha(Q_n)}{\gamma(Q_n)} = -1, \quad \lim_{n\to+\infty} \frac{\beta(Q_n)}{\gamma(Q_n)} = 0, \quad \lim_{n\to+\infty} \frac{\delta(Q_n)}{\gamma(Q_n)} = 0 \quad \textit{for all } n \in \mathbb{N},\ \gamma(Q_n) > 0.$$

Proof Let $L > 0$, and let $Q = ([ABCD], [AC])$ be a hinge of type (x, L) such that $AC = l$. Without loss of generality, we may choose Cartesian coordinates of $\mathbb{E}^2$ such that $A : (0, 0)$ is the origin, $C : (0, l)$, $x_B > 0$, and $x_D < 0$.

There exists some $\lambda > 0$ such that

$$\beta(Q) = \lambda \frac{l}{|x_B|}, \quad \alpha(Q) = \lambda\left(\frac{l-y_B}{|x_B|} + \frac{l-y_D}{|x_D|}\right), \quad \delta(Q) = \lambda \frac{l}{|x_D|}, \quad \gamma(Q) = \lambda\left(\frac{y_B}{|x_B|} + \frac{y_D}{|x_D|}\right).$$

We have $|x_B| \le x$, $|x_D| \le x$, $y_B \ge \sqrt{L^2 - x^2}$, and $y_D \ge \sqrt{L^2 - x^2}$; thus $\gamma(Q) > 0$ and

$$-1 \le \frac{\alpha(Q)}{f(Q)} \le -1 + \frac{l}{\sqrt{L^2-x^2}}, \quad 0 \le \frac{\beta(Q)}{\gamma(Q)} \le \frac{l}{\sqrt{L^2-x^2}}, \quad 0 \le \frac{\delta(Q)}{\gamma(Q)} \le \frac{l}{\sqrt{L^2-x^2}}. \qquad \square$$

Lemma 3.29 *Let* e *be nontrivial geodesic segment of* (Σ, S) *going from some* $\sigma_1 \in S$ *to some* $\sigma_2 \in S$ *whose relative interior is in* Σ^*.

There exists $x_0 > 0$ *such that for all* $L > 0$, *there is an immersed hinge* $Q = ([ABCD], [AC], \eta)$ *of type* (x_0, L) *such that* $\eta([AC]) = e$.

Proof Let $M := \max_{x\in\Sigma} \mathrm{d}(x, S)$ and $m := \min_{s\in S} \min_{s'\in S\setminus\{s\}} \mathrm{d}(s, s')$.

Define $\Phi\colon \mathcal{U} \to \Sigma$ as the exponential map at σ_1 defined on some maximal star-shaped open neighborhood $\mathcal{U}$ of 0 in the tangent plane $T_{\sigma_1}\Sigma$ above σ_1 such that $\Phi(\mathcal{U} \setminus \{0\}) \subset \Sigma \setminus S$. We identify $T_{\sigma_1}\Sigma$ with $\mathbb{E}^2_\alpha$, where α is the cone angle at σ_1, so that Φ is an isometric immersion from an open set of $\mathbb{E}^2_\alpha$ to Σ. We choose polar coordinates (r, θ) of $\mathbb{E}^2_\alpha$ so that the direction $\theta = 0$ is the initial derivative of the segment e.

With $\beta = \min\left(\frac{1}{2}\alpha, \frac{1}{6}\pi\right)$ define

$$\begin{aligned} r_{\max}\colon]-\beta, \beta[&\to \mathbb{R}_+^* \cup \{+\infty\}, & \theta &\mapsto \max\{r \in \mathbb{R}_+ \mid (r, \theta) \in \mathcal{U}\}, \\ R_\pm\colon]0, \beta[&\to \mathbb{R}_+^* \cup \{+\infty\}, & \theta &\mapsto \min_{\theta' \in]0,\theta]} r_{\max}(\pm\theta'). \end{aligned}$$

For any given $\theta \in]-\beta, \beta[$, if $r_{\max}(\theta) < +\infty$ we extend Φ continuously to $(r_{\max}(\theta), \theta)$; note that in this case $\Phi(r_{\max}(\theta), \theta) \in S$.

Claim $$\limsup_{\theta\to 0^+} \theta R_\pm(\theta) \le 2M.$$

Let $\theta \in]0, \beta[$. Φ is defined on the interior of the triangle $[OAB] \subset \mathbb{E}^2_\alpha$ with $A = (R_+(\theta), 0)$ and $B = (R_+(\theta), \theta)$ in polar coordinates. The inscribed circle of $[OAB]$ bounds an open disc whose image by Φ does not contain any element of S, and hence the radius $\frac{1}{2}R_+(\theta)(\cos(\theta) + \sin(\theta) - 1)$ of this inscribed circle is less than M. One easily checks that $\cos(\theta) + \sin(\theta) - 1 \sim_{\theta\to 0^+} \theta$. The result follows for R_+, and one may proceed the same way for R_-; see Figure 6.

Claim $$\lim_{\theta\to 0^+} R_\pm(\theta) = +\infty.$$

The function R_+ is nondecreasing by definition, so the limit is well defined. Define a sequence $(\theta_n)_{n\in\mathbb{N}}$ as follows: choose some $\theta_0 \in]0, \beta[$ such that $r_{\max}(\theta_0) = R_+(\theta_0)$ and $\sin(\theta_0) \le \frac{1}{2}m$; then for all $n \in \mathbb{N}$ take $\theta_{n+1} \in \left]0, \frac{1}{2}\theta_n\right[$ such that $r_{\max}(\theta_{n+1}) = R_+(\theta_{n+1})$. The map Φ can be continuously extended to the domain

$$D := \bigcup_{n\in\mathbb{N}} \{(r, \theta) \mid \theta \in [0, \theta_n], r \le R(\theta_n)\}.$$

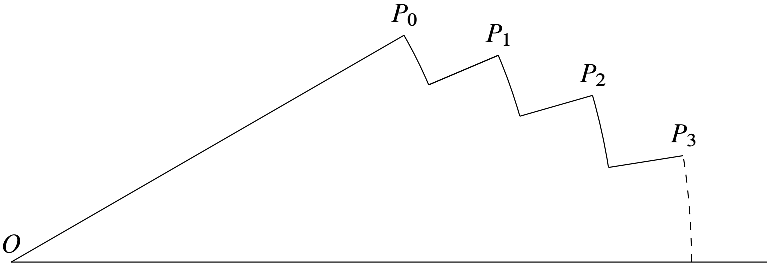


Figure 6

Write $P_n := (R_+(\theta_n), \theta_n)$; since for all $n \in \mathbb{N}$, $\Phi(P_n) \in S$, for all $n \in \mathbb{N}$,

$$R_+(\theta_{n+1}) - R_+(\theta_n) + \theta_n - \theta_{n+1} = \mathrm{d}_D(P_n, P_{n+1}) \geq \mathrm{d}_\Sigma(\Phi(P_n), \Phi(P_{n+1})) \geq m.$$

Thus

$$\text{for all } n \in \mathbb{N}, \quad R_+(\theta_n) \geq nm + R_+(\theta_0) + \theta_0 - \theta_n \xrightarrow{n \to +\infty} +\infty.$$

One may proceed the same way for R_-.

We now come back to the proof of the lemma. Take some $x_0 > M$, for any $L \in \mathbb{R}_+$. From the claims above, there exists some $\theta_+ \in]0, \beta[$ and $\theta_- \in]-\beta, 0[$ such that $|\sin(\theta_\pm) r_{\max}(\theta_\pm)| \leq x_0$ and $r_{\max}(\theta_\pm) \geq L$. Choose such a $\theta_\pm \in]-\beta, \beta[$ and notice Φ is well defined on the hinge $Q = [ABCD]$ with $A = O$, $B := (r_{\max}(\theta_-), \theta_-)$, $C := (\mathrm{length}(e), 0)$, and $D := (r_{\max}(\theta_+), \theta_+)$. The hinge Q is of type (x_0, L) and $\eta := \Phi|_Q$ is an isometric immersion. The immersed hinge (Q, η) is then of type (x_0, L) with vertices in S and such that $\Phi([AC]) = e$. □

Lemma 3.30 *There exists $C > 0$ such that for all $A, B \in S$ and all $\tau \in \mathcal{P}$,*

$$|\tau(A) - \tau(B)| \leq C.$$

Proof From Corollary 3.13, it is enough to find a $C > 0$ such that

$$\text{for all } \tau \in \mathcal{P}, \quad \min \tau = 0 \implies \max \tau \leq C.$$

From Lemmas 3.28 and 3.29 and from Proposition 3.24, for all $\varepsilon > 0$, and $A, B \in S$, if there exists a geodesic from A to B whose relative interior is in $\Sigma \setminus S$, then there exists $K > 0$ such that for all $\tau \in \mathcal{P}$, $|\tau_A - \tau_B| \leq \varepsilon \max \tau + K$. For all $A, B \in S$ there exists a geodesic from A to B possibly intersecting S in his relative interior. Hence

$$\forall \varepsilon > 0,\ \forall A, B \in S,\ \exists K > 0 \text{ such that } \forall \tau \in \mathcal{P},\ |\tau_A - \tau_B| \leq \varepsilon \max \tau + K.$$

Since S is finite,

$$\exists K > 0,\ \forall A, B \in S,\ \forall \tau \in \mathcal{P},\ |\tau_A - \tau_B| \leq \tfrac{1}{2} \max \tau + K.$$

Choose such a $K > 0$ and define $C = 2K$. Then for all $\tau \in \mathcal{P}$ such that $\min \tau = 0$,

$$\max \tau = |\max \tau - \min \tau| \leq \tfrac{1}{2} \max \tau + K.$$

Thus for such a τ

$$\max \tau \leq 2K = C. \qquad \square$$

Proof of Theorem 2 Let π be the orthogonal projection of $\mathbb{R}^S$ onto $H := \{\tau \in \mathbb{R}^S \mid \sum_{s\in S} \tau(s) = 0\}$. Note that the kernel of π is $\mathbb{R}\cdot \mathbf{1}_S$. For each triangulation $\mathcal{T}$, the set of $\tau \in \mathbb{R}_+^S$ such that $\tilde{\tau}_{\tau,\mathcal{T}}$ is Q-convex is the domain

$$\mathcal{P}_\mathcal{T} := \mathbb{R}_+^S \cap \bigcap_{e\in \mathrm{Edge}(\mathcal{T})} (Q_e^*)^{-1}(\mathbb{R}_-),$$

since $\mathbf{1}_S$ is in the kernel of the linear part of all the affine forms Q_e^*. Since the number of edges of $\mathcal{T}$ is finite, $\overline{\mathcal{P}_\mathcal{T}} := \pi(\mathcal{P}_\mathcal{T})$ is a convex polyhedron and $\mathcal{P}_\mathcal{T} = (\overline{\mathcal{P}_\mathcal{T}} + \mathbb{R}\cdot\mathbf{1}_S) \cap \mathbb{R}_+^S$.

On the one hand,

$$\mathcal{P} = \bigcup_\mathcal{T} P_\mathcal{T},$$

where $\mathcal{T}$ runs through all adapted triangulations of (Σ, S). Then defining $\overline{\mathcal{P}} := \bigcup_\mathcal{T} \overline{\mathcal{P}}_\mathcal{T}$, we have $\mathcal{P} = (\overline{\mathcal{P}} + \mathbb{R}\cdot\mathbf{1}_S) \cap \mathbb{R}_+^S$.

On the other hand, by Lemma 3.30, $\overline{\mathcal{P}} = \pi(\mathcal{P})$ is compact. Furthermore, by Proposition 3.24, $\mathcal{P}$ is convex. Hence $\overline{\mathcal{P}}$ is convex.

Then consider the set $\boldsymbol{T}$ of triangulations that are τ-Delaunay for some $\tau \in \mathcal{P}$. For any admissible $\tau \in \mathcal{P}$, it follows from Lemma 3.12 that $\tau' := \tau - \min\tau \in \mathcal{P}$ and that the set of τ-Delaunay triangulations is equal to the set of τ'-Delaunay triangulations. Therefore $\boldsymbol{T}$ is the set of triangulations that are τ-Delaunay for some $\tau \in \mathcal{P}_0 := \{\tau \in \mathcal{P} \mid \min\tau = 0\}$. By Lemma 3.30, there exists a constant C that only depends on Σ such that for all $\tau \in \mathcal{P}_0$, $\tau \leq C$. By Proposition 3.17, there thus exists a constant A that only depends on Σ such that $\tilde{\tau}_{\tau,\mathcal{T}} \leq A$ for all τ-Delaunay triangulation $\mathcal{T}$ and all $\tau \in \mathcal{P}_0$. Using notation of Lemma 3.21, we deduce that $\boldsymbol{T} \subset E_A$ is finite; hence $\boldsymbol{T}$ is finite. The domain $\overline{\mathcal{P}}$ is thus a polyhedron.

Choose any triangulation $\mathcal{T}_0$ and define $A := \sup_{\tau\in\mathcal{P}_0} \max_{x\in\Sigma} \tilde{\tau}_{\tau,\mathcal{T}_0}(x)$; by compactness of $\overline{\mathcal{P}}$, the set $\mathcal{P}_0$ is bounded. Hence $A < +\infty$. Consider the finite family $(Q_i)_{i\in[\![1,q]\!]}$ of unflippable immersed hinges around edges of triangulations in E_A and define $\mathcal{P}_A := \bigcap_{i=1}^q Q_i^{*-1}(\mathbb{R}_-)$. By Proposition 3.24 $\mathcal{P}_A \supset \mathcal{P}$. In addition, for any $\tau \in \mathcal{P}_A$ the flipping algorithm starting at $\mathcal{T}_0 \in E_A$ stops after finitely many iterations on some $\mathcal{T}_n \in E_A$; Proposition 3.17 ensures that flippable hinges of $\mathcal{T}_n$ are τ-legal and the definition of $\mathcal{P}_A$ ensures that unflippable hinges of $\mathcal{T}_n$ are also τ-legal. We deduce that $\mathcal{T}_n$ is τ-Delaunay, and hence $\tau \in \mathcal{P}$. We conclude that $\mathcal{P} = \mathcal{P}_A$, so that essential support planes of $\mathcal{P}_A$ are either

- essential support planes of $\mathbb{R}_+^S$ and thus of the form $\tau_\sigma = 0$ for some $\sigma \in S$, or
- given by "$Q_i^* = 0$" for some $i \in [\![1,q]\!]$.

Finally, since $\mathcal{P}$ is a finite union of cells, essential support planes of the second kind correspond to a facet of some cell $\mathcal{P}_\mathcal{T}$. Theorem 2(e) follows. $\square$

4 The Volkov lemma for Lorentzian convex cones

In effective methods used to prove Alexandrov-like theorems, at some point a Volkov lemma bounding the cone angle Θ around a singular line of angle κ in a Riemannian manifold is needed. This is used to exclude some positions of critical points of the Einstein–Hilbert functional introduced in the following section.

We consider spacelike convex cones in $\mathbb{E}^{1,2}_\kappa$ for $\kappa > 0$, eg the model space of the timelike singular lines of angle κ as $\mathbb{R}^3$ endowed with the metric $\mathrm{d}t^2 - \mathrm{d}r^2 - (\kappa/(2\pi))^2\,\mathrm{d}\theta^2$. There are many ways to rigorously define a spacelike cone in $\mathbb{E}^{1,2}_\kappa$. In our context, we define a cone $\mathcal{D}$ as the graph of some Lipschitz 1-homogeneous function $t\colon \mathbb{R}^2 \mapsto \mathbb{R}$. The cone is spacelike if the graph in $\mathbb{R}^3$ identified to $\mathbb{E}^{1,2}_\kappa$ is spacelike. The cone $\mathcal{D}$ is then convex if the future $J^+(\mathcal{D})$ is convex in the sense that any spacelike geodesic with extremities in $J^+(\mathcal{D})$ is in $J^+(\mathcal{D})$. The Lorentzian structure of $\mathbb{E}^{1,2}_\kappa$ induces complete metric space structure on the cone, which is locally Euclidean except possibly at $\{r = 0\}$. In other words, $\mathcal{D}$ is isometric to $\mathbb{E}^2_\Theta$ for some $\Theta > 0$; this Θ is its so-called cone angle.

Let $\mathcal{D}$ be a cone defined as the graph of $t\colon \mathbb{R}^2 \to \mathbb{R}$. A *wedge* is the graph of t on some domain $\{\theta \in I,\, r \geq 0\}$ with I an interval; Such a wedge is *coplanar* if it is totally geodesic. A wedge is isometric to some domain $\{(r,\theta) \mid r \geq 0,\, 0 \leq \theta \leq \pi\}$ in $(\mathbb{R}^2, \mathrm{d}r^2 + (\alpha/\pi)^2 r^2\,\mathrm{d}\theta^2)$; the value of α is unique and we refer to it as the *Euclidean angle* of the wedge.

Theorem 3 *Let $\Theta > 0$ and $\kappa > 0$. Let $\mathcal{D}$ be a convex spacelike cone in $\mathbb{E}^{1,2}_\kappa$ of cone angle Θ whose vertex is on the singular line of $\mathbb{E}^{1,2}_\kappa$.*

Assuming $\mathcal{D}$ has a coplanar wedge of Euclidean angle at least $\min(\pi, \Theta)$,

- *if $\Theta > 2\pi$ then $\kappa > 2\pi$,*
- *if $\Theta = 2\pi$ then $\kappa = 2\pi$,*
- *if $\Theta \in\,]\pi, 2\pi[$ then $\kappa \geq \Theta$,*
- *if $\Theta = \pi$ then $\kappa = \pi$,*
- *if $\Theta < \pi$ then $\kappa \in\,]0, \Theta]$ with $\kappa = \Theta$ if and only if $\mathcal{D}$ is the horizontal plane,*

and all the bounds above are sharp.

Remark Though results such as stated above are used one way or another in [3; 6; 18; 19; 22; 26], to our knowledge, a complete proof of the bounds we use is not available in English (one may appear in the original thesis of Volkov which is in Russian, and only a summary is available in English [34]). We thus provide a complete proof.

Remark In Minkowski, a convex cone always has a cone angle bigger than 2π. One may expect this to be carried out in $\mathbb{E}^{1,2}_\kappa$ for arbitrary $\kappa \geq 0$. Theorem 3 shows this intuition is valid for $\kappa \in [0, \pi] \cup \{2\pi\}$ but not for $\kappa \in\,]\pi, 2\pi[$.

When considering a cone $\mathcal{D}$ in $\mathbb{E}^3$, an elementary remark is that the angle of the conical singularity is, in fact, the length of its *stalk*: the curve given by the intersection $\mathcal{D}\cap\mathbb{S}^2$. By extension "stalk" refers to curves in $\mathbb{S}^2$ or $\mathbb{S}^{1,1}$ that are graphs over the "equator". As in the Euclidean case, we may notice that the angle Θ of the conical singularity of a spacelike cone in $\mathbb{E}^{1,2}_\kappa$ is given by the length of the spacelike curve induced on $\mathbb{S}^{1,1}_\kappa := \{(t,r,\theta)\in\mathbb{E}^{1,2}_\kappa \mid r^2-t^2=1\}$. However, the relation between κ and Θ is far from trivial, and the Lorentzian nature of $\mathbb{S}^{1,1}$ does not help. One may devise an analytical proof of the needed Volkov lemma [9], but a more geometrical one is provided based on a suggestion of Graham Smith.

The key idea developed in Section 4.1 is that to each cone stalk $\rho\colon\mathbb{R}/\kappa\mathbb{Z}\to\mathbb{S}^{1,1}_\kappa$ corresponds a dual stalk $\gamma\colon\mathbb{R}/\Theta\mathbb{Z}\to\mathbb{S}^2_\Theta$. The length of ρ is the Euclidean cone angle Θ while the length of γ is the Lorentzian cone angle κ.

4.1 Stalks of Lorentzian cones

Although we defined cones with Lipschitz regularity, we will focus our argumentation on polyhedral cones that are simpler to describe. A density argument allows us to generalize to lower regularity.

Definition 4.1 (stalk of a spacelike cone) Let $\kappa>0$, $\Theta>0$ and $\mathcal{D}$ be a spacelike cone of $\mathbb{E}^{1,2}_\kappa$ of cone angle Θ. In cylindrical coordinates (r,θ,t), the set $\mathbb{S}^{1,1}_\kappa\cap\mathcal{D}$ can be parametrized by arc length with increasing θ coordinate:

$$\mathcal{D}\cap\mathbb{S}^{1,1}_\kappa=\left\{\begin{pmatrix}t(s)\\ r(s)\\ \theta(s)\end{pmatrix}\;\middle|\; s\in\mathbb{R}\right\}.$$

The stalk $\rho_{\mathcal{D}}$ of $\mathcal{D}$ is the function $t\colon\mathbb{R}\to\mathbb{R}$ of this parametrization.

Remark The stalk ρ of a cone is unique up to precomposition by an affine transformation of slope ±1.

Proposition 4.2 *Let* $\kappa>0$, $\Theta>0$ *and* $\mathcal{D}$ *be a cone of* $\mathbb{E}^{1,2}_\kappa$ *of cone angle* Θ *whose vertex is on the origin and of stalk* $\rho:=\rho_{\mathcal{D}}$. *We have the following*:

(1) $\rho\colon\mathbb{R}\to\mathbb{R}$ *is* Θ*-periodic and Lipschitz continuous.*

(2) $\mathcal{D}$ *is polyhedral if and only if* ρ *is piecewise trigonometric* (*piecewise of the form* $\theta\mapsto A\cos(\theta+\varphi)$).

(3) *If* $\mathcal{D}$ *is polyhedral then*
$$\mathcal{D}\text{ is convex}\iff\rho\text{ is }Q\text{-convex.}$$

(4) $\kappa=\int_0^\Theta\sqrt{1+\rho(\theta)^2+\rho'(\theta)^2}/(1+\rho(\theta)^2)\,\mathrm{d}\theta.$

Proof The first three points are simple enough. To obtain the last item, we first choose a parametrization by arc length $s\mapsto(t,r,\theta)$ of $\mathcal{D}\cap\mathbb{S}^{1,1}$ with θ increasing and notice

$$2\pi=\int_0^\Theta\theta'(s)\,\mathrm{d}s,\quad r^2-\rho^2=1,\quad -(\rho')^2+(r')^2+\left(\frac{\kappa}{2\pi}\right)^2r^2(\theta')^2=1.$$

Therefore $rr' = \rho\rho'$ and

$$\begin{aligned}(\theta')^2 &= \Big(\frac{2\pi}{\kappa}\Big)^2 \frac{1+(\rho')^2-(r')^2}{r^2} = \Big(\frac{2\pi}{\kappa}\Big)^2 \frac{1+(\rho')^2-(\rho\rho'/r)^2}{1+\rho^2} = \Big(\frac{2\pi}{\kappa}\Big)^2 \frac{(1+(\rho')^2)(1+\rho^2)-\rho^2(\rho')^2}{(1+\rho^2)^2} \\ &= \Big(\frac{2\pi}{\kappa}\Big)^2 \frac{1+(\rho')^2+\rho^2}{(1+\rho^2)^2},\end{aligned}$$

and so

$$\theta' = \frac{2\pi}{\kappa}\frac{\sqrt{1+(\rho')^2+\rho^2}}{1+\rho^2}.$$

Insert the last line in $2\pi = \int_0^\Theta \theta'$ to get the result. □

Remark For $\rho\colon I \to \mathbb{R}$ continuous piecewise trigonometric, ρ is Q-convex if and only if $s \mapsto \rho(-s)$ is Q-convex.

Definition 4.3 (mass of a stalk) For $\rho\colon [a,b] \to \mathbb{R}$ (resp. $\rho\colon \mathbb{R}/\Theta\mathbb{Z} \to \mathbb{R}$), define

$$\kappa(\rho) := \int_a^b \frac{\sqrt{1+\rho^2+\rho'^2}}{1+\rho^2} \quad \left(\text{resp. } \int_0^\Theta \frac{\sqrt{1+\rho^2+\rho'^2}}{1+\rho^2}\right).$$

Remark Every $\rho\colon \mathbb{R} \to \mathbb{R}$ piecewise trigonometric Q-convex and Θ-periodic induces a convex polyhedral embedding of $\mathbb{E}^2_\Theta$ into $\mathbb{E}^{1,2}_{\kappa(\rho)}$. Furthermore, this embedding is essentially unique: from Proposition 4.2, the mass κ is given by ρ; there is thus no choice for the space $\mathbb{E}^{1,2}_\kappa$ and two embeddings of the same germ only differ by a rotation or a symmetry.

Corollary 4.4 *Let $\kappa > 0$, $\Theta > 0$ and $\mathcal{D}$ be a spacelike polyhedral cone in $\mathbb{E}^{1,2}_\kappa$ of cone angle Θ.*

If its stalk ρ is $\mathscr{C}^1$ then $\kappa(\rho) = \Theta$. Furthermore, if κ is not a multiple of 2π then $\rho = 0$.

Proof To begin with, since ρ is piecewise trigonometric and continuously differentiable, ρ is in fact trigonometric. Then either 2π is the minimal period of ρ or $\rho = 0$. If $\rho = 0$, the result follows from Proposition 4.2(4). Otherwise $\Theta \in 2\pi\mathbb{N}$ and we notice that for any A and φ we have $\kappa(s \mapsto A\cos(s+\varphi)) = 2k\pi$ if $\Theta = 2k\pi$. □

Lemma 4.5 *Let I be an interval, $\rho\colon I \to \mathbb{R}$ be piecewise trigonometric Q-convex, and let $\theta_0 \in I$. Let $\bar\rho$ be the unique trigonometric function such that $\rho(\theta_0) = \bar\rho(\theta_0)$ and $\rho'(\theta_0^+) = \bar\rho'(\theta_0)$. Then for all $\theta \in I \cap [\theta_0, \theta_0+\pi]$,*

$$\rho(\theta) \geq \bar\rho(\theta).$$

Furthermore:

$$\textit{there exists } \theta \in \,]\theta_0, \theta_0+\pi[\, \textit{ such that } \bar\rho(\theta) = \rho(\theta) \iff \textit{for all } \theta \in [\theta_0, \theta_0+\pi],\ \rho(\theta) = \bar\rho(\theta).$$

Proof Let $(\theta_0, \theta_1, \dots, \theta_n = \theta_0 + \pi)$ be subdivision adapted to ρ. For $k \in [\![1, n]\!]$, denote by $\rho_k\colon \mathbb{R} \to \mathbb{R}$ the unique trigonometric function such that $\rho|_{k[\theta_{k-1},\theta_k]} = \rho|_{[\theta_{k-1},\theta_k]}$ and define $\rho_0 = \bar\rho$.

For $k \in [\![0, n-1]\!]$, we have $\rho_k(\theta_k) = \rho_{k+1}(\theta_k)$. If $\rho'_k(\theta_k) = \rho'_{k+1}(\theta_k)$ then $\rho_k = \rho_{k+1}$. Otherwise $\rho'_k(\theta_k) < \rho'_{k+1}(\theta_k)$; thus $\rho_k < \rho_{k+1}$ on a nontrivial interval $[\theta_k, \theta_k + \varepsilon]$. These two trigonometric functions are in particular distinct and intersect each other on the set $\theta_k + \pi\mathbb{Z}$. Hence $\rho_k - \rho_{k+1}$ has constant sign on the interval $[\theta_k, \theta_k + \pi]$ and $\rho_k \le \rho_{k+1}$ on $[\theta_k, \theta_k + \pi]$. By induction, the result follows. □

Definition 4.6 Let $\mathbb{S}^2_\infty$ be the universal covering of the round sphere branched over its north and south poles, eg $\left[-\frac{1}{2}\pi, \frac{1}{2}\pi\right] \times \mathbb{R}/\sim$ endowed with the metric

$$\mathrm{d}s^2 = \mathrm{d}\phi^2 + \cos(\phi)^2\,\mathrm{d}\theta^2,$$

where $\sim$ identifies all points such that $\phi = \frac{1}{2}\pi$ together as the north pole N and all points such that $\phi = -\frac{1}{2}\pi$ as the south pole S.

Definition 4.7 A piecewise geodesic curve $\gamma\colon I \to \mathbb{S}^2_\infty$ is Q-convex if $\theta \circ \gamma$ is injective and $\phi \circ \gamma$ is Q-convex.

Lemma 4.8 *Let $\rho\colon [a,b] \to \mathbb{R}$ be Lipschitz continuous and define $\gamma\colon [a,b] \to \mathbb{S}^2_\infty$, $\theta \mapsto (\arctan \rho(\theta), \theta)$. Then*

$$\kappa(\rho) = \mathrm{length}(\gamma).$$

Furthermore, γ is a piecewise geodesic Q-convex curve if and only if ρ is piecewise trigonometric Q-convex.

Proof By direct computation:

$$\begin{aligned}\mathrm{length}(\gamma) &= \int_a^b \sqrt{\frac{(\rho')^2(\theta)}{(1+\rho^2(\theta))^2} + \cos^2(\arctan \circ \rho(\theta))}\,\mathrm{d}\theta = \int_a^b \sqrt{\frac{(\rho')^2(\theta)}{(1+\rho^2(\theta))^2} + \frac{1}{1+\rho^2(\theta)}}\,\mathrm{d}\theta \\ &= \int_a^b \sqrt{\frac{(\rho')^2(\theta) + 1 + \rho^2(\theta)}{(1+\rho^2(\theta))^2}}\,\mathrm{d}\theta = \kappa(\rho).\end{aligned}$$

Then it suffices to note that curves of the form $t \mapsto (\phi(t), \theta(t))$ with $\phi(t) = \arctan(\alpha\cos(\theta(t) + \phi_0) \notin \left\{-\frac{1}{2}\pi, \frac{1}{2}\pi\right\}$ and $\theta(t) = t$ are exactly the nonmeridional geodesic segment of $\mathbb{S}^2_\infty$. □

Proposition 4.9 *Let S_Θ be the set of Lipschitz stalks of convex spacelike cones of cone angle Θ admitting a coplanar wedge of Euclidean angle $\min(\Theta, \pi)$ endowed with the Lipschitz norm*

$$\|\rho\|_{\mathrm{Lip}} := \sup_{s \in I} |\rho(s)| + \sup_{s_1 \neq s_2} \left| \frac{\rho(s_1) - \rho(s_2)}{s_1 - s_2} \right|.$$

Then the subspace of piecewise trigonometric Q-convex functions is dense in S_Θ.

Sketch of proof Consider the stalk $\rho_{\mathcal{D}}\colon \mathbb{R} \to \mathbb{R}$ of a (Lipschitz) spacelike convex cone $\mathcal{D}$ and its associated geodesics γ in $\mathbb{S}^2_\infty$. Consider $[a, a+\alpha] + \Theta\mathbb{Z}$ with $\alpha \ge \min(\Theta, \pi)$ and $\rho|_{[a+k\Theta, a+k\Theta+\alpha)}$ trigonometric for all $k \in \mathbb{Z}$.

The curve γ divides $\mathbb{S}^2_\infty$ into two parts (north and south); the epigraph of γ is the northern domain. Convexity of the spacelike cone translates in $\mathbb{S}^2_\infty$ into the local convexity of the epigraph of γ. We may construct an approximating sequence $(\gamma_n)_{n\in\mathbb{N}}$ of γ interpolating by geodesics, say between points of the form $(s_k, \gamma(s_k))$ with $(s_k)_{k\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$ increasing such that $|s_{k+1}-s_k| \leq 1/(1+n)$, $\lim_{\pm\infty} s_k = \pm\infty$, and $\{a+k\Theta, a+\alpha+k\Theta \mid k\in\mathbb{Z}\} \subset \{s_k \mid k\in\mathbb{Z}\}$. Then notice that for n big enough, the geodesics are not meridional and thus correspond to a piecewise trigonometric Θ-periodic stalk ρ_n. By convexity of γ, each γ_n is Q-convex for n big enough, and so are the ρ_n. We note that γ is Lipschitz and that the sequence γ_n converges in Lipschitz norm to γ. □

4.2 Lower bounds

Lemma 4.8 provides a neat geometrical translation from Lorentzian to Riemannian. Indeed, an issue with the geometry of Lorentzian manifolds is that spacelike geodesics are not characterized as minimizers of the usual energy Lagrangian $\int g(\dot\gamma,\dot\gamma)$. The description of convex polyhedral cones as Q-convex piecewise geodesics in $\mathbb{S}^2_\infty$ allows us to leverage the usual Riemannian theory of geodesics.

Proposition 4.10 *Let $\Theta \geq \pi$. Then*

$$\inf_{\rho\in S_\Theta} \kappa(\rho) = \min(2\pi, \Theta),$$

the infimum being taken over the set S_Θ of stalks $\rho\colon \mathbb{R}/\Theta\mathbb{Z} \to \mathbb{R}$ of convex spacelike cones admitting a coplanar wedge of Euclidean angle at least π. Furthermore, the infimum is achieved if and only if $\Theta \leq 2\pi$.

Proof Note that by Proposition 4.9, piecewise trigonometric elements of S_Θ form a dense subspace for a norm for which κ is continuous. By abuse of language, we say that "ρ is a Q-convex stalk", meaning that ρ is a Lipschitz limit of piecewise trigonometric Q-convex stalks.

- Assume $\Theta \geq 2\pi$. Consider for $\alpha > 0$ the stalk
$$\rho\colon \mathbb{R}/\Theta\mathbb{Z} \to \mathbb{R}, \quad \theta \mapsto \begin{cases} \sinh(\alpha)\sin(\theta) & \text{for } \theta \in \left[-\frac{3}{2}\pi, \frac{1}{2}\pi\right], \\ \sinh(\alpha) & \text{otherwise,} \end{cases}$$
so that ρ is Q-convex and $\kappa(\rho) = 2\pi + (\Theta - 2\pi)/\cosh(\alpha)$. As a result, $\inf_{\rho\in S_\Theta} \kappa(\rho) \leq 2\pi$.

- Assume $\Theta \leq 2\pi$. Then the stalk $\rho = 0$ is such that $\kappa(\rho) = \Theta$.

- Let $\gamma\colon [0, \Theta-\pi] \to \mathbb{S}^2_\infty$ be a Lipschitz curve from $(\phi_0, 0)$ to $(-\phi_0, \Theta-\pi)$ minimizing the length with $\gamma(0), \gamma(\Theta-\pi) \notin \{N, S\}$. The curve γ is a geodesic with possibly intermediate points in $\{N, S\}$.
  - Assume γ does not intersect $\{N, S\}$. Then up to reparametrization $\phi\circ\gamma$ is of the form $\theta \mapsto \arctan(\alpha\cos(\theta+\theta_0))$. If $\Theta \geq 2\pi$, then the length of such a curve is at least π. Otherwise, since $\phi\circ\gamma(0) = -\phi\circ\gamma(\Theta-\pi)$ up to reparametrization, $\rho(\theta) = \sinh(\alpha)\sin(\theta)$ with $\theta \in \left[m-\frac{1}{2}\pi, \frac{1}{2}\pi - m\right]$ and $m = \pi - \frac{1}{2}\Theta$. Then
  $$\kappa(\rho) = \pi - 2\arctan\left(\frac{\tan(m)}{\cosh(\alpha)}\right),$$
  which is minimal if and only if $\alpha = 0$, in which case the length of γ is $\Theta - \pi$.

- Assume γ intersects $\{N, S\}$ exactly once. Then γ is formed of a geodesic from $\gamma(0)$ to N (resp. S) followed by a geodesic from N (resp. S) to $\gamma(\Theta)$. Such geodesics are meridional; hence the length of γ is exactly π.
- Assume γ intersects $\{N, S\}$ at least twice. Then it contains a meridional geodesic from N to S and its length is strictly bigger than π.

In any case, the length of the curve associated by Lemma 4.8 to a stalk ρ in S_Θ is bounded from below by π plus the length of such a minimizing curve γ. Hence $\inf_{\rho\in S_\Theta}\kappa(\rho)=\min(\Theta,2\pi)$. Furthermore, the infimum is achieved if and only if the minimizing geodesic can be associated with a stalk, which is only possible if the geodesic γ considered above reaches neither N nor S; this is possible only if $\Theta\leq 2\pi$. Reciprocally, if $\Theta\leq 2\pi$, then the stalk $\rho=0$ achieves the infimum. □

Proposition 4.11 *Let $\Theta<\pi$. Then*
$$\inf_{S_\Theta}\kappa(\rho)=0,$$
the infimum being taken over the set S_Θ of stalks $\rho\colon \mathbb{R}/\Theta\mathbb{Z}\to\mathbb{R}$ of convex spacelike cones admitting a coplanar wedge of Euclidean angle at least Θ.

Furthermore, among such stalks, $\kappa(\rho)\leq\Theta$ with equality if and only if $\rho=0$.

Proof Any element of S_Θ is of the form
$$\rho_\alpha\colon \mathbb{R}/\Theta\mathbb{Z}\to\mathbb{R},\quad \theta+\Theta\mathbb{Z},\theta\in[-\Theta/2,\Theta/2]\mapsto\sinh(\alpha)\cos(\theta+\theta_0),$$
for some $\alpha\in\mathbb{R},\theta_0\in\,]-\pi,\pi[$. If $\alpha=0$ we may choose $\theta_0=0$; otherwise, up to translation, we may assume that $\rho\big(-\tfrac{1}{2}\Theta\big)=\rho\big(\tfrac{1}{2}\Theta\big)$, which implies that $\big(\tfrac{1}{2}\Theta+\theta_0\big)=\pm\big(-\tfrac{1}{2}\Theta+\theta_0\big)+2k\pi$ for some $k\in\mathbb{Z}$. Therefore either $\Theta=2k\pi$ or $\theta_0=k\pi$; since $0<\Theta<\pi$ and $|\theta_0|<\pi$, it follows that $\theta_0=0$.

In particular, all elements of S_Θ are piecewise trigonometric. On the one hand, since ρ_α is Q-convex only for $\alpha\geq 0$,
$$\text{for all }\alpha\in\mathbb{R},\quad \rho_\alpha\in S_\Theta\iff\alpha\geq 0.$$
On the other hand
$$\text{for all }\alpha\geq 0,\quad \kappa(\rho_\alpha)=2\arctan\left(\frac{\tan\big(\tfrac{1}{2}\Theta\big)}{\cosh(\alpha)}\right)\xrightarrow{\alpha\to+\infty}0.$$
It follows that $\inf_{\rho\in S_\Theta}\kappa(\rho)=0$. Note that $\alpha\mapsto\kappa(\rho_\alpha)$ is decreasing; the maximum is thus reached for $\alpha=0$; hence $\rho=0$. The formula above gives $\kappa(0)=\kappa(\rho_0)=2\arctan\big(\tan\big(\tfrac{1}{2}\Theta\big)\big)=\Theta$. □

4.3 Proof of the Volkov lemma

We now compile and complete the elements proven in the previous section.

Proof of Theorem 3 Proposition 4.10 implies the first and third claims, and partially the second. The fifth claim is a consequence of Proposition 4.11

To complete the second consider the stalk ρ of a convex spacelike cone of Euclidean angle 2π having a coplanar wedge of angle π. By Proposition 4.9 we may assume ρ is piecewise trigonometric. Using the remark just before Definition 4.3, assume without loss of generality that $\rho(0) = -\rho(\pi) \geq 0$ and $\rho|_{[\pi,2\pi]}$ is trigonometric. Using Lemma 4.5 we see that if ρ is not trigonometric on $[0,\pi]$ then $-\rho(0) = \rho(\pi) > \bar{\rho}(\pi) = -\bar{\rho}(0) = -\rho(0)$ for some trigonometric function $\bar{\rho}$, a contradiction. Therefore ρ is trigonometric on $[0,\pi]$ and on $[\pi, 2\pi]$ so that $\kappa(\rho) = 2\pi$.

The same argument allows us to prove the fourth claim. □

5 The Einstein–Hilbert functional

We give ourselves a Euclidean surface Σ with conical singularities and marked points $S \supset \mathrm{Sing}(\Sigma)$; we will keep this surface fixed in the whole section.

To sum up the results of the preceding sections, we have a construction that associates to any $\tau \in \mathcal{P}$ a radiant spacetime $M(\tau)$ and a convex polyhedral embedding ι_τ of (Σ, S) into $M(\tau)$. We know from Proposition 2.23 this construction reaches every equivalence classes of such a couple (M, ι) and is injective. By Theorem 2, $\mathcal{P}$ is a convex domain of $\mathbb{R}_+^S$ and is the union of finitely many convex cells, each corresponding to a triangulation of (Σ, S).

The objective is now to construct polyhedral embeddings (M, ι) such that the singularities of M have cone angles we gave ourselves a priori.

Definition 5.1 (mass function) Let $\tau \in \mathcal{P}$ and $(M(\tau), \iota_\tau)$ be its associated polyhedral embedding of (Σ, S). For $\sigma \in S$ define $\kappa_\sigma(\tau)$ the (Lorentzian) cone angle of $M(\tau)$ at $\iota_\tau(\sigma) \in M(\tau)$.

We define $\kappa \colon \mathcal{P} \to \mathbb{R}_+^S$ the map that associates to τ the vector $(\kappa_\sigma(\tau))_{\sigma \in S}$.

Remark On each cell $\mathcal{P}_\mathcal{T} := \{\tau \in \mathcal{P} \mid \mathcal{T}_\tau = \mathcal{T}\}$, the function $\tau \mapsto \kappa(\tau)$ is continuous and furthermore $\mathscr{C}^1$. Since we will actually compute the derivative later on, we do not prove it now.

Furthermore, if $\tau \in \mathcal{P}_\mathcal{T} \cap \mathcal{P}_{\mathcal{T}'}$ the triangulations $\mathcal{T}$ and $\mathcal{T}'$ are τ-equivalent; κ computed with either triangulation yields the same result since $M(\tau)$ may be constructed using $\mathcal{T}$ or $\mathcal{T}'$. The map $\tau \mapsto \kappa$ is thus continuous on $\mathcal{P}$.

Remark As a complement to the previous remark, we do not neglect the limit case $\tau_\sigma = 0$ for which $\kappa_\sigma = 0$ by convention. One may check directly that $\lim_{\tau_\sigma \to 0^+} \kappa_\sigma(\tau) = 0$.

Reformulating with this notation, we thus aim to solve the following:

Problem *Let $\bar{\kappa} \in \mathbb{R}_+^S$. Is there some $\tau \in \mathcal{P}$ such that $\kappa(\tau) = \bar{\kappa}$, and if so, is it unique*?

There is a restriction on the possible $\bar{\kappa}$. Indeed, for any $\tau \in \mathcal{P}$, the spacetime $M(\tau)$ is the suspension of some marked closed hyperbolic surface with conical singularities and cusp[4] $(\Sigma_{\mathbb{H}^2}, S')$ marked by (Σ, S), and the

[4] See Definitions A.6 and A.7 in the appendix.

cone angles at S' are $\kappa(\tau)$. Therefore, by the Gauss–Bonnet formula, $\sum_{\sigma\in S}(2\pi-\kappa(\tau)_\sigma)-\mathrm{Area}(\Sigma_{\mathbb{H}^2}) = 2\pi\chi(\Sigma) = \sum_{\sigma\in S}(2\pi-\theta_\sigma)$. Hence

$$\text{for all } \tau\in\mathcal{P},\quad \sum_{\sigma\in S}\theta_\sigma > \sum_{\sigma\in S}\kappa(\tau)_\sigma.$$

In addition to this global constraint, there are local constraints due to upper bounds in the Volkov lemma. We do not systematically explore the local upper bounds and only provide the one that is consistent with the boundary condition (ie the last item of Theorem 3). We settle for an incomplete statement.

Theorem 4 *Let (Σ, S) be a closed locally Euclidean surface of genus g with conical singularities of angles $(\theta_\sigma)_{\sigma\in S}$. Using notation of the previous sections,*

$$\textit{for all } \bar\kappa \in \Bigl(\prod_{\sigma\in S}[0,\min(\theta_\sigma,2\pi)]\Bigr)\setminus\{(\theta_\sigma)_{\sigma\in S}\} \textit{ there exists } \tau\in\mathcal{P} \textit{ such that } \kappa(\tau)=\bar\kappa.$$

Furthermore, if for all $\sigma\in S, \bar\kappa_\sigma<\theta_\sigma$, then such a τ is unique. Finally, if $\theta_\sigma\le\pi$ for some $\sigma\in S$ then for all $\tau\in\mathcal{P}, \kappa_\sigma(\tau)\le\theta_\sigma$.

The proof relies on the analysis of a so-called Einstein–Hilbert functional; the first step is to define a functional $\mathcal{H}_{\bar\kappa}$ on $\mathcal{P}$ for a given $\bar\kappa$ whose critical points are solution to the problem before Theorem 4. In fact, one could check that such a functional exists by checking $\partial\kappa_{\sigma_1}/\partial h_{\sigma_2} = \partial\kappa_{\sigma_2}/\partial h_{\sigma_1}$.

For technical reasons which will shortly make themselves clear, it will be more appropriate to define such a functional on the domain $\mathcal{P}^{1/2} := \{h\in\mathbb{R}_+^S \mid h^2\in\mathcal{P}\}$. Elements of $\mathcal{P}^{1/2}$ will be denoted systematically by h, while elements of $\mathcal{P}$ will be denoted by τ. Going from the one to the other being simple, we extend all definitions to $\mathcal{P}^{1/2}$: $M(h) := M(h^2)$, etc.

A standard analysis of the critical points of $\mathcal{H}_{\bar\kappa}$ as well as its gradient on the boundary of $\mathcal{P}^{1/2}$ follows. Under the assumption that for all $\sigma\in S$, $\bar\kappa_\sigma$ is no greater than 2π and less than the cone angle of Σ at σ, we show that critical points of $\mathcal{H}_{\bar\kappa}$ are positive definite and that the gradient of $\mathcal{H}$ on the boundary of $\mathcal{P}$ is homotopic to an outward vector field.

5.1 Reminders on Lorentzian angles and Schläffli's Formula

The following is an adaptation of the exposition of Rabah Souam [32].

To begin with, the modulus $|u|$ of a vector u of $\mathbb{E}^{1,2}$ is

$$|u| = \sqrt{\langle u \mid u\rangle},$$

with the convention that when $\langle u\mid u\rangle<0$ we have that $|u|=\lambda i$ with $\lambda>0$ and $i^2=-1$. Let u and v be two vectors of $\mathbb{E}^{1,2}$. Then the angle $\angle uv$ is defined so that it satisfies the following properties:

(1) For all vectors u and v, $\angle uv\in\mathbb{R}+i\mathbb{R}/(2\pi\mathbb{Z})$.

(2) For all vectors u and v, $\langle u\mid v\rangle = |u||v|\cosh(\angle uv)$.

(3) For all vectors u, v and w coplanar, $\angle uv+\angle vw=\angle uv$.

Beware that if u and v are spacelike, $\angle uv$ is not the usual angle $\widehat{uv}$ but actually $\widehat{uv}\cdot i$. Angles are well defined only if neither u nor v are lightlike.

Definition 5.2 (type of a vector of $\mathbb{E}^{1,1}$) Choose a direct Cartesian coordinate system (t,x) of the vector space underlying $\mathbb{E}^{1,1}$. Let u be a nonlightlike vector of $\mathbb{E}^{1,1}$. The type $k_u \in \mathbb{Z}/4\mathbb{Z}$ of u is defined as follows:

- $k_u = 0$ if u is future timelike.
- $k_u = 1$ if u is spacelike with negative spacelike coordinate.
- $k_u = 2$ if u is past timelike.
- $k_u = 3$ if u is spacelike with positive spacelike coordinate.

Definition 5.3 Define $\mathbb{H}^1_+$ as the Riemannian submanifold of unit future timelike vectors in $\mathbb{E}^{1,1}$. We choose the orientation $\vec{x}$ of $\mathbb{H}^1_+$ so that $(\vec{x},\vec{n})$ induces the same orientation as $\mathbb{E}^{1,1}$ for any future timelike vector $\vec{n}$.

Definition 5.4 Let u and v be two linearly independent nonlightlike unit vectors in $\mathbb{E}^{1,2}$ and let Π the vectorial plane generated by u and v.

- If Π is spacelike,
$$\angle uv = i\theta,$$
with θ the angle from u to v in Π oriented by the future timelike normal.

- If Π is timelike and u and v of types k_u and k_v in Π are identified with $\mathbb{E}^{1,1}$ and oriented by the basis (u,v), then
$$\angle uv = \alpha + i(k_v - k_u)\tfrac{1}{2}\pi,$$
with α the (oriented) length of the geodesics from u' to v' in $\mathbb{H}^1_+$, where u' (resp. v') is the unique future unit timelike vector of Π orthogonal or colinear to u (resp. v).

Definition 5.5 (dihedral angle) Let Π_1 and Π_2 be two vectorial half-planes that intersect along their common boundary Δ. Assume none of Π_1, Π_2, and Δ are lightlike and write $v_i = \Delta^\perp \cap \Pi_i$ for $i \in \{1,2\}$. We choose some $u \in \Delta$ and for $i \in \{1,2\}$ define n_i, the unique unit vector normal to Π_i such that (u, v_i, n_i) is a direct basis. The dihedral angle $\angle \Pi_1\Pi_2$ between the planes Π_1 and Π_2 is then defined as
$$\angle \Pi_1\Pi_2 := \begin{cases} \mathrm{Real}(\angle n_1 n_2) & \text{if } \Delta^\perp \text{ is Lorentzian,} \\ \mathrm{Im}(\angle n_1 n_2) \in \,]-\pi,\pi] & \text{if } \Delta^\perp \text{ is Riemannian.} \end{cases}$$

Remark In the definition above, the dihedral angle does not depend on the choice of u.

Definition 5.6 (1-parameter family of oriented polyhedra) A 1-parameter family of oriented locally Minkowski polyhedra is the data of an oriented simplicial complex $\mathcal{K}$ and a map $\psi\colon [0,1]\times\mathcal{K}\to\mathbb{E}^{1,2}$ such that

(1) for all simplices P of $\mathcal{K}$ and all $t\in[0,1]$, $\psi|_{\{t\}\times P}$ is an orientation-preserving smooth embedding and $\psi(t,P)$ a polyhedron of $\mathbb{E}^{1,2}$,

(2) for all simplices P the restriction of ψ to $[0,1]\times P$ is smooth.

Let $(\mathcal{K},\psi)$ be a 1-parameter family of locally Minkowski polyhedra. If e is an edge of $\mathcal{K}$, then for all $t\in[0,1]$, we write $l_{e,t}\geq 0$ for the length of the edge $\psi(e,t)\subset\mathbb{E}^{1,2}$ and $\theta_{e,t}$ for the sum of the dihedral angles between the faces of the simplices of $\mathcal{K}$ around the edge e.

We will also have to assume that adjacent 2-facets never change convexity. This can be made rigorous by saying that the family $\{u,v_1,v_2\}$ used in the definition of the dihedral angle above always is such that $\det(uv_1v_2)$ has constant sign (but can be 0).

Theorem (Schläffli's formula [32]) *Let $(\mathcal{K},\psi)$ be a 1-parameter family of oriented locally Minkowski polyhedra such that none of its faces or edges are lightlike and such that adjacent 2-facets never change convexity. Denoting by $\mathcal{E}$ the set of edges of $\mathcal{K}$, we have*

$$\sum_{e\in\mathcal{E}} l_{e,t}\frac{\mathrm{d}\theta_{e,t}}{\mathrm{d}t}=0.$$

Remark The convexity condition is always satisfied by construction for the polyhedra we consider.

5.2 Kites and angles

Consider an adapted triangulation $\mathcal{T}$ of (Σ,S) and consider a cell $\mathcal{P}^{1/2}_{\mathcal{T}}$ of $h\in\mathcal{P}^{1/2}$ of nonempty interior.

For $h\in\mathcal{P}^{1/2}_{\mathcal{T}}$, the past of Σ in $M(h)$ is a locally Minkowski polyhedron with each simplex being a pyramid of $\mathbb{E}^{1,2}$ as represented in Figure 7, the notation of which we give a more precise meaning. If T is a triangle of $\mathcal{T}$ of vertices σ_1, σ_2, and σ_3, while $e=\overrightarrow{\sigma_1\sigma_2}$ and $e'=\overrightarrow{\sigma_1\sigma_3}$ are two edges on the boundary of T, define ρ_e the real part of the angle from $\overrightarrow{\sigma_1 O}$ to $\overrightarrow{\sigma_1\sigma_2}$, $\theta_{ee'}$ the real part of the angle from $\overrightarrow{\sigma_1\sigma_2}$ to $\overrightarrow{\sigma_1\sigma_3}$ and α_e the real part of the dihedral angle from the plane $(O\sigma_1\sigma_1)$ to the plane $(\sigma_1\sigma_2\sigma_3)$. In this section, edges are oriented so that we distinguish α_e and α_{-e}: the angle α_e is on the left of e, and thus α_{-e} is the angle on the right of e.

We aim at proving κ is continuous and computing the partial derivatives

$$\frac{\partial\kappa_{\sigma_1}}{\partial h_{\sigma_2}}\quad\text{for } \sigma_1,\sigma_2\in S.$$

If there is no edge from σ_1 to σ_2, then this derivative is null. If there is an edge e from σ_1 to σ_2, then in both pyramids P_+ and P_- on both sides of e, we need to study the variations of the dihedral angle

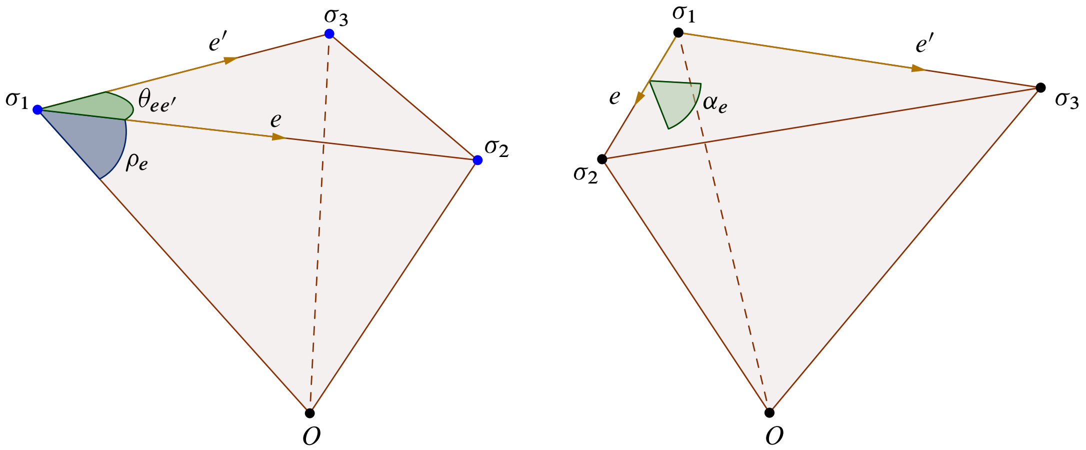


Figure 7: The simplex cell of the past of Σ in $M(h)$. The following angles are represented: ρ_e the angle from $\overrightarrow{\sigma_1 O}$ to $\overrightarrow{\sigma_1\sigma_2}$, $\theta_{ee'}$ the angle from $\overrightarrow{\sigma_1\sigma_2}$ to $\overrightarrow{\sigma_1\sigma_3}$ and α_e the angle from the plane $(0\sigma_1\sigma_1)$ to the plane $(\sigma_1\sigma_2\sigma_3)$.

on the edge $[O\sigma_1]$ with respect to h_{σ_2} and h_{σ_1}. Since the algebraic relationship between κ_σ and h_σ is complicated, a key to obtaining meaningful relations is to draw the kite associated with each embedded triangle $\mathbb{E}^{1,2}$.

Definition–Proposition 5.7 (kite, [14, pages 90–91]) *A hyperbolic kite* (*resp. Euclidean kite*) *is a quadrangle $ABCD$ in $X = \mathbb{H}^2$* (*resp. in $X = \mathbb{E}^2$*) *with two opposite right angles and possibly with self-intersections. We parametrize kites by fixing a convex quadrangle decorated as in Figure 8 and constructing it as follows*:

(1) *choose some point A in X and some direction $\vec{u} \in T_A X$,*

(2) *move ρ_1 along the oriented line $(A\vec{u})$ to reach at $B = \exp_A(\rho_1\vec{u})$,*

(3) *turn $\frac{1}{2}\pi$ (counterclockwise) to obtain the new direction $\vec{v} \in T_B X$,*

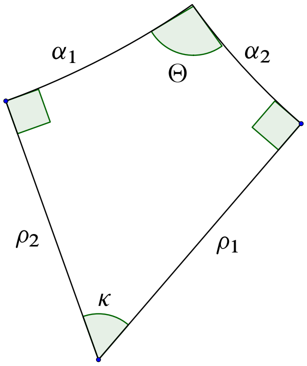


Figure 8

(4) *move α_2 on the oriented line $(B\vec{v})$ to reach $C=\exp_B(\alpha_2\vec{v})$,*

(5) *turn $\pi-\Theta$ to obtain $\vec{w}$,*

(6) *move a distance α_1 on the oriented line $(C\vec{w})$ to reach D,*

(7) *turn $\frac{1}{2}\pi$ to obtain $\vec{k}$,*

(8) *move a distance ρ_2 on the oriented line $(D\vec{k})$ to reach A',*

(9) *turn $\pi-\kappa$ to obtain $\vec{u}'$.*

For any choices of three out of the six parameters $\alpha_1, \alpha_2, \rho_1, \rho_2, \kappa$, and Θ, there exists a unique choice for the three others so that the construction above yields a hyperbolic kite, ie $X=\mathbb{H}^2$, $A'=A$, and $\vec{u}'=\vec{u}$. Furthermore, for such six parameters,

$$\cos(\kappa)=\frac{\sinh(\rho_1)\sinh(\rho_2)-\cos(\Theta)}{\cosh(\rho_1)\cosh(\rho_2)},\qquad \sinh(\rho_2)=\frac{\cos(\kappa)\sinh(\alpha_1)+\sinh(\alpha_2)}{\sin(\kappa)\cosh(\alpha_1)},$$
$$\frac{\sin(\kappa)}{\sin(\Theta)}=\frac{\cosh(\alpha_2)}{\cosh(\rho_2)}=\frac{\cosh(\alpha_1)}{\cosh(\rho_1)}.$$

Consider P_+ and use the notation of Figure 7. Then consider the quadrilateral of $\mathbb{H}^2$ given by the sequence of geodesics in the set of future unit timelike vectors identified with the hyperbolic plane $\mathbb{H}^2$:

$$(O\sigma_1)\to(O\sigma_1\sigma_2)\cap(\sigma_1\sigma_2)^{\perp}\to(\sigma_1\sigma_2)^{\perp}\cap(\sigma_1\sigma_3)^{\perp}\to(\sigma_1\sigma_3)^{\perp}\cap(O\sigma_1\sigma_3)\to(O\sigma_1).$$

To identify the parameters $(\rho_1,\alpha_2,\alpha_1,\rho_2,\kappa,\Theta)$ as in Figure 9, we note the following.

- $\rho_e:=\mathrm{Real}(\angle\overrightarrow{\sigma_1 O}\overrightarrow{\sigma_1\sigma_2}):=d_{\mathbb{H}}((O\sigma_1),(O\sigma_1\sigma_2)\cap(\sigma_1\sigma_2)^{\perp})$ since $(O\sigma_1\sigma_2)$ is the (timelike) vectorial plane containing both vectors and $\overrightarrow{\sigma_1\sigma_2}$ is spacelike, and $\overrightarrow{\sigma_1 O}$ is past timelike. So the (oriented) length of the geodesic $(O\sigma_1)\to(O\sigma_1\sigma_2)\cap(\sigma_1\sigma_2)^{\perp}$ is ρ_e, and thus $\rho_1=\rho_e$.

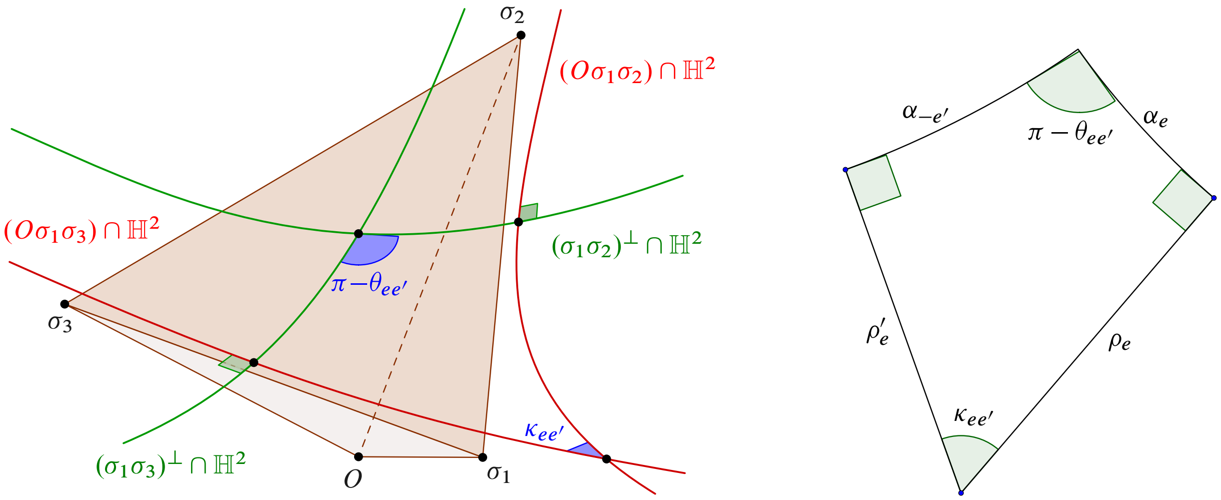

σ_2
$(O\sigma_1\sigma_2)\cap\mathbb{H}^2$
$(O\sigma_1\sigma_3)\cap\mathbb{H}^2$
$(\sigma_1\sigma_2)^{\perp}\cap\mathbb{H}^2$
$\pi-\theta_{ee'}$
σ_3
$\kappa_{ee'}$
$(\sigma_1\sigma_3)^{\perp}\cap\mathbb{H}^2$
O
σ_1
$\alpha_{-e'}$
$\pi-\theta_{ee'}$
α_e
ρ'_e
ρ_e
$\kappa_{ee'}$

Figure 9: The kite associated to an edge, with e the edge $\overrightarrow{\sigma_1\sigma_2}$ and e' the edge $\overrightarrow{\sigma_1\sigma_3}$.

Mutatis mutandis, we show the same way $-\rho_2 = -\rho_{e'}$; thus $\rho_2 = \rho_{e'}$.

- $\alpha_e := \mathrm{Real}(\angle(O\sigma_1\sigma_2)(\sigma_1\sigma_2\sigma_3)) = \mathrm{Real}(\angle n_1 n_2)$, where n_1 and n_2 are respectively the normals to the planes $(O\sigma_1\sigma_2)$ and $(\sigma_1\sigma_2\sigma_3)$ such that $(\overrightarrow{\sigma_1\sigma_2}, \overrightarrow{\sigma_1 O}, n_1)$ and $(\overrightarrow{\sigma_1\sigma_2}, \overrightarrow{\sigma_1\sigma_3}, n_2)$ are direct bases. We thus have $n_2 \in (\sigma_1\sigma_2\sigma_3)^\perp = (\sigma_1\sigma_2)^\perp \cap (\sigma_1\sigma_3)^\perp$ future timelike and $n_1 \in (O\sigma_1\sigma_2)^\perp$ spacelike, so $\angle n_1 n_2 = \angle n_3 n_2$ with $n_3 \in ((O\sigma_1\sigma_2)^\perp)^\perp \cap (n_1 n_2) = (O\sigma_1\sigma_2) \cap (\sigma_1\sigma_2)^\perp$. Therefore α_e is the (oriented) length of the geodesic
$$(O\sigma_1\sigma_2) \cap (\sigma_1\sigma_2)^\perp \to (\sigma_1\sigma_2)^\perp \cap (\sigma_1\sigma_3)^\perp.$$

We thus have shown that $\alpha_2 = \alpha_e$.

Mutatis mutandis, we show the same way that $-\alpha_{-e'} = -\alpha_1$; thus $\alpha_{e'} = \alpha_1$.

- The parameter κ is given by the dihedral angle $\angle(O\sigma_1\sigma_2)(O\sigma_1\sigma_3)$; thus $\kappa = \kappa_{ee'}$.

- Finally, to compute Θ, notice that the radial projection of the hyperbolic kite on the spacelike plane $(\sigma_1\sigma_2\sigma_3)$ yields a Euclidean kite with the same signs of oriented lengths of sides and whose κ parameter is $\theta_{ee'}$. Furthermore, the plane $(\sigma_1\sigma_2\sigma_3)$ is orthogonal to the timelike line from O to $(\sigma_1\sigma_2)^\perp \cap (\sigma_1\sigma_3)^\perp$, so the angle in $\mathbb{H}^2$ at $\mathbb{H}^2 \cap (\sigma_1\sigma_2)^\perp \cap (\sigma_1\sigma_3)^\perp$ is the same as the Euclidean angle in $(\sigma_1\sigma_2\sigma_3)$ at $(\sigma_1\sigma_2\sigma_3) \cap (\sigma_1\sigma_2)^\perp \cap (\sigma_1\sigma_3)^\perp$. We deduce that $\Theta = \pi - \theta_{ee'}$.

Corollary 5.8 *Using the same notation as in Definition–Proposition 5.7 and choosing Θ, ρ_1, and ρ_2 as parameters,*
$$\frac{\partial\kappa}{\partial\rho_1} = -\frac{\tanh(\alpha_2)}{\cosh(\rho_1)}.$$

We thus need to compute the derivative of ρ_e with respect to the heights $(h_\sigma)_{\sigma\in S}$ for each edge e.

Lemma 5.9 *Using the notation of Figure 7,*
$$\mathrm{d}\rho_e = -\frac{(h_{\sigma_1}^2 + h_{\sigma_2}^2 + l_e^2)\,\mathrm{d}h_{\sigma_1} - 2h_{\sigma_1}h_{\sigma_2}\,\mathrm{d}h_{\sigma_2}}{2l_e h_{\sigma_1}^2 \cosh(\rho_e)}.$$

Proof From the cosine law in $\mathbb{E}^{1,2}$:
$$\begin{aligned} -h_{\sigma_2}^2 &= -h_{\sigma_1}^2 + l_e^2 - 2l_e h_{\sigma_1}\sinh(\rho_e),\\ \cosh(\rho_e)\,\mathrm{d}\rho_e &= \frac{h_{\sigma_1}(-2h_{\sigma_1}) - (h_{\sigma_2}^2 + l_e^2 - h_{\sigma_1}^2)}{2l_e h_{\sigma_1}^2}\,\mathrm{d}h_{\sigma_1} + \frac{2h_{\sigma_2}h_{\sigma_1}}{2l_e h_{\sigma_1}^2}\,\mathrm{d}h_{\sigma_2},\\ \mathrm{d}\rho_e &= -\frac{(h_{\sigma_2}^2 + l_e^2 + h_{\sigma_1}^2)\,\mathrm{d}h_{\sigma_1} - 2h_{\sigma_1}h_{\sigma_2}\,\mathrm{d}h_{\sigma_2}}{2l_e h_{\sigma_1}^2\cosh(\rho_e)}. \end{aligned}$$ □

5.3 The Einstein–Hilbert functional

We give ourselves some $Z \subset S$, and define $z := |Z|$ and $s := |S|$. Define $\mathcal{P}_Z := \{\tau \in \mathbb{R}^S \mid \forall\sigma \in Z,\ \tau_\sigma = 0\}$ as well as $\mathcal{P}_Z^{1/2} := \{\bar\kappa \in \mathbb{R}_+^S \mid \forall\sigma \in Z,\ \bar\kappa_\sigma = 0\}$. Recall that we set $\kappa_\sigma(\tau) = 0$ if $\tau_\sigma = 0$.

Definition 5.10 (Einstein–Hilbert functional) Let $\bar{\kappa} \in K_Z$. For $h \in \mathcal{P}_Z^{1/2}$ and for an edge e of $\mathcal{T}_h$, we denote by l_e the length of e and by θ_e the dihedral angle of the embedding ι_h at the edge e.

The Einstein–Hilbert functional is defined as follows:

$$\mathcal{H}_{\bar{\kappa}} \colon \mathcal{P}_Z^{1/2} \to \mathbb{R}, \quad h \mapsto \sum_{\sigma \in S} h_\sigma(\kappa_\sigma - \bar{\kappa}_\sigma) + \sum_{e \in \mathrm{Edge}(\mathcal{T}_h)} l_e \theta_e.$$

Lemma 5.11 *Letting $\sigma \in S$, the map $h \mapsto \kappa_\sigma(h)$ is continuous on $\mathcal{P}^{1/2}$ and $\mathscr{C}^1$ on each cell $\mathcal{P}_{\mathcal{T}}^{1/2}$ of $\mathcal{P}^{1/2}$.*

Proof From $-h_{\sigma_2}^2 = -h_{\sigma_1}^2 + l_e^2 - 2l_e h_{\sigma_1}\sinh(\rho_e)$ — the cosine law in $\mathbb{E}^{1,2}$ — together with the first equality of Definition–Proposition 5.7, the restriction of κ_σ to $H_{\mathcal{T}}^{\sigma+} := \{h \in \mathcal{P}_{\mathcal{T}}^{1/2} \mid h(\sigma) > 0\}$ for each triangulation $\mathcal{T}$ is $\mathscr{C}^1$. Let $h \in \partial\mathcal{P}_{\mathcal{T}}^{1/2}$. Then for all cells $\mathcal{P}_{\mathcal{T}'}^{1/2}$ containing h, the triangulations $\mathcal{T}$ and $\mathcal{T}'$ are equivalent by Proposition 2.20, so $M_{h,\mathcal{T}} \simeq M_{h,\mathcal{T}'}$. Hence $\kappa_{\sigma,\mathcal{T}}(h) = \kappa_{\sigma,\mathcal{T}'}(h)$, and we deduce that κ_σ is continuous on $H^{\sigma+} := \{h \in \mathcal{P}^{1/2} \mid h(\sigma) > 0\}$. Again using the cosine law, for any edge $e = [\sigma\sigma']$ in some triangulation $\mathcal{T}$ and any $\bar{h} \in \mathcal{P}_{\mathcal{T}}^{1/2}$ such that $\bar{h}(\sigma) = 0$, we have

$$\lim_{\substack{h \to \bar{h} \\ h \in H_{\mathcal{T}}^{\sigma+}}} \rho_e = \lim_{\substack{h \to \bar{h} \\ h \in H_{\mathcal{T}}^{\sigma+}}} \sinh^{-1} \frac{l_e^2 - h_\sigma^2 + h_{\sigma'}^2}{l_e h_\sigma} = +\infty.$$

The first equality of Definition–Proposition 5.7 yields

$$\lim_{\substack{h \to \bar{\kappa} \\ h \in H_{\mathcal{T}}^{\sigma+}}} \cos\kappa_\sigma(h) = 1,$$

and since $0 \le \kappa \le \pi$ we deduce that $\lim_{h \to \bar{h}, h \in H_{\mathcal{T}}^{\sigma+}} \kappa_\sigma(h) = 0$. Hence κ_σ is continuous on $\mathcal{P}_{\mathcal{T}}^{1/2}$. From Corollary 5.8 and $\lim_{h \to \bar{h}, h \in H_{\mathcal{T}}^{\sigma+}} \rho_e = +\infty$, we also obtain that

$$\lim_{\substack{h \to \bar{h} \\ h \in H_{\mathcal{T}}^{\sigma+}}} \frac{\partial\kappa_\sigma}{\partial h_{\sigma''}}(h) = 0$$

for any $\sigma'' \in S$, and obviously $(\partial\kappa/\partial h_{\sigma''})(h) = 0$ if $h(\sigma) = 0$ and $\sigma'' \neq \sigma$. We deduce that κ is continuously differentiable on $\mathcal{P}_{\mathcal{T}}^{1/2}$.

Finally, since $\mathcal{P}^{1/2}$ is the union of finitely many such cells $\mathcal{P}_{\mathcal{T}}^{1/2}$, we get continuity on $\mathcal{P}^{1/2}$. □

Lemma 5.12 *Let $\mathcal{T}$ be a triangulation associated with a cell $\mathcal{P}_{\mathcal{T}}^{1/2}$ of $\mathcal{P}^{1/2}$. For any edge e of $\mathcal{T}$, the map $h \mapsto \theta_e(h)$ is $\mathscr{C}^1$ on $\mathcal{P}_{\mathcal{T}}^{1/2}$.*

Proof By construction of the polyhedral embedding, it suffices to show the "half" dihedral angle α_e is $\mathscr{C}^1$, effectively reducing the problem to the embedding of a fixed triangle $T = [\sigma_1\sigma_2\sigma_3] \in \mathcal{T}$. Note that although the edge $[O\sigma_i]$ may become lightlike when $h_{\sigma_i} \to 0$, the planes $(O\sigma_i\sigma_j)$ (resp. $(\sigma_1\sigma_2\sigma_3)$) are never degenerated and stay timelike (resp. spacelike) for $i, j \in \{1,2,3\}$. Therefore the angle α_e is well defined and depends in a $\mathscr{C}^1$ manner in the coordinates of the embeddings of the σ_i for $i \in \{1,2,3\}$.

Using notation of Lemma 2.3, the center ω is the orthogonal projection of O on $(\sigma_1\sigma_2\sigma_3)$. We may choose the embedding of T in such a way that $(\sigma_1\sigma_2\sigma_3)$ is the plane $\{t = \sqrt{\tau_0}\}$, ie $\omega = (t,0,0)$. Recall that $\tau_0(h)$ is positive and depends polynomially on h. We may in addition fix the embedding ι so that $\iota(\sigma_1) = (h, x, 0)$ with $x > 0$. Then elementary trigonometry in the spacelike plane $(\sigma_1\sigma_2\sigma_3)$ yields that the coordinates of $\iota(\sigma_i)$ are $\mathscr{C}^1$ functions in h. □

Proposition 5.13 *Let $\bar\kappa \in \mathbb{R}_+^S$. The functional $\mathcal{H}_{\bar\kappa}$ is well defined, $\mathscr{C}^1$ on $\mathcal{P}_Z^{1/2}$, and*

$$\mathrm{d}\mathcal{H}_{\bar\kappa} = \sum_{\sigma\in S\setminus Z} (\kappa_\sigma - \bar\kappa_\sigma)\,\mathrm{d}h_\sigma.$$

Proof We prove the proposition for $Z = \varnothing$; the other cases are corollaries.

Consider the family of compact locally Minkowski polyhedra $(Q_h)_{h\in\mathcal{P}^{1/2}}$ given by the past of the polyhedral Cauchy surface $\iota_h(\Sigma) \subset M(h)$. For any triangulation $\mathcal{T}$ defining a cell $\mathcal{P}_\mathcal{T}$ of $\mathcal{P}$, the underlying simplicial complex $\mathcal{K}_h$ of Q_h is constant on $\mathcal{P}_\mathcal{T}$. The edges of $\mathcal{T}$ are always spacelike, κ is well defined and continuous on $\mathcal{P}_\mathcal{T}^{1/2}$, and $\mathcal{K}_h$ is a continuous family of polyhedra. All the angles in the definition of $\mathcal{H}_{\bar\kappa}$ are $\mathscr{C}^1$ on each cell $\mathcal{P}_\mathcal{T}^{1/2}$ by Lemmas 5.11 and 5.12. In addition, Lemma 5.11 gives continuity of $h \mapsto \sum_{\sigma\in S} h_\sigma(\kappa_\sigma - \bar\kappa_\sigma)$ on the whole $\mathcal{P}^{1/2}$. Continuity of $\sum_{e\in\mathrm{Edge}(\mathcal{T}_h)} l_e\theta_e$ follows from the remark that at some h on the interface of adjacent cells $\mathcal{P}_\mathcal{T}^{1/2}$ and $\mathcal{P}_{\mathcal{T}'}^{1/2}$, one obtains $\mathcal{T}'$ from $\mathcal{T}$ by flipping h-critical edges. On such edges e one has $\theta_e = 0$, so the sums $\sum_{e\in\mathrm{Edge}(\mathcal{T}_h)} l_e\theta_e$ and $\sum_{e\in\mathrm{Edge}(\mathcal{T}_h')} l_e\theta_e$ only differ by null terms. We conclude that $\mathcal{H}_{\bar\kappa}$ is continuous on $\mathcal{P}^{1/2}$ and its restriction to any cell is $\mathscr{C}^1$.

Schläffli's formula thus applies to the interior of any cell $\mathcal{P}_\mathcal{T}^{1/2}$ where $h > 0$ and gives

$$\sum_{\sigma\in S} h_\sigma\,\mathrm{d}\kappa_\sigma + \sum_{e\in\mathcal{A}_h} l_e\,\mathrm{d}\theta_e = 0.$$

Hence

$$\mathrm{d}\mathcal{H}_{\bar\kappa} = \sum_{\sigma\in S}(\kappa_\sigma - \bar\kappa_\sigma)\,\mathrm{d}h_\sigma + \sum_{\sigma\in S} h_\sigma\,\mathrm{d}\kappa_\sigma + \sum_{e\in\mathcal{A}_h} l_e\,\mathrm{d}\theta_e = \sum_{\sigma\in S}(\kappa_\sigma - \bar\kappa_\sigma)\,\mathrm{d}h_\sigma.$$

We have thus proved the result for the restriction to the interior of any cell $\mathcal{P}_\mathcal{T}^{1/2}$, and hence on a dense subset of $\mathcal{P}^{1/2}$. Finally, by continuity of $h \mapsto \sum_{\sigma\in S}(\kappa_\sigma - \bar\kappa_\sigma)\,\mathrm{d}h_\sigma$ and $d\mathcal{H}_{\bar\kappa}$ on $\mathcal{P}^{1/2}$, the result follows. □

We now study the Hessian of the Einstein–Hilbert functional on the interior of the domain of admissible times $\mathcal{P}^{1/2}$.

Lemma 5.14 *The map* κ *is* $\mathscr{C}^1$ *on* $\mathcal{P}_Z^{1/2}$*, and for all* h *in* $\mathcal{P}_Z^{1/2}$ *and all* $\sigma \in S \setminus Z$*, we have*

$$\mathrm{d}_h\kappa_\sigma = \sum_{\substack{e\in\mathcal{E}_h,e\colon\sigma\rightsquigarrow\sigma'\\ \sigma'\in S\setminus Z}} (\tanh(\alpha_e)+\tanh(\alpha_{-e}))\frac{(h_{\sigma'}^2+l_e^2+h_\sigma^2)\,\mathrm{d}h_\sigma - 2h_\sigma h_{\sigma'}\,\mathrm{d}h_{\sigma'}}{2l_e c_e^2},$$

where $\mathcal{E}_h$ *is the set of edges of any* h*-Delaunay triangulation and where*

$$c_e = \begin{cases}\cosh(\rho_e)h_\sigma & \text{if } h_\sigma \neq 0,\\ (l_e^2+h_{\sigma'}^2)/l_e & \text{if } h_\sigma = 0.\end{cases}$$

Proof By Lemma 5.11, the restriction of κ is $\mathscr{C}^1$ on each cell $\mathcal{P}_{Z,\mathcal{T}}^{1/2}$. To prove κ is $\mathscr{C}^1$ on the whole $\mathcal{P}_Z^{1/2}$, it suffices to show that the equality holds on the relative interiors of cells and that the right-hand side is well defined and continuous on the whole $\mathcal{P}_Z^{1/2}$.

As argued in the proof of Lemma 5.12, for any edge $e\colon\sigma\rightsquigarrow\sigma'$, the angles α_e and α_{-e} are well defined and continuous even at h with null coordinates. In addition, by the cosine law, when $h\to\bar h$ for some $\bar h$ such that $\bar h_\sigma = 0$, we have $\rho_e \to +\infty$ and

$$\cosh(\rho_e)h_\sigma \sim \sinh(\rho_e)h_\sigma \sim \frac{l_e^2+h_{\sigma'}^2}{l_e}.$$

The right-hand side is then well defined and continuous when restricted to a given cell $\mathcal{P}_{Z,\mathcal{T}}^{1/2}$.

As before, critical edges e in the sum yield zero terms as $0=\theta_e=\alpha_e+\alpha_{-e}$, ie $\alpha_e=-\alpha_{-e}$, so that $\tanh\alpha_e = -\tanh\alpha_{-e}$. We conclude that the right-hand side does not depend on the h-Delaunay triangulation and is thus well defined and continuous on the whole $\mathcal{P}_Z^{1/2}$.

For h in the relative interior of a cell $\mathcal{P}_{Z,\mathcal{T}}^{1/2}$ associated to a triangulation $\mathcal{T}$ and for $\sigma\in S\setminus Z$, denote by $(e_i)_{i\in\mathbb{Z}/n\mathbb{Z}}$ the family of outgoing edges from σ enumerated coherently with the orientation of Σ. Define $\sigma_i\in S$ the other end of e_i so that

$$\begin{aligned}\mathrm{d}_h\kappa_\sigma &= \sum_{i\in\mathbb{Z}/n\mathbb{Z}} \mathrm{d}_h\kappa_{e_ie_{i+1}} = \sum_{i\in\mathbb{Z}/n\mathbb{Z}}\left(-\frac{\tanh(\alpha_{e_i})}{\cosh(\rho_{e_i})}\,\mathrm{d}\rho_{e_i} - \frac{\tanh(\alpha_{-e_{i+1}})}{\cosh(\rho_{e_{i+1}})}\,\mathrm{d}\rho_{e_{i+1}}\right)\\ &= -\sum_{i\in\mathbb{Z}/n\mathbb{Z}}\left(\frac{\tanh(\alpha_{e_i})}{\cosh(\rho_{e_i})}+\frac{\tanh(\alpha_{-e_i})}{\cosh(\rho_{e_i})}\right)\mathrm{d}\rho_{e_i} = -\sum_{i\in\mathbb{Z}/n\mathbb{Z}}\frac{\tanh(\alpha_{e_i})+\tanh(\alpha_{-e_i})}{\cosh(\rho_{e_i})}\,\mathrm{d}\rho_{e_i}\\ &= \sum_{i\in\mathbb{Z}/n\mathbb{Z}}\frac{\tanh(\alpha_{e_i})+\tanh(\alpha_{-e_i})}{\cosh(\rho_{e_i})}\frac{(h_\sigma^2+h_{\sigma_i}^2+l_e^2)\,\mathrm{d}h_\sigma - 2h_\sigma h_{\sigma_i}\,\mathrm{d}h_{\sigma_i}}{2l_eh_\sigma^2\cosh(\rho_e)}\\ &= \sum_{i\in\mathbb{Z}/n\mathbb{Z}}\frac{\tanh(\alpha_{e_i})+\tanh(\alpha_{-e_i})}{\cosh^2(\rho_{e_i})}\frac{(h_\sigma^2+h_{\sigma_i}^2+l_e^2)\,\mathrm{d}h_\sigma - 2h_\sigma h_{\sigma_i}\,\mathrm{d}h_{\sigma_i}}{2l_eh_\sigma^2}.\end{aligned}$$ □

Proposition 5.15 *For* $\bar\kappa\in\mathbb{R}_+^S$*, the functional* $\mathcal{H}_{\bar\kappa}$ *is convex on* $\mathcal{P}_Z^{1/2}$ *and strictly convex on the relative interior of* $\mathcal{P}_Z^{1/2}$*.*

Proof From Proposition 5.13 and Lemma 5.14, $\mathcal{H}_{\bar\kappa}$ is $\mathscr{C}^2$ on $\mathcal{P}_Z^{1/2}$ and its Hessian matrix H has the following coefficients for $h \in \mathcal{P}_Z^{1/2}$, an h-Delaunay triangulation being chosen, for all $\sigma, \sigma' \in S \setminus Z$ with $\sigma \neq \sigma'$:

$$H_{\sigma,\sigma'} = -\sum_{e\colon \sigma \rightsquigarrow \sigma'} (\tanh(\alpha_e) + \tanh(\alpha_{-e})) \frac{2h_\sigma h_{\sigma'}}{2l_e c_e^2} \leq 0$$

$$H_{\sigma,\sigma} = \sum_{\sigma' \in S} \sum_{e\colon \sigma \rightsquigarrow \sigma'} (\tanh(\alpha_e) + \tanh(\alpha_{-e})) \frac{h_\sigma^2 + h_{\sigma'}^2 + l_e^2}{2l_e c_e^2} - \sum_{e\colon \sigma \rightsquigarrow \sigma} (\tanh(\alpha_e) + \tanh(\alpha_{-e})) \frac{2h_\sigma h_\sigma}{2l_e c_e^2}.$$

Since the embedding of Σ into $M(h)$ is convex, $\tanh(\alpha_e) + \tanh(\alpha_{-e}) \geq 0$ with equality if and only if the edge is h-critical. Therefore, for all $\sigma \in S$,

$$\begin{aligned} H_{\sigma,\sigma} + \sum_{\sigma' \neq \sigma} H_{\sigma,\sigma'} &= \sum_{\sigma' \in S} \sum_{e\colon \sigma \rightsquigarrow \sigma'} (\tanh(\alpha_e) + \tanh(\alpha_{-e})) \frac{h_{\sigma'}^2 + l_e^2 + h_\sigma^2 - 2h_\sigma h_{\sigma'}}{2l_e c_e^2} \\ &= \sum_{\sigma' \in S} \sum_{e\colon \sigma \rightsquigarrow \sigma'} (\tanh(\alpha_e) + \tanh(\alpha_{-e})) \frac{(h_{\sigma'} - h_\sigma)^2 + l_e^2}{2l_e c_e^2} \geq 0. \end{aligned}$$

The Hessian matrix of $\mathcal{H}_{\bar\kappa}$ is thus diagonally dominant on $\mathcal{P}_Z^{1/2}$.

Consider some h in $\mathcal{P}_Z^{1/2}$ and $\sigma \in S \setminus Z$ such that $H_{\sigma,\sigma} - \sum_{\sigma' \neq \sigma} |H_{\sigma,\sigma'}| = 0$. Then all outgoing edges from σ are h-critical. We build a hinge as follows.

(1) Take any h-Delaunay triangulation of Σ and enumerate counterclockwise the p vertices $(\sigma_k)_{k \in \mathbb{Z}/p\mathbb{Z}}$ of the neighborhood of σ.

(2) Consider the hinge $Q = ([\sigma\sigma_{-2}\sigma_{-1}\sigma_0], [\sigma\sigma_{-1}])$.

(3) If Q is unflippable return Q.

(4) Otherwise, flip Q; the neighborhood vertices of σ are now $(\sigma_k)_{k \in \mathbb{Z}/(p-1)\mathbb{Z}}$. Then return to step (2).

Since at each step, the number of neighbors of σ decreases, the algorithm eventually stops after finitely many iterations and thus returns an unflippable immersed hinge in the neighborhood of σ. Such a hinge is h-critical and unflippable; hence h^2 is in a boundary facet of $\mathcal{P}$ not of the type $h_{\sigma'} = 0$. We conclude that h is not in the relative interior of $\mathcal{P}_Z^{1/2}$. Finally, the Hessian matrix H is strictly diagonally dominant on the relative interior of $\mathcal{P}_Z^{1/2}$. □

5.4 Proof of the main theorem

Theorem 5 *Let Σ be a closed locally Euclidean surface of genus g with s marked conical singularities of angles $(\theta_i)_{i \in [\![1,s]\!]}$. For all*

$$\bar\kappa \in \left(\prod_{i=1}^{s} [0, \min(\theta_i, 2\pi)] \right) \setminus \{(\theta_1, \dots, \theta_s)\},$$

there is a radiant singular flat spacetime M homeomorphic to $\Sigma \times \mathbb{R}$ with exactly s marked lines $\Delta_1, \dots, \Delta_s$ of respective cone angles $\bar\kappa_1, \dots, \bar\kappa_s$ and a convex polyhedral embedding $\iota\colon (\Sigma, S) \to (M, (\Delta_i)_{i \in [\![1,s]\!]})$.

Furthermore, if for all $i \in [\![1,s]\!], \bar{\kappa}_i < \theta_i$, then such a couple (M,ι) is unique up to equivalence.

Finally, if for some $i \in [\![1,s]\!], \theta_i \leq \pi$, there is no such convex polyhedral embedding such that $\kappa_i > \theta_i$.

Denoting by $\kappa(x)$ the cone angle at x if x is a point in an $\mathbb{H}^2_{\geq 0}$-manifold, in view of Theorem 6 the main case of the theorem can also be stated as follows:

Corollary 5.16 *Let Σ be a closed locally Euclidean surface of genus g with s marked conical singularities of angles $(\theta_\sigma)_{\sigma\in S}$. For all $\bar{\kappa} \in \prod_{\sigma\in S}[0,2\pi]\cap[0,\theta_\sigma[$, there exists a closed $\mathbb{H}^2_{\geq 0}$-manifold $\Sigma_{\bar{\kappa}}$ together with a homeomorphism $h\colon \Sigma \to \Sigma_{\bar{\kappa}}$ and a convex polyhedral embedding $\iota\colon (\Sigma, S) \to \mathrm{susp}(\Sigma_{\bar{\kappa}})s$ such that*

- *for all $\sigma \in S, \bar{\kappa}_\sigma = \kappa(h(\sigma))$,*
- *with $\mathrm{susp}(\Sigma_{\bar{\kappa}}) \xrightarrow{\pi} \Sigma_{\bar{\kappa}}$ the natural projection, we have $\pi \circ \iota = h$.*

Furthermore, such a triple $(\Sigma_{\bar{\kappa}}, h, \iota)$ is unique up to equivalence.

Remark Equivalence between triples $(\Sigma^{(i)}_{\bar{\kappa}}, h^{(i)}, \iota^{(i)})$ for $i \in \{1,2\}$ is understood as an isomorphism $\varphi\colon \Sigma^{(1)}_{\bar{\kappa}} \to \Sigma^{(2)}_{\bar{\kappa}}$ such that $\iota^{(2)} = \hat{\varphi} \circ \iota^{(1)}$ with $\hat{\varphi}\colon \mathrm{susp}(\Sigma^{(1)}_{\bar{\kappa}}) \xrightarrow{\sim} \mathrm{sup}(\Sigma^{(2)}_{\bar{\kappa}})$ the isomorphism induced by φ.

Let us prove a last lemma:

Lemma 5.17 *With $\theta = (\theta_\sigma)_{\sigma\in S}$ the cone angles of Σ, we have*

$$\lim_{\substack{\tau\in\mathcal{P}\\ \tau\to+\infty}} \kappa(\tau) = \theta.$$

Proof We use the same notation as in the preceding section. In a given cell $\mathcal{P}_\mathcal{T}$ of $\mathcal{P}$, for each vertex $\sigma \in S$ and for all edges e of $\mathcal{T}$ outgoing from σ to some σ_2, by the cosine law

$$-\tau_{\sigma_2} = -\tau_\sigma + l_e^2 - 2l_e\sqrt{\tau_\sigma}\sinh(\rho_e).$$

Since $|\tau_{\sigma_1} - \tau_{\sigma_2}|$ is uniformly bounded on $\mathcal{P}$ and l_e is constant, $\rho_e \xrightarrow{\tau\to+\infty} 0$. Then from Definition–Proposition 5.7, with e' the subsequent edge around σ, we have $\kappa_{ee'} \xrightarrow{\tau\to+\infty} \theta_{ee'}$. Hence,

$$\kappa_\sigma(\tau) \xrightarrow{\tau\in\mathcal{P}_\mathcal{T},\tau\to+\infty} \theta_\sigma.$$

Finally, there are only finitely many cells $\mathcal{P}_\mathcal{T}$, and S is finite. □

Proof of Theorem 5 Let $Z \subset S$. We prove the theorem for $\bar{\kappa}$ such that $\{\sigma \in S \mid \bar{\kappa}_\sigma = 0\} = Z$. It suffices to show that for such $\bar{\kappa}$ the Einstein–Hilbert functional $\mathcal{H}_{\bar{\kappa}}$ has exactly one critical point in $\mathcal{P}_Z^{1/2}$. Define $K_Z := \bigl\{\bar{\kappa} \in \prod_{\sigma\in S}[0,2\pi]\cap[0,\theta_\sigma[\mid \forall\sigma\in Z, \bar{\kappa}_\sigma = 0\bigr\}$. We need to prove the existence and uniqueness of critical points of $\mathcal{H}_{\bar{\kappa}}$ for any $\bar{\kappa} \in K_Z$.

If $z = s$ then $K_Z = \{0\}$ and $\mathcal{P}_Z = \{0\}$ by Theorem 2(c), and there is nothing else to prove. Otherwise, we proceed as follows.

By Proposition 5.15 the functional $\mathcal{H}_{\bar\kappa}$ is strictly convex in the relative interior of $\mathcal{P}_Z^{1/2}$; thus the critical points are of index 1 when considered as a function on the relative interior of $\mathcal{P}_Z$.

Let $\tau \in \partial\mathcal{P}_Z$, the relative boundary of $\mathcal{P}_Z$, and let $\bar\kappa \in K_Z$. By Theorem 2(e), on $\partial\mathcal{P}_Z$ there exists $\sigma \in S \setminus Z$ such that either $\tau_\sigma = 0$ or τ is in the kernel of the affine form of an unflippable immersed hinge. In the former situation, $0 = \kappa_\sigma < \bar\kappa_\sigma$. In the latter situation, consider such a hinge (Q, η) with $Q = ([ABCD], [AC])$.

- If (Q, η) is embedded, then Q is unflippable. Without loss of generality, we may assume $C \in [ABD]$, the cone around $\sigma = \eta(C)$ is then convex and contains a coplanar wedge of Euclidean angle at least π; in particular $\theta_\sigma \geq \pi$. By the Lorentzian Volkov's lemma (Theorem 3),
  - if $\theta_\sigma > 2\pi$ we have $\kappa_\sigma > 2\pi \geq \bar\kappa_\sigma$,
  - if $\pi \leq \theta_\sigma \leq 2\pi$ we have $\kappa_\sigma \geq \theta_\sigma > \bar\kappa_\sigma$.
- If η is not an embedding, then without loss of generality we may assume $\eta(A) = \eta(B) = \eta(D)$; being h-critical, all edges have null dihedral angles so that the stalk of the cone around $\sigma := \eta(C)$ is trigonometric (without breaking point). In particular, $\theta_\sigma = \kappa_\sigma > \bar\kappa$.

Either way, $\kappa_\sigma > \bar\kappa_\sigma$. Together with Proposition 5.13 this implies that $\mathcal{H}_{\bar\kappa}$ has no critical points on $\partial\mathcal{P}_Z^{1/2}$.

If $z = s - 1$, then κ is a function defined on an interval, and is continuous and increasing from 0 to some $\kappa_{\max} > \bar\kappa$. The result follows.

We now assume $z \leq s-2$. Define $\overline{\mathcal{P}}_Z^{1/2} := \mathcal{P}_Z^{1/2}$ if $Z \neq \varnothing$ and $\overline{\mathcal{P}}_Z^{1/2} := \mathcal{P}_Z^{1/2} \cup \{\infty\}$ if $Z = \varnothing$. This way $\overline{\mathcal{P}}_Z^{1/2}$ is homeomorphic to an $s - z$ dimensional closed ball and its boundary $\partial\overline{\mathcal{P}}_Z^{1/2}$ is homeomorphic to an $(s-z-1)$-dimensional sphere. The homeomorphism may be made explicit by the radial map from some $\tau_0 \in \mathrm{Int}(\mathcal{P}_Z^{1/2})$, the relative interior of $\mathcal{P}_Z^{1/2}$. Consider the family of vector fields indexed on K_Z,

$$X_{\bar\kappa} : \overline{\mathcal{P}}_Z^{1/2} \to \mathbb{R}^{s-z}, \quad h \neq \infty \mapsto (\kappa_\sigma(h) - \bar\kappa_\sigma)_{\sigma\in S\setminus Z}, \quad \infty \mapsto (\theta_\sigma - \bar\kappa_\sigma)_{\sigma\in S\setminus Z},$$

and notice that $X|_{\bar\kappa\,\mathrm{Int}(\mathcal{P}_Z^{1/2})}$ is the gradient of $\mathcal{H}|_{\bar\kappa\,\mathrm{Int}(\mathcal{P}_Z^{1/2})}$ for $\bar\kappa \in K_Z$ by Proposition 5.13. By Lemma 5.17, X is continuous at ∞ if $Z = \varnothing$; thus $\bar\kappa, h \mapsto X_{\bar\kappa}(h)$ are continuous on $K_Z \times \overline{\mathcal{P}}_Z^{1/2}$ and, from the discussion above, nonsingular on the boundary of $\overline{\mathcal{P}}_Z^{1/2}$. By Proposition 5.15 and the Poincaré-Hopf theorem [8, Theorem 12.13], the number of singular points of the vector field $X_{\bar\kappa}$ in the interior of $\mathcal{P}_Z^{1/2}$ is equal to the index of $X_{\bar\kappa}/\|X_{\bar\kappa}\|$ on $\partial\overline{\mathcal{P}}_Z^{1/2}$. Since $\bar\kappa \mapsto X(\bar\kappa, \cdot)$ is continuous and K_Z is connected, the index of $X_{\bar\kappa}/\|X_{\bar\kappa}\|$ is independent from $\bar\kappa$.

Finally, take some $\bar\kappa \in K_Z$ and $\bar h$ in the interior of $\mathcal{P}_Z^{1/2}$ close enough to 0 that $\prod_{\sigma\in S\setminus Z}[0, 2\bar h_\sigma] \subset \mathcal{P}_Z^{1/2}$ and consider the vector field $Y : h \to (h-\bar h)/\|h-\bar h\|$ on $\partial\mathcal{P}_Z^{1/2}$, which can be continuously extended to the whole $\partial\overline{\mathcal{P}}_Z^{1/2}$ since $\lim_{h\to+\infty} Y(h) = \mathbf{1}_S$. On the one hand, for h on an "$h_\sigma = 0$" boundary component, $Y(h)_\sigma < 0$ while $\kappa_\sigma(h) = 0$; on the other hand, for h on a "$Q^*(h) = 0$" boundary component, there is a $\sigma \in S \setminus Z$ such that $\kappa_\sigma - \bar\kappa_\sigma > 0$, and on such a component, for all $\sigma' \in S \setminus Z$, $(h - \bar h)_{\sigma'} > 0$. At infinity, both X and Y have positive coordinates. In any case for all $h \in \partial\overline{\mathcal{P}}_Z^{1/2}$, $Y \neq -X_{\bar\kappa}/\|X_{\bar\kappa}\|$; thus $X_{\bar\kappa}/\|X_{\bar\kappa}\|$ is homotopic to Y among nonsingular vector fields on $\partial\overline{\mathcal{P}}_Z^{1/2}$. The latter has index 1; thus so does the former. Finally, for all $\bar\kappa \in K_Z$, $\mathcal{H}_{\bar\kappa}$ has exactly one critical point on $\mathcal{P}_Z^{1/2}$. Existence and uniqueness follow for $\bar\kappa \in K_Z$.

By continuity of κ and compactness of $\overline{\mathcal{P}}_Z^{1/2}$, any $\bar\kappa \in \prod_{\sigma\in S}[0,2\pi] \cap [0,\theta_\sigma]$ is in the image of $h \mapsto (\kappa_\sigma(h))_{\sigma\in S}$, except possibly $(\bar\kappa_\sigma)_{\sigma\in S} = (\theta_\sigma)_{\sigma\in S}$, which is the limit at ∞.

Finally, the last point follows from the case $\Theta \le \pi$ of Theorem 3. □

Appendix Radiant $2+1$ singular spacetimes

Before providing a more thorough description of our singularities, allow us to stress that there is a subtle point one needs to be aware of. We construct 3-manifolds with a geometric structure locally modeled on the Minkowski space $\mathbb{E}^{1,2}$ except on a discrete family of lines we deem reasonable to call "singular". The geometric $\mathbb{E}^{1,2}$-structure (in a sense described below) on the complement of the singular lines is easily defined, but our manifolds are not naturally metric spaces; they are spacetimes and come with a natural local order relation: the causal order. As a consequence, characterizing the isomorphism classes of the singular lines requires some care in general, especially for lightlike lines. We refer to [4] for the zoology of Lorentzian singular lines obtained via finite polyhedra gluings in dimension $2+1$, which should convince the reader that one should be slightly careful.

The causal structure is a tool to characterize lightlike singularities; furthermore, the boundary of the polyhedron we will construct has a special role with respect to this structure: it is a *Cauchy surface*, as defined below.

In this section, we discuss the isomorphisms classes of singularities in our manifolds: their local description as well as constructions with the addition of some more general background.

A.1 Singular (G,X)-manifolds

Let (G,X) be an analytical structure, ie a group G acting on a locally connected Hausdorff space X by homeomorphisms so that any element $g \in G$ is completely determined by its action on a nontrivial open subset. Following [11], we define a singular (G,X)-manifold as a Hausdorff second countable topological M space endowed with a (G,X)-structure on an open and dense subset $\mathcal{U}$ locally connected in M. There exists a unique maximal extension of this (G,X)-structure to a maximal open and dense subset $\mathrm{Reg}(M)$ locally connected in M called the regular locus of M. An a.e. (G,X)-morphism is a continuous map sending regular locus to regular locus and which is a (G,X)-morphism on the regular locus.

A singular (G,X)-manifold is locally modeled on a family $(X_\alpha)_{\alpha\in A}$ if for all $\alpha \in A$, X_α is a singular (G,X)-manifold and for all $x \in M$, there exists a neighborhood $\mathcal{U}$ of x and an open $\mathcal{V}$ of some X_α such that $\mathcal{U}$ is isomorphic to $\mathcal{V}$.

In our situation, the singular locus is a union of 1-dimensional submanifolds of a 3-manifold. The hypotheses of [11] are then satisfied, and the isomorphism class of a singular point is thus well defined.

A.2 Local models of singular lines

We now introduce the local models of the singular $\mathcal{F}$-manifolds we will consider.

Definition A.1 (massive particles model space) Let $\alpha \in \mathbb{R}_+^*$. The manifold $\mathbb{E}_\alpha^{1,2}$ is $\mathbb{R}^3$ endowed with the flat Lorentzian metric

$$\mathrm{d}s_\alpha^2 = -\mathrm{d}t^2 + \mathrm{d}r^2 + \left(\frac{\alpha}{2\pi} r\right)^2 \mathrm{d}\theta^2$$

on $\operatorname{Reg}(\mathbb{E}_\alpha^{1,2}) := \{r > 0\}$, the complement of the line $\operatorname{Sing}(\mathbb{E}_\alpha^{1,2}) := \{r = 0\}$, where (t, r, θ) are cylindrical coordinates of $\mathbb{R}^3$.

For $\alpha > 0$, the metric on $\mathbb{E}_\alpha^{1,2}$ induces a unique $(\operatorname{Isom}_0(\mathbb{E}^{1,2}), \mathbb{E}^{1,2})$-structure on $\operatorname{Reg}(\mathbb{E}_\alpha^{1,2})$ such that the curves $t \mapsto c(t) = (t, r_0, \theta_0)$ are future causal for $r_0 > 0$ and all $\theta_0 \in \mathbb{R}/2\pi\mathbb{Z}$.

Definition A.2 (BTZ line model space) The manifold $\mathbb{E}_0^{1,2}$ is $\mathbb{R}^3$ endowed with the flat Lorentzian metric

$$\mathrm{d}s_0^2 = -2\,\mathrm{d}\tau\,\mathrm{d}\mathfrak{r} + \mathrm{d}\mathfrak{r}^2 + \mathfrak{r}^2\,\mathrm{d}\theta^2$$

on $\operatorname{Reg}(\mathbb{E}_0^{1,2}) := \{\mathfrak{r} > 0\}$, the complement of the line $\operatorname{Sing}(\mathbb{E}_0^{1,2}) := \{\mathfrak{r} = 0\}$, where $(\tau, \mathfrak{r}, \theta)$ are cylindrical coordinates of $\mathbb{R}^3$.

The metric on $\mathbb{E}_0^{1,2}$ induces a unique $(\operatorname{Isom}_0(\mathbb{E}^{1,2}), \mathbb{E}^{1,2})$-structure on $\operatorname{Reg}(\mathbb{E}_0^{1,2})$ such that the curves $\tau \mapsto c(\tau) = (\tau, \mathfrak{r}_0, \theta_0)$ are future causal for $\mathfrak{r}_0 > 0$ and all $\theta_0 \in \mathbb{R}/2\pi\mathbb{Z}$. The model spaces $\mathbb{E}_{\geq 0}^{1,2}$ are singular $\mathbb{E}^{1,2}$-manifolds but not singular $\mathcal{F}$-manifolds. We thus introduce the following:

Definition A.3 For $\alpha \geq 0$ define $\mathcal{F}_\alpha := \operatorname{Int}(J^+(O))$ with $O = (0,0,0) \in \mathbb{E}_\alpha^{1,2}$.

By [10, Proposition 1.3], if $\varphi\colon \mathcal{U}_\alpha \to \mathcal{U}_\beta$ is an a.e. $\mathrm{SO}_0(1,2)$-isomorphism between neighborhoods of singular points in $\mathcal{F}_\alpha$ and $\mathcal{F}_\beta$, then $\alpha = \beta$ and φ is induced by an element of $\mathrm{SO}_0(1,2)$. The local models are thus nonisomorphic as singular $\mathcal{F}$-manifolds. Note that the singular line of a massive particle is timelike while the singular line of $\mathbb{E}_0^{1,2}$ is lightlike.

A.3 Causal structure

An $\mathcal{F}$-manifold M comes with a causal structure, eg a family $(\leq_{\mathcal{U}}, \ll_{\mathcal{U}})_{\mathcal{U}}$ of transitive relations, each defined on an open subset $\mathcal{U}$ of M which is inherited from the causal and chronological relation of $\mathcal{F}$. The causal structure on $\operatorname{Reg}(\mathcal{F}_\alpha)$ can be extended to $\mathcal{F}_\alpha$ so that any $\mathcal{F}_{\geq 0}$-manifold M comes with a causal structure. A future causal curve is then a curve in M, which is locally increasing for $\leq$. The causal past/future of a point p can then be defined accordingly, and we denote them by $J^-(p)$ and $J^+(p)$, respectively.

Note that $\leq_{\mathcal{U}}$ is an order relation for $\mathcal{U}$ small enough, but this is not necessarily the case for $\leq_M$. We say that an $\mathcal{F}_{\geq 0}$-manifold M is *causal* if $\leq_M$ is an order relation; we say furthermore that M is *globally hyperbolic* if it is causal and for any $p, q \in M$, $J^+(p) \cap J^-(q)$ is compact. A *Cauchy surface* of M is a topological 2-dimensional submanifold Σ in M which intersects every future causal curve exactly once. One can prove a version of the Geroch theorem valid for $\mathcal{F}_{\geq 0}$-manifolds [5] which states that an $\mathcal{F}_{\geq 0}$-manifold M admits a Cauchy surface if and only if it is globally hyperbolic. An $\mathcal{F}_{\geq 0}$-manifold is *Cauchy-compact* if it admits a compact Cauchy surface.

A morphism $M_1 \to M_2$ between globally hyperbolic $\mathcal{F}_{\geq 0}$-manifolds is a Cauchy-embedding if it is injective and sends a Cauchy surface of M_1 to a Cauchy surface of M_2; the latter is then called a Cauchy-extension of M_1. A manifold M_1 is *Cauchy-maximal* if, for any Cauchy-embedding $M_1 \xrightarrow{\varphi} M_2$, the map φ is an isomorphism. One can prove [9; 10] a version of the Choquet–Bruhat–Geroch theorem for $\mathcal{F}_{\geq 0}$-manifolds following the lines of [30], which states that any $\mathcal{F}_{\geq 0}$-manifold admits a unique Cauchy-maximal Cauchy-extension.

A.4 Rays, suspensions, and the structure theorem

Letting M be an $\mathcal{F}_{\geq 0}$-manifold, $\mathrm{Reg}(M)$ admits a natural causal geodesic foliation, the leaves of which we call *rays*. We notice that in the model spaces, $\mathcal{F}_\alpha$ the foliation can be extended to the whole $\mathcal{F}_\alpha$; furthermore, the extended foliation to the whole $\mathcal{F}_\alpha$ induces a causal foliation on M.

Definition A.4 For $\alpha \in \mathbb{R}_+$, define $\mathbb{H}^2_\alpha$ as the space of ray of $\mathcal{F}_\alpha$ and define the natural projection $\pi_\alpha \colon \mathcal{F}_\alpha \to \mathbb{H}^2_\alpha$.

Proposition A.5 *For $\alpha \geq 0$, $\mathbb{H}^2_\alpha$ is homeomorphic to $\mathbb{R}^2$ and comes with a natural singular $\mathbb{H}^2$-structure whose singular locus contains at most one point. Furthermore,*

- *if $\alpha = 2\pi$, $\mathbb{H}^2_\alpha$ is regular and isomorphic to $\mathbb{H}^2$,*
- *if $2\pi \neq \alpha > 0$, the singular point is a conical singularity of angle α,*
- *if $\alpha = 0$, the singular point is a cusp.*

Proof • To begin with, in $\mathcal{F}_\alpha$, define the plane $\Pi := \{t = 1\}$ if $\alpha > 0$ and $\Pi := \{\tau = 1\}$ if $\alpha = 0$. The plane Π intersects each ray exactly once and $\pi|_\Pi$ is a homeomorphism.

• Define the surface $\mathcal{H}^* := \{\tau = (1+\mathfrak{r}^2)/(2\mathfrak{r}),\ \mathfrak{r} > 0\}$ if $\alpha = 0$ and $\mathcal{H}^* := \{t^2 - r^2 = 1,\ r > 0\}$ if $\alpha > 0$. The Lorentzian metric of $\mathcal{F}_\alpha$ induces a hyperbolic metric on $\mathcal{H}^*$ which intersects each ray of $\mathrm{Reg}(\mathcal{F}_\alpha)$ exactly once, and the projection $\mathcal{F}_\alpha \to \mathbb{H}^2_\alpha$ induces a homeomorphism $\mathcal{H}^* \simeq (\mathbb{H}^2_\alpha \setminus \mathrm{Sing}(\mathcal{F}_\alpha))$. Hence $\mathbb{H}^2_\alpha$ has an $\mathbb{H}^2$-structure defined on the complement of $\mathrm{Sing}(\mathcal{F}_\alpha)$, eg on the complement of a subset containing at most one point.

• If $\alpha = 2\pi$ then $\mathcal{F}_\alpha \simeq \mathcal{F}$ and the result follows.

• If $\alpha = 0$, one can check that $\mathcal{H}^*$ is complete and that the singular point of $\mathbb{H}^2_\alpha$ has a neighborhood of finite volume. The singular point is thus a cusp.

- If $2\pi \neq \alpha > 0$, then one can check that the length of the circle of radius $r > 0$ in $\mathbb{H}^2_\alpha$ around the singular point is αr. The singular point is a conical singularity of angle α. □

Definition A.6 An $\mathbb{H}^2_{\geq 0}$-manifold is a singular $\mathbb{H}^2$-manifold whose singular locus is locally modeled on $\mathbb{H}^2_\alpha$ for some $\alpha \geq 0$.

Definition A.7 Let Σ be an $\mathbb{H}^2_{\geq 0}$-manifold, let $(\mathcal{U}_i, \mathcal{V}_i, \varphi_i, \alpha_i)_{i \in I}$ be an $\mathbb{H}^2_{\geq 0}$-atlas of Σ with $\mathcal{V}_i \subset \mathbb{H}^2_{\alpha_i}$, and let $\mathcal{U}_{ij} := \mathcal{U}_i \cap \mathcal{U}_j$ and $\mathcal{V}_{ij} := \varphi_i(\mathcal{U}_i \cap \mathcal{U}_j)$ for $i, j \in I$ such that $\mathcal{U}_i \cap \mathcal{U}_j \neq \varnothing$. We add the convention that $\alpha_i \neq 2\pi$ if and only if $\mathcal{V}_i$ contains a neighborhood of the singular point of $\mathbb{H}^2_{\alpha_i}$ such that for any $i, j \in I$ where $\mathcal{U}_{ij} \neq \varnothing$ and $\mathcal{U}_i$ contains a singular point, $\alpha_i = \alpha_j$ and the change of charts $\mathcal{V}_{ij} \xrightarrow{\varphi_{ij}} \mathcal{V}_{ji}$ comes from some $\phi_{ij} \in \mathrm{SO}_0(1,2)$ acting both on $\mathbb{H}^2_{\alpha_i}$ and $\mathcal{F}_{\alpha_i}$.

Define the suspension $\operatorname{susp}(\Sigma)$ of Σ as the gluing of $(\pi_{\alpha_i}^{-1}(\mathcal{V}_i))_{i \in I}$ via $(\pi_{\alpha_i}^{-1}(\mathcal{V}_{ij}) \xrightarrow{\phi_{ij}} \pi_{\alpha_j}^{-1}(\mathcal{V}_{ji}))_{i,j \in I}$.

Remark The suspension susp is a functor from the category of $\mathbb{H}^2_{\geq 0}$-manifolds to the category of $\mathcal{F}_{\geq 0}$-manifolds.

Remark By construction, $\operatorname{susp}(\Sigma)$ is an $\mathcal{F}_{\geq 0}$-manifold with a natural projection $\operatorname{susp}(\Sigma) \to \Sigma$. One can check that diamonds $J^+(p) \cap J^-(q)$ are compact and that $\operatorname{susp}(\Sigma)$ is causal, and hence globally hyperbolic. Furthermore, the natural projection induces a homeomorphism $\pi \colon \Sigma_0 \to \Sigma$ for any Cauchy surface Σ_0.

Remark Be wary that the following simpler construction might be deceptively wrong. Start from $(\mathbb{H}^2_\alpha, \boldsymbol{h}_\alpha)$ as a hyperbolic conical singularity (or a cusp) with $\boldsymbol{h}_\alpha$ its Riemannian metric; then define the suspension as

$$\mathcal{F}'_\alpha := (\mathbb{R}^*_+ \times \mathbb{H}^2_\alpha, \boldsymbol{g}_\alpha), \quad \boldsymbol{g}_\alpha := -\mathrm{d}t^2 + t^2 \boldsymbol{h}_\alpha.$$

Though one indeed obtains $\mathcal{F}'_\alpha \simeq \mathcal{F}_\alpha$ for $\alpha > 0$ as well as $\operatorname{Reg}(\mathcal{F}_0) \simeq \operatorname{Reg}(\mathcal{F}'_0)$, note that $\mathcal{F}'_0$ is not isomorphic to $\mathcal{F}_0$ and not isomorphic to a neighborhood of a singular point of $\mathbb{E}^{1,2}_0$. To see this, notice that the past causal geodesics in $\operatorname{Reg}(\mathcal{F}'_\alpha)$ that "should" hit the singular line all converge to the same ideal point in the past (the origin) but never actually hit the singular line.

Definition A.8 A radiant spacetime is a Cauchy-compact Cauchy-maximal globally hyperbolic $\mathcal{F}_{\geq 0}$-manifold M.

We have a structure theorem for radiant spacetimes. This result is in the line of Mess's theorem [25] and is akin to previous results by Bonsante and Seppi [7], or the author [10] though in a much simpler context. To the author's knowledge, while this result is expected and "folkloric", there is no existing reference to point to. We therefore provide a proof.

Theorem 6 *Let M be a radiant spacetime. There exists a compact singular $\mathbb{H}^2_{\geq 0}$-manifold Σ such that $M \simeq \operatorname{susp}(\Sigma)$.*

Proof Let Σ_0 be a Cauchy surface of M and consider the natural projections $\pi_\alpha\colon \mathcal{F}_\alpha \to \mathbb{H}^2_\alpha$. Consider an $\mathcal{F}$-atlas $(\varphi_i, \mathcal{U}_i, \mathcal{V}_i)_{i\in I}$ of $\operatorname{Reg}(M)$ such that each $\mathcal{V}_i$ is causally convex in $\mathcal{F}$. Write $\mathcal{U}_{ij} := \mathcal{U}_i \cap \mathcal{U}_j$ for $i \in I$, and for $i, j \in I$ such that $\mathcal{U}_i \cap \mathcal{U}_j \neq \varnothing$ write $\mathcal{V}_{ij} := \varphi_i(\mathcal{U}_i \cap \mathcal{U}_j)$ as well as $\mathcal{W}_{ij} := \pi(\mathcal{V}_{ij}) \subset \mathbb{H}^2$. We then have a unique $\phi \in \mathrm{SO}_0(1,2)$ such that for all $x \in \mathcal{V}_{ij}$, $\varphi_j \circ \varphi_i(x) = \phi \cdot x$. Hence, for any $i, j \in I$ such that $\mathcal{U}_i \cap \mathcal{U}_j \neq \varnothing$, we have the following commutative diagram:

$$\begin{tikzcd}
\Sigma_0 & & & M & & & \Sigma_0 \\
\Sigma_0 \cap \mathcal{U}_i \arrow[u] \arrow[ddd, bend right, "\simeq"'] & \mathcal{U}_i \arrow[l, "\supset"'] \arrow[urr, "\subset"] \arrow[dd, "\varphi_i"'] & & \mathcal{U}_{ij} \arrow[u] \arrow[dll, "\simeq"', "\varphi_i"] \arrow[drr, "\simeq", "\varphi_j"'] & & \mathcal{U}_j \arrow[r, "\subset"] \arrow[ull, "\supset"'] \arrow[dd, "\varphi_j"] & \Sigma_0 \cap \mathcal{U}_j \arrow[u] \arrow[ddd, bend left, "\simeq"] \\
 & \mathcal{V}_{ij} \arrow[r, "\subset"'] \arrow[d, equal] & \mathcal{F} \arrow[rr, dashed, "\exists!\phi\in\mathrm{SO}_0(1,2)"] & & \mathcal{F} & \mathcal{V}_{ji} \arrow[l, "\supset"] \arrow[d, equal] & \\
\mathcal{V}_i \arrow[d, "\pi"] & \mathcal{V}_{ij} \arrow[l] \arrow[r, "\subset"'] \arrow[d, "\pi"] & \pi^{-1}(\mathcal{W}_{ij}) \arrow[u] \arrow[rr, dashed, "\phi"] \arrow[d, "\pi"] & & \pi^{-1}(\mathcal{W}_{ji}) \arrow[u] \arrow[d, "\pi"] & \mathcal{V}_{ji} \arrow[l, "\supset"] \arrow[r] \arrow[d, "\pi"] & \mathcal{V}_j \arrow[d, "\pi"] \\
\mathcal{W}_i & \mathcal{W}_{ij} \arrow[l] \arrow[r, equal] & \mathcal{W}_{ij} \arrow[rr, dashed, "\phi"] & & \mathcal{W}_{ji} \arrow[r, equal] & \mathcal{W}_{ij} \arrow[r] & \mathcal{W}_j
\end{tikzcd}$$

Since Σ_0 is acausal, the projection the maps $\Sigma_0 \cap \mathcal{U}_i \to \mathcal{W}_i$ are injective and by definition surjective; Σ_0 as well as all the $\mathcal{W}_i$ are 2-dimensional manifolds; by invariance of domain, the maps $\Sigma_0 \cap \mathcal{U}_i \to \mathcal{W}_i$ are then homeomorphisms. The $\mathcal{F}$-structure on M thus induces on Σ_0 a singular $\mathbb{H}^2$-structure; we call this singular $\mathbb{H}^2$-manifold Σ. Proceeding the same way around singular points of M, the local models $\mathcal{F}_\alpha$ of M induce a local model $\mathbb{H}^2_\alpha$ for each singular point of Σ. The suspension $\operatorname{susp}(\Sigma)$ of Σ is then given by the induced gluing of the cones $\pi_{\alpha_i}^{-1}(\mathcal{W}_i)$ along the $\pi_\alpha^{-1}(\mathcal{W}_{ij})$.

One can then define a natural map $M \xrightarrow{\iota} \operatorname{susp}(\Sigma)$ on each chart $(\mathcal{U}, \mathcal{V}, \varphi)$ of the $(\mathcal{F}_\alpha)_{\alpha\geq 0}$-atlas of M with $\mathcal{V} \subset \mathcal{F}_\alpha$ as $\iota\colon \mathcal{U} \to \pi_\alpha^{-1}(\pi_\alpha(\mathcal{V}))$, $x \mapsto \varphi(x)$. By construction, the map ι is an injective a.e. $\mathcal{F}$-morphism. Since M is Cauchy-maximal and Cauchy-compact, it follows from [10, Proposition 2.20] that the map ι is surjective, and thus an isomorphism. □

Corollary A.9 *Any radiant spacetime admits an embedded natural* $\mathbb{H}^2_{>0}$*-surface which is a Cauchy surface of its* $\mathcal{F}_{>0}$ *part.*

Another way to construct the suspension of an $\mathbb{H}^2_{\geq 0}$-surface Σ (and hence a radiant spacetime) is to choose a geodesic cellulation of Σ such that each cell is a polygon of $\overline{\mathbb{H}}^2$. The surface Σ can thus be seen as a gluing of a family of cells $\mathcal{P} = (P_i)_{i\in I}$ along their edges $\mathcal{E} = (e_i^{(j)})_{i\in I, j\in J_i}$ (where J_i parametrizes the edges of P_i) via isometries $\phi_{e,e'} \in \mathrm{SO}_0(1,2)$ sending the edge e to the edge e'. We denote by $\mathcal{G}$ the set of couples $(e, e') \in \mathcal{E}$ such that e is glued to e'. We can then construct $\operatorname{susp}(\Sigma)$ by gluing the cones $C_i := \pi^{-1}(P_i)$ for $i \in I$ along their faces $(\pi^{-1}(e))_{e\in\mathcal{E}}$ via the isometries $(\phi_{e,e'})_{(e,e')\in\mathcal{G}}$. We thus have the following:

Proposition A.10 *Any gluing of cones of $\overline{\mathcal{F}} = J^+(O)$ with polygonal basis, gluing couples of distinct 2-facets together via elements of $\mathrm{SO}_0(1,2)$ and without leaving unglued 2-facets, is a radiant spacetime.*

References

[1] **S Alexander**, **V Kapovitch**, **A Petrunin**, *An invitation to Alexandrov geometry:* CAT(0) *spaces*, Springer (2019) MR Zbl

[2] **A Alexandrov**, *Existence of a convex polyhedron and of a convex surface with a given metric*, Mat. Sb. 53 (1942) 15–65 MR Zbl In Russian

[3] **A D Alexandrov**, *Convex polyhedra*, Springer (2005) MR

[4] **T Barbot**, **F Bonsante**, **J-M Schlenker**, *Collisions of particles in locally AdS spacetimes, I: Local description and global examples*, Comm. Math. Phys. 308 (2011) 147–200 MR Zbl

[5] **P Bernard**, **S Suhr**, *Lyapounov functions of closed cone fields: from Conley theory to time functions*, Comm. Math. Phys. 359 (2018) 467–498 MR Zbl

[6] **A I Bobenko**, **I Izmestiev**, *Alexandrov's theorem, weighted Delaunay triangulations, and mixed volumes*, Ann. Inst. Fourier (Grenoble) 58 (2008) 447–505 MR Zbl

[7] **F Bonsante**, **A Seppi**, *Spacelike convex surfaces with prescribed curvature in* $(2{+}1)$*-Minkowski space*, Adv. Math. 304 (2017) 434–493 MR Zbl

[8] **G E Bredon**, *Topology and geometry*, Graduate Texts in Math. 139, Springer (1993) MR Zbl

[9] **L Brunswic**, *Surfaces de Cauchy polyédrales des espaces temps-plats singuliers*, PhD thesis, Université d'Avignon (2017) Available at `https://theses.hal.science/tel-01818016`

[10] **L Brunswic**, *Cauchy-compact flat spacetimes with extreme BTZ*, Geom. Dedicata 214 (2021) 571–608 MR Zbl

[11] **L Brunswic**, *On branched coverings of singular* (G, X)*-manifolds*, Geom. Dedicata 218 (2024) art. id. 43 MR Zbl

[12] **C Ehresmann**, *Œuvres complètes et commentées, I-1,2: Topologie algébrique et géométrie différentielle*, Cahiers Topologie Géom. Différentielle 24 (suppl. 1), Cahiers Topologie Géom Différentielle, Amiens (1983) MR Zbl

[13] **D B A Epstein**, **R C Penner**, *Euclidean decompositions of noncompact hyperbolic manifolds*, J. Differential Geom. 27 (1988) 67–80 MR Zbl

[14] **W Fenchel**, *Elementary geometry in hyperbolic space*, De Gruyter Studies in Math. 11, de Gruyter, Berlin (1989) MR Zbl

[15] **F Fillastre**, *Polyhedral realisation of hyperbolic metrics with conical singularities on compact surfaces*, Ann. Inst. Fourier (Grenoble) 57 (2007) 163–195 MR Zbl

[16] **F Fillastre**, *Existence and uniqueness theorem for convex polyhedral metrics on compact surfaces*, from "Conference on metric geometry of surfaces and polyhedra", Current Prob. Math. Mech. 6, Moscow State Univ. (2010) 208–223

[17] **F Fillastre**, *Fuchsian polyhedra in Lorentzian space-forms*, Math. Ann. 350 (2011) 417–453 MR Zbl

[18] **F Fillastre**, **I Izmestiev**, *Hyperbolic cusps with convex polyhedral boundary*, Geom. Topol. 13 (2009) 457–492 MR Zbl

[19] **F Fillastre**, **I Izmestiev**, *Gauss images of hyperbolic cusps with convex polyhedral boundary*, Trans. Amer. Math. Soc. 363 (2011) 5481–5536 MR Zbl

[20] **W M Goldman**, *Geometric structures on manifolds*, Graduate Studies in Math. 227, Amer. Math. Soc., Providence, RI (2022) MR Zbl

[21] **C D Hodgson**, **I Rivin**, *A characterization of compact convex polyhedra in hyperbolic* 3*-space*, Invent. Math. 111 (1993) 77–111 MR Zbl

[22] **I Izmestiev**, *A variational proof of Alexandrov's convex cap theorem*, Discrete Comput. Geom. 40 (2008) 561–585 MR Zbl

[23] **D Kane**, **G N Price**, **E D Demaine**, *A pseudopolynomial algorithm for Alexandrov's theorem*, from "Algorithms and data structures", Lecture Notes in Comput. Sci. 5664, Springer (2009) 435–446 MR Zbl

[24] **H Masur**, **J Smillie**, *Hausdorff dimension of sets of nonergodic measured foliations*, Ann. of Math. 134 (1991) 455–543 MR Zbl

[25] **G Mess**, *Lorentz spacetimes of constant curvature*, Geom. Dedicata 126 (2007) 3–45 MR Zbl

[26] **A D Milka**, *Space-like convex surfaces in pseudo-Euclidean spaces*, from "Some questions of differential geometry in the large", Amer. Math. Soc. Transl. Ser. 2 176, Amer. Math. Soc., Providence, RI (1996) 97–150 MR Zbl

[27] **R C Penner**, *The decorated Teichmüller space of punctured surfaces*, Comm. Math. Phys. 113 (1987) 299–339 MR Zbl

[28] **R C Penner**, *Decorated Teichmüller theory*, Eur. Math. Soc., Zürich (2012) MR Zbl

[29] *SageMath*, *version* 9.4 (2022) Available at `https://www.sagemath.org`

[30] **J Sbierski**, *On the existence of a maximal Cauchy development for the Einstein equations: a dezornification*, Ann. Henri Poincaré 17 (2016) 301–329 MR Zbl

[31] **J-M Schlenker**, *Hyperbolic manifolds with polyhedral boundary*, preprint (2001) arXiv math/0111136

[32] **R Souam**, *The Schläfli formula for polyhedra and piecewise smooth hypersurfaces*, Differential Geom. Appl. 20 (2004) 31–45 MR Zbl

[33] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `https://url.msp.org/gt3m`

[34] **Y A Volkov**, *Existence of a polyhedron with prescribed development*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) 476 (2018) 50–78 MR Zbl In Russian; translated in J. Math. Sci. (N.Y.) 251 (2020) 462–479

Centre de recherche astrophysique de Lyon
Lyon, France

Current address: *Huawei, Noah Ark Laboratories*
Montreal QC, Canada

`leo@brunswic.fr`

`https://leo.brunswic.fr`

Received: 4 February 2022 Revised: 17 November 2023

Algebraic & Geometric Topology **25**:3 (2025) 1377–1409
DOI: 10.2140/agt.2025.25.1377
Published: 20 June 2025

Real algebraic overtwisted contact structures on 3-spheres

ŞEYMA KARADERELI
FERIT ÖZTÜRK

A real algebraic link in the 3-sphere is defined as the zero locus in the 3-sphere of a real algebraic function from $\mathbb{R}^4$ to $\mathbb{R}^2$ with an isolated singularity at the origin. A real algebraic open book decomposition on the 3-sphere is by definition the Milnor fibration of such a real algebraic function. We prove that every overtwisted contact structure on the 3-sphere with positive three-dimensional invariant d_3 (apart from 13 exceptions) are real algebraic via functions of the form $f\bar{g}$ with f, g complex algebraic and with the pages of the associated open books planar.

32S55, 57K33; 32C05

1 Introduction

A Milnor fillable 3-manifold is a connected closed oriented contact 3-manifold which is contact isomorphic to the contact link manifold of a complex analytic surface with isolated singularity. We know that any such manifold admits a unique Milnor fillable contact structure up to contactomorphism — see Caubel, Némethi and Popescu-Pampu [5] — and moreover a Milnor fillable contact structure is tight. For instance there is a unique tight contact structure on the 3-sphere S^3 and it is Milnor fillable (by eg the nonsingularity 0 in $\mathbb{C}^2$).

Here we ask a similar question regarding overtwisted contact structures. We confine ourselves to S^3 although the definitions and questions below can be easily generalized. We investigate fibered links in S^3 which are given real algebraically (or more generally real analytically). Let us call an oriented link in S^3 weakly real algebraic if it is isotopic to the link of a real algebraic surface with an isolated singularity at 0 (ie it is the zero locus of an algebraic map $h\colon \mathbb{R}^4 \to \mathbb{R}^2$ with an isolated critical point on its zero locus). It is well known that every link in S^3 is weakly real algebraic; see Akbulut and King [1]. Nevertheless the map h may have singularities outside its zero locus arbitrarily close to 0. If 0 is an isolated critical point of h, we call the associated oriented link in S^3 real algebraic. This condition of isolatedness is called the Milnor condition. In such a case there is a Milnor fibration on the link exterior in S^3 over S^1; see Milnor [16, Section 11]. In other words the real algebraic link is the binding of an (in general rational) open book with the open book decomposition given as the Milnor fibration (see eg Baker and Etnyre [2] for rational open books). If moreover the Milnor fibration is given by $h/\|h\|$ we call the associated open book (and the supported contact structure) on S^3 real algebraic.

Although the fibration is given by $h/\|h\|$ in a tubular neighborhood of the zero set of h and that fibration can always be inflated to a Milnor fibration on S^3 (see eg the survey of Seade [21]), it is not always true that this Milnor fibration coincides with the one given by $h/\|h\|$ on S^3. A quite simple counterexample is given in [16, Section 11].

On the other hand, compared to weakly real algebraic ones it is rather hard to construct examples of real algebraic maps with an isolated singularity and this issue has been long studied. For example it is known that the fibered figure-eight knot is not complex algebraic but is real algebraic; see Perron [18]. Meanwhile since every real algebraic link is fibered, a nonfibered weakly real algebraic link cannot be real algebraic. We believe it is still unknown whether every fibered link is real algebraic (see eg Bode [4]).

An obvious way to produce real algebraic links in S^3 is as follows. Take two nonconstant complex algebraic maps $f, g\colon \mathbb{C}^2 \to \mathbb{C}$ and consider the real algebraic map $h = f\bar{g}$. The oriented link L that is the zero locus of h in S^3 has components $\{f=0\} \cap S^3$ with canonical orientations and $\{g=0\} \cap S^3$ with the reverse orientations. Such links are special examples of graph links, ie spliced Seifert links; see Eisenbud and Neumann [7]. Moreover h has an isolated singularity at 0 if and only if L is fibered, and in that case the Milnor fibration is given by $h/\|h\|$; see Pichon [19]. Now, as a corollary to Ishikawa [15] the real algebraic open book corresponding to such h determines an overtwisted contact structure on S^3.

We recall that there are countably infinite number of overtwisted contact structures in S^3. They are distinguished by the half-integer-valued d_3 invariant (see eg Ding, Geiges and Stipsicz [6]) or equivalently the Hopf invariant H of the monodromy vector field; on S^3 these two invariants satisfy $H = -d_3 - \frac{1}{2}$ (see eg Tagami [22]). They are also related to the enhanced Milnor number λ of the binding of an open book that supports the contact structure: $\lambda = -H$ (see eg Hedden [13]; for the introduction of λ see Neumann and Rudolph [17]). Inaba [14] has already proven that all overtwisted structures in S^3 are real algebraic, by explicitly constructing real algebraic maps for any given $\lambda \in \mathbb{Z}$. More precisely these maps are mixed polynomials of the form $f\bar{g}$, are polar weighted homogenous and conveniently strongly nondegenerate. The computation of λ uses the ideas introduced in [17] for multilinks that are given by splice diagrams. The constructed open books have pages with varying genera.

In this article we are interested in the genera of the pages of the real algebraic open books. Recall that any overtwisted contact structure is planar, ie it is supported by a planar open book; see Etnyre [9]. Here we prove the following planarity result in the real algebraic setup.

Theorem 1.1 *All overtwisted contact structures on S^3 with $d_3 > 0$ and*

$$d_3 + \tfrac{1}{2} \notin \{4, 5, 9, 11, 17, 19, 25, 37, 47, 61, 79, 95, 109\}$$

are real algebraic, with the associated real algebraic open books having planar pages. These planar, real algebraic overtwisted structures are exactly the ones which can be obtained by functions of the form $f\bar{g}$ with $f, g\colon \mathbb{C}^2 \to \mathbb{C}$ complex algebraic.

We remark that the polynomials $f\bar{g}$ that we construct have real coefficients. Also recall the supporting genus results for tight contact structures: not only a tight structure may have positive minimal supporting genus among supporting open books, it has been also shown that the Milnor fillable (tight) contact structures may have Milnor genus strictly greater than the support genus; see Bhupal and Ozbagci [3].

In order to build the overtwisted structures in the theorem we consider all fibered Seifert/graph multilinks with planar fibers; these turn out to be exactly the ones that appear in [7, page 123] and their possible splicings. Going through all these fibered links which are also known to be real algebraic, we prove the theorem. In this way we exhaust all Seifert/graph multilinks that are given by real analytic functions of the form $f\bar{g}$. To come up with new real algebraic planar open books one has to use real analytic functions of different forms.

We believe that the 13 sporadic exceptions that appear in the theorem are real algebraic, planar as well, although the families of real algebraic Milnor fibrations that we have produced via functions $f\bar{g}$ miss them. The nonnegativity that emerges might be more resilient. Thus we ask

Question 1.2 *Is there a real algebraic, planar overtwisted contact structure on S^3 with negative d_3? The supporting real algebraic open book is rational in general; ie the fibered link is a multilink. Can the open book be made an integral open book? That is, can the binding be a simple link which is not a multilink?*

Generalizing our definitions we ask

Question 1.3 *Is it true that every overtwisted contact structure on a Milnor fillable* 3*-manifold is real algebraic? Can the associated real algebraic open books have planar pages?*

To proceed towards the proof of Theorem 1.1, we recall in Section 2 the Seifert and graph multilinks and the splicing operation. There we also give our families of fibered graph multilinks in S^3 and compute the associated monodromy maps. In Section 3 we demonstrate that those families of graph multilinks and the corresponding open book decompositions are real algebraic via functions of the form $f\bar{g}$. In Section 4 we briefly recall a way to compute the d_3 invariant, by constructing almost complex 4-manifolds that fill the given open book decompositions in S^3. Finally in Section 5 we prove Theorem 1.1 by computing the d_3 invariants explicitly for our families of examples. It turns out that one of our families of graph multilinks exhausts all the overtwisted structures with $d_3 > 461$. Then by computer aid we show that those with $0 < d_3 < 461$ (except the 13 values given in the theorem) are realized by our families of graph multilinks as well. In the computation of d_3 the constructed 4-manifolds have large intersection matrices. For the clarity of the exposition, those intersection matrices are presented in Appendix A and the tedious computations regarding those matrices are given in Appendix B.

Acknowledgements The final revisions for this article were completed during the visit of Karadereli to the Erwin Schrödinger International Institute for Mathematics and Physics, University of Vienna. Karadereli would like to express gratitude to ESI and Vera Vertesi for their hospitality during the visit.

2 Seifert multilinks and splicing

In this section we recall introductory information on Seifert and graph multilinks and present several families of examples which, as to be argued in the next sections, are planarly fibered and real algebraic via functions of the form $f\bar{g}$. Our discussion here is based on [7].

2.1 Seifert multilinks

A Seifert fibered manifold is a closed 3-manifold given as an S^1-bundle with the orbit space a 2-orbifold. A Seifert multilink in a Seifert fibered 3-sphere is an oriented link L that is constituted of a finite number of Seifert fibers S_i and an integer multiplicity m_i assigned to each component. In this work we are solely interested in Seifert multilinks in S^3. We are going to denote a Seifert multilink with n components by $L(m_1, \dots, m_n)$. L is canonically oriented by the sign of the multiplicities m_i. In this setup the homology class $\underline{m} = (m_1, \dots, m_n) \in H_1(L) \simeq \mathbb{Z}^n$ determines a cohomology class in the link complement as well, since $H_1(L) \simeq H^1(M - L)$. That class is given by

$$\underline{m}(\gamma) = \mathrm{lk}(L, \gamma) = \sum_{i=1}^{n} m_i \cdot \mathrm{lk}(S_i, \gamma).$$

Let μ_i denote the meridian of the i^{th} link component. Then we have $\underline{m}(\mu_i) = m_i$. Moreover we can realize the Seifert surface of the multilink as an embedded oriented surface whose intersection with the boundary of a tubular neighborhood of S_i is $(\delta_i \cdot (m_i/\delta_i, -(m_i)'/\delta_i))$-cable of S_i, where $(m_i)' = \sum_{j \neq i}^{n} m_j \,\mathrm{lk}(S_i, S_j)$ and $\delta_i = \gcd(m_i, m_i')$ [7, page 30].

Multilinks are represented by splice diagrams as exemplified in Figure 1. The central node represents the ambient Seifert manifold. The numbers adjacent to the node for each branch are called the weights and the numbers next to the arrowheads are the multiplicities m_i.

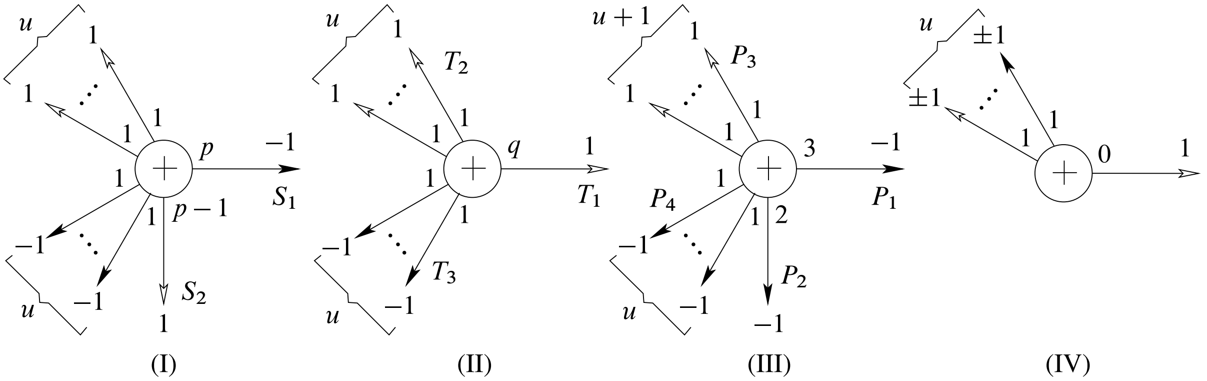


Figure 1: Splice diagrams for Seifert multilinks of type (I), (II), (III) and (IV). These are exactly all fibered Seifert multilinks with trivial geometric monodromy.

An arrowhead with weight $+1$ (respectively > 1) corresponds to a regular (respectively singular) Seifert fiber. The multilink (I) in Figure 1 has $2u+2$ connected components in the underlying manifold S^3 on which the Seifert fibration is given by the S^1-action $(x,y) \mapsto (t^{p-1}x, t^p y)$ for $t \in S^1$. Here the orbit $\{x=0\}$ corresponds to the singular fiber S_1 with weight p and $\{y=0\}$ corresponds to the singular fiber S_2 with weight $p-1$. The linking numbers of link components can be computed easily using the splice diagram [7, Proposition 7.4]. For instance, the linking number of any nonsingular fiber with the singular fiber S_1 (respectively with S_2) is the product of weights of the remaining vertices, which equals $p-1$ (respectively p). The linking number of S_1 and S_2 is 1. Thus the multilink (I) is isotopic to the negative Hopf link union u positively oriented and u negatively oriented isotopic copies of the $(p, p-1)$ torus knot cabled around S_1.

A multilink $L(\underline{m})$ is fibered if there exists a locally trivial fibration $M - L \to S^1$ in the homotopy class corresponding to $\underline{m}$, whose fibers are minimal Seifert surfaces for the multilink. Using the analytic description of the Seifert fibration of the link exterior, it can be easily seen that a Seifert multilink is fibered if and only if the linking number of any nonsingular fiber γ with the multilink does not vanish [7, page 90]. In other words, denoting by α_i the weight of the i^{th} link component S_i, the integer

$$l = \underline{m}(\gamma) = \sum_{i=1}^{n} m_i \operatorname{lk}(\gamma, S_i) = \sum_{i=1}^{n} m_i \alpha_1 \cdots \hat{\alpha}_i \cdots \alpha_n$$

is nonzero. Moreover if $l=1$ then the pages of the corresponding open book are planar. The families of diagrams in Figure 1 are exactly those Seifert multilinks with $l=1$ [7, page 123].

A fibered multilink determines a rational open book decomposition for the ambient Seifert manifold. If each $m_i = \pm 1$ then the open book is an integral open book.

The monodromy of the fibration can be represented as the flow along the Seifert fibers. Thus in the interior of the pages it is isotopic to a homeomorphism of order l. On the other hand the monodromy flow near each boundary component is computed as a $(-(\delta_i/m_i l)\alpha_i)$-worth (in general rational) twist along a boundary parallel curve [7, page 108].

Example 2.1 For the multilinks of type (I) given in Figure 1, the multilink is fibered since we have $l = (-1)\cdot(p-1) + 1\cdot p + u\cdot(1)\cdot p(p-1) + u\cdot(-1)\cdot p(p-1) = 1 \neq 0$. The pages are $(2u+2)$-punctured spheres. The monodromy flow is trivial in the interior of the pages. However near the boundary components corresponding to the singular fibers, the flow is given as $-\frac{1}{-1\cdot 1}p = p$ and $-\frac{1}{1\cdot 1}(p-1) = -(p-1)$ twists. Along the boundary components corresponding to the nonsingular fibers with positive and negative multiplicities, the flow is -1 and $+1$ twist respectively. Therefore the monodromy is given as

$$\phi = a^p \cdot b^{-(p-1)} \cdot c_1^{-1} \cdots c_u^{-1} \cdot d_1^1 \cdots d_u^1. \tag{2-1}$$

Here, a and b denote Dehn twists along curves parallel to the boundary components $\{x=0\}$ and $\{y=0\}$ respectively; c_i and d_i are twists along curves parallel to the nonsingular components with positive and negative multiplicities respectively.

Similarly, as noted above, the multilinks of type (II) and (III) in Figure 1 are fibered multilinks in S^3 with $l=1$ too. The pages of the multilink of type (II) are $(2u+1)$-punctured spheres and the monodromy is

$$\phi = a^{-q}\cdot b^{-1}\cdot c_1^{-1}\cdots c_{u-1}^{-1}\cdot d_1^1\cdots d_u^1. \tag{2-2}$$

The pages of the multilink of type (III) are $(2u+3)$-punctured spheres and the monodromy is

$$\phi = a^{3}\cdot b^{2}\cdot c_1^{-1}\cdots c_{u+1}^{-1}\cdot d_1^1\cdots d_u^1. \tag{2-3}$$

2.2 Splicing multilinks

The splice of two multilinks along a specified pair of link components is constructed topologically by excising tubular neighborhoods of the given link components and gluing the remaining manifolds in a meridian-to-longitude fashion. Note that topologically splicing multilinks in S^3 produces a multilink still in S^3. Moreover a cohomology class is determined by the multiplicities of the components of the resulting multilink. For the splicing operation we require that the restriction of this cohomology class on each manifold gives the cohomology class of the splice component. This condition is equivalent to the following. Let S_0 and $\widetilde{S}_0$ with multiplicities m_0 and $\widetilde{m}_0$ be the spliced link components; (μ_0,λ_0) and $(\tilde{\mu}_0,\tilde{\lambda}_0)$ be the meridians and longitudes on the tori on which the splicing occurs. Then we must have

$$\begin{aligned} m_0 &= \underline{m}(\mu_0) = \underline{\widetilde{m}}(\tilde{\lambda}_0) = (\widetilde{m}_0)', \\ \widetilde{m}_0 &= \underline{\widetilde{m}}(\tilde{\mu}_0) = \underline{m}(\lambda_0) = (m_0)', \end{aligned}$$

where $(\widetilde{m}_0)'$ and $(m_0)'$ are defined as in Section 2.1. Observe that these requirements are exactly the conditions for the Seifert surfaces in each splice component to glue together along the splicing tori. Moreover since Seifert surfaces approach the spliced link components as $\delta_0=\gcd(m_0,(m_0)')$ copies of the $(m_0/\delta_0,(m_0)'/\delta_0)$ curve, the Seifert surfaces are pasted together along δ_0 tori.

Splicing of two multilinks is represented by a splice diagram (with more than one node) obtained by joining the two diagrams along the arrowheads corresponding to the link components at which splicing occurs. A multilink with such a splice diagram is called a graph multilink.

As an example, consider the multilink (I) in Figure 1 and there the link component S_1 of weight p. Since $\underline{m}(\lambda_1)=(1)\cdot 1\cdots 1+u\cdot(1)\cdot 1\cdots 1\cdot(p-1)+u\cdot(-1)\cdot 1\cdots 1\cdot(p-1)=1$ and $\underline{m}(\mu_1)=-1$, one can splice S_1 only with a link component whose multiplicity is $\widetilde{m}_1=1$ and $(\widetilde{m}_1)'=-1$, ie the pages must approach the link component as $(1,-1)$ curves. Similarly for the link component S_2 of weight $p-1$, we have $\underline{m}(\lambda_2)=(-1)\cdot 1\cdots 1+u\cdot(1)\cdot 1\cdots 1\cdot p+u\cdot(-1)\cdot 1\cdots 1\cdot p=-1$ and $\underline{m}(\mu_2)=1$. Therefore, given two multilinks of type (I) one can only splice S_1 in one with S_2 in the other.

Another possible splicing occurs between the splice multilink (II) and the multilink (I) in a single case; that occurs when $q=2$. In fact, computing $\underline{m}(T_j)$ for T_j as in Figure 1 we obtain 0, $1-q$ and $1+q$ for $j=1,2,3$ respectively. Thus splicing is only possible when $q=2$ and the splicing occurs between the knot S_1 of type (I) and T_2 of type (II).

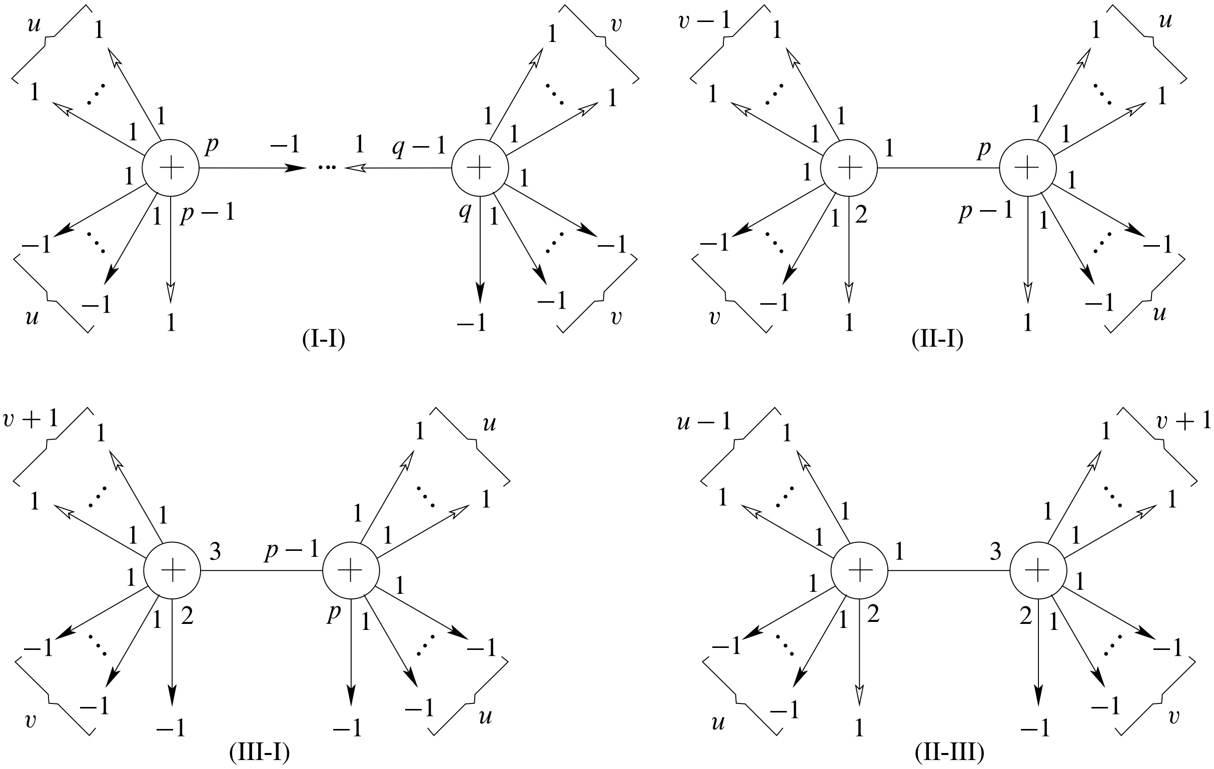


Figure 2: All possible splice diagrams consisting of (I), (II) and (III) are made up of these pieces.

Similarly splicing is possible between the knot S_2 of type (I) and P_1 of type (III), and between the knot T_2 of type (II) and P_1 of type (III). Here P_1 is as in Figure 1. Going through all possible cases we obtain the following list.

Lemma 2.2 *All possible splice diagrams in S^3 that can be obtained via the multilinks* (I), (II) *and* (III) *are trees where each splicing is one of those in Figure 2.*

A graph multilink is fibered if and only if it is an irreducible link and each of its splice components is fibered [7, Theorem 4.2]. The monodromy is pieced together from the monodromy maps of the splice components. In each splice component the monodromy is given by the flow along the corresponding Seifert fibers whereas on the tubular neighborhoods of the separating tori, it has two different flows in each end given by the Seifert fibration of each Seifert component. Therefore after splicing, the Dehn twists corresponding to glued boundaries become trivial and on the separating annuli the monodromy acts as a twist map which measures the difference between the two flows of Seifert fibers. In [7, Theorem 13.1] the monodromy flow on a separating annulus is computed as a τ-worth twist along the core of the annulus with

$$\tau = \frac{-\delta_0}{l_1 \cdot l_2}(\alpha_0\beta_0 - \alpha_1 \cdots \alpha_n \cdot \beta_1 \cdots \beta_m), \tag{2-4}$$

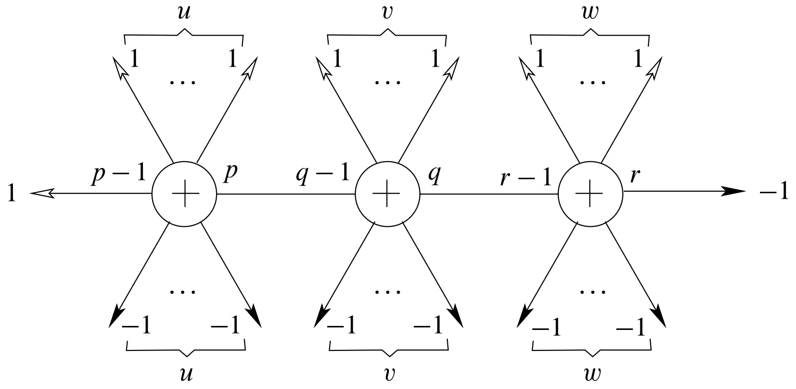


Figure 3: Splice diagram for (I-I-I).

where α_0, β_0 are the weights of the spliced components and α_i, β_j are the weights of the remaining link components around the two nodes.

Example 2.3 Consider the multilink (I-I) given in Figure 2. Note that when $q = p$ the graph multilink is simply a Seifert multilink [7, Theorem 8.1(6)]. So let us consider the case $q > p$.

By the previous discussion we know that $l_1 = l_2 = 1$; also $\delta = \gcd(-1, 1) = 1$. Thus (2-4) gives

$$\tau = -\frac{1}{1\cdot 1}(p(q-1) - q(p-1)) = p - q.$$

Since $\delta = 1$, we glue the pages of the spliced components, which are $(2u+2)$- and $(2v+2)$-punctured spheres respectively, along a single annulus neighborhood of the spliced boundary components. Consequently the pages of the spliced multilink are $(2u+2v+2)$-punctured spheres.

As given in (2-1) the splice components have monodromies $\phi_1 = \alpha^p \cdot a^{-(p-1)} \cdot c_1^{-1} \cdots c_u^{-1} \cdot d_1^1 \cdots d_u^1$ and $\phi_2 = b^q \cdot \beta^{-(q-1)} \cdot e_1^{-1} \cdots e_v^{-1} \cdot f_1^1 \cdots f_v^1$. The monodromy flow is $q - p$ negative Dehn twists about the core circle, say γ, in the annulus. Therefore the monodromy of the spliced multilink is

$$\phi = a^{-(p-1)} \cdot c_1^{-1} \cdots c_u^{-1} \cdot d_1^1 \cdots d_u^1 \cdot \gamma^{-(q-p)} \cdot b^q \cdot e_1^{-1} \cdots e_v^{-1} \cdot f_1^1 \cdots f_v^1. \tag{2-5}$$

Example 2.4 Similarly let us consider a graph multilink of the form (I-I-I) as in Figure 3. Recall that we splice the knot with weight q of the first splice component to the knot with weight $(r-1)$ of the second splice component.

As in the previous examples $l_1 = l_2 = 1$ and $\delta = \gcd(m_1, m_2) = 1$. Assuming $r > q$, we have

$$\tau = -\frac{\delta}{l_1 l_2}(q(r-1) - r(q-1) = q - r < 0.$$

The page of the splice multilink is a union of the pages of the splice components joined together along a boundary by a $(q-r)$-twisted annulus (since $\delta = 1$). Since the splice components have $(2u+2v+2)$- and $(2w+2)$-punctured sphere pages, the pages for the splice link are $(2u+2v+2w+2)$-punctured spheres.

The monodromy of the new fibration is

$$\phi = a^{-(p-1)} c_1^{-1} \cdots c_u^{-1} d_1^1 \cdots d_u^1 \gamma^{-(q-p)} e_1^{-1} \cdots e_v^{-1} f_1^1 \cdots f_v^1 \theta^{-(r-q)} b^r g_1^{-1} \cdots g_w^{-1} h_1^1 \cdots h_w^1 \tag{2-6}$$

where θ denotes the Dehn twist about the core circle in the latter annulus.

Example 2.5 As in the previous example one can compute the monodromies of the other multilinks given in Figure 2. Among these we will need the monodromy of the splicing (III-I),

$$\phi = a^{(p)} \cdot c_1^{-1} \cdots c_u^{-1} \cdot d_1^1 \cdots d_u^1 \cdot \gamma^{-(p-3)} \cdot b^2 \cdot e_1^{-1} \cdots e_{v+1}^{-1} \cdot f_1^1 \cdots f_v^1. \tag{2-7}$$

Here, we assume that $p \geq 4$ because the graph multilink is simply a Seifert multilink when $p = 3$ [7, Theorem 8.1(6)].

3 Real algebraic singularities and associated contact structures

In this section we assert that the graph multilinks and the associated open books that have been considered in the previous section with explicit monodromy can be realized real algebraically via functions of the form $f\bar{g}$.

For an isolated singularity of a holomorphic (or a complex algebraic) function from $\mathbb{C}^2$ to $\mathbb{C}$, the corresponding Milnor fibration defines an open book structure on S^3, whose binding is isotopic to the singularity link. In such a setup we call the singularity link and the open book and the supported tight contact structure complex analytic/algebraic. Any complex algebraic link in S^3 is a graph multilink and the corresponding splice diagram can be deduced from the Puiseux pairs [7, Appendix 1]. Of course not all the graph multilinks in S^3 are complex algebraic. Eisenbud and Neumann [7, Theorem 9.4] gave the precise condition for a graph multilink to be complex algebraic.

Similarly an isolated singularity of a real analytic function $h\colon \mathbb{R}^4 \to \mathbb{R}^2$ determines a Milnor fibration in S^3 under the condition that the Jacobian matrix of h has rank 2 on an open neighborhood of the origin, except the origin. This is the Milnor condition. A link is said to be real analytic/algebraic if it is the singularity link of a real analytic/algebraic map $h\colon \mathbb{R}^4 \to \mathbb{R}^2$ that satisfies the Milnor condition. In the absence of the Milnor condition, there might not even exist a Milnor fibration. In the particular case $h = f\bar{g}$ where f and g are holomorphic functions, [19] and [20] discuss the Milnor fibration in the link exterior and the geometry of the fibration near the singularity link.

The isotopy class of a multilink is encoded in a plumbing tree that is decorated with arrows having multiplicities for the link components. When a multilink is isotopic to the singularity of a holomorphic germ, the plumbing tree for the multilink can be obtained as the dual tree of any normal crossing resolution of the function. Since $L_{f\bar{g}}$ as an unoriented link is $L_f \cup L_g$, it follows that the resolution graph of a real algebraic germ of the form $f\bar{g}$ is nothing but the resolution graph of fg with negative signs for the multiplicities of the link components corresponding to g. Passing to the corresponding splice diagram as described in [7, Section 20], we conclude that the conditions in [7, Theorem 9.4] are necessary for real algebraicity via $f\bar{g}$. Namely these conditions are:

(i) the weights of all vertices are positive;

(ii) for every splicing $\alpha_0\beta_0 > \alpha_1\cdots\alpha_n\cdot\beta_1\cdots\beta_m$ where α_0, β_0 are the weights of the spliced components and α_i, β_j are the remaining weights around the two nodes.

Thus we immediately conclude that (IV) in Figure 1 fails (i) for real algebraicity via $f\bar{g}$, and the splicings (II-I) and (II-III) fail (ii). Moreover any splicing involving (IV) either fails (i) or (ii). So the only cases in the previous section that satisfy the necessary conditions (i) and (ii) are (I), (II), (III) and any segment of (III-I-I-…).

Having said these, the following theorem explains exactly when the singularity link of a real algebraic germ of the form $f\bar{g}$ has a real algebraic open book.

Theorem 3.1 [19, Theorem 5.1] *Let $f\colon(\mathbb{C}^2,0)\to(\mathbb{C},0)$ and $g\colon(\mathbb{C}^2,0)\to(\mathbb{C},0)$ be two holomorphic germs with isolated singularities and having no common branches. Then the real analytic germ $f\bar{g}$ has an isolated singularity at 0 if and only if the link $L_f - L_g$ is fibered.*

Moreover, if this condition holds, then the Milnor fibration of the link $L_f - L_g$ is given by $f\bar{g}/\|f\bar{g}\|$.

Let us elaborate in our running examples.

Example 3.2 For $\eta^{2u+1} = 1$ consider the functions

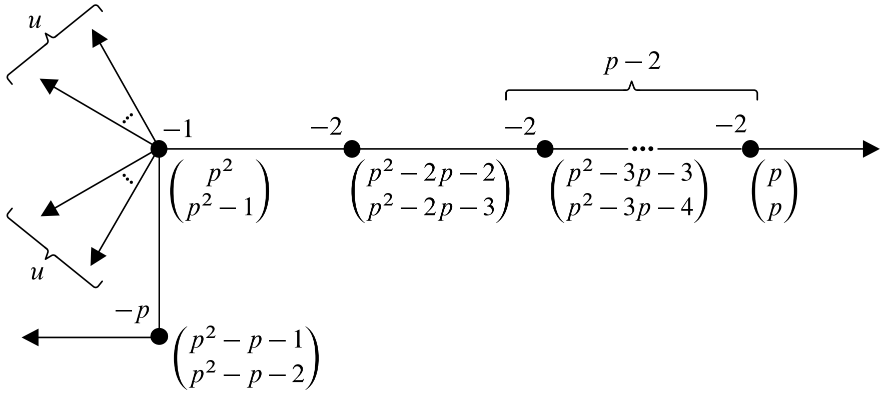
$$f(x,y) = y\prod_{i=1}^{u}(x^p + \eta^i y^{p-1}) \quad\text{and}\quad g(x,y) = x\prod_{j=u+1}^{2u}(x^p + \eta^j y^{p-1}).$$

After resolving the germ of fg, we obtain the plumbing diagram of $L_{f\bar{g}}$ given in Figure 4. As in [7, Section 20], we can obtain the splice diagram of the singularity link from the plumbing diagram and see that it is isotopic to the multilink of type (I) in Figure 1. Since we have already noted that the multilink is fibered, it follows from Theorem 3.1 that $f\bar{g}$ has an isolated singularity and the fibration of the multilink which we investigated in the previous section is the Milnor fibration of the germ. Observe also that the branch $\{\bar{x} = 0\}$ corresponds to the singular link component of weight p, $\{y = 0\}$ corresponds to the singular component of weight $p-1$ and the positively (respectively negatively) oriented u copies of $(p, p-1)$ cables around $\{x = 0\}$ component correspond to the branches $\bigl\{\prod_{i=1}^{u}(x^p + \eta^i y^{p-1}) = 0\bigr\}$ (respectively $\bigl\{\prod_{i=1}^{u}\overline{(x^p + \eta^i y^{p-1})} = 0\bigr\}$).

Example 3.3 Similarly we observe that the singularity links of the real algebraic germs

$$\Bigl(xy\prod_{i=1}^{u}(x^q + \eta^i y)\Bigr)\cdot\Bigl(\prod_{j=1}^{u-1}\overline{(x^q + \eta^{u+j} y)}\Bigr) \quad\text{and}\quad \Bigl(\prod_{i=1}^{u+1}(x^3 + \eta^i y^2)\Bigr)\cdot\Bigl(\bar{x}\bar{y}\prod_{j=1}^{u}\overline{(x^3 + \eta^{u+j+1} y^2)}\Bigr)$$

are isotopic to the fibered multilinks of type (II) and (III) in Figure 1 respectively; therefore have isolated singularities at the origin and engender Milnor fibrations.



u
u
$p-2$
-1
-2
-2
-2
$-p$
$\begin{pmatrix} p^2 \\ p^2-1 \end{pmatrix}$
$\begin{pmatrix} p^2-2p-2 \\ p^2-2p-3 \end{pmatrix}$
$\begin{pmatrix} p^2-3p-3 \\ p^2-3p-4 \end{pmatrix}$
$\begin{pmatrix} p \\ p \end{pmatrix}$
$\begin{pmatrix} p^2-p-1 \\ p^2-p-2 \end{pmatrix}$

Figure 4: Dual tree of a resolution π of fg with associated multiplicities given in the parentheses which are the multiplicities m_i^f and m_i^g of $f\circ\pi$ and $g\circ\pi$, respectively, along the irreducible component for the i^{th} exceptional divisor. As a side remark we recall that $L_f - L_g$ is fibered if and only if $m_i^f \neq m_i^g$ at the rupture vertices [20, Corollary 2.2].

As for the graph multilinks obtained via splicing in the previous section, a priori they might not be algebraic. Consider the positively oriented graph multilink isotopic to the multilink (I-I). This multilink is complex algebraic when $q > p$ [7, Theorem 9.4]. The corresponding holomorphic function can be easily deduced from the holomorphic germs related to the spliced components as follows. Recall that we splice the component corresponding to the branch $\{x=0\}$ of a multilink L_1 of type (I) with weights for singular fibers $p, p-1$ with the component $\{y=0\}$ of a multilink L_2 of type (I) with weights $q, q-1$. By isotopy, the nonsingular link components of L_1 which are $(p, p-1)$ cables of $\{x=0\}$ can be realized as $(p-1, p)$ cables of the $\{y=0\}$ component of L_1. As we splice, we remove the spliced link components and keep the remaining ones. The resulting multilink is a positive Hopf link with $2u$ many $(p-1, p)$ cables around the link component $\{y=0\}$ (coming from L_1) and $2v$ many $(q, q-1)$ cables around the link component $\{x=0\}$ (coming from L_2). Again by isotopy, $(p-1, p)$ cables around the former component can be seen as $(p, p-1)$ cable around the latter. The resulting multilink is the union of all components of the spliced multilinks except the ones we spliced. Thus the corresponding holomorphic function is nothing but the product of the algebraic functions corresponding to branches. Since the spliced multilink (I-I) is the above multilink where some of the link components are oriented negatively, it becomes real algebraic when $q > p$ and the corresponding real algebraic map is the map where we take the conjugate of the algebraic functions corresponding to the branches that are oriented negatively. The real algebraic map corresponding to this graph multilink is of the form $f\bar{g}$ and is given by

$$\bar{x}y\prod_{i=1}^{u}(x^p+\eta^i y^{p-1})\prod_{j=u+1}^{2u}\overline{(x^p+\eta^j y^{p-1})}\prod_{i=1}^{v}(x^q+\eta^i y^{q-1})\prod_{j=u+1}^{2v}\overline{(x^q+\eta^j y^{q-1})}. \tag{3-1}$$

Thus Theorem 3.1 assures real algebraicity of the open book. Similarly, the graph multilink (I-I-I) is real algebraic when $p < q < r$ and the multilink (III-I) is real algebraic when $p > 3$.

In [15] it is proven that if the link components of a fibered multilink in a homology 3-sphere are canonically oriented (or all those orientations are reversed), then the multilink is the binding of an open book which supports a tight contact structure; otherwise the supported contact structure is overtwisted. So one can conclude that the Milnor open books of the real algebraic links we have constructed so far support overtwisted contact structures in S^3.

4 Calculation of the 3-dimensional invariant from open books

In this section we recall how to detect the overtwisted contact structures compatible with the Milnor fibered multilinks constructed in the previous sections using the monodromy data.

Recall that two overtwisted contact structures on S^3 are contact isotopic if and only if they are homotopic as 2-plane fields [8]. Moreover the homotopy class of a 2-plane field is determined by the induced spin^c structure and the d_3 invariant (see [12; 23]). Since S^3 has a unique spin^c structure, the overtwisted structures on S^3 are classified by their d_3 invariants, which take values in $\mathbb{Z}+\frac{1}{2}$ (see eg [6]). There may be various ways to compute the d_3 invariant of a given contact structure. One can even compute the enhanced Milnor number as explained in [17] or in a way similar to [14] (in the latter the real algebraic functions are so-called "convenient" while ours in Section 3 are not). Here, bearing in mind the fillings of contact 3-manifolds, we will use the method in [11] to calculate d_3 from the monodromy data of the compatible open book.

It is known that given an achiral Lefschetz fibration on a 4-manifold W with fibers F with boundary, W can be described as $F\times D^2$ with 2-handles attached to some vanishing cycles γ_i with appropriate framings. The Lefschetz fibration on W induces an open book decomposition and hence a contact structure on ∂W. The contact structure induced on ∂W is obtained by contact $(+1)/(-1)$-surgeries on the Legendrian realizations of the vanishing cycles of respectively negative/positive critical points, each embedded in distinct fibers of the open book; the contribution to the monodromy is respectively a left/right handed Dehn twist about the vanishing cycle. In the reverse direction given a 3-manifold with an open book decomposition, the monodromy data determines an achiral Lefschetz fibration on a 4-manifold which on the boundary gives the given open book.

It should be noted that 2-handle attachments with (-1) framing result in an honest Lefschetz fibration carrying a natural almost complex structure which is the extension of the one on $D^2\times F$. However, attaching a 2-handle with $(+1)$ framing gives an achiral Lefschetz fibration which does not have a natural almost complex structure that comes from extending the older one. It is shown in [6] that if W_0 is the handlebody decomposition of the 4-manifold admitting the Lefschetz fibration constructed via k $(+1)$-surgeries, $W=W_0\,\#\,k\mathbb{C}P^2$ (with the same boundary) has a natural almost complex structure. When the second cohomology has no torsion (where W is assumed to have no 1-handles) one has the following formula (see [10] or [11]) which is the generalization of the similar statement in [6]:

$$d_3(\xi)=\tfrac{1}{4}(c^2(W)-2\chi(W)-3\sigma(W))+k. \tag{4-1}$$

Here $\sigma(W)$ and $\chi(W)$ are the signature and the Euler characteristic of W. The Chern class $c \in H^2(W;\mathbb{Z})$ is the Poincaré dual to $\sum_{i=1}^{n} r(\gamma_i)C_i$ where C_i is the cocore of the 2-handle attached along the vanishing cycle γ_i, and $r(\gamma_i)$ is the rotation number of γ_i. Since $c(W)|_{\partial W} = c(\xi)$ is zero, $c(W) \in H^2(W)$ comes from a class in $H^2(W, \partial W)$ thus can be squared. A way to calculate $r(\gamma_i)$ on a page is explained in [11] in detail. The rotation number is equal to the winding number of the projection of the curve to a page with respect to the orientation on the Kirby diagram obtained by the usual orientation of D^2 extended over 1-handles.

5 Proof of Theorem 1.1

We have seen that the multilinks (I), (II) and (III) in Figure 1 are fibered with planar pages (see Section 2.1) and are real algebraic via functions of the form $f\bar{g}$ while the multilink (IV) is not (see Section 3). Splicing together these multilinks in the forms (III-I), (I-I), (I-I-...), (III-I-I-...) leads wider families of planarly fibered multilinks (Section 2.2) which are also real algebraic via functions of the form $f\bar{g}$ (Section 3). Our ongoing discussion shows that these are all possible fibered multilinks which are real algebraic via functions of the form $f\bar{g}$. Moreover there is no other fibered multilink in S^3 with planar pages. In fact, for a fibered Seifert multilink with n components the Euler number of a page F is $\chi(F) = |l| \cdot \bigl(2-k+\sum_{j=n+1}^{k} 1/\alpha_j\bigr)$ with $k \geq n$ and $\alpha_j \geq 1$ [7, page 91]. In order to have F planar, $\chi(F)$ must equal $2-n$. Equating, we get either $n=2$ or $|l|=1$. In both cases k is arbitrary and $\alpha_j = 1$ for all $n < j \leq k$. The case $n=2$ gives nothing but a Hopf link in S^3. The latter case where $|l|=1$ is all that appear in Figure 1.

Furthermore we have noted that the corresponding contact structures are overtwisted (see Section 3). In this section, we calculate their d_3 invariants and show what overtwisted contact structures on S^3 are supported by those real algebraic planar open books. We only focus on the graph multilinks (I-I), (III-I) and (I-I-I) as the families (III-I-I-...) and (I-I-...) with larger d_3 invariants do not provide different contact structures. This discussion will be tied in Section 5.7 to prove Theorem 1.1.

5.1 Overtwisted structures via (I)

We first consider the family of multilinks of type (I). Recall that the open books that they determine have pages $(2u+2)$ times punctured spheres (denoted by $\Sigma_{0,2u+2}$). Moreover the monodromy (2-1) of the open book is

$$\phi = a^p \cdot b^{-(p-1)} \cdot c_1^{-1} \cdots c_u^{-1} \cdot d_1^1 \cdots d_u^1,$$

where a, b and c are boundary parallel curves. Observe that the number of negative Dehn twists in this expression is $p+u-1$.

As we discussed in Section 4, via the monodromy information of the given open book decomposition we can construct a 4-manifold with boundary S^3 as the underlying space of an achiral Lefschetz fibration. In that way we can calculate the d_3 invariant of the overtwisted contact structure on S^3 supported by the open book. Now, since the pages have $(2u+2)$ boundary components, we first attach $(2u+1)$ 1-handles

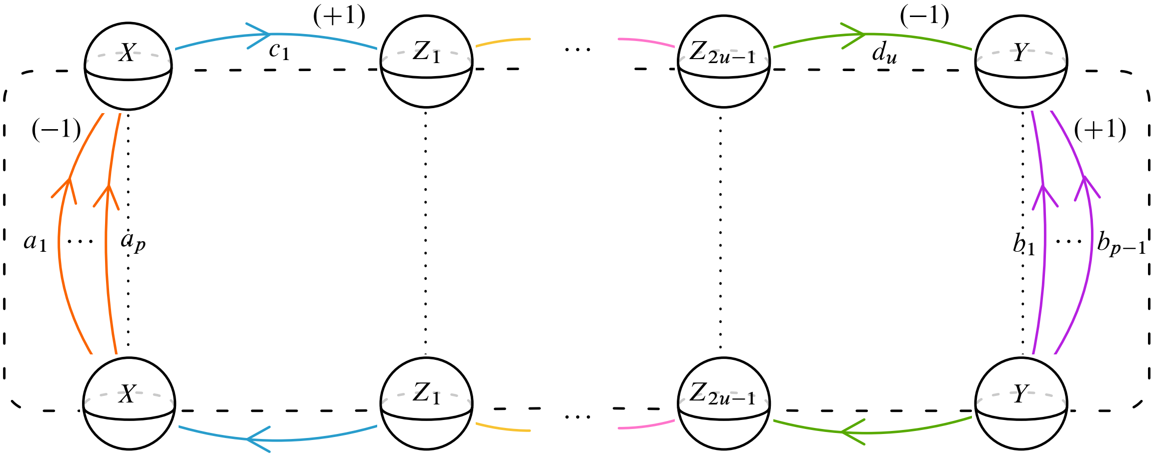


Figure 5: Kirby diagram for the 4-manifold corresponding to (I).

to D^4 to get $D^2 \times \Sigma_{0,2u+2}$. Then, we attach 2-handles along Legendrian copies of boundary parallel curves on $\Sigma_{0,2u+2}$ with framing ± 1, depending on the parity of the Dehn twist. The resulting 4-manifold W is given in Figure 5.

The 1-chain group $C_1(W)$ of W has a basis $\{X, Y, Z_1, \dots, Z_{2u-1}\}$ and $C_2(W)$ has a basis

$$\{a_1, \dots, a_p, b_1, \dots, b_{p-1}, c_1, \dots, c_u, d_1, \dots, d_u\}.$$

The boundary map $D\colon C_2(W) \to C_1(W)$ is given by

$$\begin{aligned}
& & D(a_j) &= X, & j &= 1, \dots p,\\
& & D(b_j) &= Y, & j &= 1, \dots p-1,\\
D(c_1) &= Z_1 - X, & D(c_i) &= Z_i - Z_{i-1}, & i &= 2, \dots u,\\
D(d_u) &= Y - Z_{2u-1}, & D(d_i) &= Z_{u+i} - Z_{u+i-1}, & i &= 1, \dots u-1.
\end{aligned}$$

Thus, $H_2(W)$ has a basis with generators

$$\left\{ a_1 - a_2, \dots, a_{p-1} - a_p, b_1 - b_2, \dots, b_{p-2} - b_{p-1}, b_1 - \sum_{i=1}^{u} (c_i + d_i) - a_p \right\}.$$

Since rank $H_0 = 1$, rank $H_1 = 0$ and rank $H_2 = 2p-2$, we get $\chi(W) = 2p-1$.

Note that $a_j^2 = -1 = d_j^2$ and $b_j^2 = 1 = c_j^2$. So the squares of the basis elements are $(a_j - a_{j+1})^2 = -2$, $(b_j - b_{j+1})^2 = 2$ and $\bigl(b_1 - \sum_{i=0}^{u}(c_i + d_i) - a_p\bigr)^2 = 0$. Thus in this basis the intersection matrix is Q_{I} as given in Appendix A. We also compute in Appendix B that $\sigma(W) = \sigma(Q_{\mathrm{I}}) = 0$, and $\det Q_{\mathrm{I}} = (-1)^{p-1}$.

To calculate the square of the first Chern class, we chose an orientation of the curves and compute the rotation numbers of the curves with respect to the orientation induced from blackboard. Thus we get $r(a) = 0 = r(b)$, $r(c_i) = -1$ and $r(d_i) = -1$. Note that the calculation of c^2 is independent of the chosen orientations. Let us denote the cocores of the 2-handles attached along a_i, b_j, c_k and d_l by A_i,

B_j, C_k and D_l respectively. Then $c(W)$ is Poincaré dual to $-\bigl(\sum_{i=1}^{u} C_i + \sum_{j=1}^{u} D_j\bigr)$. This evaluates on the basis above as $w = (0, \dots, 2u)^T$. Hence,

$$c^2(W) = Q_W(PD(c(W))) = w^T Q^{-1} w = \frac{4u^2 \cdot (-1)^{p-1} \cdot (p-1) \cdot p}{(-1)^{p-1}} = 4u^2 p(p-1).$$

Inserting the results of the previous steps in (4-1) we get

$$d_3(\xi) = \tfrac{1}{4}\bigl(4u^2(p-1)p - 2(2p-1) - 3 \cdot 0\bigr) + (p+u-1) = u^2 p(p-1) + u - \tfrac{1}{2}. \tag{5-1}$$

5.2 Overtwisted structures via (II)

We perform similar calculation for the multilinks (II) given in Figure 1. The associated monodromy (2-2) has $q+u$ negative Dehn twists. After following the same steps to construct the 4-manifold W we find $\chi(W) = q+1$, and as pointed out in Appendix B, $\sigma(W) = q$. Similarly as before, we have

$$c^2(W) = (2u-1)^2 q.$$

Inserting in (4-1) we get

$$d_3(\xi) = \tfrac{1}{4}\bigl((2u-1)^2 q - 2(q+1) - 3q\bigr) + (q+u) = u(u-1)q + u - \tfrac{1}{2}. \tag{5-2}$$

5.3 Overtwisted structures via (III)

As for the multilinks (III) in Figure 1, the associated monodromy (2-3) has $u+1$ negative Dehn twists. The constructed 4-manifold W has $\chi(W) = 5$, and as pointed out in Appendix B, $\sigma(W) = -2$. Moreover,

$$c^2(W) = \frac{(2u+1)^2 \cdot -6}{-1} = 6(2u+1)^2.$$

Inserting in (4-1) we get

$$d_3(\xi) = \tfrac{1}{4}\bigl(6(2u+1)^2 - 2 \cdot 5 - 3 \cdot (-2)\bigr) + (u+1) = 6u(u+1) + u + 2 - \tfrac{1}{2}. \tag{5-3}$$

5.4 Overtwisted structures via (I-I)

We consider the graph multilinks (I-I) obtained by splicing two multilinks of type (I), as we have constructed in Figure 2, top left. The monodromy (2-5) of the associated open book has $q+u+v-1$ negative Dehn twists.

Since the monodromy is obtained by the monodromies of the splice components, to construct the 4-manifold, we can use the Kirby diagrams for the splice components. One can see that the Kirby diagram of the spliced multilink can be constructed as follows. We identify the 1-handles corresponding to the spliced boundary components, thus the 2-handles whose attaching circles corresponds to the Dehn twists along that boundary components cancel. By means of the new Dehn twist contributions to the monodromy, we

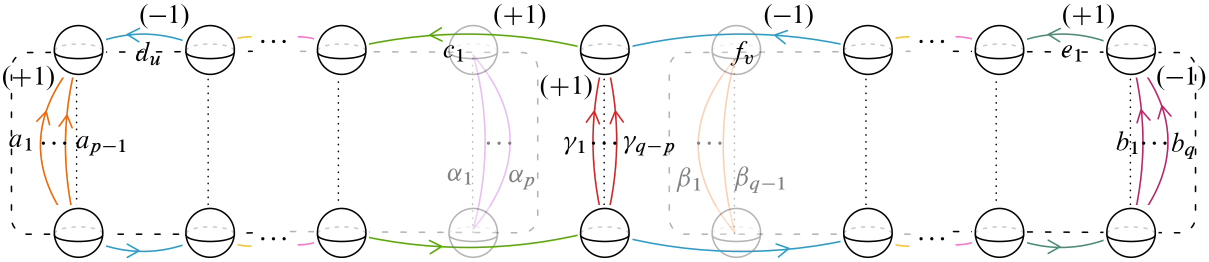


Figure 6: Kirby diagram for the 4-manifold corresponding to (I-I). The faded ends of the previous diagrams are the deleted blocks.

add new 2-handles whose attaching circles are along the identified boundary component. Consequently, we see that the corresponding 4-manifold has the Kirby diagram given in Figure 6.

Furthermore, $H_2(W)$ has a basis with generators

$$a_1-a_2,\dots,a_{p-2}-a_{p-1},\gamma_1-\gamma_2,\dots,\gamma_{k-1}-\gamma_k,b_1-b_2,\dots,b_{q-1}-b_q,$$
$$\gamma_1+\left(\sum_{i=1}^{u}c_i+d_i\right)-a_{p-1},b_1+\left(\sum_{i=1}^{v}e_i+f_i\right)-\gamma_k.$$

Since rank $H_0=1$, rank $H_1=0$ and rank $H_2=2q-2$, we have $\chi(W)=2q-1$.

Note that, $a_j^2=c_j^2=e_j^2=\gamma_j^2=1$ and $b_j^2=d_j^2=f_j^2=-1$. So the squares of the basis elements are $(a_j-a_{j+1})^2=2$, $(\gamma_j-\gamma_{j+1})^2=2$, $(b_j-b_{j+1})^2=-2$, $\big(\gamma_1+\big(\sum_{i=1}^{u}c_i+d_i\big)-a_{p-1}\big)^2=2$ and $\big(b_1+\big(\sum_{i=1}^{v}e_i+f_i\big)-\gamma_k\big)^2=0$. In this basis the intersection matrix is $Q_{\text{I-I}}$ as given in Appendix A. We compute in Appendix B that $\sigma(W)=\sigma(Q_{\text{I-I}})=0$, and det $Q_{\text{I-I}}=(-1)^{q-1}$.

Note that, $r(a)=r(\gamma)=r(b)=0$, $r(c_i)=-1$, $r(d_i)=-1$, $r(e_i)=-1$ and $r(f_i)=-1$. Therefore,

$$c(W)=-\sum_{i=1}^{u}(C_i+D_i)-\sum_{j=1}^{v}(E_j+F_j).$$

This evaluates on the basis above as $w=(0,\dots,-2u,-2v)^T$. In order to calculate c^2, it is sufficient to calculate the inverse of last 2×2 block of $Q_{\text{I-I}}$. We deduce that

$$c^2(W)=4u^2p(p-1)+8uvq(p-1)+4v^2q(q-1).$$

Explicit calculations can be found in Appendix B.

Inserting all these results in (4-1) we get

$$\begin{aligned}d_3(\xi)&=\tfrac{1}{4}(4u^2p(p-1)+8uvq(p-1)+4v^2q(q-1)-2(2q-1)-3\cdot 0)+q+u+v-1\\&=u^2p(p-1)+v^2q(q-1)+2uvq(p-1)+u+v-\tfrac{1}{2}.\end{aligned}$$

As we have seen, the information about the resulting graph link and its fibration can be deduced from the splice components easily. In the next example, we will construct a wider family of overtwisted contact structures and observe how the procedure goes on.

5.5 Overtwisted structures via (I-I-I)

We consider the graph multilinks (I-I-I) obtained by splicing three multilinks of type (I), as we have constructed in Figure 3. The monodromy (2-6) of the associated open book has $r+u+v+w-1$ negative Dehn twists. By the same arguments as in the previous example, the corresponding 4-manifold has the Kirby diagram given in Figure 7.

Then $H_2(W)$ has a basis with generators

$$a_1-a_2,\dots,a_{p-2}-a_{p-1},\gamma_1-\gamma_2,\dots,\gamma_{q-p-1}-\gamma_{q-p},$$
$$\theta_1-\theta_2,\dots,\theta_{r-q-1}-\theta_{r-q},b_1-b_2,\dots,b_{r-1}-b_r,\gamma_1+\Bigl(\sum_{i=1}^{u}c_i+d_i\Bigr)-a_{p-1},$$
$$\theta_1+\Bigl(\sum_{i=1}^{v}e_i+f_i\Bigr)-\gamma_{q-p},b_1+\Bigl(\sum_{i=1}^{w}g_i+h_i\Bigr)-\theta_{r-q}.$$

Since rank $H_0=1$, rank $H_1=0$ and rank $H_2=2r-2$, we have $\chi(W)=2r-1$.

Note that $a_j^2=c_j^2=e_j^2=g_j^2=\gamma_j^2=\theta_j^2=1$ and $b_j^2=d_j^2=f_j^2=h_j^2=-1$. So the squares of the basis elements are

$$(a_j-a_{j+1})^2=2,\quad (\gamma_j-\gamma_{j+1})^2=2,\quad (\theta_j-\theta_{j+1})^2=2,\quad (b_j-b_{j+1})^2=-2,$$
$$\Bigl(\gamma_1-\Bigl(\sum_{i=1}^{u}c_i+d_i\Bigr)-a_{p-1}\Bigr)^2=2,\quad \Bigl(\theta_1-\Bigl(\sum_{i=1}^{v}e_i+f_i\Bigr)-\gamma_{q-p}\Bigr)^2=2,\quad \Bigl(b_1-\Bigl(\sum_{i=1}^{w}g_i+h_i\Bigr)-\theta_{r-q}\Bigr)^2=0.$$

In this basis the intersection matrix is $Q_{\text{I-I-I}}$ as given in Appendix A. We compute in Appendix B that $\det Q_{\text{I-I-I}}=(-1)^{q-1}$.

As we discussed in the previous example the number of positive eigenvalues is

$$(p-2)+(q-p-1)+(r-q-1)+3=r-1$$

and the number of negative eigenvalues is $(r-1)$. Thus, $\sigma(W)=0$.

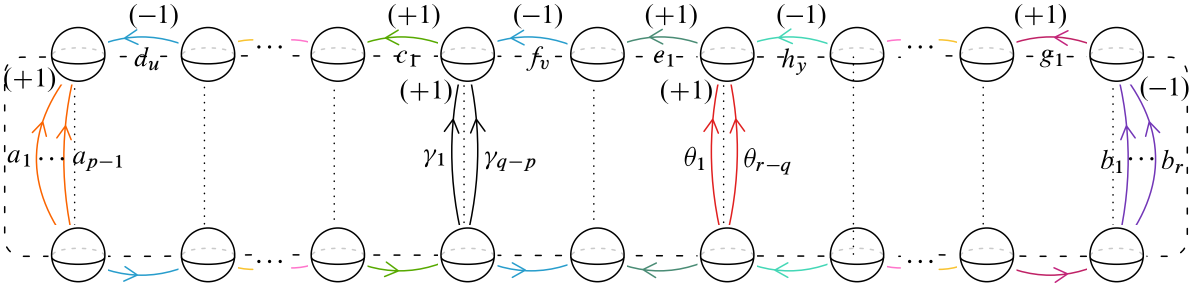


Figure 7: Kirby diagram for the 4-manifold corresponding to (I-I-I).

Note that, $r(a) = r(b) = r(\gamma) = r(\theta) = 0$, whereas $r(c_i) = -1$, $r(d_i) = -1$, $r(e_i) = -1$, $r(f_i) = -1$, $r(g_i) = -1$ and $r(h_i) = -1$. Therefore, we have

$$c(W) = -\sum_{i=1}^{u}(C_i + D_i) - \sum_{j=1}^{v}(E_j + F_j) - \sum_{j=1}^{w}(G_j + H_j).$$

This evaluates on the basis above as $w = (0, \dots, -2u, -2v, -2w)^T$. In order to calculate c^2, it is sufficient to calculate the inverse of the last 3×3 block of $Q_{\text{I-I-I}}$. The calculations in Appendix B show

$$c^2(W) = 4u^2 p(p-1) + 4v^2 q(q-1) + 4w^2 r(r-1) + 8uvq(p-1) + 8uwr(p-1) + 8vwr(q-1).$$

Inserting in (4-1) we get

$$\begin{aligned} \text{(5-4)} \quad d_3(\xi) &= \tfrac{1}{4}\big(4u^2p(p-1)+4v^2q(q-1)+4w^2r(r-1)+8uvq(p-1) \\ &\qquad +8uwr(p-1)+8vwr(q-1)-2\cdot(2r-1)-3\cdot(0)\big)+(r+u+v+w-1) \\ &= u^2p(p-1)+v^2q(q-1)+w^2r(r-1) \\ &\qquad +2uvq(p-1)+2uwr(p-1)+2vwr(q-1)+u+v+w-\tfrac{1}{2}. \end{aligned}$$

5.6 Overtwisted structures via (III-I)

We consider the graph multilinks (III-I) obtained by splicing two multilinks of type (III) and (I), as constructed in Figure 2, bottom left. The monodromy (2-7) of the associated open book has $p+u+v-2$ negative Dehn twists. $H_2(W)$ has a basis with generators

$$a_1 - a_2, \dots, a_{p-1} - a_p, \gamma_1 - \gamma_2, \dots, \gamma_{p-4} - \gamma_{p-3}, b_1 - b_2,$$
$$\gamma_1 - \left(\sum_{i=1}^{u} c_i + d_i\right) - a_p, b_1 - \sum_{i=1}^{v+1} e_i - \sum_{i=1}^{v} f_i - \gamma_{p-3}.$$

In this basis the intersection matrix is $Q_{\text{III-I}}$ as given in Appendix A. Similar calculations as before show that $\chi(W) = 2p-1$, $\sigma(W) = -2$, $\det Q_{\text{III-I}} = (-1)^p$ and

$$c^2(W) = (2u, 2v+1)\begin{pmatrix} p(p-1) & 2p \\ 2p & 6 \end{pmatrix}(2u, 2v+1)^T = 4u^2p(p-1) + 24v^2 + 8up(2v+1) + 24v + 6.$$

Inserting the results of the previous steps into the formula of d_3 invariant, we obtain

$$\begin{aligned} d_3(\xi) &= \tfrac{1}{4}(4u^2p(p-1) + 24v^2 + 8up(2v+1) + 24v + 6 - 2(2p-1) - 3(-2)) + p + u + v - 2 \\ &= u^2p(p-1) + 6v^2 + 2pu(2v+1) + 6v + u + v + 2 - \tfrac{1}{2}. \end{aligned}$$

5.7 Proof of the main theorem

Finally here we prove our main theorem by showing first that the family of fibered multilinks we obtained by splicing (I-I-I) gives us all the overtwisted contact structures with $d_3 + \frac{1}{2} \geq 431$ except $d_3 + \frac{1}{2} = 461$. Then we show that all the remaining ones, except for the ones with

$$d_3 + \tfrac{1}{2} \in \{4, 5, 9, 11, 17, 19, 25, 37, 47, 61, 79, 95, 109\},$$

are obtained by the other ways of splicing that we have presented in the previous paragraphs of the present section. We will give a list for that at the end of the section. We do not know yet if the 13 overtwisted structures that we have missed are real algebraic.

Let $d \in \mathbb{Z}$ denote the sum $d_3 + \frac{1}{2}$ in (5-4),

$$d = u^2p(p-1) + v^2q(q-1) + w^2r(r-1) + 2uvq(p-1) + 2uwr(p-1) + 2vwr(q-1) + u + v + w,$$

where the variables are positive integers with the algebraicity condition $p < q < r$. We fix $v = w = 1$ once and for all. We will use the three moves below:

(i) Replacing q and r with $(q+1)$ and $(r-1)$; this increases d by 2.

(ii) As long as $p = 2$ replacing u with $(u+2)$ and r with $(r-2)$; this increases d by $4u + 12$.

(iii) When $p = 2$, increasing r by 1; this increases d by $2(r + u + q - 1)$.

We start from the *state* $(p, q, r, u, v, w) = (2, 3, r, 1, 1, 1)$. These values give $d = r^2 + 5r + 17$, which is odd. Any application of the moves above produces an odd number. First we will tell how to obtain all odd integers greater than 431 (except 461) via these moves.

Starting from the initial state and applying the move (iii) for each r increases the sum by $2r + 6$. We discuss how to obtain any odd number between $d = r^2 + 5r + 17$ and $d + 2r + 6 = (r+1)^2 + 5(r+1) + 17$ using the first two moves, provided that r is large enough.

Now starting from the initial state the application of (ii) k times increases d by $4k^2 + 12k$ Let k be the largest integer satisfying $4k^2 + 12k < 2r + 6$. Note that we have $k = 1$ for $5 < r \leq 17$, $k = 2$ for $17 < r \leq 33$ and $k = 3$ for $33 < r \leq 53$.

Furthermore any odd number between $d + 4c^2 + 12c$ and $d + 4(c+1)^2 + 12(c+1)$ for $0 \leq c < k$ can be obtained by applying move (i) $\frac{1}{2}(8c + 16) - 1 = 4c + 7$ times. Recall that we have the restriction $q < r$ and that application of moves (i) and (ii) decreases r. Hence in order to obtain all the values in between we must have $q + 4c + 7 < (r - 2c) - 4c - 7$, ie $r > 10c + 17$ for any $0 \leq c < k$.

When $c = 0$, any odd number between d and $d + 16$ can be obtained by applying move (i) 7 times. Therefore, we have the restriction that $q + 7 < r - 7$, hence $r > 17$.

We have observed above that for $17 < r \leq 33$, the move (ii) is applied twice. Hence for $c = 1$ any odd number between $d + 16$ and $d + 40$ can be obtained for $r > 27$. For $17 < r \leq 27$, there are few values less than $d + 40$ that we cannot obtain in this way.

As for $33 < r \leq 53$, we can apply move (ii) thrice. Since $r > 27$, we have observed above that any sum between d and $d + 40$ can be obtained. For $c = 2$ for the numbers between $d + 40$ and $d + 72$, we must have $r > 37$. Hence whenever $37 < r \leq 53$ we can obtain any odd number between d and $d + 72$. For $33 < r \leq 37$ we cannot obtain all the numbers in between though. For larger r (more precisely for $r > 37$) the inequality $r > 10c + 17$ is always satisfied so that we can obtain any odd number between $d + 4c^2 + 12c$ and $d + 4(c+1)^2 + 12(c+1)$.

Finally in order to obtain any odd number between $d + 4k^2 + 12k$ and $d + 2r + 6$ via move (i), we must have

$$r - 2k - \left(\frac{2r + 6 - 4k^2 - 12k}{2} - 1\right) > 3 + \frac{2r + 6 - 4k^2 - 12k}{2} - 1,$$

ie $r < 4k^2 + 10k - 7$. Recall that k is the largest integer satisfying $4k^2 + 12k < 2r + 6$. Comparing these inequalities, one can see that when $r > 33$ any odd number in between can be obtained via move (i). As a result we conclude that for $r \geq 18$, ie starting from $d = 431$ all the odd integers are obtained, except some finitely many missed ones for $29 \leq r \leq 37$. Precisely the number of these missed ones is 45.

Here one can find the exact states that give these missing numbers on a computer. Instead we try to enrich our set of moves in order to obtain most of these 45 numbers. Indeed, at the state $(2, 3, r - 2, 3, 1, 1)$ when we have the sum $d + 16$, we increase p and q by 1, decrease r by 4 to get to the state $(3, 4, r - 6, 3, 1, 1)$ and the sum $d + 36$. Then applying the move (i) successively produces the missing numbers between $d + 36$ and $d + 40$. Thereby, we can obtain 18 out of 30 missing odd numbers between $17 < r \leq 27$. For the 15 missing odd numbers between $29 \leq r \leq 37$, we replace p, q, r by $p + 2$, $q + 5$ and $r - 11$ at the state $(2, 3, r - 2, 3, 1, 1)$ to get to the state $(4, 8, r - 13, 3, 1, 1)$ and the sum $d + 62$. Again, the application of the move (i) successively produces all the odd numbers between $d + 62$ and $d + 2r + 6$. For the remaining 12 missing odd numbers smaller than $d + 36$, at the state $(2, 3, r - 2, 3, 1, 1)$ we replace u with $u + 4$, and r with $r - 5$. Application of this move increases the sum by $6u - 2r + 30$, thus we can obtain all the odd numbers except $d = 461$.

To obtain even numbers, we start from the state $(2, 3, r, 2, 1, 1)$ that gives the even integer $d = r^2 + 7r + 30$. Then move (iii) increases the sum by $2r + 8$. We will now obtain any even number between d and $d + 2r + 8 = (r + 1)^2 + 7(r + 1) + 30$ by applying the first two moves. Let k be the largest integer satisfying $4k^2 + 16k < 2r + 8$. Applying (ii) k times takes us to the state $(2, 3, r - 2k, 2 + 2k, 1, 1)$ and increases the value by $4k^2 + 16k$. Each application of (ii), while passing from the step $u + 2k$ to $u + 2(k + 1)$, increases the value by $8k + 20$. Note that $k = 1$ for $6 < r \leq 20$, we have $k = 2$ for $20 < r \leq 38$ and $k = 3$ for $38 < r \leq 60$.

Any number between $d + 4c^2 + 16c$ and $d + 4(c + 1)^2 + 16(c + 1)$ for $0 \leq c < k$ can be obtained by applying move (i) $4c + 9$ times. In order to obtain all the sums in between we must have $r > 10c + 21$ for any $0 \leq c < k$. When $c = 0$ any even number between d and $d + 20$ can be obtained by applying move (i) 9 times for $r > 21$. We observed above that for $20 < r \leq 38$ we apply (ii) twice. Hence any even number between $d + 20$ and $d + 48$ (ie for $c = 1$) can be obtained whenever $r > 31$. For $20 < r \leq 31$ there are few values less than $d + 48$ that we cannot obtain in this way.

For $38 < r \leq 60$ we can apply (ii) thrice. Any sum between d and $d + 48$ can be obtained as discussed in the previous arguments. For $c = 2$, for the numbers between $d + 48$ and $d + 84$ we must have $r > 41$. As a result, when $41 < r \leq 60$, we can obtain any even number between d and $d + 84$. Moreover, larger r values always satisfy $r > 10c + 21$ and we can obtain any even number between $d + 4c^2 + 16c$ and $d + 4(c + 1)^2 + 16(c + 1)$.

	state	$d_3+\frac{1}{2}$	state	$d_3+\frac{1}{2}$	state	$d_3+\frac{1}{2}$	state	$d_3+\frac{1}{2}$
Type I, (p,u)	(2,1)	3	(5,1)	21	(2,5)	55	(10,1)	91
	(3,1)	7	(6,1)	31	(8,1)	57	(11,1)	111
	(4,1)	13	(7,1)	43	(9,1)	73	(12,1)	133
Type II, (q,u)	(2,1)	1	(5,3)	33	(7,3)	45	(12,3)	75
	(4,3)	27	(6,3)	39	(10,3)	63	(15,3)	93
Type III, (u)	(0)	2	(1)	15	(3)	77		
Type III-I, (p,u,v)	(4,1,0)	23	(2,2,1)	49	(10,1,0)	113	(4,5,0)	347
	(2,3,0)	35	(7,1,0)	59				
Type I-I, (p,q,u,v)	(2,3,2,1)	29	(3,5,2,1)	87	(4,5,2,1)	131	(2,6,3,2)	215
	(2,4,2,1)	39	(3,4,1,2)	89	(2,7,4,1)	135	(2,14,2,1)	249
	(2,3,1,2)	41	(2,5,4,1)	97	(4,6,2,1)	153	(2,7,3,2)	275
	(2,5,2,1)	51	(3,6,2,1)	105	(4,5,1,2)	155	(2,4,3,4)	313
	(2,6,2,1)	65	(2,6,4,1)	115	(2,8,4,1)	157	(3,11,4,1)	387
	(2,4,1,2)	69	(2,9,2,1)	119	(2,11,2,1)	165	(7,8,1,2)	461
	(3,4,2,1)	71	(2,3,1,4)	127	(3,6,1,2)	177		
	(2,7,2,1)	81	(3,5,1,2)	129	(2,9,4,1)	181		
Type I-I-I, (p,q,r,u,v,w)	(2,3,4,1,1,1)	53	(3,5,8,1,1,1)	201	(2,4,6,2,1,2)	281	(2,6,13,1,1,1)	359
	(2,3,5,1,1,1)	67	(3,6,7,1,1,1)	203	(2,3,14,1,1,1)	283	(2,7,12,1,1,1)	361
	(2,3,6,1,1,1)	83	(4,5,7,1,1,1)	205	(2,4,13,1,1,1)	285	(2,8,11,1,1,1)	363
	(2,4,5,1,1,1)	85	(2,3,6,5,1,1)	207	(2,5,12,1,1,1)	287	(2,9,10,1,1,1)	365
	(2,3,4,3,1,1)	99	(2,3,9,3,1,1)	209	(2,6,11,1,1,1)	289	(2,3,9,3,3,1)	367
	(2,3,7,1,1,1)	101	(2,4,8,3,1,1)	211	(2,3,9,5,1,1)	291	(2,3,14,3,1,1)	369
	(2,4,6,1,1,1)	103	(2,5,7,3,1,1)	213	(2,4,8,5,1,1)	293	(2,4,13,3,1,1)	371
	(3,4,5,1,1,1)	107	(2,3,4,3,3,1)	217	(2,5,7,5,1,1)	295	(2,5,12,3,1,1)	373
	(2,3,5,3,1,1)	117	(2,3,7,1,3,1)	219	(4,6,9,1,1,1)	297	(2,6,11,3,1,1)	375
	(2,3,8,1,1,1)	121	(2,3,12,1,1,1)	221	(2,3,12,3,1,1)	299	(2,7,10,3,1,1)	377
	(2,3,4,1,1,1)	123	(2,4,11,1,1,1)	223	(2,4,11,3,1,1)	301	(2,8,9,3,1,1)	379
	(2,5,6,1,1,1)	125	(2,5,10,1,1,1)	225	(2,5,10,3,1,1)	303	(2,4,5,3,2,2)	381
	(2,3,6,3,1,1)	137	(2,6,9,1,1,1)	227	(2,6,9,3,1,1)	305	(2,3,4,3,5,1)	383
	(2,4,5,3,1,1)	139	(2,7,8,1,1,1)	229	(2,7,8,3,1,1)	307	(2,3,7,1,5,1)	385
	(2,3,5,2,2,1)	141	(3,6,8,1,1,1)	231	(2,4,9,2,2,1)	309	(3,4,10,3,1,1)	389
	(2,3,9,1,1,1)	143	(2,3,7,5,1,1)	233	(2,3,7,2,1,2)	311	(2,3,17,1,1,1)	391
	(2,4,8,1,1,1)	145	(2,4,6,5,1,1)	235	(2,5,7,2,2,1)	315	(2,4,16,1,1,1)	393
	(2,5,7,1,1,1)	147	(2,3,10,3,1,1)	237	(2,3,15,1,1,1)	317	(2,5,15,1,1,1)	395
	(3,4,7,2,1,1)	149	(2,4,9,3,1,1)	239	(2,4,14,1,1,1)	319	(2,6,14,1,1,1)	397
	(3,5,6,1,1,1)	151	(2,5,8,3,1,1)	241	(2,5,13,1,1,1)	321	(2,7,13,1,1,1)	399
	(2,3,7,2,1,1)	159	(2,6,7,3,1,1)	243	(2,6,12,1,1,1)	323	(2,3,4,8,1,2)	401
	(2,4,6,3,1,1)	161	(2,3,4,3,2,2)	245	(2,7,11,1,1,1)	325	(2,9,11,1,1,1)	403
	(2,3,6,2,2,1)	163	(2,3,6,2,1,2)	247	(2,8,10,1,1,1)	327	(2,3,4,7,3,1)	405
	(2,3,10,1,1,1)	167	(2,3,13,1,1,1)	251	(2,6,7,5,1,1)	329	(2,3,15,3,1,1)	407
	(2,4,9,1,1,1)	169	(2,4,12,1,1,1)	253	(3,8,9,1,1,1)	331	(2,4,14,3,1,1)	409
	(2,5,8,1,1,1)	171	(2,5,11,1,1,1)	255	(2,3,13,3,1,1)	333	(2,5,13,3,1,1)	411
	(2,6,7,1,1,1)	173	(2,6,10,1,1,1)	257	(2,4,12,3,1,1)	335	(2,6,12,3,1,1)	413
	(3,5,7,1,1,1)	175	(2,7,9,1,1,1)	259	(2,5,11,3,1,1)	337	(2,7,11,3,1,1)	415
	(4,5,6,1,1,1)	179	(2,3,8,5,1,1)	261	(2,6,10,3,1,1)	339	(2,8,10,3,1,1)	417
	(2,3,8,3,1,1)	183	(2,4,7,5,1,1)	263	(2,7,9,3,1,1)	341	(2,3,4,7,3,1)	419
	(2,4,7,3,1,1)	185	(2,5,6,5,1,1)	265	(2,4,10,2,2,1)	343	(2,3,8,1,5,1)	421
	(2,5,6,3,1,1)	187	(2,3,11,3,1,1)	267	(2,4,5,3,3,1)	345	(2,3,5,2,2,3)	423
	(2,3,5,2,1,2)	191	(2,4,10,3,1,1)	269	(2,5,8,2,2,1)	349	(2,3,10,7,1,1)	425
	(2,3,11,1,1,1)	193	(2,5,9,3,1,1)	271	(2,3,7,1,2,2)	351	(2,4,9,7,1,1)	427
	(2,4,10,1,1,1)	195	(2,6,8,3,1,1)	273	(2,3,16,1,1,1)	353	(2,4,10,1,3,1)	429
	(2,5,9,1,1,1)	197	(2,3,9,1,3,1)	277	(2,4,15,1,1,1)	355		
	(2,6,8,1,1,1)	199	(2,3,5,2,4,1)	279	(2,5,14,1,1,1)	357		

Table 1: How to obtain the overtwisted structures with $d_3+\frac{1}{2}\leq 461$ (except 4, 5, 9, 11, 17, 19, 25, 37, 47, 61, 79, 95 and 109).

Finally in order to obtain any even number between $d+4k^2+16k$ and $d+2r+8$ via move (i), we must have $r<4k^2+14k-9$. Checking for the values of k, one can see that the above condition is satisfied for $r>41$. As a result we conclude that for $r\geq 17$, ie starting from $d=438$ all the even integers are obtained, except some finitely many missed ones. Precisely the number of these missed ones is 78.

We have the following additional operations to produce the missed even numbers. To obtain the ones between $d+14$ and $d+20$, at the state $(2,3,r,2,1,1)$ we increase p and q by 1 and decrease r by 3. The new state $(3,4,r-3,2,1,1)$ gives the sum $d+12$. Then we apply the move (i) successively to produce all the missing even numbers in between. For the missed even numbers between $d+48$ and $d+84$, at the state $(2,3,r-4,6,1,1)$ we decrease u by 4, r by 14 and increase p and q by 6 to get the state $(8,9,r-18,2,1,1)$ and the sum $d+72$. Then applying the move (i) successively produces all the missing even integers between $d+78$ and $d+84$.

Moreover, at the state $(2,3,r-2,4,1,1)$ with $d+20$, in order to obtain the missed ones between $d+20$ and $d+48$, we decrease u by 2, r by 7 and increase p and q by 3 which increases the sum by 16. As before, we can successively apply the move (i) to obtain the missing ones between $d+36$ and $d+48$. However, for the small values of r, 18 of the missing even numbers cannot be obtained because of the restriction $q<r$ in each step. We have realized that 12 of these 18 missing numbers can be produced by the application of the move (i) successively at the states $(2,3,r-7,8,1,1)$ with the sum $d+38$. Lastly, one can see that the remaining sums 520, 558, 714, 766, 820, 876 can be obtained by the states $(5,6,8,1,2,1)$, $(4,6,10,1,2,1)$, $(5,7,10,1,2,1)$, $(5,7,11,1,2,1)$, $(5,7,12,1,2,1)$, $(5,7,13,1,2,1)$.

Up to now we have proved that any overtwisted structure with $d_3+\frac{1}{2}\geq 431$ (except 461) can be obtained by (I-I-I) splicing. Note that for the multilinks (II), the supported contact structures have $d_3+\frac{1}{2}=2q+2$ whenever $u=2$. Therefore the even values of $d_3+\frac{1}{2}$ which are between 6 and 431 can be obtained by the multilinks of type (II). Via computer assistance we find that the ones with all the other smaller d_3's (except $4,5,9,11,17,19,25,37,47,61,79,95,109$) are obtained via splicings as shown in Table 1 (note that usually there is more than one way to construct each case; here we give single samples).

Appendix A Intersection matrices

Here, we give the intersection matrices of the 4-manifolds that we have constructed in Section 5, in the bases we presented there.

Let J_n and $\tilde{J}_n$ be the matrices

$$J_n=\begin{pmatrix} 2 & -1 & 0 & \cdots & 0\\ -1 & 2 & -1 & 0 & \vdots\\ 0 & -1 & \ddots & \cdots & 0\\ \vdots & 0 & \cdots & 2 & -1\\ 0 & \cdots & 0 & -1 & 2\end{pmatrix}_{n\times n}, \qquad \tilde{J}_n=\begin{pmatrix} -2 & 1 & 0 & \cdots & 0\\ 1 & -2 & 1 & 0 & \vdots\\ 0 & 1 & \ddots & \cdots & 0\\ \vdots & 0 & \cdots & -2 & 1\\ 0 & \cdots & 0 & 1 & -2\end{pmatrix}_{n\times n}.$$

Then the intersection matrices for (I), (II) and (III) are

$$Q_{\mathrm{I}} = \left(\begin{array}{c|c|c} \tilde{J}_{p-1} & & \\ & & -1 \\ \hline & & 1 \\ & J_{p-2} & \\ \hline -1 & 1 & 0 \end{array}\right), \quad Q_{\mathrm{II}} = \left(\begin{array}{c|c} J_{q-1} & \\ & 1 \\ \hline 1 & 1 \end{array}\right), \quad Q_{\mathrm{III}} = \begin{pmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 0 & -1 \\ 0 & 0 & -2 & -1 \\ 0 & -1 & -1 & -1 \end{pmatrix}.$$

The intersection matrices for the 4-manifolds obtained for the spliced graph multilinks (I-I) and (I-I-I) are respectively as follows; here $a = -1$ if $q = p+1$ and $a = 0$ for $q > p+1$; $b = -1$ if $r = q+1$ and $b = 0$ for $r > q+1$:

$$Q_{\text{I-I}} = \left(\begin{array}{c|c|c|cc} J_{p-2} & & & & \\ & & & 1 & \\ \hline & & & 1 & \\ & J_{q-p-1} & & & \\ & & & & 1 \\ \hline & & & & -1 \\ & & \tilde{J}_{q-1} & & \\ \hline 1 & 1 & & 2 & a \\ & 1 & -1 & a & 0 \end{array}\right), \quad Q_{\text{I-I-I}} = \left(\begin{array}{c|c|c|c|ccc} J_{p-2} & & & & & & \\ & & & & 1 & & \\ \hline & & & & 1 & & \\ & J_{q-p-1} & & & & & \\ & & & & & 1 & \\ \hline & & & & & 1 & \\ & & J_{r-q-1} & & & & \\ & & & & & & 1 \\ \hline & & & & & & -1 \\ & & & \tilde{J}_{r-1} & & & \\ \hline 1 & 1 & & & 2 & & a \\ & 1 & 1 & & & 2 & b \\ & & 1 & -1 & a & b & 0 \end{array}\right).$$

Finally here is the intersection matrix for the 4-manifold obtained for the spliced graph multilink (III-I); here $a = -1$ if $p = 4$ and $a = 0$ for $p \geq 4$:

$$Q_{\text{III-I}} = \left(\begin{array}{c|c|c|cc} \tilde{J}_{p-1} & & & & \\ & & & -1 & \\ \hline & & & 1 & \\ & J_{p-4} & & & \\ & & & & 1 \\ \hline & & -2 & & -1 \\ \hline -1 & 1 & & 0 & a \\ & 1 & -1 & a & 1 \end{array}\right).$$

Appendix B Determinant, signature and c^2 computation

Here, we give detailed calculations for the results about the intersection matrices we used in Section 5. For practical reference we summarize these results in Table 2.

B.1 The matrix Q_{I}

First we compute the diagonalization of the intersection matrix Q_{I} above. Consider the lower triangular matrix S_n with its ij entry $(i \geq j)$ being equal to j/i. It can be easily seen that $J_n = S_n D_n S_n^T$ and $\tilde{J}_n = S_n(-D_n)S_n^T$ where $D_n = \operatorname{diag}\bigl(2, \frac{3}{2}, \dots, \frac{n+1}{n}\bigr)$. It follows that $\det J_n = n+1$ and $\det \tilde{J}_n = (-1)^n(n+1)$.

Thus we see that $Q_{\mathrm{I}} = SDS^T$ where

$$D = \operatorname{diag}\left(-2, -\frac{3}{2}, \dots, -\frac{p}{p-1}, 2, \frac{3}{2}, \dots, \frac{p-1}{p-2}, \frac{1}{p(p-1)}\right)$$

and

$$S = \left(\begin{array}{ccc|ccc|c} & S_{p-1} & & & & & \\ & & & & & & -1 \\ \hline & & & & & & 1 \\ & & & & S_{p-2} & & \\ & & & & & & \\ \hline -\frac{1}{p} & \cdots & -\frac{p-1}{p} & -\frac{p-2}{p-1} & \cdots & -\frac{1}{p-1} & 1 \end{array}\right).$$

We conclude that $\sigma(Q_{\mathrm{I}}) = 0$ and $\det Q_{\mathrm{I}} = (-1)^{p-1}$.

Similarly, signatures and determinants of the intersection matrices for (II) and (III) can be calculated as given in Table 2.

B.2 The matrix $Q_{\text{I-I}}$

As for the intersection matrix $Q_{\text{I-I}}$ above for the splicing (I-I), we first show that $\det Q_{\text{I-I}} = (-1)^{q-1}$. Then we compute the signature and c^2.

	σ	det	c^2
Q_{I}	0	$(-1)^{p-1}$	$4u^2p(p-1)$
Q_{II}	q	1	$(2u-1)^2q$
Q_{III}	-2	-1	$6(2u+1)^2$
$Q_{\text{I-I}}$	0	$(-1)^{q-1}$	$4u^2p(p-1)+8uvq(p-1)+4v^2q(q-1)$
$Q_{\text{I-I-I}}$	0	$(-1)^{q-1}$	$4u^2p(p-1)+4v^2q(q-1)+4w^2r(r-1)+8uvq(p-1)+8uwr(p-1)+8vwr(q-1)$
$Q_{\text{III-I}}$	0	1	$8u^2+24v^2+32uv+8u+8v$

Table 2: Signatures and determinants of the intersection matrices and the corresponding c^2.

Let J'_n (respectively J''_n) be the matrix obtained by removing the last (respectively the first) column of the matrix J_n.

We assume that $q > p+1$, ie $a = 0$ in $Q_{\text{I-I}}$; it can be shown that the results are the same in the case $q = p+1$. Now we calculate $\det Q_{\text{I-I}}$ via its last row. We observe that it is equal to $(-1)^{q-1}$ times the sum of the determinants

$$\left|\begin{array}{c|c|c|c}
J_{p-2} & & & \begin{matrix} \\ 1\end{matrix} \\ \hline
 & J'_{q-p-1} & & \begin{matrix}1 \\ \\ \hfill 1\end{matrix} \\ \hline
 & & \tilde{J}_{q-1} & \begin{matrix}\hfill -1 \\ \\ \end{matrix} \\ \hline
\hfill 1 & 1 \hfill & & 2
\end{array}\right| + \left|\begin{array}{c|c|c|c}
J_{p-2} & & & \begin{matrix} \\ 1\end{matrix} \\ \hline
 & J_{q-p-1} & & \begin{matrix}1 \\ \\ \hfill 1\end{matrix} \\ \hline
 & & \tilde{J}''_{q-1} & \begin{matrix}\hfill -1 \\ \\ \end{matrix} \\ \hline
\hfill 1 & 1 \hfill & & 2
\end{array}\right|.$$

Now we move the last column of each matrix above to the positions of the removed columns, ie in the first matrix we move the $(2q-3)^{\text{rd}}$ column to the $(q-3)^{\text{rd}}$ position and in the second matrix to the $(q-2)^{\text{nd}}$ position. These row exchanges multiply the determinants by $(-1)^{(2q-3-q+3)}$ and $(-1)^{(2q-3-q+2)}$, respectively. Since

$$\left|\begin{array}{c|c} \multicolumn{2}{c}{J'_n} \\ \hline & 1 \end{array}\right| = \det J_{n-1}, \qquad \left|\begin{array}{c|c} -1 & \\ \hline & \tilde{J}''_n \end{array}\right| = -\det \tilde{J}_{n-1} = (-1)^n n,$$

we have

$$\begin{aligned}
\det Q_{\text{I-I}} &= (-1)^{q-1}\big((-1)^{q-1}\cdot\det J'_{p-2}\cdot\det J'_{q-p-1}\cdot\tilde{J}_{q-1} + (-1)^{q-1}\cdot\det J_{p-2}\cdot\det J'''_{q-p-1}\cdot\det\tilde{J}_{q-1} \\
&\qquad\qquad + (-1)^{q}\cdot 2\cdot\det J_{p-2}\cdot\det J'_{q-p-1}\cdot\det\tilde{J}_{q-1}\big) \\
&\quad + (-1)^{q-1}\big((-1)^{q}\cdot\det J'_{p-2}\cdot\det J_{q-p-1}\cdot\tilde{J}''_{q-1} + (-1)^{q}\cdot\det J_{p-2}\cdot\det J''_{q-p-1}\cdot\det\tilde{J}''_{q-1} \\
&\qquad\qquad + (-1)^{q-1}\cdot 2\cdot\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det\tilde{J}''_{q-1}\big) \\
&= (-1)^q q(q-2) - (-1)^q (q-1)^2 \\
&= (-1)^{q-1}.
\end{aligned}$$

Now we compute the signature of $Q_{\text{I-I}}$. This matrix is of the form

$$\left[\begin{array}{c|c} A & B^T \\ \hline B & C \end{array}\right]$$

where A, B and C are $(2q-4)\times(2q-4)$, $2\times(2q-4)$ and 2×2 symmetric matrices respectively. Let S_1 and S_2 be the orthogonal matrices that diagonalize A and $C - BA^{-1}B^T$ respectively. Define

$$S = \left[\begin{array}{c|c} S_1 & 0 \\ \hline -S_2BA^{-1} & S_2 \end{array}\right].$$

It can be easily seen that

$$SQS^T = \left[\begin{array}{c|c} S_1AS_1^T & 0 \\ \hline 0 & S_2(C-BA^{-1}B^T)S_2^T \end{array}\right].$$

Hence

$$\begin{aligned}\sigma(Q_{\text{I-I}}) = \sigma(SQ_{\text{I-I}}S^T) &= \sigma(S_1AS_1^T) + \sigma(S_2(C-BA^{-1}B^T)S_2^T) \\ &= \sigma(A) + \sigma(C-BA^{-1}B^T).\end{aligned}$$

We know that J_n and $\tilde{J}_n$ are diagonalizable and are positive definite and negative definite respectively. Therefore, A has $(p-2)+(q-p-1) = q-3$ positive and $q-1$ negative eigenvalues. Moreover it is easy to observe that $C-BA^{-1}B^T$ is positive definite, hence has 2 positive eigenvalues. Thus $\sigma(Q_{\text{I-I}}) = 0$.

Now we compute c^2. As we have observed, the basis of $H_2(W;\mathbb{Z})$ given in Section 5.1, $c(W)$ evaluates on as $w = (0,\dots,-2u,-2v)^T$. Hence, in order to calculate c^2, it is sufficient to calculate the inverse of the last 2×2 block of $Q_{\text{I-I}}$. Let D denote this matrix and d_{ij} be its (i,j) entry. We claim

$$d_{11} = \frac{\text{cofac}_{11}}{\det Q_{\text{I-I}}} = p(p-1), \quad d_{12} = d_{21} = q(p-1), \quad d_{22} = q(q-1).$$

To prove these we compute the cofactors explicitly. First,

$$\begin{aligned}\text{cofac}_{11} &= \left|\begin{array}{c|c|c|c} J_{p-2} & & & \\ \hline & J_{q-p-1} & & 1 \\ \hline & & \tilde{J}_{q-1} & -1 \\ \hline & 1 & -1 & 0 \end{array}\right| \\ &= (-1)^q(-1)^{q-1}\det J_{p-2}\cdot\det J'_{q-p-1}\cdot\det\tilde{J}_{q-1} + (-1)^q(-1)^{q-2}\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det\tilde{J}''_{q-1} \\ &= (-1)^q q(p-1)(q-p-1) + (-1)^{q-1}(q-1)(p-1)(q-p) \\ &= (-1)^{q-1}p(p-1).\end{aligned}$$

The following determinants are evaluated by induction:

$$\bar{J}_n = \left|\begin{array}{c|c} & J''_n \\ \hline 1 & \end{array}\right| = (-1)^2\bar{J}_{n-1} = 1 \quad \text{and} \quad \left|\begin{array}{c|c} & J_n \\ \hline 1 & \end{array}\right| = 0.$$

Therefore,

$$\begin{aligned}
\mathrm{cofac}_{12} &= (-1)^{p-1} \left| \begin{array}{c|c|c|c} J'_{p-2} & & & \\ \hline & J_{q-p-1} & & 1 \\ \hline & & \tilde{J}_{q-1} & -1 \end{array} \right| + (-1)^{p} \left| \begin{array}{c|c|c|c} J_{p-2} & & & \\ \hline & J''_{q-p-1} & & 1 \\ \hline & & \tilde{J}_{q-1} & -1 \end{array} \right| \\
&= 0 + (-1)^{p}(-1)^{p-1}(-1)^{q-1} q(p-1) \\
&= (-1)^{q} q(p-1), \\
\mathrm{cofac}_{22} &= \left| \begin{array}{c|c|c|c} J_{p-2} & & & 1 \\ \hline & J_{q-p-1} & & 1 \\ \hline & & \tilde{J}_{q-1} & \\ \hline 1 & 1 & & 2 \end{array} \right| \\
&= (-1)^{p-1}(-1)^{p} \det J'_{p-2} \cdot \det J_{q-p-1} \cdot \det \tilde{J}_{q-1} \\
&\qquad + (-1)^{p}(-1)^{p-1} \det J_{p-2} \cdot \det J''_{q-p-1} \cdot \det \tilde{J}_{q-1} + 2 \det J_{p-2} \cdot \det J_{q-p-1} \cdot \det \tilde{J}_{q-1} \\
&= (-1)^{q-1} q(q-1).
\end{aligned}$$

As a result, we conclude that

$$c^2(W) = (-2u, -2v) \begin{pmatrix} p(p-1) & q(p-1) \\ q(p-1) & q(q-1) \end{pmatrix} (-2u, -2v)^T = 4u^2 p(p-1) + 8uvq(p-1) + 4v^2 q(q-1).$$

B.3 The matrix $Q_{\text{I-I-I}}$

Now we consider the splicing (I-I-I) and the associated intersection matrix $Q_{\text{I-I-I}}$ given in Appendix A. We omit the calculations of the cases when a and b are nonzero since they give the same results. Calculating the determinant of the intersection matrix with $a = b = 0$ with respect to its last three rows as in the previous example we have

$$\begin{aligned}
\det Q_{\text{I-I-I}} &= (-1)^{r-1} r(r-q-1)(q-2) + (-1)^{r-1} r(r-q-2)(q-1) - 2(-1)^{r-1} r(r-q-1)(q-1) \\
&\qquad + (-1)^{r}(r-1)(r-q)(q-2) + (-1)^{r}(r-1)(r-q-1)(q-1) - 2(-1)^{r}(r-1)(r-q)(q-1) \\
&= (-1)^{r-1}.
\end{aligned}$$

Moreover $c(W)$ evaluates on the given basis of $H_2(W;\mathbb{Z})$ as $w = (0,\dots,-2u,-2v,-2y)^T$. In order to calculate c^2, it is sufficient to calculate the inverse of the last 3×3 block D of $Q_{\text{I-I-I}}$. We have

$$D = \begin{pmatrix} p(p-1) & q(p-1) & r(p-1) \\ q(p-1) & q(q-1) & r(q-1) \\ r(p-1) & r(q-1) & r(r-1) \end{pmatrix}.$$

We see that

$$\mathrm{cofac}_{11} = \left|\begin{array}{c|c|c|c|cc} J_{p-2} & & & & & \\ \hline & J_{q-p-1} & & & & \\ & & & & 1 & \\ \hline & & & & 1 & \\ & & J_{r-q-1} & & & \\ & & & & & 1 \\ \hline & & & & & -1 \\ & & & \tilde{J}_{r-1} & & \\ \hline & 1 & 1 & & 2 & \\ & & 1 & -1 & & 0 \end{array}\right|,$$

which can be calculated with respect to the last two rows, yielding

$$\begin{aligned} \mathrm{cofac}_{11} &= \det J_{p-2}\cdot\det J'_{q-p-1}\cdot\det J'_{r-q-1}\cdot\det \tilde{J}_{r-1} + \det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J'''_{r-q-1}\cdot\det \tilde{J}_{r-1} \\ &\quad -2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J'_{r-q-1}\cdot\det \tilde{J}_{r-1} + \det J_{p-2}\cdot\det J'_{q-p-1}\cdot\det J_{r-q-1}\cdot\det \tilde{J}''_{r-1} \\ &\quad +\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J''_{r-q-1}\cdot\det \tilde{J}''_{r-1} - 2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J_{r-q-1}\cdot\det \tilde{J}''_{r-1} \\ &= (-1)^{r-1} r(p-1)(p-r+1) + (-1)^r (r-1)(p-q)(p-r) \\ &= (-1)^{r-1} p(p-1). \end{aligned}$$

Meanwhile,

$$\mathrm{cofac}_{22} = \left|\begin{array}{c|c|c|c|cc} J_{p-2} & & & & & \\ & & & & 1 & \\ \hline & & & & 1 & \\ & J_{q-p-1} & & & & \\ \hline & & J_{r-q-1} & & & \\ & & & & & 1 \\ \hline & & & & & -1 \\ & & & \tilde{J}_{r-1} & & \\ \hline 1 & 1 & & & 2 & \\ & & 1 & -1 & & 0 \end{array}\right|,$$

which can be calculated with respect to the last two rows, yielding

$$\begin{aligned}\operatorname{cofac}_{22} &= \det J'_{p-2}\cdot\det J_{q-p-1}\cdot\det J'_{r-q-1}\cdot\det\tilde J_{r-1}+\det J_{p-2}\cdot\det J''_{q-p-1}\cdot\det J'_{r-q-1}\cdot\det\tilde J_{r-1}\\ &\qquad -2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J'_{r-q-1}\cdot\det\tilde J_{r-1}+\det J'_{p-2}\cdot\det J_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J''_{r-1}\\ &\qquad\quad +\det J_{p-2}\cdot\det J''_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J''_{r-1}-2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J''_{r-1}\\ &= (-1)^r q((p-2)(q-p)+(p-1)(q-p-1)-2(p-1)(q-p)) = (-1)^{r-1}q(q-1).\end{aligned}$$

Similarly,

$$\operatorname{cofac}_{33}=\left|\begin{array}{c|c|c|c|cc}
J_{p-2} &&&&& \\
&&&& 1 & \\ \hline
&&&& 1 & \\
& J_{q-p-1} &&&& \\
&&&&& 1 \\ \hline
&&&&& 1 \\
&& J_{r-q-1} &&& \\ \hline
&&& \tilde J_{r-1} && \\ \hline
1 & 1 &&& 2 & \\
& 1 & 1 &&& 2
\end{array}\right|,$$

which can be calculated with respect to the last two rows, yielding

$$\begin{aligned}\operatorname{cofac}_{33} &= \det J'_{p-2}\cdot\det J'_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J_{r-1}+\det J_{p-2}\cdot\det J'''_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J_{r-1}\\ &\qquad -2\det J_{p-2}\cdot\det J'_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J_{r-1}+\det J'_{p-2}\cdot\det J_{q-p-1}\cdot\det J''_{r-q-1}\cdot\det\tilde J_{r-1}\\ &\qquad +\det J_{p-2}\cdot\det J''_{q-p-1}\cdot\det J''_{r-q-1}\cdot\det\tilde J_{r-1}-2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J''_{r-q-1}\cdot\det\tilde J_{r-1}\\ &\qquad\quad +2\det J'_{p-2}\cdot\det J_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J_{r-1}+2\det J_{p-2}\cdot\det J''_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J_{r-1}\\ &\qquad\qquad +4\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J_{r-q-1}\cdot\det\tilde J_{r-1}\\ &= (-1)^{r-1}r(r-1).\end{aligned}$$

Note also that

$$\bar J_m=\left|\begin{array}{c|c} 1 & \\ \hline & J'_m\end{array}\right| = (-1)^{m-1},$$

which follows by induction. Thus

$$\operatorname{cofac}_{12}=(-1)^{r-1}(-1)^{q-1}\left|\begin{array}{c|c|c|c|cc}
J_{p-2} &&&&& \\
&&&& 1 & \\ \hline
&&&& 1 & \\
& J'_{q-p-1} &&&& \\ \hline
&& J'_{r-q-1} &&& \\
&&&&& 1 \\ \hline
&&&&& -1 \\
&&& \tilde J_{r-1} &&
\end{array}\right.$$

$$
+(-1)^{r-1}(-1)^{q}\left|\begin{array}{c|c|c|c|c}
J_{p-2} & & & & \\
 & & & & 1 \\
\hline
 & & & & 1 \\
 & J_{q-p-1} & & & \\
\hline
 & & J'''_{r-q-1} & & \\
 & & & & 1 \\
\hline
 & & & & -1 \\
 & & & \tilde{J}_{r-1} & \\
\end{array}\right|
$$

$$
+(-1)^{r-1}(-1)^{q-1}\left|\begin{array}{c|c|c|c|c}
J_{p-2} & & & & \\
 & & & & 1 \\
\hline
 & & & & 1 \\
 & J'_{q-p-1} & & & \\
\hline
 & & J_{r-q-1} & & \\
 & & & & 1 \\
\hline
 & & & & -1 \\
 & & & \tilde{J}''_{r-1} & \\
\end{array}\right|
$$

$$
+(-1)^{r-1}(-1)^{q}\left|\begin{array}{c|c|c|c|c}
J_{p-2} & & & & \\
 & & & & 1 \\
\hline
 & & & & 1 \\
 & J_{q-p-1} & & & \\
\hline
 & & J''_{q-r-1} & & \\
 & & & & 1 \\
\hline
 & & & & -1 \\
 & & & \tilde{J}''_{r-1} & \\
\end{array}\right|
$$

$$
=(-1)^{r-1}r(p-1)(r-q-1)+0+(-1)^{r}(r-1)(p-1)(r-q)+0=-(-1)^{r-1}q(p-1).
$$

Similarly,

$$
\mathrm{cofac}_{13}=\left|\begin{array}{c|c|c|c|cc}
J_{p-2} & & & & & \\
 & & & & 1 & \\
\hline
 & & & & 1 & \\
 & J_{q-p-1} & & & & \\
 & & & & & 1 \\
\hline
 & & & & & 1 \\
 & & J_{r-q-1} & & & \\
\hline
 & & & \tilde{J}_{r-1} & & \\
\hline
1 & 1 & & & 2 & \\
 & & 1 & -1 & & 2 \\
\end{array}\right|,
$$

which can be calculated with respect to the last two rows, yielding

$$\begin{aligned}
\operatorname{cofac}_{13} = &-(-1)^{r-1}(-1)^{p-1}\det J_{p-2}\cdot\det \bar{J}'_{q-p-1}\cdot\det \bar{J}'_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
&+(-1)^{r-1}(-1)^{p-1}\det J_{p-2}\cdot\det \bar{J}_{q-p-1}\cdot\det \bar{J}'''_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
&-(-1)^{r-1}(-1)^{q}2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det \bar{J}'_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
&-(-1)^{r-1}(-1)^{p-1}\det J_{p-2}\cdot\det \bar{J}'_{q-p-1}\cdot\det \bar{J}_{r-q-1}\cdot\det \tilde{J}''_{r-1}\\
&+(-1)^{r-1}(-1)^{p-1}\det J_{p-2}\cdot\det \bar{J}_{q-p-1}\cdot\det J''_{r-q-1}\cdot\det \tilde{J}''_{r-1}\\
&-(-1)^{r-1}(-1)^{q}2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det \bar{J}_{r-q-1}\cdot\det \tilde{J}''_{r-1}\\
&+2(-1)^{q-1}(-1)^{p-1}\det J_{p-2}\cdot\det \bar{J}'_{q-p-1}\cdot\det J_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
&+2(-1)^{q}(-1)^{p-1}\det J_{p-2}\cdot\det \bar{J}_{q-p-1}\cdot\det J''_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
&+4\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det J_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
= &\,(-1)^{r-1}r(p-1).
\end{aligned}$$

Finally,

$$\operatorname{cofac}_{23} = \left|\begin{array}{c|c|c|c|cc}
J_{p-2} & & & & & \\
 & & & & 1 & \\
\hline
 & & & & 1 & \\
 & J_{q-p-1} & & & & \\
 & & & & & 1 \\
\hline
 & & & & & 1 \\
 & & J_{r-q-1} & & & \\
 & & & & & \\
\hline
 & & & & & \\
 & & & \tilde{J}_{r-1} & & \\
 & & & & & \\
\hline
1 & 1 & & & 2 & \\
 & & 1 & -1 & & 0
\end{array}\right|$$

which can be calculated with respect to the last two rows, yielding

$$\begin{aligned}
\operatorname{cofac}_{23} = &\,(-1)^{r-1}(-1)^{q-1}\det J'_{p-2}\cdot\det J_{q-p-1}\cdot\det \bar{J}'_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
&+(-1)^{r-1}(-1)^{q-1}\det J_{p-2}\cdot\det J''_{q-p-1}\cdot\det \bar{J}'_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
&-(-1)^{r-1}(-1)^{q-1}2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det \bar{J}'_{r-q-1}\cdot\det \tilde{J}_{r-1}\\
&+(-1)^{r-1}(-1)^{q-1}\det J'_{p-2}\cdot\det J_{q-p-1}\cdot\det \bar{J}_{r-q-1}\cdot\det \tilde{J}''_{r-1}\\
&+(-1)^{r-1}(-1)^{q-1}\det J_{p-2}\cdot\det J''_{q-p-1}\cdot\det \bar{J}_{r-q-1}\cdot\det \tilde{J}''_{r-1}\\
&-(-1)^{r-1}(-1)^{q-1}2\det J_{p-2}\cdot\det J_{q-p-1}\cdot\det \bar{J}_{r-q-1}\cdot\det \tilde{J}''_{r-1}\\
= &-(-1)^{r-1}r(q-1).
\end{aligned}$$

As a result, we conclude that

$$\begin{aligned}c^2(W) &= (-2u,-2v,-2y)\begin{pmatrix} p(p-1) & q(p-1) & r(p-1)\\ q(p-1) & q(q-1) & r(q-1)\\ r(p-1) & r(q-1) & r(r-1)\end{pmatrix}(-2u,-2v,-2y)^T\\ &= 4u^2p(p-1)+4v^2q(q-1)+4y^2r(r-1)+8uvq(p-1)+8uyr(p-1)+8vyr(q-1).\end{aligned}$$

References

[1] **S Akbulut**, **H King**, *All knots are algebraic*, Comment. Math. Helv. 56 (1981) 339–351 MR Zbl

[2] **K Baker**, **J Etnyre**, *Rational linking and contact geometry*, from "Perspectives in analysis, geometry, and topology", Progr. Math. 296, Birkhäuser, New York (2012) 19–37 MR Zbl

[3] **M Bhupal**, **B Ozbagci**, *Milnor open books of links of some rational surface singularities*, Pacific J. Math. 254 (2011) 47–65 MR Zbl

[4] **B Bode**, *Constructing links of isolated singularities of polynomials* $\mathbb{R}^4 \to \mathbb{R}^2$, J. Knot Theory Ramifications 28 (2019) art. id. 1950009 MR Zbl

[5] **C Caubel**, **A Némethi**, **P Popescu-Pampu**, *Milnor open books and Milnor fillable contact* 3*-manifolds*, Topology 45 (2006) 673–689 MR Zbl

[6] **F Ding**, **H Geiges**, **A I Stipsicz**, *Surgery diagrams for contact* 3*-manifolds*, Turkish J. Math. 28 (2004) 41–74 MR Zbl

[7] **D Eisenbud**, **W Neumann**, *Three-dimensional link theory and invariants of plane curve singularities*, Ann. of Math. Stud. 110, Princeton Univ. Press (1985) MR Zbl

[8] **Y Eliashberg**, *Classification of overtwisted contact structures on* 3*-manifolds*, Invent. Math. 98 (1989) 623–637 MR Zbl

[9] **J B Etnyre**, *Planar open book decompositions and contact structures*, Int. Math. Res. Not. 2004 (2004) 4255–4267 MR Zbl

[10] **J B Etnyre**, **T Fuller**, *Realizing* 4*-manifolds as achiral Lefschetz fibrations*, Int. Math. Res. Not. 2006 (2006) art. id. 70272 MR Zbl

[11] **J B Etnyre**, **B Ozbagci**, *Invariants of contact structures from open books*, Trans. Amer. Math. Soc. 360 (2008) 3133–3151 MR Zbl

[12] **R E Gompf**, *Handlebody construction of Stein surfaces*, Ann. of Math. 148 (1998) 619–693 MR Zbl

[13] **M Hedden**, *Some remarks on cabling, contact structures, and complex curves*, from "Proceedings of Gökova Geometry-Topology Conference 2007", GGT, Gökova, Turkey (2008) 49–59 MR Zbl

[14] **K Inaba**, *On the enhancement to the Milnor number of a class of mixed polynomials*, J. Math. Soc. Japan 66 (2014) 25–36 MR Zbl

[15] **M Ishikawa**, *Compatible contact structures of fibered Seifert links in homology* 3*-spheres*, Tohoku Math. J. 64 (2012) 25–59 MR Zbl

[16] **J Milnor**, *Singular points of complex hypersurfaces*, Ann. of Math. Stud. 61, Princeton Univ. Press (1968) MR Zbl

[17] **W D Neumann**, **L Rudolph**, *Difference index of vectorfields and the enhanced Milnor number*, Topology 29 (1990) 83–100 MR Zbl

[18] **B Perron**, *Le nœud “huit” est algébrique réel*, Invent. Math. 65 (1982) 441–451 MR Zbl

[19] **A Pichon**, *Real analytic germs $f\bar{g}$ and open-book decompositions of the* 3*-sphere*, Int. J. Math. 16 (2005) 1–12 MR Zbl

[20] **A Pichon**, **J Seade**, *Fibred multilinks and singularities $f\bar{g}$*, Math. Ann. 342 (2008) 487–514 MR Zbl

[21] **J Seade**, *On Milnor’s fibration theorem and its offspring after* 50 *years*, Bull. Amer. Math. Soc. 56 (2019) 281–348 MR Zbl

[22] **K Tagami**, *A note on stabilization heights of fiber surfaces and the Hopf invariants*, Bull. Korean Math. Soc. 58 (2021) 1097–1107 MR Zbl

[23] **V G Turaev**, *Euler structures*, *nonsingular vector fields*, *and Reidemeister-type torsions*, Izv. Akad. Nauk SSSR Ser. Mat. 53 (1989) 607–643 MR Zbl In Russian; translated in Math. USSR-Izv. 34 (1990) 627–662

Department of Mathematics, Boğaziçi University
Istanbul, Turkey

seyma.karadereli@boun.edu.tr, ferit.ozturk@boun.edu.tr

Received: 2 June 2022 Revised: 19 May 2023

Algebraic & Geometric Topology **25**:3 (2025) 1411–1432
DOI: 10.2140/agt.2025.25.1411
Published: 20 June 2025

Fully augmented links in the thickened torus

ALICE KWON

We study the geometry of fully augmented link complements in the thickened torus and describe their geometric properties, generalizing the study of fully augmented links in S^3. We classify which fully augmented links in the thickened torus are hyperbolic, and show that their complements in the thickened torus decompose into ideal right-angled torihedra. We also study volume density of fully augmented links in S^3, defined to be the ratio of its volume and the number of augmentations. We prove the volume density conjecture for fully augmented links, which states that the volume density of a sequence of fully augmented links in S^3 which diagrammatically converges to a biperiodic link converges to the volume density of that biperiodic link. Furthermore, we show that the complement of a sequence of these links approaches the complement of the biperiodic link as a geometric limit.

57K10, 57K32, 57M50

1 Introduction

We study a class of links called *fully augmented links*. Fully augmented links in S^3 are obtained from diagrams of links in S^3 as follows. Let K be a link in S^3 with a given planar link diagram $D(K)$. We encircle each twist region (a maximal string of bigons) of $D(K)$ with a single unknotted component, called a *crossing circle*. The complement of the resulting link is homeomorphic to the link obtained by removing all *full-twists*, ie pairs of crossings from each twist region. Therefore a diagram of the fully augmented link contains a finite number of crossing circles, each encircling two strands of the link. These crossing circles are perpendicular to the projection plane and the other link components are embedded on the projection plane, except possibly for a finite number of single crossings, called *half-twists*, which are adjacent to the crossing circles; see Figure 1.

The geometry of fully augmented link complements in S^3 can be explicitly described in terms of an ideal right-angled polyhedral decomposition which is closely related to the link diagram. This geometry has been studied in detail by Adams [2], Agol and D Thurston [15, Appendix], Purcell [17] and Chesebro, Deblois and Wilton [12]. In [11] Champanerkar, Kofman and Purcell studied the geometry of alternating link complements in the thickened torus and described their decompositions into torihedra, which are toroidal analogs of polyhedra. We combine the methods used to study fully augmented links in S^3 and alternating links in the thickened torus to study the geometry of fully augmented link complements in the thickened torus. We generalize many geometric properties of fully augmented links in S^3 to those in the thickened torus $T^2 \times I$, where $I = (-1, 1)$.

A *biperiodic link* $\mathcal{L}$ is an infinite link in $\mathbb{R}^2 \times I$ with a projection on $\mathbb{R}^2 \times \{0\}$ which is invariant under an action of a two-dimensional lattice Λ by translations. The quotient $L = \mathcal{L}/\Lambda$ is a link in $T^2 \times I$ with a projection on $T^2 \times \{0\}$. This projection on $T^2 \times \{0\}$ is the link diagram of L.

Volume density of a link K was first introduced by Champanerkar, Kofman and Purcell in [10] as the ratio of its hyperbolic volume, $\mathrm{vol}(K)$, and its crossing number, $c(K)$. In [10; 11] they studied volume densities of sequences of alternating links in S^3 which diagrammatically converge to two specific biperiodic links called the square weave and the triaxial link. They proved that volume density of such a sequence of alternating links converges to that of the corresponding biperiodic link. In general, they conjectured the following:

Conjecture 1.1 (volume density conjecture [11]) *Let $\mathcal{L}$ be any biperiodic alternating link with alternating quotient link L. Let $\{K_n\}$ be a sequence of alternating hyperbolic links which Følner converges to $\mathcal{L}$. Then*

$$\lim_{n\to\infty} \frac{\mathrm{vol}(K_n)}{c(K_n)} = \frac{\mathrm{vol}((T^2 \times I) - L)}{c(L)}.$$

Definition 1.2 A *fully augmented biperiodic link* $\mathcal{L}$ is a fully augmented infinite link in $\mathbb{R}^2 \times I$ with a projection on $\mathbb{R}^2 \times \{0\}$ which is invariant under an action of a two-dimensional lattice Λ by translations. The quotient $L = \mathcal{L}/\Lambda$ is a fully augmented link in $T^2 \times I$ with a projection on $T^2 \times \{0\}$.

We define the *volume density* of a fully augmented link in S^3 (with or without half-twists) to be the ratio of its volume and the number of augmentations. We similarly define volume density of fully augmented links in the thickened torus. Using the geometry of fully augmented link complements in S^3 studied previously, and our results on the geometry of fully augmented link complements in the thickened torus, we prove the volume density conjecture for fully augmented links.

In Section 2 we classify hyperbolic fully augmented links in the thickened torus.

Theorem 2.11 *Let K be a link in $T^2 \times I$ with a weakly prime, twist-reduced cellular link diagram D. Let L be a link obtained by fully augmenting D. Then $T^2 \times I - L$ decomposes into two isometric totally geodesic right-angled torihedra, and hence L is hyperbolic.*

Remark 1.3 Augmented link diagrams are link diagrams obtained by adding crossing circles to some of the twist sites of a given link diagram, and are different from fully augmented links. Kwon and Tham [14] proved that augmented links in the thickened torus are hyperbolic. A generalization to thickened surfaces was also proved by Adams, Capovilla-Searle, D Li, L Q Li, McErlean, Simons, Stewart and Wang [4]. Theorem 2.11 gives a much stronger result for fully augmented links, as it describes the right-angled geometry of the complement and uses very different proof techniques than [4; 14]. The decomposition of the link L in Theorem 2.11 into right-angled torihedra (see Definition 2.6) is very important for Theorem 3.20, which investigates limit points of volume densities of fully augmented links.

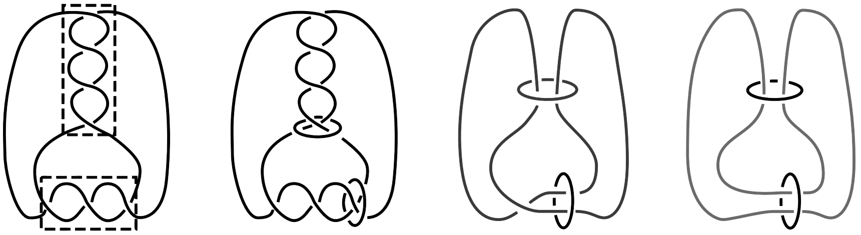


Figure 1: Left: link diagram of K. Center left: crossing circles added to each twist region. Center right: a fully augmented link diagram with all full-twists removed. Right: fully augmented link diagram with no half-twists.

In Section 3 we discuss volume density and the volume density spectrum of fully augmented links in S^3, and give many examples. In Section 3.2 we define Følner convergence for fully augmented links and prove the volume density conjecture for fully augmented links. Følner convergence for links was first defined by Champanerkar, Kofman and Purcell [10] for alternating links; we adapt the definition of Følner convergence for sequences of fully augmented links.

Theorem 3.20 *Let $\mathcal{L}$ be a biperiodic fully augmented link with quotient link L. Let $\{K_n\}$ be a sequence of hyperbolic fully augmented links in S^3 such that K_n Følner converges to $\mathcal{L}$ geometrically. Then*

$$\lim_{n\to\infty} \frac{\mathrm{vol}(K_n)}{a(K_n)} = \frac{\mathrm{vol}((T^2\times I)-L)}{a(L)},$$

where $a(K)$ denotes the number of augmentations of a fully augmented link K.

As an application in Corollary 3.23 we show that the endpoint $10v_{\mathrm{tet}}$ of the volume density spectrum of fully augmented links in S^3 is a limit point, by constructing a sequence of hyperbolic fully augmented links in S^3 which Følner converge everywhere to a fully augmented biperiodic link whose volume density is $10v_{\mathrm{tet}}$.

Acknowledgments I would like to thank my advisor Abhijit Champanerkar for guidance in this paper. I would also like to thank Ilya Kofman and Jessica Purcell for helpful conversations in regards to this project.

2 Hyperbolicity of fully augmented links in the thickened torus and volume bounds

To define fully augmented links in the thickened torus we first need to define twist-reduced diagrams for links in $T^2\times I$. Howie and Purcell defined twist-reduced diagrams for links in thickened surfaces in [13]. However for links in the thickened torus we can also define twist-reduced diagrams using the biperiodic link diagram in $\mathbb{R}^2$:

Definition 2.1 A *twist region* in the biperiodic link diagram $\mathcal{L}$ is a maximal string of bigons, or a single crossing. A *twist region* in the link diagram $L=\mathcal{L}/\Lambda$ is a quotient of a twist region in $\mathcal{L}$.

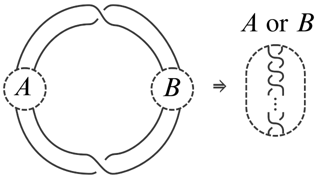


Figure 2: Twist-reduced diagram.

A biperiodic link $\mathcal{L}$ is called *twist-reduced* if for any simple closed curve on the plane that intersects $\mathcal{L}$ transversely in four points, with two points adjacent to one crossing and the other two points adjacent to another crossing, the simple closed curve bounds a subdiagram consisting of a (possibly empty) collection of bigons strung end-to-end between these crossings; see Figure 2. We say L is *twist-reduced* if it is the quotient of a twist-reduced biperiodic link.

Definition 2.2 A *fully augmented link diagram in $T^2 \times I$* is a diagram of a link L that is obtained from a twist-reduced diagram K in $T^2 \times I$ as follows: augment every twist region with a circle component, called a *crossing circle*, and get rid of all full-twists; see Figure 3. A *fully augmented link in $T^2 \times I$* is a link which has a fully augmented link diagram in $T^2 \times I$.

Remark 2.3 For fully augmented links in S^3, depending on the parity of the number of crossings in a twist region, the fully augmented link may or may not have a half-twist at that crossing circle; see Figure 1, center right. Similarly, depending on the parity of the number of crossings at a twist region, a fully augmented link in the thickened torus may or may not have a half-twist at that crossing circle.

Definition 2.4 A graph $G = (V, E)$ on the torus is *cellular* if its complement is a collection of open disks.

Torihedra were first defined in [11] and play the role of polyhedra in polyhedral decompositions of link compliments in S^3, eg it is proved in [11] that a complement of a link in the thickened torus decomposes into torihedra. Here we recall the definition of a torihedron.

2.1 Torihedral decomposition

Definition 2.5 A *torihedron* is a cone on the torus, $T^2 \times [0,1]/(T^2 \times \{1\})$, with a cellular graph G on $T^2 \times \{0\}$. The edges and faces of G are called the edges and faces of the torihedron. An *ideal torihedron*

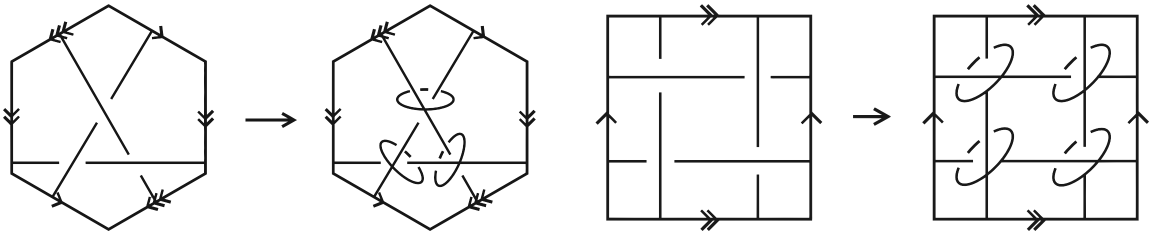


Figure 3: Left: a fully augmented triaxial link. Right: a fully augmented link on the square weave.

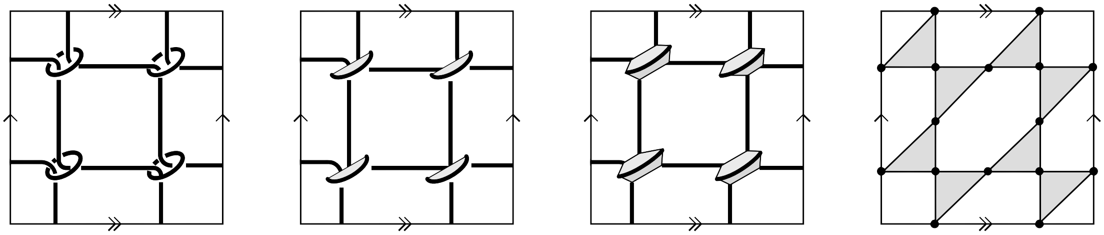


Figure 4: Left: a fundamental domain for a fully augmented square weave, L. Center left: disks cut in half at each crossing circle. Center right: sliced and flattened half-disks at each crossing circle. Right: collapsing the strands of the link and parts of the augmented circles (shown in bold) to ideal points gives the bowtie graph Γ_L. The disks become shaded bowties and the white regions become hexagons.

is a torihedron with the vertices of G and the vertex $T^2 \times \{1\}$ removed. Hence, an ideal torihedron is homeomorphic to $T^2 \times [0,1)$ with a finite set of points (ideal vertices) removed from $T^2 \times \{0\}$. The graph G is called the *graph of the torihedron*.

Definition 2.6 An *angled torihedron* is a torihedron with an angle assignment on each edge of the graph of the torihedron. An assignment of the angle $\frac{1}{2}\pi$ on each edge is called a *right-angled torihedron*.

Proposition 2.7 *Let L be a fully augmented link in $T^2 \times I$. Then there is a decomposition of the link complement $(T^2 \times I) - L$ into two combinatorially isomorphic torihedra such that*

(i) *the faces of each torihedron can be checkerboard colored so that the shaded faces are triangular and arise from the bowties corresponding to crossing circles*,

(ii) *the graph of each torihedron is* 4*-valent.*

Proof We follow the cut-slice-flatten construction described in [15]. Let L be a fully augmented link in $T^2 \times I$. We begin by assuming that there are no half-twists, the crossing circles are lateral to $T^2 \times \{0\}$ and the components of L that are not crossing circles lie flat on $T^2 \times \{0\}$. There are twice-punctured disks bounded by the crossing circles which are perpendicular to the projection plane.

(i) Cut $T^2 \times I$ along the projection surface $T^2 \times \{0\}$ into two pieces. This cuts each of the twice-punctured disks bounded by a crossing circle in half; see Figure 4, center left.

(ii) For each of the two pieces resulting from (i), slice the middle of the halves of twice-punctured disks and flatten the half-disks out; see Figure 4, center right.

(iii) Collapse strands of the link and parts of the augmented circles to ideal vertices in each of the two pieces; see Figure 4, right.

It follows from (i)–(iii) that each piece of the decomposition is homeomorphic to $T^2 \times [0,1)$, with the same graph on $T^2 \times \{0\}$ with vertices deleted. Hence $(T^2 \times I) - L$ decomposes into two identical ideal torihedra.

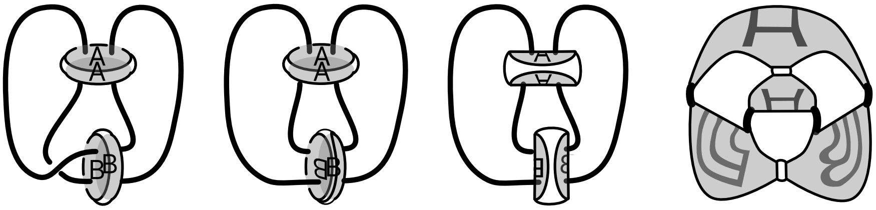


Figure 5: The gluing of the torihedra when a half-twist is present (disk B) and when a half-twist is absent (disk A). This figure, adapted from [18], is for links in S^3, but since this is a local move the same gluing works for links in $T^2 \times I$.

After (ii), the cut-sliced-flattened half-disks become a hexagon with an edge in the middle corresponding to the strand of half of a crossing circle. Upon collapsing the crossing circle this becomes a bowtie; see Figure 4, center right and far right. Each vertex of the graph is 4-valent since it is shared by two triangles of either two different bowties or one bowtie. Again by construction, each edge is shared by a triangle of a bowtie and a polygon that does not come from a bowtie; see Figure 4, right. Hence we can shade each triangle of the bowtie to get a checkerboard coloring on the graph of the torihedron such that the shaded faces are bowties.

The two torihedra are glued together as follows: the white faces are glued to the corresponding white faces, and the bowties are glued as shown in Figure 6, left.

In the case when there is a half-twist at a crossing circle, we split the whole twice-punctured disk into two copies, and flip one of the disks to remove the half-twist. This only affects the gluing of faces of the torihedra. Hence if there are half-twists, then we get the same torihedra but with a different gluing pattern on the bowties as shown in Figure 5 and Figure 6, right. □

Definition 2.8 For a fully augmented link L in the thickened torus, the decomposition of $T^2 \times I - L$ described above is called the *bowtie torihedral decomposition* of L. We call the graph of the torihedra the *bowtie graph* of L and denote it by Γ_L.

Lemma 2.9 *Let L be a hyperbolic fully augmented link in $T^2 \times I$. The following surfaces are embedded totally geodesic surfaces in the hyperbolic structure on the link complement*:

(i) *each twice-punctured disk bounded by a crossing circle*,

(ii) *each connected component of the projection surface.*

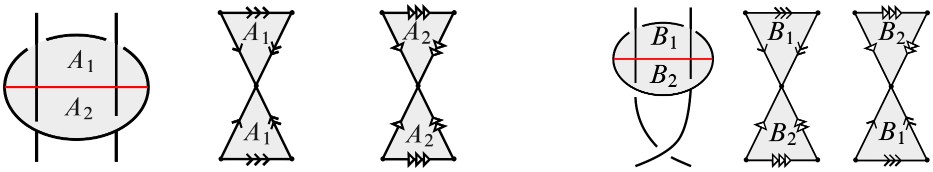


Figure 6: Left: gluing information on the edges of the bowtie without half-twists. Right: gluing information on the edges of the bowtie a with half-twist.

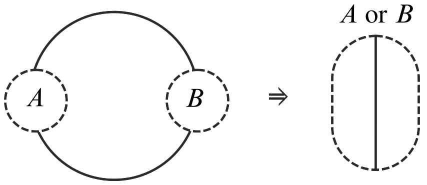


Figure 7: Prime diagram.

Proof (i) The disk E bounded by a crossing circle is punctured by two arcs of the link diagram lying on the projection plane. Adams [1] showed that any incompressible twice-punctured disk properly embedded in a hyperbolic 3-manifold is totally geodesic. Hence it suffices to show that E is incompressible. Let L be a hyperbolic fully augmented link in $T^2 \times I$. Since $T^2 \times I \simeq S^3 - H$, where H is the Hopf link, $L \cup H$ is a hyperbolic link in S^3.

Suppose there is a compressing disk D with $\partial D \subset E$. Since ∂D is an essential closed curve on E, it must encircle one or two punctures of E. Suppose it encircles only one puncture. This means that the union of D and the disk bounded by ∂D inside the closure of E forms a sphere in S^3 met by the link exactly once. This is a contradiction to the generalized Jordan curve theorem. Hence ∂D must bound a twice-punctured disk E' on E. This means $\overline{(E - E')} \cup D$ is a boundary-compressing disk for the crossing circle, contradicting the boundary irreducibility of $S^3 - (L \cup H)$.

(ii) Notice that the reflection through the projection surface ($T^2 \times \{0\}$) preserves the link complement, fixing the plane pointwise. Then it is a consequence of Mostow–Prasad rigidity that such a surface must be totally geodesic; see [17, Lemma 2.1]. □

2.2 Hyperbolicity

Definition 2.10 Let $\mathcal{L}$ be a biperiodic link with diagram $D(\mathcal{L})$. We say $D(\mathcal{L})$ is prime if whenever a disk embedded in $\mathbb{R}^2 \times \{0\}$ meets $D(\mathcal{L})$ transversely in exactly two edges, then the disk contains a simple edge of the diagram and no crossings; see Figure 7.

A diagram of a link L in $T^2 \times I$, denoted by $D(L)$, is *weakly prime* if $D(L)$ is a quotient of a prime biperiodic link diagram $D(\mathcal{L})$ in $\mathbb{R}^2 \times \{0\}$.

Theorem 2.11 *Let K be a link in $T^2 \times I$ with a weakly prime twist-reduced cellular link diagram D. Let L be a link obtained by fully augmenting D. Then $T^2 \times I - L$ decomposes into two isometric totally geodesic right-angled torihedra, and hence L is hyperbolic.*

The proof of Theorem 2.11 relies on a result about the existence of certain circle patterns on the torus due to Bobenko and Springborn [7]. We use similar ideas from [11] to prove Theorem 2.11

Theorem 2.12 [7] *Suppose G is a 4-valent graph on the torus T^2, and $\theta \in (0, 2\pi)^E$ is a function on edges of G that sums to 2π around each vertex. Let G^* denote the dual graph of G. Then there exists* *a circle pattern on T^2 with circles circumscribing faces of G (after isotopy of G) and having exterior intersection angles θ if and only if the following condition is satisfied:*

Suppose we cut the torus along a subset of edges of G^*, obtaining one or more pieces. For any piece that is a disk, the sum of θ over the edges in its boundary must be at least 2π, with equality if and only if the piece consists of only one face of G^* (only one vertex of G).

The circle pattern on the torus is uniquely determined up to similarity.

Proof of Theorem 2.11 Decompose $(T^2 \times I) - L$ into two torihedra using Proposition 2.7. Let Γ_L be the bowtie graph on $T^2 \times \{0\}$. Assign angles $\theta(e) = \frac{1}{2}\pi$ for every edge e in Γ_L. We now verify the condition of Theorem 2.12. This will prove the existence of an orthogonal circle pattern (circle pattern whose angle at the intersection of any two circles is orthogonal) circumscribing the faces of Γ_L.

Let C be a loop of edges of Γ_L^* enclosing a disk D. Suppose C intersects n edges of Γ_L transversely. Let V denote the number of vertices of Γ_L that lie in D, and let E denote the number of edges of Γ_L inside D disjoint from C. Because the vertices of Γ_L are 4-valent and since the edges inside D which are disjoint from C get counted twice for each of its end vertices, $n + 2E = 4V$. This implies n is even. Since K is weakly prime and C is made up of edges dual to Γ_L this implies $n > 2$. Since n is even, $n \geq 4$. Hence the sum of the angles for all edges of C must be at least 2π.

We now show that this is an equality if and only if C consists of one face of Γ_L^*, ie C encloses only one vertex. Suppose that $\sum_{e \in C} \theta(e) > 2\pi$. Since $\theta(e) = \frac{1}{2}\pi$ for every $e \in \Gamma_L$, and n is even, $n \geq 6$. Moreover

$$n \geq 6 \implies 4V - 2E \geq 6 \implies 2V - E \geq 3 \implies V \geq 2.$$

Hence C encloses more than one vertex.

Conversely, let $\sum_{e \in C} \theta(e) = 2\pi$. This implies $n = 4$.

Let the edges of C be e_i for $0 \leq i \leq 3$, with e_i incident to vertices v_i and v_{i+1}, and $v_0 = v_4$. Let the faces dual to v_i be F_{v_i}. Without loss of generality, let F_{v_0} be a shaded triangular face. Since Γ_L is checkerboard colored, F_{v_2} is also a shaded triangular face.

Suppose $F_{v_0} \cap F_{v_2} = \varnothing$. Then the edge e_2 must enter a white face F_{v_3} which has empty intersection with F_{v_0}; see Figure 8, left.

Since the bowties correspond to crossing circles (see Figure 9, left) the loop C gives a loop which intersects L. At the vertex v_0, which is in the shaded bowtie, at least one of the edges incident to v_0 has to intersect L. If only one edge at v_0 intersects L, since C bounds a disk, only one edge at v_2 intersects L, giving the case shown in Figure 9, center. Similarly if both edges incident to v_0 intersect L, since C bounds a disk, then the same is true for both edges incident at v_2, giving the case shown in Figure 9, right. If C intersects two strands of L as in Figure 9, center, since C bounds a disk, this contradicts the weakly prime condition of K. If C intersects two strands on each side as in Figure 9, right, this will contradict the twist-reduced condition on K.

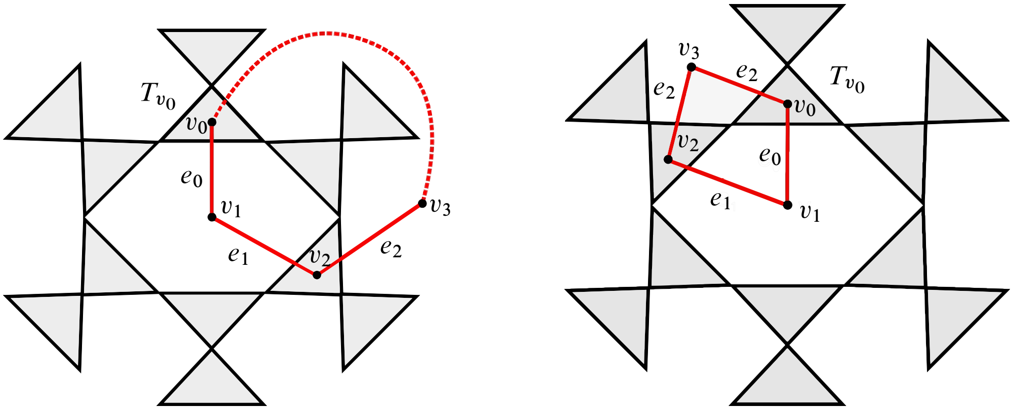


Figure 8: Left: when $n \geq 5$ and C closes with ≥ 5 edges. Right: when $n = 4$ and C closes with four edges.

Therefore $F_{v_0} \cap F_{v_2} \neq \phi$. Since both faces are triangles, they can only intersect in a vertex. This implies that C encloses a single vertex; see Figure 8, right.

Now, since we showed that Γ_L is a graph on the torus which satisfies the conditions of Theorem 2.12, there exists an orthogonal circle pattern on the torus with circles circumscribing the faces of Γ_L. Since a white face of the decomposition intersects any other white face only at ideal vertices, the circles which circumscribe the white faces create a circle packing, where the points of tangency are those corresponding to the associated ideal vertices. Since Γ_L is 4-valent and every edge has been assigned an angle of $\frac{1}{2}\pi$, the circles of the shaded faces meet orthogonally.

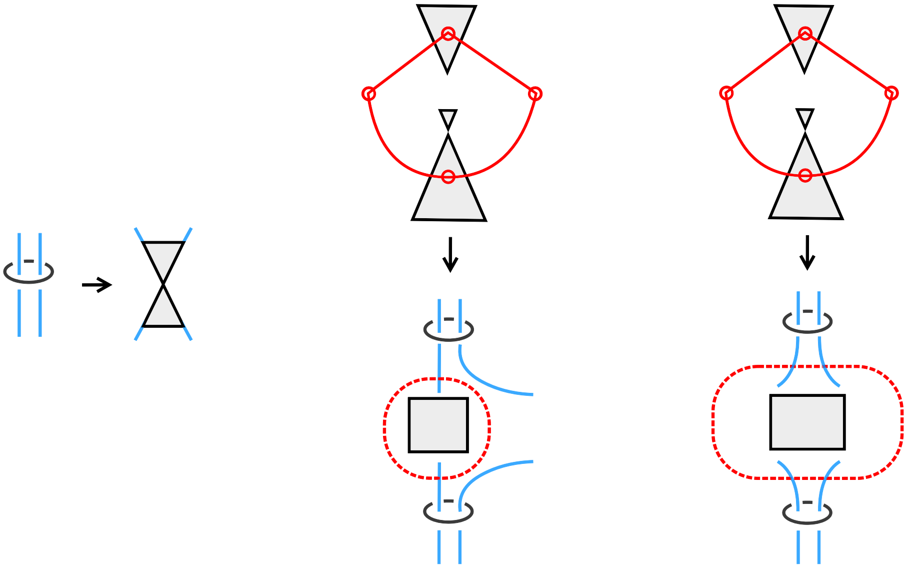


Figure 9: Left: the crossing circle splits into a bowtie. Center: C is in red. C intersects the original link in two points and hence must be a trivial edge. Right: C is in red. C can intersect the original link at four points and therefore must bound a twist region on one side.

Lifting the circle pattern to the universal cover of the torus defines an orthogonal biperiodic circle pattern on the plane. Considering the plane $z = 0$ as a part of the boundary of $\mathbb{H}^3$, this circle pattern defines a right-angled biperiodic ideal hyperbolic polyhedron in $\mathbb{H}^3$. The torihedron of the decomposition of $(T^2 \times I) - L$ is the quotient of $\mathbb{H}^3$ by $\mathbb{Z} \times \mathbb{Z}$ which is now realized as a right-angled hyperbolic torihedron. It follows from [13, Theorem 1.1] that $(T^2 \times I) - L$ is hyperbolic. □

Remark 2.13 Adams [2] proved that fully augmented link complements in S^3 are hyperbolic. We have proved an analogous result for fully augmented link complements in $T^2 \times I$. Our method of finding an orthogonal circle pattern which circumscribed the faces of the bowtie graph can also be applied to the case of fully augmented links in S^3. In this we have to use Andreev's theorem [19] to ensure a totally geodesic right-angled polyhedra.

2.3 Volume bounds

We show that a hyperbolic fully augmented link with c crossings in the thickened torus has an upper volume bound of $10cv_{\text{tet}}$. In the next section we show volume density convergence of fully augmented links. This means if we can find a link in the thickened torus whose volume is exactly $10cv_{\text{tet}}$ the corresponding biperiodic link will have volume density $10v_{\text{tet}}$. We will use this to show that an endpoint of the volume density spectrum of fully augmented links can be obtained as a limit.

Proposition 2.14 *Let L be a hyperbolic fully augmented link with c crossing circles. Then*

$$2cv_{\text{oct}} \le \operatorname{vol}(T^2 \times I - L) \le 10cv_{\text{tet}},$$

where $v_{\text{oct}} = 3.66386\ldots$ is the volume of a regular ideal octahedron and $v_{\text{tet}} = 1.01494\ldots$ is the volume of a regular ideal tetrahedron.

Proof We will first prove the lower bound. By work of Adams [2], the volume of the complement of L in $T^2 \times I$ agrees with that of the fully augmented link with no half-twists. This means a lower volume bound for the complement of L in $T^2 \times I$ with half-twists will be a lower volume bound of the fully augmented link with no half-twists. Hence we will assume L has no half-twists and obtain a lower bound for $T^2 \times I - L$.

Cut $T^2 \times I - L$ along the reflection plane $T^2 \times \{0\}$, dividing it into two isometric hyperbolic manifolds. The boundary of each of these consists of the regions of L on the projection surface with punctures for the crossing circles. By Lemma 2.9 these regions are geodesic. Hence cutting along the projection surface divides $T^2 \times I - L$ into isometric hyperbolic manifolds with totally geodesic boundary.

Miyamoto showed that if N is a hyperbolic 3-manifold with totally geodesic boundary, then $\operatorname{vol}(N) \ge -v_{\text{oct}}\chi(N)$ [16], with equality exactly when N decomposes into regular ideal octahedra. In our case, the manifold N consists of two copies of $T^2 \times [0, 1)$ with half-annuli removed for half the crossing circles.

For every half a crossing circle removed, we are removing one edge and two vertices. Hence for each crossing circle removed the Euler characteristic changes by -1. Since there are c crossing circles, the Euler characteristic would be $-c$ for each half-cut $T^2 \times [0,1)$. The lower bound now follows.

We now prove the upper bound. The torihedral decomposition of the link complement gives a decomposition into two identical ideal torihedra. Every triangular shaded face which comes from a bowtie corresponding to a crossing circle gives a tetrahedron when coned to the ideal vertex $T^2 \times \{1\}$ on each torihedra. Since there are c crossing circles, this gives c bowties; hence this gives $2c$ triangular shaded faces, and hence $4c$ tetrahedra. The cones on the white faces in each torihedra can be glued to make bipyramids on the white faces. These bipyramids can then be stellated into tetrahedra. Hence the number of tetrahedra coming from stellated bipyramids equals the number of edges of all the white faces. Since an edge of a white face is shared with an edge of a black triangle, this equals the number of edges of the torihedral graph, which has $6c$ edges. Hence the bipyramids on the white faces decompose into $6c$ tetrahedra. Thus the total count of tetrahedra is $4c + 6c = 10c$. Since the volume of an ideal tetrahedron is bounded by the volume of the regular ideal tetrahedron v_{tet}, the upper bound now follows. □

Remark 2.15 In Proposition 3.7 below we show that our upper bound is sharp by showing that the fully augmented square weave achieves the upper bound.

3 The volume density convergence conjecture

3.1 Volume density and its spectrum

In this section we discuss volume density of fully augmented links in S^3, its spectrum and asymptotic behavior. Champanerkar, Kofman and Purcell [10] defined volume density of a hyperbolic link in S^3 as the ratio of the volume of the link complement to its crossing number, and studied the asymptotic behavior of the volume density for sequences of alternating links which diagrammatically converge to a biperiodic alternating link.

For a hyperbolic link L in S^3, let $\mathrm{vol}(L)$ denote the hyperbolic volume of $S^3 - L$. In this section we assume that all links are hyperbolic.

Definition 3.1 Let L be a fully augmented link in S^3 with or without half-twists. The *volume density of* L is defined to be the ratio of the volume of L and the number of augmentations, ie $\mathrm{vol}(L)/a(L)$ where $a(L)$ is the number of augmentations of the link L. We similarly define the volume density of a fully augmented link in $T^2 \times I$.

Remark 3.2 Adams [2] showed that the volume of an augmented link with a half-twist at the crossing circle of the augmentation is equal to the volume without a half-twist. However, fully augmented links with and without half-twists have different crossing numbers. Hence in our definition above we divide by the number of augmentations rather than the number of crossings.

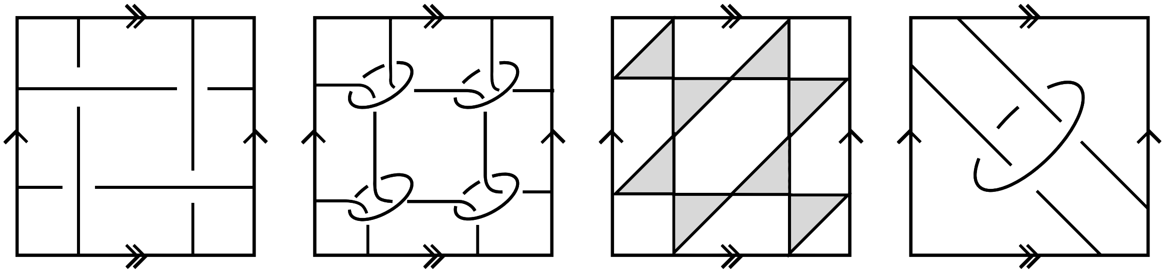


Figure 10: Left: the fundamental domain of the square weave $\mathcal{W}$. Center left: the fundamental domain of the fully augmented square weave, denoted by W_f. Center right: the bowtie graph Γ_{W_f} of the square weave on the left. Right: a quotient of W_f with same volume as the triaxial link.

Remark 3.3 For a fully augmented link without half-twists, the crossing number of the diagram is $4a(L)$. Thus the volume density of such a fully augmented link L is related to the volume density of L as defined in [10] by a factor of 4.

Throughout this section and the next we consider fully augmented links without half-twists.

Example 3.4 The Borromean rings B has $\mathrm{vol}(B) = 2v_{\mathrm{oct}}$ and $a(L) = 2$, and hence the volume density $\mathrm{vol}(B)/a(B)$ equals v_{oct}.

Definition 3.5 The volume density spectrum of fully augmented links in S^3 is defined as $\mathcal{S}_{\mathrm{aug}} = \{\mathrm{vol}(L)/a(L) : L \text{ is a fully augmented link in } S^3\}$.

Proposition 3.6 *The volume density spectrum* $\mathcal{S}_{\mathrm{aug}}$ *is a subset of* $[v_{\mathrm{oct}}, 10v_{\mathrm{tet}})$.

Proof Let L be a fully augmented link. Then by [17, Proposition 3.8] the volume of L is at least $2v_{\mathrm{oct}}(a(L)-1)$. Since L is hyperbolic, $a(L) \geq 2$, which implies

$$\frac{\mathrm{vol}(L)}{a(L)} \geq \frac{2v_{\mathrm{oct}}a(L)}{a(L)} - \frac{2v_{\mathrm{oct}}}{a(L)} > 2v_{\mathrm{oct}}\left(1 - \frac{1}{a(L)}\right) \geq v_{\mathrm{oct}}.$$

Since the volume density of the Borromean rings is v_{oct}, the lower bound is realized. Agol and D Thurston [15, Appendix] showed that $\mathrm{vol}(L) \leq 10v_{\mathrm{tet}}(a(L)-1)$. Hence the volume density of L is at most $10v_{\mathrm{tet}}$. □

We show below that $10v_{\mathrm{tet}}$ occurs as a volume density of the fully augmented square weave. Let $\mathcal{W}_f$ denote the fully augmented square weave as in Figure 10, center left.

Proposition 3.7

$$\frac{\mathrm{vol}(T^2 \times I - W_f)}{a(W_f)} = 10v_{\mathrm{tet}}.$$

Proof A fourfold quotient of $\mathcal{W}_f$ as shown in Figure 10, right, was studied in [8]. The authors proved that the volume of this link complement in the thickened torus is $10v_{\mathrm{tet}}$. Hence $\mathrm{vol}(T^2 \times I - W_f) = 40v_{\mathrm{tet}}$, and its volume density is $10v_{\mathrm{tet}}$. □

Remark 3.8 The quotient of W_f as in Figure 10, right, has the same volume as that of a quotient of a triaxial link which is not a fully augmented link; see Figure 3, left. However the two links are not the same, as they have different numbers of cusps. The triaxial link has five cusps — three from each link component in the thickened torus and two from each link component of the Hopf Link — whereas the quotient of W_f in Figure 10, right, has four cusps — two from each component of the link in the thickened torus (which includes the crossing circle) and two from each link component of the Hopf Link.

3.2 Følner convergence

The volume density of the fully augmented square weave is $10v_{\text{tet}}$. We will prove below that $10v_{\text{tet}}$ is also a limit point of the $\mathcal{S}_{\text{aug}}$ by investigating the asymptotic behavior of volume density of a sequence of fully augmented links in S^3 which diagrammatically converge to the biperiodic fully augmented square weave, as defined below. We use the notion of Følner convergence, which was first introduced in [10]. We begin by modifying its definition.

In [10] the authors used the Tait graph (checkerboard graph) of alternating links to define Følner convergence. We will use bowtie graphs to define Følner convergence for fully augmented links; see Definition 2.8 and [17, Proposition 2.2].

Definition 3.9 Let $\mathcal{L}$ be a biperiodic fully augmented link. We will say that a sequence of fully augmented links $\{K_n\}$ in S^3 *Følner converges almost everywhere geometrically* to $\mathcal{L}$, denoted by $K_n \xrightarrow{\text{GF}} \mathcal{L}$, if the respective bowtie graphs $\{\Gamma_{K_n}\}$ and $\Gamma_{\mathcal{L}}$ satisfy the following: there are subgraphs $G_n \subset \Gamma_{K_n}$ such that

(i) $G_n \subset G_{n+1}$, and $\bigcup G_n = \Gamma_{\mathcal{L}}$,

(ii) $\lim_{n\to\infty} |\partial G_n|/|G_n| = 0$, where $|\cdot|$ denotes the number of vertices and $\partial G_n \subset \Gamma_{\mathcal{L}}$ consists of the vertices of G_n that share an edge in $\Gamma_{\mathcal{L}}$ with a vertex not in G_n,

(iii) $G_n \subset \Gamma_{\mathcal{L}} \cap (n\Lambda)$, where $n\Lambda$ represents n^2 copies of the fundamental domain for the lattice Λ such that $L = \mathcal{L}/\Lambda$,

(iv) $\lim_{n\to\infty} |G_n|/3a(K_n) = 1$.

Remark 3.10 The number 3 appears in the denominator in the last condition for the definition of Følner convergence because the number of vertices of the bowtie polyhedron for K_n equals three times the number of augmentations. To see this, note that every bowtie shares two vertices with another bowtie and hence contributes three vertices to the graph. Since each bowtie corresponds to a crossing circle, the number of vertices of the graph is $3a(K)$.

Remark 3.11 Many fully augmented links can have the same bowtie graph. For example, a fully augmented link with and without half-twists have the same bowtie graph but different gluing; see Figures 11 and 12. Another example of this is when the bowtie graphs are same but with different pairing

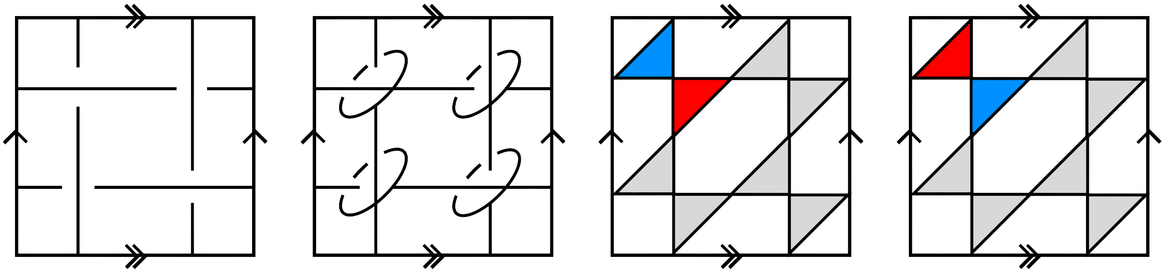


Figure 11: Left: the quotient of the square weave. Center left: W_f with half-twists at each crossing circle. Center right and far right: the bowtie graph with blue (red) face bowtie of the top torihedron being glued to a blue (red) face of the bottom torihedron

of triangles; see Figure 13 for an example of two links with same bowtie graphs but different pairings. In our definition above, we are using only the polyhedral graphs but not the pairing information of the bowties. Hence we call our Følner convergence geometric. This has the advantage of having many more sequences converging to a given biperiodic fully augmented link.

3.3 Volume density conjecture

Conjecture 3.12 (volume density conjecture) *Let $\mathcal{L}$ be any biperiodic alternating link with alternating quotient link L. Let $\{K_n\}$ be a sequence of alternating hyperbolic links such that K_n Følner converges to $\mathcal{L}$. Then*

$$\lim_{n\to\infty}\frac{\operatorname{vol}(K_n)}{c(K_n)}=\frac{\operatorname{vol}((T^2\times I)-L)}{c(L)}.$$

Champanerkar, Kofman and Purcell proved this conjecture when $\mathcal{L}$ is the square weave [10] and the triaxial link [9] by finding upper and lower bounds on $\operatorname{vol}(K_n)$ such that for a sequence of alternating links $K_n \xrightarrow{F} \mathcal{L}$, these bounds are equal in the limit. One of the key tools in their proof is the use of right-angled circle patterns. Using the right-angled decomposition of fully augmented link complements in S^3, we construct right-angled circle patterns, and use these to prove the volume density conjecture for fully augmented links in S^3.

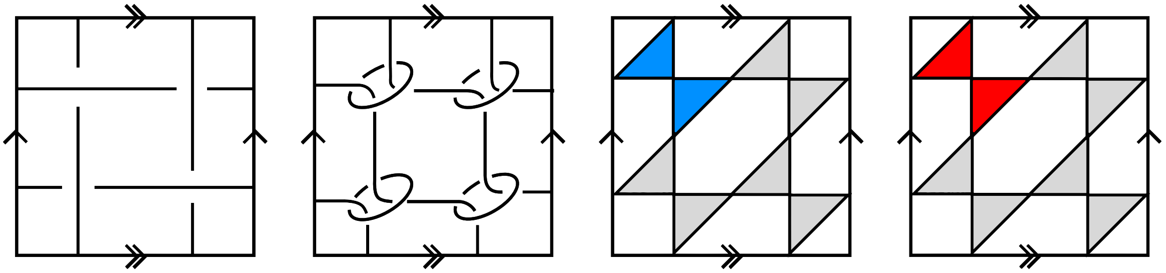


Figure 12: Left: the quotient of the square weave. Center left: W_f with no half-twists at each crossing circle. Center right and far right: the bowtie graph with blue (red) face bowtie of the top torihedron being glued to a blue (red) face of the bottom torihedron

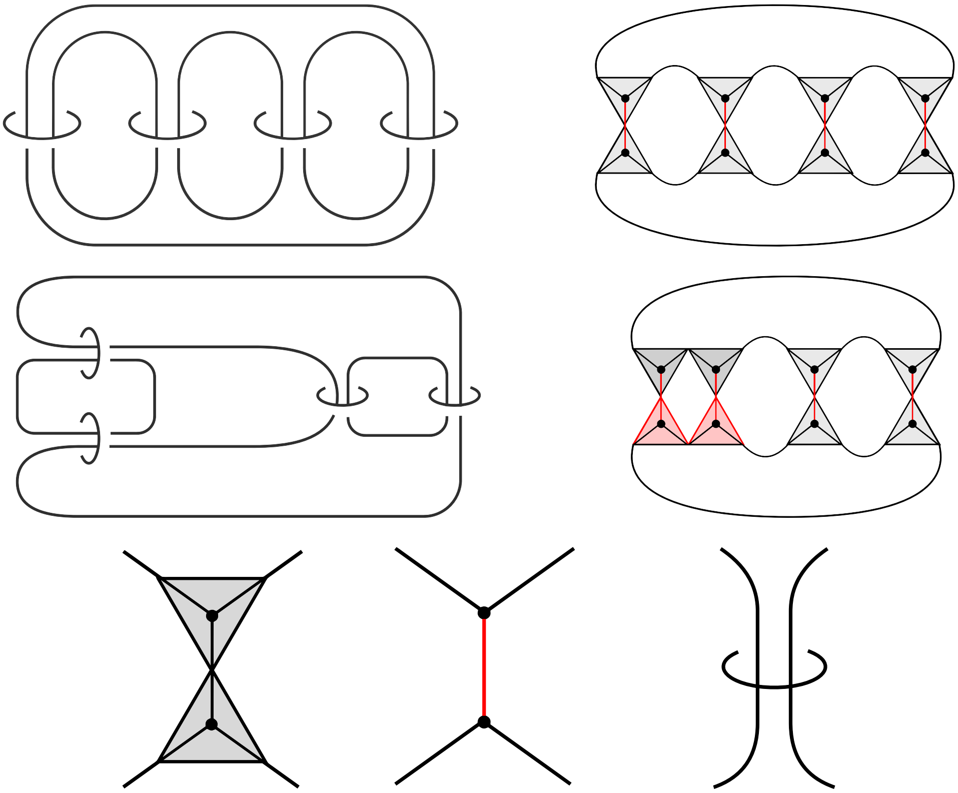


Figure 13: The fully augmented links above and below are two different links with the same bowtie graph with different pairing information.

The idea is as follows: As described in [17], each hyperbolic fully augmented link complement in S^3 can be decomposed into two right-angled ideal polyhedra which are described by a right-angled circle pattern. By Theorem 2.11 each torihedra of the bowtie torihedral decomposition is right-angled and described by another right-angled circle pattern on the torus. The $\mathbb{Z} \times \mathbb{Z}$ lift of this circle pattern is the circle pattern associated to $\mathcal{L}$. We show below that when a sequence of fully augmented links K_n converges to $\mathcal{L}$, $K_n \xrightarrow{\text{GF}} \mathcal{L}$, the circle pattern for K_n converges to an infinite circle pattern for $\mathcal{L}$. As a consequence we obtain the volume density convergence.

In order to work with circle patterns and convergence of circle patterns, we recall the following definitions from [5]:

Definition 3.13 A *disk pattern* is a collection of closed round disks in the plane such that no disk is the Hausdorff limit of a sequence of distinct disks and such that the boundary of any disk is not contained in the union of two other disks.

Definition 3.14 A *simply connected* disk pattern is a disk pattern in the plane such that the union of the disks is simply connected.

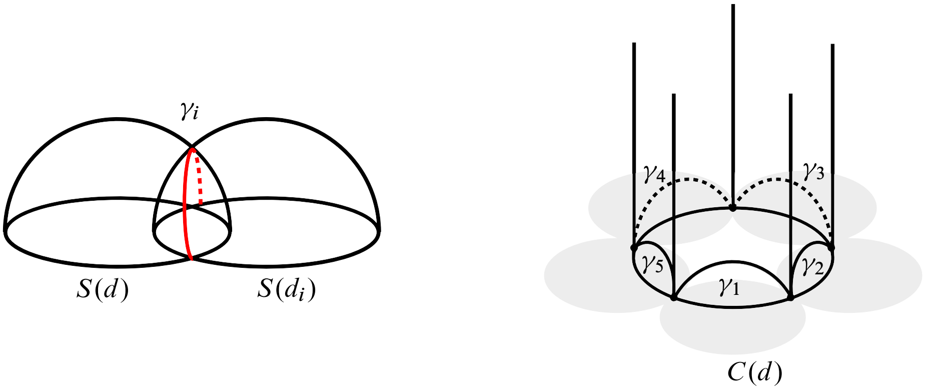


Figure 14: Left: $S(d) \cap S(d')$. Right: C(d).

Let D be a disk pattern in $\mathbb{C}$. Let $G(D)$ be the graph with a vertex for each disk and an edge between any two vertices when the corresponding disks overlap. The graph $G(D)$ inherits an embedding in the plane from the disk pattern and we will identify $G(D)$ with its plane embedding. A face of $G(D)$ is an unbounded component of the complement of $G(D)$ in the plane. We can label the edges of $G(D)$ with the angles between the intersecting disks.

Definition 3.15 A disk pattern D is called an *ideal disk pattern* if the labels of edges of $G(D)$ are in the interval $\left(0, \frac{1}{2}\pi\right]$ and the labels around each triangle or quadrilateral in $G(D)$ sum to π or 2π, respectively.

It is clear that ideal disk patterns in $\mathbb{C}$ correspond to ideal polyhedra in $\mathbb{H}^3$, with the disks corresponding to the faces of the ideal polyhedron.

Definition 3.16 Let D and D' be disk patterns. Give $G(D)$ and $G(D')$ the path metric in which each edge has length 1. For disks d in D and d' in D', we say (D, d) and (D', d') *agree to generation* n if the balls of radius n centered at vertices corresponding to d and d' admit a graph isomorphism with labels on edges preserved.

Definition 3.17 For a disk d in a disk pattern D, we let $S(d)$ be the geodesic hyperplane in $\mathbb{H}^3$ whose boundary agrees with that of d. That is, $S(d)$ is the Euclidean hemisphere in $\mathbb{H}^3$ with boundary coinciding with the boundary of d. For a disk pattern coming from a right-angled ideal polyhedron, the planes $S(d)$ form the boundary faces of the polyhedron. In this case, the disk pattern D is *simply connected* and *ideal*, since it corresponds to an ideal polyhedron.

Similarly, for a disk d in D with intersecting neighboring disks $d_1, \dots, d_m$, the intersection $S(d) \cap S(d_i)$ is a geodesic γ_i in $\mathbb{H}^3$. The geodesics γ_i for $i = 1, \dots, m$ on $S(d)$ bound an ideal polygon in $\mathbb{H}^3$. The cone of this polygon to the point at infinity is denoted by $C(d)$; see Figure 14.

Definition 3.18 A disk pattern D is said to be *rigid* if $G(D)$ has only triangular and quadrilateral faces, and each quadrilateral face has the property that the four corresponding disks of the disk pattern intersect in exactly one point.

Lemma 3.19 (Atkinson [5]) *Let D_∞ be an infinite rigid disk pattern. Then there exists a bounded sequence $0 \le \epsilon_l \le b < \infty$ converging to zero such that if D is a simply connected ideal rigid finite disk pattern containing a disk d such that (D_∞, d_∞) and (D, d) agree to generation l, then*

$$|\mathrm{vol}(C(d)) - \mathrm{vol}(C(d_\infty))| \le \epsilon_l.$$

Note that the sequence $\{\epsilon_l\}$ in above lemma only depends on D_∞.

Theorem 3.20 (volume density conjecture for fully augmented links) *Let $\mathcal{L}$ be a biperiodic fully augmented link with quotient link L. Let $\{K_n\}$ be a sequence of hyperbolic fully augmented links in S^3. Then*

$$K_n \xrightarrow{\mathrm{GF}} \mathcal{L} \implies \lim_{n\to\infty} \frac{\mathrm{vol}(K_n)}{a(K_n)} = \frac{\mathrm{vol}((T^2 \times I) - L)}{a(L)}.$$

Proof Let P_L be the bowtie torihedron with bowtie graph Γ_L of L. Let P_∞ be the infinite polyhedron in $\mathbb{H}^3$ which is the biperiodic lift of P_L with its cone vertex taken to be ∞. P_∞ can be seen to be made up of $\mathbb{Z}^2$ copies of an embedding of P_L in $\mathbb{H}^3$ with its cone vertex taken to be ∞, glued according to the biperiodic lift. Note that since the graph of P_L is the bowtie graph Γ_L of L, which is toroidal, the graph of P_∞ is a biperiodic lift of Γ_L and is isomorphic to the bowtie graph $\Gamma_{\mathcal{L}}$ coming from $\mathcal{L}$. Let D_∞ be the infinite disk pattern coming from the infinite polyhedron P_∞. Since P_L is a right-angled torihedron, P_∞ is also right-angled, and hence D_∞ is a right-angled disk pattern.

Since $\{K_n\}$ is a sequence of fully augmented links, each K_n is a fully augmented hyperbolic link in S^3. The bowtie polyhedron of K_n is a right-angled ideal hyperbolic polyhedron with the same graph as the bowtie graph Γ_{K_n}. The assumption that the sequence $\{K_n\}$ Følner converges almost everywhere geometrically to $\mathcal{L}$ implies that there are subgraphs $G_n \subset \Gamma_{K_n}$ which satisfy the conditions of Følner convergence in Definition 3.9. Hence we can embed bowtie polyhedra of K_n in $\mathbb{H}^3$ so that a vertex in $\Gamma_{K_n} - G_n$ is sent to infinity, and $G_n \subset G_{n+1}$. We denote this polyhedron in $\mathbb{H}^3$ by P_n. First note that $\mathrm{vol}(K_n) = 2\,\mathrm{vol}(P_n)$. Let $v(P_n)$ denote the number of vertices of P_n. Since P_n is a 4-valent checkerboard graph whose shaded faces are triangles coming from the bowties, one for each augmentation, every vertex is shared by two triangles. Hence $v(P_n) = 3 \cdot 2a(K_n)\frac{1}{2} = 3a(K_n)$. Therefore,

$$\frac{\mathrm{vol}(K_n)}{3a(K_n)} = 2\frac{\mathrm{vol}(P_n)}{v(P_n)}.$$

Let D_n be the disk pattern of the polyhedron P_n. It follows that D_n is a right-angled simply connected disk pattern. Since D_n corresponds to a disk pattern arising from a fully augmented link, D_n is rigid (see Definition 3.18 and Figure 15). We will now use Følner convergence to relate D_n and D_∞.

Let F_l^n be the set of disks d in D_n such that (D_n, d) agrees to generation l but not to generation $l+1$ with (D_∞, d_∞). For every positive integer k, let $|f_k^n|$ denote the number of faces of P_n with k sides that are not contained in $\bigcup_l F_l^n$ and do not meet the point at infinity. By counting vertices we obtain

$$\sum_k k|f_k^n| \le 4|\Gamma_{K_n} - G_n|.$$

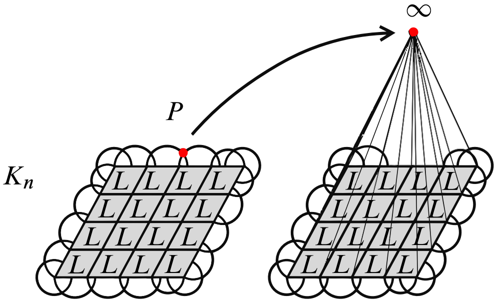


Figure 15: Left: an $n \times n$ copy of the fundamental domain of Λ with an arbitrary closure, and a marked point P on the crossing of the closure. Right: the point P moved to the cone point at ∞.

The term $|\Gamma_{K_n} - G_n|$ counts the number of vertices that are in Γ_{K_n} but not in G_n. Since all the vertices of the graph Γ_{K_n} are 4-valent we get a factor of 4. Hence $|\Gamma_{K_n}| = v(P_n) = 3a(K_n)$, and

$$\text{(1)} \qquad \lim_{n\to\infty} \frac{|G_n|}{3a(K_n)} = 1 \implies \lim_{n\to\infty} \frac{4|\Gamma_{K_n} - G_n|}{v(P_n)} = 0 \implies \lim_{n\to\infty} \frac{\sum_k k|f_k^n|}{v(P_n)} = 0.$$

Let $d \in F_l^n$ and let $v_1, \dots, v_m$ be the vertices of G_n which lie on the boundary of d; see Figure 16. Let $B(v,r) \subset G_n$ denote the ball centered at vertex v of radius r in the path metric on G_n. It follows from the definition of F_l^n and the fact that G_n is the planar dual of the graph of the disk pattern $G(D_n)$ — without the vertex corresponding to the unbounded face — that $d \in F_l^n$ implies $B(v_i, l) \subset G_n$ but $B(v_i, l+1) \not\subset G_n$ for $i = 1, \dots, m$. Hence the distance from v_i to ∂G_n is l, ie $v_i \in \partial B(x,l)$ for some $x \in \partial G_n$ for all $i = 1, \dots, m$. Hence $F_l^n \subset \bigcup_{x \in \partial G_n} \partial B(x,l)$.

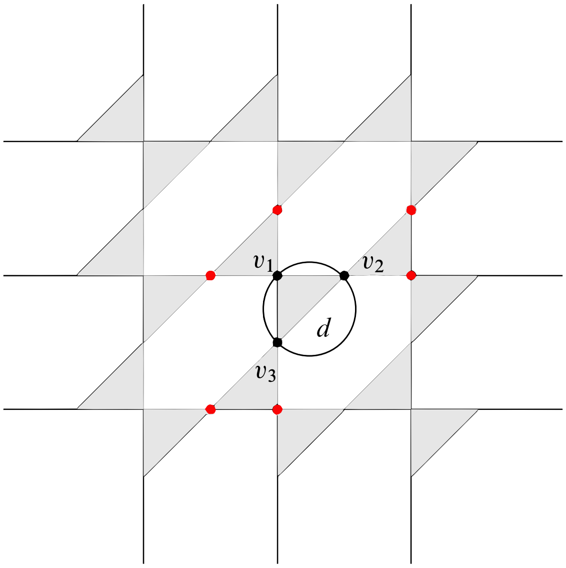


Figure 16: The circle in black is an example of $B(v_i, 1)$, and the boundary of the union over all i of $B(v_i, 1)$ is colored in red.

Lemma 3.21
$$\lim_{n\to\infty}\frac{\left|\bigcup_l F_l^n\right|}{v(P_n)}=1.$$

Proof We begin by showing that there exists $m>0$ such that $|\partial B(x,l)|\le ml$ for any $x\in G_n$. By definition of Følner convergence, $G_n\subset\Gamma_{\mathcal{L}}$. Babai [6] showed that the growth rate for almost vertex-transitive graphs with one end is quadratic, that is, growth of $|B(x,l)|$ is quadratic in l. Since $\Gamma_{\mathcal{L}}$ is a biperiodic 4-valent planar graph, it satisfies the conditions of Babai's theorem, and hence has quadratic growth rate. By definition, the vertices in $\partial B(x,l)$ are incident to vertices in $B(x,l-1)$, and hence $|\partial B(x,l)|$ has linear growth rate in l.

Thus $|F_l^n|\le ml|\partial G_n|$ and we obtain
$$\lim_{n\to\infty}\frac{|F_l^n|}{v(P_n)}\le\lim_{n\to\infty}\frac{ml|\partial G_n|}{3a(K_n)}=\frac{ml}{3}\lim_{n\to\infty}\frac{|\partial G_n|}{|G_n|}\frac{|G_n|}{a(K_n)}=0.$$
Since $G_n\subset G(\mathcal{L})$, every vertex of G_n lies on a disk in F_l^n for some l, and for every disk in F_l^n there are no vertices in $G(K_n)-G_n$ which lie on the disk. Now, by assumption, $\lim_{n\to\infty}|G_n|/(3a(K_n))=1$. Hence $\lim_{n\to\infty}\left|\bigcup_l F_l^n\right|/v(P_n)=\lim_{n\to\infty}|G_n|/(3a(K_n))=1$. □

Let f_k^n be the face with k sides that is not contained in $\bigcup_l F_l^n$ which does not meet the point at infinity. For each n, $\operatorname{vol}(C(f_k^n))\le k\lambda\left(\frac{1}{6}\pi\right)$, where $\lambda(\theta)$ is the Lobachevsky function defined as
$$\lambda(\theta)=-\int_0^\theta\log|2\sin(t)|\,dt,$$
whose maximum value is $\lambda\left(\frac{1}{6}\pi\right)$ [19]; see also [3].

Let E^n denote the sum of the actual volumes of all the cones over the faces f_k^n, for every integer k. Then
$$E^n\le\sum_k\sum_{f_k^n}k\lambda\left(\tfrac{1}{6}\pi\right)=\sum_k k|f_k^n|\lambda\left(\tfrac{1}{6}\pi\right). \tag{2}$$

As mentioned before, every vertex of G_n lies on a disk in F_l^n for some l, and for every disk in F_l^n there are no vertices in $\Gamma_{K_n}-G_n$ which lie on the disk. By assumption $G_n\subset\Gamma_{\mathcal{L}}\cap(n\Lambda)$, where $n\Lambda$ represents n^2 copies of the fundamental domain for the lattice Λ such that $L=\mathcal{L}/\Lambda$.

Since the cone vertex of the torihedron for $T^2\times I-L$ is at infinity, the disk pattern obtained from taking n^2 copies of L just extends the disk pattern from one copy of L to $n\times n$ grid, as in Figure 17. The graph for the disk pattern for n^2 copies of L intersects Γ_{K_n} in G_n, as in Figure 15.

For any face f in F_l^n, let δ_l^n be a positive number such that $\operatorname{vol}(C(f))=\operatorname{vol}(C(f'))\pm\delta_l^n$, where f' is a face in the disk pattern of $\mathcal{L}$ such that the graph isomorphism between $G(D_n)$ and $G(D_\infty)$ sends f to f'. Furthermore, we choose δ_l^n so that we can bound the sequence of δ_l^n by a sequence which will converge to zero, as in Lemma 3.19.

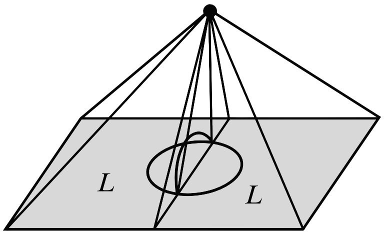


Figure 17: Two copies of the link L coned to the point at infinity. The disk pattern from one copy of L extends to the next copy.

Then

$$\operatorname{vol}(P_n) = \sum_l \sum_{f \in F_l^n} (\operatorname{vol}(f') \pm \delta_l^n) + E^n. \tag{3}$$

By (3) we get

$$\operatorname{vol}(P_n) = \tfrac{1}{2} n^2 \operatorname{vol}((T^2 \times I) - L) + \sum_l \sum_{f \in F_l^n} (\pm \delta_l^n) + E^n. \tag{4}$$

We divide each term by $a(K_n)$ and take the limit. For the first term of (4) we obtain

$$\lim_{n\to\infty} \frac{1}{2} \frac{n^2 \operatorname{vol}((T^2 \times I) - L)}{a(K_n)} = \frac{1}{2} \frac{n^2 \operatorname{vol}((T^2 \times I) - L)}{n^2 a(L)} = \frac{1}{2} \frac{\operatorname{vol}((T^2 \times I) - L)}{a(L)}.$$

From our assumption of Følner convergence, the last condition gives us

$$\lim_{n\to\infty} \frac{a(K_n)}{n^2 a(L)} = 1.$$

By Lemma 3.19 there are positive numbers ϵ_l such that $\delta_l^n \le \epsilon_l$, so the second term of (4) becomes

$$\lim_{n\to\infty} \frac{\left|\sum_l \sum_{f \in F_l^n} (\pm \delta_l^n)\right|}{a(K_n)} \le \lim_{n\to\infty} \frac{\sum_l |F_l^n| \epsilon_l}{a(K_n)}.$$

Lemma 3.22

$$\lim_{n\to\infty} \frac{\sum_l |F_l^n| \epsilon_l}{a(K_n)} = 0.$$

Proof Fix any $\epsilon > 0$. Because $\lim_{l\to\infty} \epsilon_l = 0$, there is K large enough that $\epsilon_l < \frac{1}{3}\epsilon$ for $l > K$. Then $\sum_{l=1}^K \epsilon_l$ is a finite number, say M. Since we've seen above that $\lim_{n\to\infty} \bigcup_l |F_l^n| / v(P_n) = 1$ and $\lim_{n\to\infty} |F_l^n| / v(P_n) = 0$, there exists N such that if $n > N$ then $\max_{l \le L} |F_l^n| / v(P_n) < \epsilon/(3MK)$ and $\left|\bigcup_l F_l^n\right| / v(P_n) < (1+\epsilon)$. Then for $n > N$,

$$\frac{\sum_l |F_l^n| \epsilon_l}{v(P_n)} = \frac{\sum_{l=1}^K |F_l^n| \epsilon_l}{v(P_n)} + \frac{\sum_{l > K} |F_l^n| \epsilon_l}{v(P_n)} < \frac{\epsilon K}{3MK} + (1+\epsilon)\frac{\epsilon}{3} < \epsilon. \qquad \square$$

Now setting $v(P_n) = 3a(K_n)$ we get that the limit of the second term is zero.

Finally, by (1) and (2) we get that the third term of (4) equals zero:

$$\lim_{n\to\infty}\frac{E^n}{a(K_n)}\le\lim_{n\to\infty}\frac{\sum_k k|f_k^n|\lambda\left(\frac{1}{6}\pi\right)}{a(K_n)}=0.$$

Therefore $\lim_{n\to\infty}\operatorname{vol}(P_n)/a(K_n)=\frac{1}{2}\operatorname{vol}(T^2\times I-L)/a(L)$, which means $\lim_{n\to\infty}\operatorname{vol}(K_n)/a(K_n)=\operatorname{vol}(T^2\times I-L)/a(L)$. □

Recall that $\mathcal{W}_f$ denotes the fully augmented square weave link whose quotient is W_f with volume $10cv_{\text{tet}}$.

Corollary 3.23 *Let K_n be any sequence of hyperbolic fully augmented links such that K_n Følner converges everywhere to $\mathcal{W}_f$. Then*

$$\lim_{n\to\infty}\frac{\operatorname{vol}(K_n)}{a(K_n)}=10v_{\text{tet}}.$$

Proof This follows from Proposition 3.7 and Theorem 3.20. □

References

[1] **C C Adams**, *Thrice-punctured spheres in hyperbolic* 3*–manifolds*, Trans. Amer. Math. Soc. 287 (1985) 645–656 MR Zbl

[2] **C C Adams**, *Augmented alternating link complements are hyperbolic*, from "Low-dimensional topology and Kleinian groups", Lond. Math. Soc. Lect. Note Ser. 112, Cambridge Univ. Press (1986) 115–130 MR Zbl

[3] **C Adams**, **A Calderon**, **N Mayer**, *Generalized bipyramids and hyperbolic volumes of alternating k–uniform tiling links*, Topology Appl. 271 (2020) art. id. 107045 MR Zbl

[4] **C Adams**, **M Capovilla-Searle**, **D Li**, **L Q Li**, **J McErlean**, **A Simons**, **N Stewart**, **X Wang**, *Augmented cellular alternating links in thickened surfaces are hyperbolic*, Eur. J. Math. 9 (2023) art. id. 100 MR Zbl

[5] **C K Atkinson**, *Volume estimates for equiangular hyperbolic Coxeter polyhedra*, Algebr. Geom. Topol. 9 (2009) 1225–1254 MR Zbl

[6] **L Babai**, *The growth rate of vertex-transitive planar graphs*, from "Proceedings of the eighth annual ACM–SIAM symposium on discrete algorithms", ACM, New York (1997) 564–573 MR Zbl

[7] **A I Bobenko**, **B A Springborn**, *Variational principles for circle patterns and Koebe's theorem*, Trans. Amer. Math. Soc. 356 (2004) 659–689 MR Zbl

[8] **A Champanerkar**, **D Futer**, **I Kofman**, **W Neumann**, **J S Purcell**, *Volume bounds for generalized twisted torus links*, Math. Res. Lett. 18 (2011) 1097–1120 MR Zbl

[9] **A Champanerkar**, **I Kofman**, *Determinant density and biperiodic alternating links*, New York J. Math. 22 (2016) 891–906 MR Zbl

[10] **A Champanerkar**, **I Kofman**, **J S Purcell**, *Geometrically and diagrammatically maximal knots*, J. Lond. Math. Soc. 94 (2016) 883–908 MR Zbl

[11] **A Champanerkar**, **I Kofman**, **J S Purcell**, *Geometry of biperiodic alternating links*, J. Lond. Math. Soc. 99 (2019) 807–830 MR Zbl

[12] **E Chesebro**, **J DeBlois**, **H Wilton**, *Some virtually special hyperbolic* 3*–manifold groups*, Comment. Math. Helv. 87 (2012) 727–787 MR Zbl

[13] **J A Howie**, **J S Purcell**, *Geometry of alternating links on surfaces*, Trans. Amer. Math. Soc. 373 (2020) 2349–2397 MR Zbl

[14] **A Kwon**, **Y H Tham**, *Hyperbolicity of augmented links in the thickened torus*, J. Knot Theory Ramifications 31 (2022) art. id. 2250025 MR Zbl

[15] **M Lackenby**, *The volume of hyperbolic alternating link complements*, Proc. Lond. Math. Soc. 88 (2004) 204–224 MR Zbl With an appendix by I Agol and D Thurston

[16] **Y Miyamoto**, *Volumes of hyperbolic manifolds with geodesic boundary*, Topology 33 (1994) 613–629 MR Zbl

[17] **J S Purcell**, *An introduction to fully augmented links*, from "Interactions between hyperbolic geometry, quantum topology and number theory", Contemp. Math. 541, Amer. Math. Soc., Providence, RI (2011) 205–220 MR Zbl

[18] **J S Purcell**, *Hyperbolic knot theory*, Graduate Studies in Math. 209, Amer. Math. Soc., Providence, RI (2020) MR Zbl

[19] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `https://url.msp.org/gt3m`

Science Department, SUNY Maritime
Throggs Neck, NY, United States

akwon@sunymaritime.edu

Received: 19 August 2022 Revised: 22 May 2023

Algebraic & Geometric Topology **25**:3 (2025) 1433–1500
DOI: 10.2140/agt.2025.25.1433
Published: 20 June 2025

Unbounded $\mathfrak{sl}_3$-laminations and their shear coordinates

TSUKASA ISHIBASHI
SHUNSUKE KANO

Generalizing the work of Fock and Goncharov on rational unbounded laminations, we give a geometric model of the tropical points of the cluster variety $\mathcal{X}_{\mathfrak{sl}_3,\Sigma}$, which we call *unbounded* $\mathfrak{sl}_3$*-laminations*, based on Kuperberg's $\mathfrak{sl}_3$-webs. We introduce their tropical cluster coordinates as an $\mathfrak{sl}_3$-analogue of Thurston's shear coordinates associated with any ideal triangulation. As a tropical analogue of gluing morphisms among the moduli spaces $\mathcal{P}_{\mathrm{PGL}_3,\Sigma}$ of Goncharov and Shen, we describe a geometric gluing procedure of unbounded $\mathfrak{sl}_3$-laminations with pinnings via "shearings". We also investigate a relation to the graphical basis of the $\mathfrak{sl}_3$-skein algebra of Ishibashi and Yuasa (2023), which conjecturally leads to a quantum duality map.

13F60, 57K20, 57K31

1 Introduction

1.1 Background

The notion of measured geodesic laminations (or its equivalent, measured foliations) on a surface was first introduced by W Thurston [43], as a powerful geometric tool to study the mapping class groups and the large-scale geometry of the Teichmüller space. After a couple of decades, Fock and Goncharov [11] studied Thurston's *shear coordinates* on the space $\widehat{\mathcal{ML}}(\Sigma)$ of (enhanced) measured geodesic laminations on a marked surface Σ, which gives a global coordinate system parametrized by the interior edges of an ideal triangulation $\triangle$ of Σ: $\widehat{\mathcal{ML}}(\Sigma) \xrightarrow{\sim} \mathbb{R}^{e_{\mathrm{int}}(\triangle)}$. Moreover, they observed that these coordinates can be viewed as a "tropical analogue" of the cross-ratio coordinates[1] on the enhanced Teichmüller space $\widehat{\mathcal{T}}(\Sigma) \xrightarrow{\sim} \mathbb{R}_{>0}^{e_{\mathrm{int}}(\triangle)}$ studied by Fock and Chekhov [8], as their coordinate transformation rule is exactly the tropical analogue of that for the latter. These facts indicate that there would be a universal algebraic object behind the Teichmüller and lamination spaces: this idea leads to the theory of *cluster varieties* developed by Fock and Goncharov [13]. In their terms, there is a cluster $\mathcal{X}$-variety[2] $\mathcal{X}_\Sigma^{\mathrm{uf}}$ associated with Σ such that the spaces $\widehat{\mathcal{T}}(\Sigma)$ and $\widehat{\mathcal{ML}}(\Sigma)$ are naturally identified with the spaces $\mathcal{X}_\Sigma^{\mathrm{uf}}(\mathbb{R}_{>0})$, $\mathcal{X}_\Sigma^{\mathrm{uf}}(\mathbb{R}^T)$ of positive real points and the real tropical points, respectively. We call the latter space $\mathcal{X}_\Sigma^{\mathrm{uf}}(\mathbb{R}^T)$ the *tropical cluster* $\mathcal{X}$*-variety* for short.

[1] The cross-ratio coordinate is an exponential version of the shear coordinate on the Teichmüller space. In this paper, we always use the term "shear coordinates" for those on the lamination spaces.

[2] Here, the superscript "uf" just indicates that it has only unfrozen coordinates. It corresponds to the situation where the shear/cross-ratio coordinates are defined only for internal edges $e_{\mathrm{int}}(\triangle)$ of an ideal triangulation $\triangle$.

In general, cluster varieties are schemes constructed from combinatorial data s (such as quivers) equipped with a birational atlas whose coordinate changes are given by specific rational transformations, called cluster transformations (see the appendix for a short review of this theory). They always come in a dual pair $(\mathcal{A}_{\mathsf{s}}, \mathcal{X}_{\mathsf{s}})$, forming a *cluster ensemble*. The *duality conjecture* is a profound conjecture of Fock and Goncharov [13] that asks for a construction of "duality maps"

$$\mathbb{I}_{\mathcal{X}}\colon \mathcal{X}_{\mathsf{s}}(\mathbb{Z}^T) \to \mathcal{O}(\mathcal{A}_{\mathsf{s}^\vee}) \quad (\text{resp. } \mathbb{I}_{\mathcal{A}}\colon \mathcal{A}_{\mathsf{s}}(\mathbb{Z}^T) \to \mathcal{O}(\mathcal{X}_{\mathsf{s}^\vee}))$$

which parametrizes a linear basis of the function ring $\mathcal{O}(\mathcal{A}_{\mathsf{s}^\vee})$ (resp. $\mathcal{O}(\mathcal{X}_{\mathsf{s}^\vee})$) of the dual cluster variety by the space $\mathcal{X}_{\mathsf{s}}(\mathbb{Z}^T) \subset \mathcal{X}_{\mathsf{s}}(\mathbb{R}^T)$ (resp. $\mathcal{A}_{\mathsf{s}}(\mathbb{Z}^T) \subset \mathcal{A}_{\mathsf{s}}(\mathbb{R}^T)$) of integral tropical points, satisfying certain strong axioms such as the positivity of structure constants.

In the surface case, the spaces $\mathcal{A}_\Sigma(\mathbb{R}_{>0})$ and $\mathcal{A}_\Sigma(\mathbb{R}^T)$ are identified with the decorated Teichmüller and lamination spaces — see Papadopoulos and Penner [39; 40] — via the λ-length and intersection coordinates [11]. The geometric realization of the tropical spaces $\mathcal{A}_\Sigma(\mathbb{Z}^T)$ $\mathcal{X}_\Sigma^{\mathrm{uf}}(\mathbb{Z}^T)$ by integral laminations [11] leads to a topological construction of the duality maps $\mathbb{I}_{\mathcal{X}}$ and $\mathbb{I}_{\mathcal{A}}$, and their required properties were proved recently by Mandel and Qin [37] based on a comparison with the *theta basis* of Gross, Hacking, Keel and Kontsevich [22]. These duality maps are two kinds of generalizations of the trace function basis for the function ring of the SL_2-character variety of a closed surface, parametrized by loops.

Strongly expected are "higher rank" generalizations of the above picture. The cluster varieties $\mathcal{X}_\Sigma^{\mathrm{uf}}$ and $\mathcal{A}_\Sigma$ are birationally isomorphic to certain generalizations of the PGL_2- and SL_2-character varieties; see Fock and Goncharov [10]. As a generalization for higher rank algebraic groups, there are cluster varieties $\mathcal{X}_{\mathfrak{g},\Sigma}^{\mathrm{uf}}$ and $\mathcal{A}_{\mathfrak{g},\Sigma}$ which are birationally isomorphic to the same kind of generalizations $\mathcal{X}_{G',\Sigma}$ and $\mathcal{A}_{G,\Sigma}$ of character varieties — see Fock, Goncharov and Shen [10; 21] and Le [35] — where the defining combinatorial data for these cluster varieties only depend on the surface Σ and a semisimple Lie algebra $\mathfrak{g}$. In particular, $\mathcal{X}_{\mathfrak{sl}_2,\Sigma}^{\mathrm{uf}} = \mathcal{X}_\Sigma^{\mathrm{uf}}$ and $\mathcal{A}_{\mathfrak{sl}_2,\Sigma} = \mathcal{A}_\Sigma$ correspond to the case mentioned above. Goncharov and Shen [21] introduced a cluster variety $\mathcal{X}_{\mathfrak{g},\Sigma}$ with frozen coordinates, which is birational to some extension $\mathcal{P}_{G',\Sigma}$ of $\mathcal{X}_{G',\Sigma}$. Hereupon, we have combinatorially defined tropical spaces $\mathcal{A}_{\mathfrak{g},\Sigma}(\mathbb{R}^T)$ and $\mathcal{X}_{\mathfrak{g},\Sigma}(\mathbb{R}^T)$, which should parametrize linear bases of the function rings of the dual varieties with good properties by the duality conjecture. The spaces $\mathcal{A}_{\mathfrak{g},\Sigma}(\mathbb{R}^T)$ and $\mathcal{X}_{\mathfrak{g},\Sigma}(\mathbb{R}^T)$ are widely expected to be certain spaces of $\mathfrak{g}$-webs on Σ, so that the duality maps are built from the web functions on the character variety. However, such a web description is still missing in general. We remark here that Le [34] gave a description of these spaces in terms of certain configurations in the affine buildings, which should be ultimately related to $\mathfrak{g}$-webs based on the geometric Satake correspondence (see, for instance, Fontaine, Kamnitzer and Kuperberg [18]).

For the first nontrivial case $\mathfrak{g} = \mathfrak{sl}_3$, major progress on the space $\mathcal{A}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T)$ has been made by Douglas and Sun [6; 7] and Kim [32]. They describe this space as an appropriate space of Kuperberg's $\mathfrak{sl}_3$-webs [33] by introducing an $\mathfrak{sl}_3$-version of the intersection coordinates with an ideal triangulation. Their coordinates can also be extended to the space $\mathcal{A}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$ by scaling equivariance.

1.2 Geometric model for the tropical space $\mathcal{X}^{\mathrm{uf}}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$

Our aim in this paper is to describe the tropical cluster variety $\mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$ on the dual side as a space of $\mathfrak{sl}_3$-webs with a different type of boundary conditions and some additional structures at punctures. We introduce the space $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ of rational unbounded $\mathfrak{sl}_3$-laminations on Σ, which are certain equivalence classes of nonelliptic *signed* $\mathfrak{sl}_3$*-webs* with positive rational weights (see Section 2.2). Then we define an $\mathfrak{sl}_3$-version of the shear coordinates of these objects with respect to an ideal triangulation $\triangle$. As in the $\mathfrak{sl}_2$-case, we need to perturb the ends incident at punctures (and thus make them spiraling) so that they intersect with $\triangle$ transversely. The spiraling directions are controlled by the signs assigned to each end of the $\mathfrak{sl}_3$-web, and this procedure leads to the notion of *spiraling diagrams* (Definition 3.8) associated with signed $\mathfrak{sl}_3$-webs. After a careful study on the "good positions" of a spiraling diagram, we obtain well-defined shear coordinates.

Theorem 1 (Theorem 3.20) *For any marked surface Σ satisfying conditions* (S1)–(S4) *in Section 2.1 and its ideal triangulation $\triangle$ without self-folded triangles, we have a bijection*

$$\mathsf{x}^{\mathrm{uf}}_{\triangle}\colon \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \xrightarrow{\sim} \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)}, \tag{1-1}$$

which we call the shear coordinate system associated with $\triangle$. Moreover, for any another ideal triangulation $\triangle'$ of Σ, the coordinate transformation $\mathsf{x}_{\triangle'}\circ\mathsf{x}_{\triangle}^{-1}$ is a composite of tropical cluster $\mathcal{X}$-transformations.

As a consequence, the shear coordinates combine to give an $\mathrm{MC}(\Sigma)$-equivariant bijection

$$\mathsf{x}^{\mathrm{uf}}_{\bullet}\colon \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \xrightarrow{\sim} \mathcal{X}^{\mathrm{uf}}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T). \tag{1-2}$$

Therefore, our space $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ of unbounded $\mathfrak{sl}_3$-laminations gives a geometric model for the tropical cluster $\mathcal{X}$-variety $\mathcal{X}^{\mathrm{uf}}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$. In other words, the space $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ can be viewed as a tropical analogue of the moduli space $\mathcal{X}_{PGL_3,\Sigma}$ of framed PGL_3-local systems [10].

In Section 3.4, we give an explicit inverse map of $\mathsf{x}^{\mathrm{uf}}_{\triangle}$ by gluing local building blocks according to the shear coordinates, in the same spirit as Fock and Goncharov. The coordinate transformation formula could be obtained by case-by-case as in [7] for the $\mathcal{A}$-side. However, in order to reduce the length of computation, we choose to derive it from the computation on the $\mathcal{A}$-side performed by Douglas and Sun after investigating their relation in detail (see Theorem 2 below). So the second statement in Theorem 1 follows from Theorem 2.

1.3 Unbounded $\mathfrak{sl}_3$-laminations with pinnings and their gluing

In order to supply the frozen coordinates, we further introduce a larger space $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ of unbounded $\mathfrak{sl}_3$-laminations *with pinnings* by attaching additional data on boundary intervals, in the same spirit as

Goncharov and Shen's construction of the moduli space $\mathcal{P}_{G',\Sigma}$ [21]. As in their work, these additional data allow us to glue the $\mathfrak{sl}_3$-laminations along boundary intervals, which leads to the *gluing map*

$$q_{E_L,E_R}\colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma',\mathbb{Q}) \tag{1-3}$$

where Σ' is the marked surface obtained from Σ by gluing two boundary intervals E_L and E_R.

The space $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ is also suited for the comparison with the works of Douglas and Sun [6; 7] and Kim [32]. Let $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ denote the space of rational bounded $\mathfrak{sl}_3$-laminations, which essentially appears in these works. See Remark 2.10. Then we define a *geometric ensemble map*

$$\tilde{p}\colon \mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \tag{1-4}$$

by forgetting the peripheral components, and assigning pinnings in a certain way. When Σ has no punctures, $\tilde{p}$ gives a bijection. For these structures, we obtain the following:

Theorem 2 (Theorems 4.7 and 4.11 and Proposition 4.10) *Under the same assumption as in Theorem 1, we have a bijection*

$$\mathsf{x}_\triangle\colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \xrightarrow{\sim} \mathbb{Q}^{I(\triangle)}, \tag{1-5}$$

whose coordinate transformations are given by tropical cluster $\mathcal{X}$-transformations (including frozen coordinates). Via these coordinate systems:

1. *The gluing map q_{E_L,E_R} coincides with the tropicalization of the amalgamation map* [9].
2. *The geometric ensemble map $\tilde{p}$ coincides with the tropicalization of the Goncharov–Shen extension of the ensemble map* (A-6).

We will also see in Section 4.4 that the shear coordinates are equivariant under the Dynkin involution $*$, which generates $\mathrm{Out}(\mathrm{SL}_3)$. In particular, we have an $MC(\Sigma)\times\mathrm{Out}(\mathrm{SL}_3)$-equivariant bijection

$$\mathsf{x}_\triangle\colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \xrightarrow{\sim} \mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T). \tag{1-6}$$

In other words, the space $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ can be viewed as a tropical analogue of the Goncharov and Shen's moduli space $\mathcal{P}_{\mathrm{PGL}_3,\Sigma}$ [21].

Property (1) allows one to reduce the computation of coordinate transformations to those for smaller surfaces. For a surface without punctures, the map $\tilde{p}$ is a bijection and property (2) shows that this map intertwines the two types of cluster transformations. This is our strategy to obtain the coordinate transformation formula for (1-5).

In our sequel paper [28], we will investigate the unbounded $\mathfrak{sl}_3$-laminations around punctures in detail, and study the tropicalizations of the cluster exact sequence of Fock and Goncharov [13] and the Weyl group actions at punctures introduced by Goncharov and Shen [20] in terms of $\mathfrak{sl}_3$-laminations. In the end, the bijections (1-2) and (1-6) turn out to be equivariant under the natural action of the group $(MC(\Sigma)\times\mathrm{Out}(\mathrm{SL}_3))\ltimes W(\mathfrak{sl}_3)^{\mathbb{M}_\circ}$.

1.4 Relation to the graphical basis of the skein algebra $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}$

As mentioned in the beginning, our space $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z}) \cong \mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T)$ is expected to parametrize a linear basis of the function ring $\mathcal{O}(\mathcal{A}_{\mathfrak{sl}_3,\Sigma})$. When the marked surface has no punctures (hence the exchange matrix has full-rank), it is also expected to parametrize a linear basis of the *quantum upper cluster algebra* $\mathcal{O}_q(\mathcal{A}_{\mathfrak{sl}_3,\Sigma})$ of Berenstein and Zelevinsky [3]. On the other hand, a skein model for $\mathcal{O}_q(\mathcal{A}_{\mathfrak{sl}_3,\Sigma})$ is investigated in [30] by the first named author and W Yuasa. They study a skein algebra $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}$ with appropriate "clasped" skein relations at marked points, and constructed an inclusion of its boundary-localization $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}]$ into the quantum cluster algebra (and hence into $\mathcal{O}_q(\mathcal{A}_{\mathfrak{sl}_3,\Sigma})$). Conjecturally these algebras coincide with each other. They give a $\mathbb{Z}_q$-basis $\mathsf{BWeb}_{\mathfrak{sl}_3,\Sigma}$ of the skein algebra $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}$ consisting of flat trivalent graphs. In this paper, we relate our integral $\mathfrak{sl}_3$-laminations with pinnings to the basis webs:

Theorem 3 (Theorem 5.2) *Assume that Σ has no punctures. Then we have an* $\mathrm{MC}(\Sigma)\times\mathrm{Out}(\mathrm{SL}_3)$*-equivariant bijection*

$$\mathbb{I}^q_{\mathcal{X}}\colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})_+ \xrightarrow{\sim} \mathsf{BWeb}_{\mathfrak{sl}_3,\Sigma} \subset \mathscr{S}^q_{\mathfrak{sl}_3,\Sigma},$$

where $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})_+ \subset \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ denotes the subspace of dominant integral $\mathfrak{sl}_3$-laminations. Moreover, it is extended to a map $\mathbb{I}^q_{\mathcal{X}}\colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z}) \hookrightarrow \mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}]$, whose image gives a $\mathbb{Z}_q$-basis of $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}]$.

The latter correspondence should be a basic ingredient for a construction of the *quantum duality map* [13] (see Qin [41, Conjecture 4.14] for a finer formulation as well as Davison and Mandel [5]). See Section 5 for a detailed discussion. Our general expectation is the following:

Conjecture 4 *The basis $\mathbb{I}^q_{\mathcal{X}}(\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z}))$ is* ***parametrized by tropical points*** *in the sense of* [41, Definition 4.13]. *Namely, for any integral $\mathfrak{sl}_3$-lamination $\hat{L} \in \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$, the quantum Laurent expression of $\mathbb{I}^q_{\mathcal{X}}(\hat{L}) \in \mathscr{A}^q_{\mathfrak{sl}_3,\Sigma}$ in the quantum cluster $\{A_i\}_{i\in I}$ associated with a vertex $\omega \in \mathbb{E}\mathrm{xch}_{\mathfrak{sl}_3,\Sigma}$ has the leading term $\bigl[\prod_{i\in I} A_i^{\mathsf{x}_i(\hat{L})}\bigr]$ with respect to the dominance order* [41, Definition 4.6], *where $\mathsf{x}^{(\omega)} = (\mathsf{x}_i)_{i\in I}$ is the shear coordinate system associated with ω.*

Currently we do not know if it gives a basis with positivity (of Laurent expressions and/or structure constants), or it requires a modification by using an $\mathfrak{sl}_3$-version of *bracelets*; see D Thurston [42]. See also Allegretti and Kim [1; 2] and Cho, Kim, Kim and Oh [4] for the progress on the positivity problem for the $\mathfrak{sl}_2$-case.

1.5 Future directions: real unbounded $\mathfrak{sl}_3$-laminations

Let $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{R})$ be the completion of the space $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ such that each shear coordinate system (1-1) extends to a homeomorphism $\mathsf{x}^{\mathrm{uf}}_{\triangle}\colon \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{R}) \xrightarrow{\sim} \mathbb{R}^{I_{\mathrm{uf}}(\triangle)}$. It is well defined since the cluster $\mathcal{X}$-transformations are Lipschitz continuous with respect to the Euclidean metrics on $\mathbb{Q}^{I_{\mathrm{uf}}(\triangle)}$. We call an element of $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{R})$ a *real unbounded $\mathfrak{sl}_3$-lamination*, which is represented by a Cauchy sequence in $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ with respect to shear coordinates. The space $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{R})$ has a natural PL structure, and is considered to be an $\mathfrak{sl}_3$-analogue of the space $\widehat{\mathcal{ML}}(\Sigma)$ of measured geodesic laminations. Recall that in the Teichmüller–Thurston theory, the latter PL manifold plays the following roles (among others):

Boundary at infinity of the Teichmüller space The *Thurston compactification* is a compactification of the Teichmüller space into a topological disk obtained by attaching the projectivization of $\widehat{\mathcal{ML}}(\Sigma)$, so that the mapping class group action is continuously extended. The measured geodesic laminations encode the "rate" of degenerations of geodesics in a divergent sequence in the Teichmüller space. The Thurston compactification is identified with the *Fock–Goncharov compactification* [14; 24; 34] $\overline{\mathcal{X}_\Sigma(\mathbb{R}_{>0})} = \mathcal{X}_\Sigma(\mathbb{R}_{>0}) \cup \mathbb{S}\mathcal{X}_\Sigma(\mathbb{R}^T)$, which is defined for any cluster $\mathcal{X}$-variety.

Place for analyzing the pseudo-Anosov dynamics The PL action of the mapping class group on $\widehat{\mathcal{ML}}(\Sigma)$ provides us rich information on the dynamics of pseudo-Anosov mapping classes. In particular, each pseudo-Anosov mapping class has the north-south dynamics on the projectivized space, and its unique attracting/repelling points are represented by a transverse pair of measured geodesic laminations. A generalization of these specific properties for elements of a general cluster modular group is proposed in [25; 26; 27], which we call the *sign stability*. The equivalence between the "uniform" sign stability and the pseudo-Anosov property is discussed in [25], based on the identification $\widehat{\mathcal{ML}}(\Sigma) \cong \mathcal{X}^{\mathrm{uf}}_\Sigma(\mathbb{R}^T)$.

It is natural to expect that the space $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{R})$ plays the same role in the $\mathfrak{sl}_3$-case. Since the positive real part $\mathcal{X}^{\mathrm{uf}}_{\mathfrak{sl}_3,\Sigma}(\mathbb{R}_{>0})$ has been identified with the moduli space of convex $\mathbb{RP}^2$-structures on Σ, the real unbounded $\mathfrak{sl}_3$-laminations are expected to encode their degenerations. The PL action of a pseudo-Anosov mapping class on the space $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{R})$ is expected to provide more rich information, which may possibly lead to a finer classification of pseudo-Anosov mapping classes. Although a concrete description of a real unbounded $\mathfrak{sl}_3$-lamination as a certain geometric object (rather than a sequence) is still missing, the cluster algebraic interpretation of Thurston's train tracks studied in [31] will be a useful tool.

Generalizations of Thurston's earthquake maps and the Hubbard–Masur theorem that relates measured foliations with quadratic differentials will be also interesting topics. A study on a cluster algebraic analogue of these theories is in progress by the authors with Takeru Asaka.

Organization of the paper

Main part (Sections 2–4) In Section 2, we introduce rational unbounded $\mathfrak{sl}_3$-laminations and briefly discuss the relation to the works of Douglas and Sun [6; 7] and Kim [32]. We study the associated spiraling diagrams and define the shear coordinates in Section 3. The bijectivity of the shear coordinate systems (1-1) is proved. In Section 4, we introduce pinnings for rational unbounded $\mathfrak{sl}_3$-laminations and discuss their gluing and the extended ensemble map. Theorem 2 is proved, and hence the proof of Theorem 1 is completed.

Relation to skein theory (Section 5) We investigate the relation to the skein algebra and quantum duality map in Section 5. Theorem 3 is proved here.

Proofs for the technical statements (Section 6) The proofs of Theorems 3.10 and 3.19 are placed in Section 6. Logically they do not depend on the contents after the places where the statements are written.

Basic terminology on the cluster varieties and the known results we need for the $\mathfrak{sl}_3$-case are collected in the appendix.

Acknowledgements

The authors thank Hyun Kyu Kim for illuminating discussion on quantum duality maps. They also appreciate the referee's careful reading and suggestions. Ishibashi is grateful to Wataru Yuasa for a valuable discussion on $\mathfrak{sl}_3$-skein algebras in the early stage of this work. He is also grateful to Zhe Sun for explaining his works on the $\mathcal{A}$-side, and giving valuable comments on this work at several stages. Kano appreciates the support by Wataru Yuasa for his visit to Kyoto University in the spring of 2021. The main part of this work was done during this visit. Ishibashi is partially supported by JSPS KAKENHI (20K22304). Kano is partially supported by scientific research support of Research Alliance Center for Mathematical Sciences, Tohoku University.

2 Unbounded $\mathfrak{sl}_3$-laminations and their shear coordinates

2.1 Marked surfaces and their triangulations

A marked surface $(\Sigma, \mathbb{M})$ is a compact oriented surface Σ together with a fixed nonempty finite set $\mathbb{M} \subset \Sigma$ of *marked points*. When the choice of $\mathbb{M}$ is clear from the context, we simply denote a marked surface by Σ. A marked point is called a *puncture* if it lies in the interior of Σ, and a *special point* otherwise. Let $\mathbb{M}_\circ = \mathbb{M}_\circ(\Sigma)$ (resp. $\mathbb{M}_\partial = \mathbb{M}_\partial(\Sigma)$) denote the set of punctures (resp. special points), so that $\mathbb{M} = \mathbb{M}_\circ \sqcup \mathbb{M}_\partial$. Let $\Sigma^* := \Sigma \setminus \mathbb{M}_\circ$. We always assume the following conditions:

(S1) Each boundary component (if exists) has at least one marked point.

(S2) $-2\chi(\Sigma^*) + |\mathbb{M}_\partial| > 0$.

(S3) $(\Sigma, \mathbb{M})$ is not a once-punctured disk with a single special point on the boundary.

We call a connected component of the punctured boundary $\partial^*\Sigma := \partial\Sigma \setminus \mathbb{M}_\partial$ a *boundary interval*. The set of boundary intervals is denoted by $\mathbb{B} = \mathbb{B}(\Sigma)$. We always endow each boundary interval with the orientation induced from $\partial\Sigma$. Then we have $|\mathbb{M}_\partial| = |\mathbb{B}|$.

Unless otherwise stated, an *isotopy* in a marked surface $(\Sigma, \mathbb{M})$ means an ambient isotopy in Σ relative to $\mathbb{M}$, which preserves each boundary interval setwise. An *ideal arc* in $(\Sigma, \mathbb{M})$ is an immersed arc in Σ with endpoints in $\mathbb{M}$ which has no self-intersection except possibly at its endpoints, and not isotopic to one point.

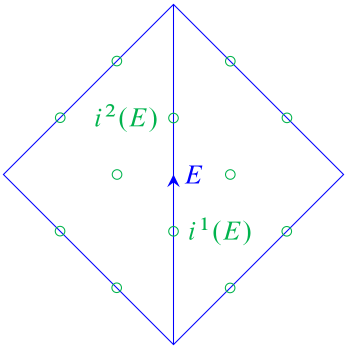
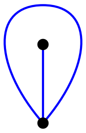


Figure 1: The set $I(\triangle)$ of distinguished points.

An *ideal triangulation* is a triangulation $\triangle$ of Σ whose set of 0-cells (vertices) coincides with $\mathbb{M}$. Conditions (S1) and (S2) ensure the existence of such an ideal triangulation, and the positive integer in (S2) gives the number of 2-cells (triangles). The 1-cells (edges) are necessarily ideal arcs. In this paper, we always consider an ideal triangulation without *self-folded triangles* of the form



Such an ideal triangulation exists by condition (S3). See, for instance, [15, Lemma 2.13]. For an ideal triangulation $\triangle$, denote the set of edges (resp. interior edges, triangles) of $\triangle$ by $e(\triangle)$ (resp. $e_{\mathrm{int}}(\triangle)$, $t(\triangle)$). Since the boundary intervals belong to any ideal triangulation, $e(\triangle) = e_{\mathrm{int}}(\triangle) \sqcup \mathbb{B}$. By a computation on the Euler characteristics, we get

$$|e(\triangle)| = -3\chi(\Sigma^*) + 2|\mathbb{M}_\partial|, \quad |e_{\mathrm{int}}(\triangle)| = -3\chi(\Sigma^*) + |\mathbb{M}_\partial|, \quad |t(\triangle)| = -2\chi(\Sigma^*) + |\mathbb{M}_\partial|.$$

It is useful to equip $\triangle$ with two distinguished points on the interior of each edge and one point in the interior of each triangle, as shown in Figure 1. The set of such points is denoted by $I(\triangle) = I_{\mathfrak{sl}_3}(\triangle)$. This set will give the vertex set of the quiver $Q^{\triangle}$ associated with $\triangle$; see Section A.3. Let $I^{\mathrm{edge}}(\triangle)$ (resp. $I^{\mathrm{tri}}(\triangle)$) denote the set of points on edges (resp. faces of triangles) so that $I(\triangle) = I^{\mathrm{edge}}(\triangle) \sqcup I^{\mathrm{tri}}(\triangle)$, where we have a canonical bijection

$$t(\triangle) \xrightarrow{\sim} I^{\mathrm{tri}}(\triangle), \quad T \mapsto i(T).$$

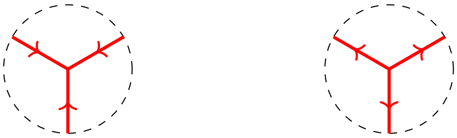
When we need to label the two vertices on an edge $E \in e(\triangle)$, we endow E with an orientation. Then let $i^1(E) \in I(\triangle)$ (resp. $i^2(E) \in I(\triangle)$) denote the vertex closer to the initial (resp. terminal) endpoint of E. Let $I(\triangle)_{\mathrm{f}} \subset I^{\mathrm{edge}}(\triangle)$ ("frozen") be the subset consisting of the points on the boundary, and let $I(\triangle)_{\mathrm{uf}} := I(\triangle) \setminus I(\triangle)_{\mathrm{f}}$ ("unfrozen"). The numbers

$$|I(\triangle)| = 2|e(\triangle)| + |t(\triangle)| = -8\chi(\Sigma^*) + 5|\mathbb{M}_\partial|,$$
$$|I(\triangle)_{\mathrm{uf}}| = 2|e_{\mathrm{int}}(\triangle)| + |t(\triangle)| = -8\chi(\Sigma^*) + 3|\mathbb{M}_\partial|$$

will give the dimensions of the PL manifolds $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{R})$ and $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{R})$ respectively.

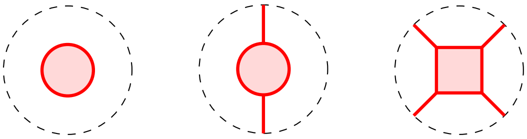
2.2 Unbounded $\mathfrak{sl}_3$-laminations

Recall that a *uni-trivalent* graph is a (possibly disconnected and/or infinite) graph whose vertices have valency either one or three. It is allowed to have a loop component (ie a connected component without vertices). An *orientation* of a uni-trivalent graph is an assignment of an orientation on each edge and loop such that any trivalent vertex is either a *sink* or a *source*, respectively:

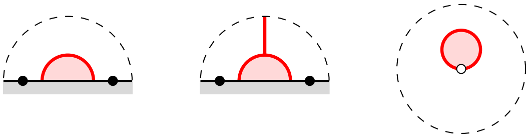


An $\mathfrak{sl}_3$-web (or simply a *web*) on a marked surface Σ is an immersed oriented uni-trivalent graph W on Σ such that each univalent vertex lie in $\mathbb{M}_\circ \cup \partial^* \Sigma$, and the other part is embedded into $\operatorname{int} \Sigma^*$. It is said to be *nonelliptic* if it has none of the following *elliptic faces*:

(2-1)



(2-2)



A web is said to be *bounded* if none of its univalent vertices lie in $\mathbb{M}_\circ$.

We will mostly deal with finite webs, while infinite ones appear when (and only when) we discuss spiraling diagrams (Definition 3.8), which are still locally finite except possibly around punctures. When we simply say an ($\mathfrak{sl}_3$-)web below, it will mean a finite web. When the web in consideration can be infinite, we will say a “possibly infinite web”.

Remark 2.1 The exclusion of the internal faces in (2-1) is usual in literature. Indeed, a web containing these faces can be written as a linear combination of nonelliptic webs in the skein algebra (see Section 5), and hence not needed as a basis element. The first two faces in (2-2) are excluded as variants of boundary skein relations [30]. It is also related to the *weakly reduced* condition in [19]. The third one can be regarded as a variant for a boundary component without marked points.

Example 2.2 (honeycomb webs) Let $T \subset \operatorname{int} \Sigma^*$ be an embedded triangle. For each positive integer n, the incoming (resp. outgoing) *honeycomb-web* (or *pyramid web*) in T of height n is the $\mathfrak{sl}_3$-web dual to the *n-triangulation* of T, oriented so that the outer-most edges are incoming to (resp. outgoing from) T. See the left picture in Figure 2 for an example. We will also use a short-hand presentation as shown in the right of Figure 2. The embedded image of a honeycomb web in Σ is simply called a *honeycomb*. The ends of a honeycomb can be connected with other oriented arcs or honeycombs on Σ.

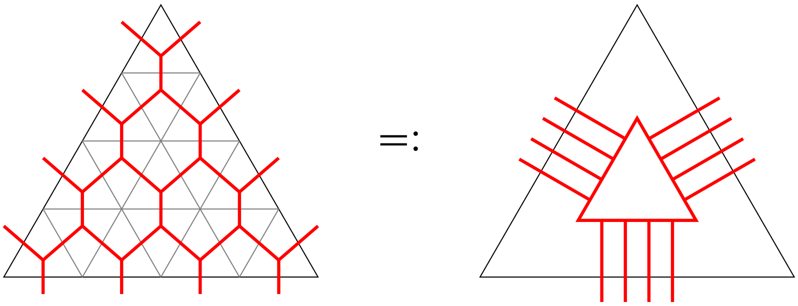
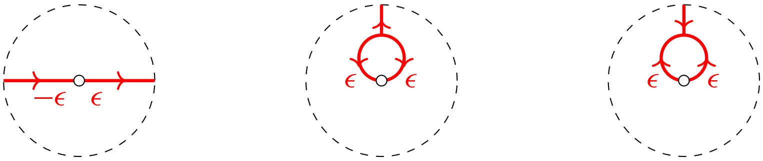


Figure 2: A honeycomb-web on a triangle T of height $n = 4$ (left) and its short-hand presentation (right).

A *signed web* is a web on Σ together with a sign ($+$ or $-$) assigned to each end incident to a puncture. The following patterns (and their orientation-reversals) of signed ends are called *bad ends*:

(2-3)



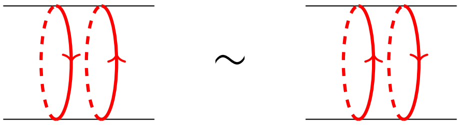
Here $\epsilon \in \{+,-\}$. A signed web is said to be *admissible* if it has no bad ends. In this paper, we always assume that the signed webs are admissible unless otherwise stated. A bounded web is naturally regarded as a signed web since we do not need to specify any signs.

Remark 2.3 The latter two types of bad ends will be excluded simply because they will not contribute to the shear coordinates. On the other hand, a pair of the first type will have nontrivial coordinates, while there is always another web that attains the same coordinates. So we only need admissible signed webs to realize the tropical space. It turns out that we need to include the bad ends of first type to define the Weyl group actions at punctures [28].

Elementary moves of signed webs We are going to introduce several elementary moves for signed webs. The first two are defined for a web without signs.

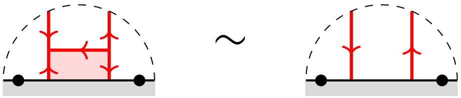
(E1) Loop parallel-move (aka *flip move* [19] or *global parallel move* [6]):

(2-4)



(E2) Boundary H-move:

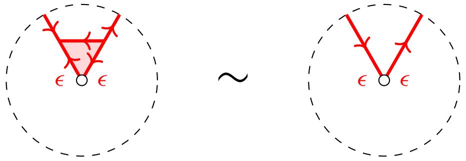
(2-5)



Similarly for the opposite orientation. We call the face in the left-hand side a *boundary H-face*.

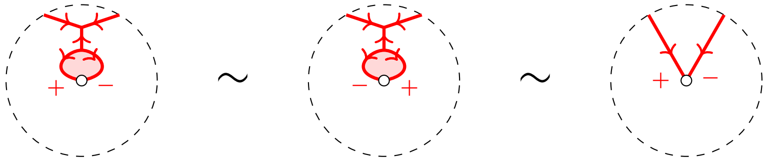
(E3) Puncture H-moves:

(2-6)



for $\epsilon \in \{+,-\}$, and

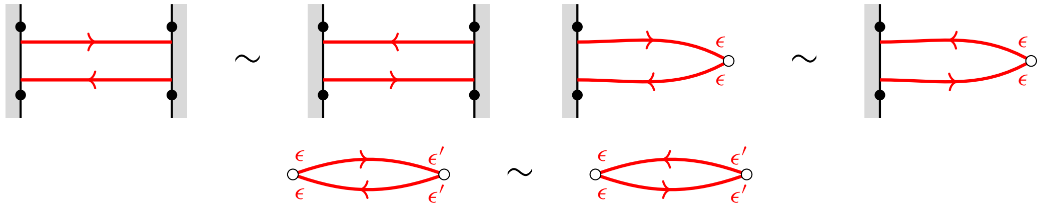
(2-7)



Similarly for the opposite orientation. We call the face in the left-hand side of (2-6) a *puncture H-face*.

The following lemma is verified by using (E2) and the first one in (E3):

Lemma 2.4 *From the boundary and puncture H-moves, we get the following "arc parallel-moves" swapping parallel arcs with opposite orientations*:

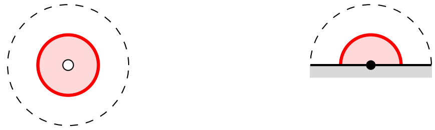


Here white (resp. black) circles stand for punctures (resp. special points), and $\epsilon, \epsilon' \in \{+,-\}$.

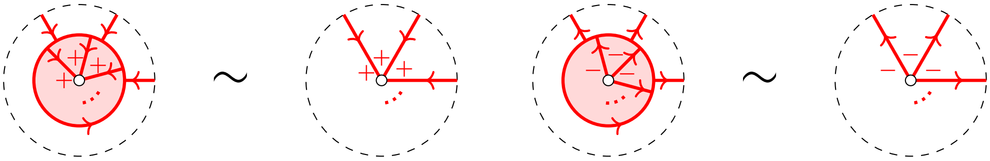
Also note that we can always transform any signed web to a signed web without boundary H-faces (resp. puncture H-faces) by applying (E2) and (E3), respectively. Slightly generalizing the terminology in [19], such a signed web is said to be *boundary-reduced* (resp. *puncture-reduced*). It is said to be *reduced* if it is both boundary- and puncture-reduced.

(E4) Peripheral move: removing or creating a peripheral component:

(2-8)



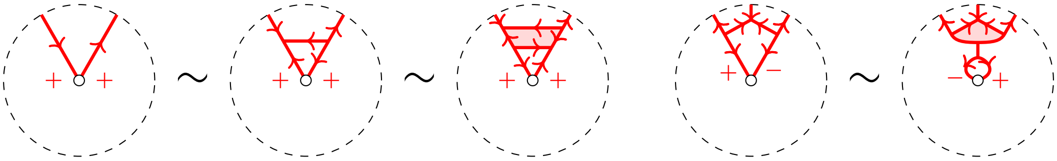
Moreover, we have the moves



Similarly for the opposite orientation.

We will consider the equivalence relation on signed webs generated by isotopies of marked surfaces and the elementary moves (E1)–(E4). Observe that the moves (E1)–(E4) preserves the admissibility. On the other hand, a nonelliptic signed web may be equivalent to an elliptic web as the following example shows.

Example 2.5 We have



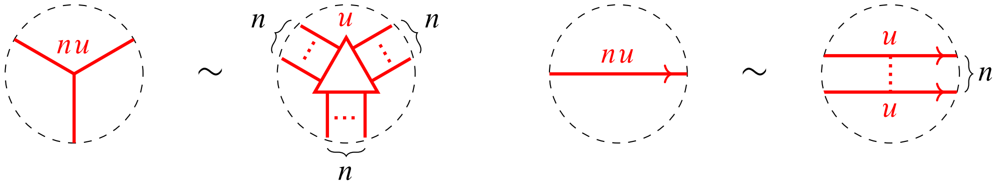
by the puncture H-moves (2-6) and (2-7), where the resulting signed webs are elliptic (having interior 4-gon faces).

Definition 2.6 (rational unbounded $\mathfrak{sl}_3$-laminations) A *rational unbounded* $\mathfrak{sl}_3$*-lamination* (or a rational $\mathfrak{sl}_3$-$\mathcal{X}$-lamination) on Σ is an admissible, nonelliptic signed $\mathfrak{sl}_3$-web W on Σ equipped with a positive rational number (called the *weight*) on each component, which is considered modulo the equivalence relation generated by isotopies and the following operations:

(1) Elementary moves (E1)–(E4) for the underlying signed webs. Here the corresponding components are assumed to have the same weights.

(2) Combine a pair of isotopic loops with the same orientation with weights u and v into a single loop with the weight $u+v$. Similarly combine a pair of isotopic oriented arcs with the same orientation (and with the same signs if some of their ends are incident to punctures) into a single one by adding their weights.

(3) For an integer $n \in \mathbb{Z}_{>0}$ and a rational number $u \in \mathbb{Q}_{>0}$, replace a component with weight nu with its n*-cabling* with weight u, which locally looks like



For a loop or arc component, it is just a successive applications of operation (2). One can also verify that the cabling operation is associative in the sense that the n-cabling followed by the m-cabling agrees with the nm-cabling, since nm-cabling is dual to the m^{th} subdivision of an n-triangulation (recall Figure 2).

See Figure 3 for a global example. Let $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ denote the set of equivalence classes of the rational unbounded $\mathfrak{sl}_3$-laminations on Σ. We have a natural $\mathbb{Q}_{>0}$-action on $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ that simultaneously rescales the weights. A rational unbounded $\mathfrak{sl}_3$-lamination is said to be *integral* if all the weights are integers. The subset of integral unbounded $\mathfrak{sl}_3$-laminations is denoted by $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$.

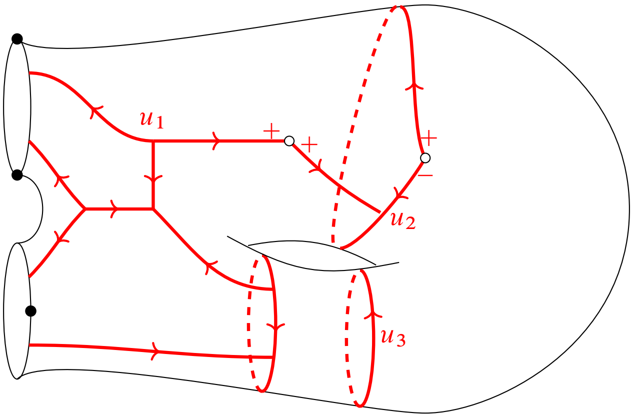
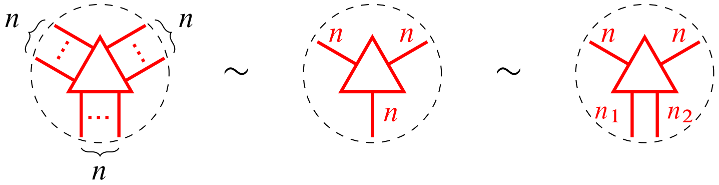


Figure 3: An example of a rational unbounded $\mathfrak{sl}_3$-lamination. Here u_1, u_2 and u_3 are arbitrary positive rational weights.

The sets $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ and $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ will be identified with the unfrozen part $\mathcal{X}^{\mathrm{uf}}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$ and $\mathcal{X}^{\mathrm{uf}}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T)$, respectively, of the tropical cluster $\mathcal{X}$-variety associated with the pair $(\mathfrak{sl}_3,\Sigma)$ (see Section A.3).

Notation 2.7 In view of the equivalence relation (4), we will occasionally use the following equivalent notations for honeycombs:



with $n_1+n_2=n$. We may also split an edge of weight n with k edges of weight $n_1,\dots,n_k$ with $n_1+\cdots+n_k=n$.

Definition 2.8 (Dynkin involution) The *Dynkin involution* is the involutive automorphism

$$*\colon \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})\to\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}),\quad \hat{L}\mapsto\hat{L}^*,$$

where $\hat{L}^*$ is obtained from $\hat{L}$ by reversing the orientation of every components of the underlying web, and keeping the signs at punctures intact. Since all the elementary moves (E1)–(E4) are equivariant under the orientation-reversion, this indeed defines an automorphism on $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$.

Bounded laminations and the ensemble map

Definition 2.9 (rational bounded $\mathfrak{sl}_3$-laminations) A *rational bounded* $\mathfrak{sl}_3$*-lamination* (or a *rational* $\mathfrak{sl}_3$*-*$\mathcal{A}$*-lamination*) on Σ is a bounded nonelliptic $\mathfrak{sl}_3$-web W on Σ equipped with a rational number (called the *weight*) on each component such that the weight on a nonperipheral component is positive. It is considered modulo the equivalence relation generated by isotopies and the operations (2)–(4) in Definition 2.6.

Let $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ denote the space of rational bounded $\mathfrak{sl}_3$-laminations. We have a natural $\mathbb{Q}_{>0}$-action on $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ that simultaneously rescales the weights. A rational bounded $\mathfrak{sl}_3$-lamination is said to be *integral* if all the weights are integers. The subset of integral bounded $\mathfrak{sl}_3$-laminations is denoted by $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$.

Remark 2.10 The space $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ is the same one as the space $\mathcal{A}_L(\Sigma;\mathbb{Z})$ that appears in Kim's work [32, Definition 3.9].[3] The space $\mathcal{W}_\Sigma$ in Douglas and Sun's work [6, Definition 6] is the subset of $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ consisting of elements with positive peripheral weights. It is straightforward to extend their coordinate systems by $\mathbb{Q}_{>0}$-equivariance to the rational case, and the space $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ is identified with the tropical cluster $\mathcal{A}$-variety $\mathcal{A}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$ [32, Theorem 3.39].[4]

By forgetting the peripheral components, we get the *geometric ensemble map*

$$p\colon \mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}). \tag{2-9}$$

We will see in Section 4 that the geometric ensemble map coincides with the cluster ensemble map (A-2) via the Douglas–Sun coordinates and our shear coordinates.

3 Shear coordinates

3.1 Essential webs on polygons

Let $\mathbb{D}_k$ denote a disk with $k \geq 2$ special points. In what follows, we simply refer to $\mathbb{D}_k$ as a *k-gon*. We say that an $\mathfrak{sl}_3$-web W on $\mathbb{D}_k$ is *taut* if for any compact embedded arc α whose endpoints lie in a common boundary interval E, the number of intersection points of W with E does not exceed that of W with α. See Figure 4. Following [6], we call a nonelliptic, taut $\mathfrak{sl}_3$-web an *essential* web. These essential webs on polygons are basic building blocks for the bounded $\mathfrak{sl}_3$-laminations studied in [6]. We recall the concrete description of the essential webs for $k = 2, 3$ following [6, Sections 2.7 and 2.8] and [19, Sections 8 and 9], including additional infinite webs needed for our purpose.

The biangle (2-gon) case Let E_L and E_R denote the boundary intervals of a biangle $\mathbb{D}_2$. A (*finite*) *symmetric strand set* on $\mathbb{D}_2$ is a pair $S = (S_L, S_R)$ of finite collections of disjoint oriented strands (ie germs of oriented arcs), where the oriented strands in S_Z are located on E_Z for $Z \in \{L, R\}$ such that the number of incoming (resp. outgoing) strands on E_L is equal to the outgoing (resp. incoming) strands on E_R. See the left-most picture in Figure 5 for an example.

[3] Indeed, an element of our space $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ can be represented by a reduced web [32, Definition 3.3] by applying the boundary H-moves, and we can rescale the weights on honeycombs to be 1 by the operation (4) in Definition 2.6.

[4] Here note that there is a subset of $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ formed by *congruent* laminations [32, Definition 3.38] which is identified with the tropical cluster $\mathcal{A}$-variety $\mathcal{A}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T)$.

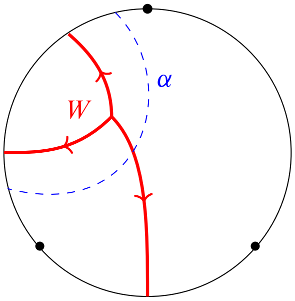
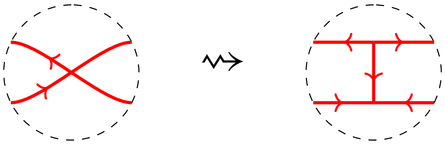


Figure 4: Example of a nontaut web in $\mathbb{D}_3$.

Given a symmetric strand set $S = (S_L, S_R)$, the associated *ladder-web* $W(S)$ on $\mathbb{D}_2$ is constructed as follows. First, let $W_{\mathrm{br}}(S)$ be the unique (up to ambient isotopy of $\mathbb{D}_2$) collection of oriented curves connecting strands in S_L with those in S_R in the order-preserving and minimally intersecting way. See the middle picture in Figure 5. It is characterized by the *pairing map* $f : S_L \to S_R$, which is an order-preserving bijection that maps each incoming (resp. outgoing) strand of S_L to an outgoing (resp. incoming) strand of S_R. The associated ladder-web $W(S)$ is obtained from $W_{\mathrm{br}}(S)$ by replacing each intersection with an H-web, as follows:

(3-1)



Conversely, the collection $W_{\mathrm{br}}(S)$ is called the *braid representation* of the ladder-web $W(S)$. It is known that all the essential webs on $\mathbb{D}_2$ arise in this way:

Proposition 3.1 [6, Proposition 19; 19, Section 8] *The ladder-web $W(S)$ is an essential web on $\mathbb{D}_2$ for any symmetric strand set S. Conversely, given an essential web W on $\mathbb{D}_2$, there exists a unique symmetric strand set S such that $W = W(S)$.*

For the study of unbounded $\mathfrak{sl}_3$-webs, we need the following infinite extension of the symmetric strand sets.

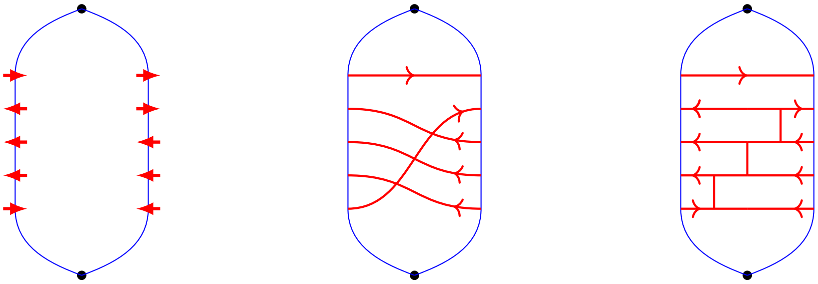


Figure 5: Construction of the ladder-webs. Left: a symmetric set S. Middle: the corresponding collection of oriented curves $W_{\mathrm{br}}(S)$. Right: the associated ladder-web $W(S)$.

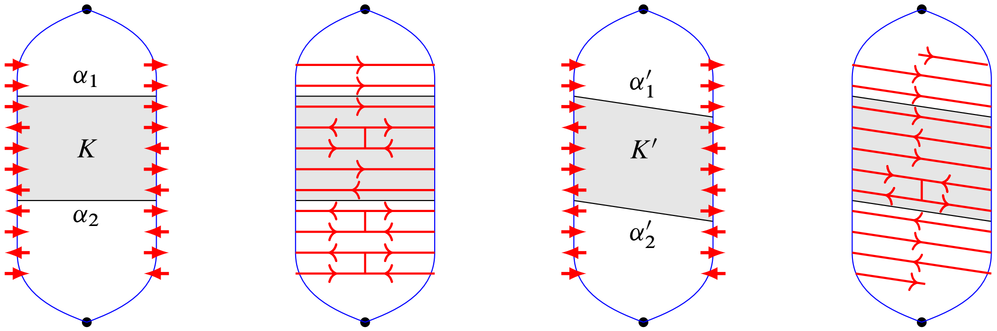


Figure 6: An asymptotically periodic symmetric strand set and the associated ladder webs corresponding to the two choices of compact strips K and K'.

Definition 3.2 (asymptotically periodic symmetric strand sets) An *asymptotically periodic symmetric strand set* $S=(S_L,S_R)$ on $\mathbb{D}_2$ consists of countable collections S_L and S_R of disjoint oriented strands, where the oriented strands in S_Z are located on E_Z without accumulation points for $Z\in\{L,R\}$. The oriented strands are required to be symmetric, and periodic away from a compact set (see Figure 6). Namely, we require that there exists a compact strip $K\subset\mathbb{D}_2\setminus\mathbb{M}$ such that

- K is bounded by two parallel arcs, α_1 and α_2, transverse to the boundary intervals of $\mathbb{D}_2$, and $\alpha_1\cup\alpha_2$ avoiding the strand sets S_L and S_R;
- the pair $(S_L\cap K, S_R\cap K)$ is a finite symmetric strand set;
- the orientation patterns of the strands in the sets S_L and S_R that belong to $\mathbb{D}_2\setminus K$ are periodic, and the pairing map $f_K\colon S_L\cap K\to S_R\cap K$ of finite symmetric strand set can be extended to an order-preserving bijection $f\colon S_L\to S_R$ that maps each incoming (resp. outgoing) strand of S_L to an outgoing (resp. incoming) strand of S_R.

Unlike the finite case, the pairing map f may not be unique, as it depends on the choice of the compact strip K. Given such a pair (S,f), we get a collection $W_{\mathrm{br}}(S,f)$ of oriented curves mutually in a minimal position, and the associated ladder-web $W(S,f)$ just in the same manner as in the finite case. We call $W(S,f)$ the ladder-web associated with the pair (S,f). It is possibly an infinite web.

Definition 3.3 An *unbounded essential web* on $\mathbb{D}_2$ is the isotopy class of the ladder-web associated with a pair (S,f) as above.

Among the others, the following way of fixing a pairing map turns out to be useful in this paper.

Definition 3.4 A *pinning* of an asymptotically periodic symmetric strand set $S=(S_L,S_R)$ is a pair $\mathsf{p}_Z=(p_Z^+,p_Z^-)$ of points in E_Z away from the set S_Z for $Z\in\{L,R\}$. The resulting tuple $\widehat{S}:=(S;\mathsf{p}_L,\mathsf{p}_R)$ is called a *pinned symmetric strand set*.

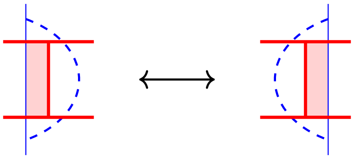


Figure 7: The H-move across an arc.

Then we define the pairing map as follows. For $Z \in \{L, R\}$, let us decompose $S_Z = S_Z^+ \sqcup S_Z^-$, where S_Z^+ (resp. S_Z^-) denotes the subset of incoming (resp. outgoing) strands. Then there exist orientation-reversing homeomorphisms $f_\pm \colon E_L \to E_R$ such that $f_\pm(S_L^\pm) = S_R^\mp$ and $f_\pm(p_L^\pm) = p_R^\mp$. Then we get the unique pairing map

$$f_{\widehat{S}} := f_+ \sqcup f_- \colon S_L^+ \sqcup S_L^- \to S_R^- \sqcup S_R^+,$$

which determines the collection $W_{\mathrm{br}}(\widehat{S}) := W_{\mathrm{br}}(S, f_{\widehat{S}})$ of oriented curves and the associated ladder-web $W(\widehat{S}) := W(S, f_{\widehat{S}})$.

The triangle (3-gon) case Let $\mathbb{D}_3$ be a triangle. Recall that we have honeycomb-webs on $\mathbb{D}_3$, which are dual to n-triangulations of $\mathbb{D}_3$.

Proposition 3.5 [6, Proposition 22; 19, Theorem 19] *A honeycomb-web is reduced* (***rung-less*** *in terms of* [6]) *and essential. Conversely, any connected reduced essential web on $\mathbb{D}_3$ having at least one trivalent vertex is a honeycomb-web.*

Consequently, any reduced essential web on $\mathbb{D}_3$ consists of a unique (possibly empty) honeycomb component together with a collection of disjoint oriented arcs located on the corners of $\mathbb{D}_3$. These oriented arcs are called *corner arcs*. Similarly to the biangle case, we may allow the collection of corner arcs to be semi-infinite and asymptotically periodic.

Definition 3.6 An *unbounded reduced essential web* on $\mathbb{D}_3$ is the isotopy class of a disjoint union of a (possibly empty) reduced essential web on $\mathbb{D}_3$ and at most one semi-infinite periodic collection of corner arcs around each corner.

3.2 Good position of an unbounded $\mathfrak{sl}_3$-lamination

Let $\triangle$ be an ideal triangulation of Σ without self-folded triangles. Recall from [6, Section 3] that a bounded $\mathfrak{sl}_3$-web W on Σ is *generic* with respect to $\triangle$ if none of its trivalent vertices intersect with the edges of $\triangle$, and W intersects with $\triangle$ transversely. A *generic isotopy* is an isotopy of webs through generic webs. Recall the *parallel-equivalence* of bounded webs, which is the equivalence relation generated by isotopies of marked surface and the loop parallel-move (E1). A generic bounded web W is said to be in *minimal position* with respect to $\triangle$ if it minimizes the sum of the intersection numbers with the edges of $\triangle$ among those parallel-equivalent to W. Then we have:

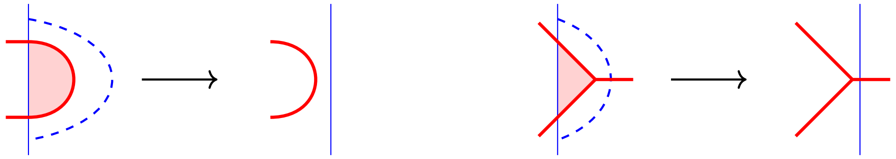


Figure 8: The intersection reduction moves across an arc.

Proposition 3.7 [6, Proposition 27; 19, Section 6] *Any parallel-equivalence class of nonelliptic bounded webs on Σ has a representative in minimal position with respect to $\triangle$. Moreover, such a representative is unique up to a sequence of H-moves across edges of $\triangle$ (Figure 7), loop parallel-moves, and generic isotopies.*

Indeed, the minimal position is realized by appropriately applying the *intersection reduction moves* (aka *tightening moves*) across edges of $\triangle$ shown in Figure 8.

The *split ideal triangulation* $\widehat{\triangle}$ is obtained from $\triangle$ by replacing each edge E into a biangle B_E. We say that a bounded web W on Σ is in *good position* with respect to $\widehat{\triangle}$ if the restrictions $W \cap B_E$ for $E \in e(\triangle)$ (resp. $W \cap T$ for $T \in t(\triangle)$) are an essential (resp. reduced essential) webs. Then it is known that any parallel-equivalence class of nonelliptic bounded webs on Σ has a representative in good position with respect to $\widehat{\triangle}$; such a representative is unique up to a sequence of modified H-moves (Figure 9), loop parallel-moves, and generic isotopies for $\widehat{\triangle}$ [6, Proposition 30; 19, Corollary 18]. Using such a representative, the Douglas–Sun coordinates are defined [6, Section 4].

Now let us consider a signed web W on Σ. In this case, W is no more parallel-equivalent to a web in good position in the above sense. To resolve this, we introduce the following notion:

Definition 3.8 (spiraling diagram) Let W be a nonelliptic signed web on Σ. Then the associated *spiraling diagram* $\mathcal{W}$ is a (possibly infinite and noncompact) $\mathfrak{sl}_3$-web obtained by the following two steps.

(1) In a small disk neighborhood D_p of each puncture $p \in \mathbb{M}_\circ$, deform each end of W incident to p into an infinitely spiraling curve, according to their signs as shown in Figure 10. Let $\mathcal{W}'$ be the resulting diagram.

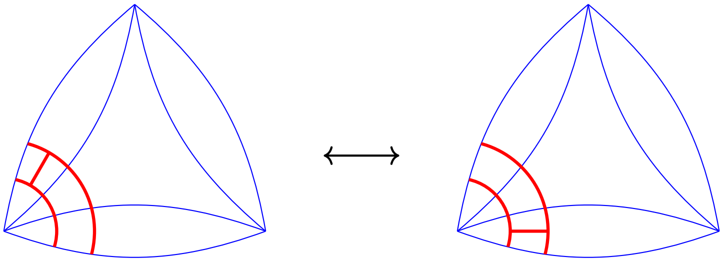


Figure 9: The modified H-move [6] (aka crossbar pass [19]) across a corner.

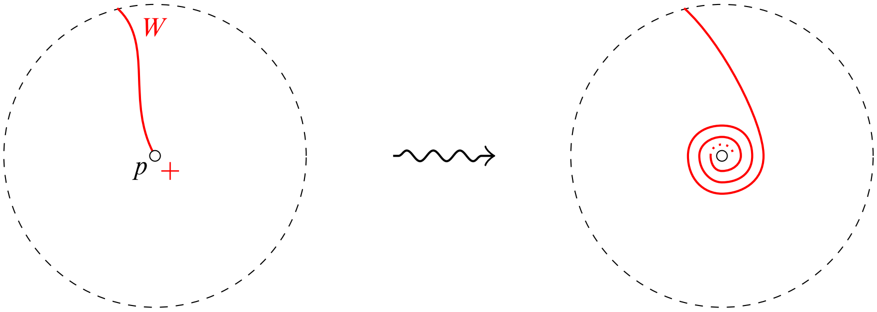


Figure 10: Construction of a spiraling diagram. The negative sign similarly produce an end spiraling counterclockwise.

(2) A pair of ends incident to a common puncture p with the opposite sign produce infinitely many intersections in $\mathcal{W}'$. We then modify these intersections into H-webs in a periodic manner, as follows. By applying an isotopy in D_p, we can make these intersections only occurring in a single half-biangle B_p in D_p with special point p, without producing additional intersections.[5] Then $\mathcal{W}' \cap B_p = W_{\mathrm{br}}(S_p)$ for an asymptotically periodic symmetric strand set S_p on B_p. By replacing the biangle part $W_{\mathrm{br}}(S_p)$ with the associated ladder-web $W(S_p)$, we get the spiraling diagram $\mathcal{W}$. Since $\mathcal{W} \cap (D_p \setminus B_p)$ consists of oriented corner arcs, the result does not depend on the choice of B_p.

See Figure 11 for a local example. A global example arising from Figure 3 is shown in Figure 12.

Definition 3.9 The spiraling diagram $\mathcal{W}$ is in a *good position* with respect to a split triangulation $\widehat{\triangle}$ if the intersection $\mathcal{W} \cap B_E$ (resp. $\mathcal{W} \cap T$) is an unbounded essential (resp. reduced essential) local web for each $E \in e(\triangle)$ and $T \in t(\triangle)$.

The loop parallel-move and the boundary H-move of a spiraling diagram are similarly defined as before, so that the construction of spiraling diagram from a signed web is equivariant under these moves. We define the *modified periodic H-move* of a spiraling diagram in a good position across a corner to be the periodic application of the modified H-move to be the periodic parts of the unbounded essential local webs on biangles. By a *strict isotopy* relative to a split triangulation $\widehat{\triangle}$, we mean an isotopy on a marked surface Σ which is the identity on each edge of $\widehat{\triangle}$ and a neighborhood of each puncture.

Theorem 3.10 (proof in Section 6.1) *Any spiraling diagram arising from a nonelliptic signed web on Σ can be isotoped into a good position with respect to $\widehat{\triangle}$ by a finite sequence of intersection reduction*

[5] Concretely, this can be done as follows. If we fix a polar coordinates (r, θ), $r < r_0$ for some $r_0 > 0$ on the punctured disk $D_p \setminus \{p\}$, each spiraling curve can be modeled by the logarithmic spiral $\ell_\pm(a)$: $\theta = \pm \log(ar)$ for some parameter $a > 0$. Then an elementary calculation shows that the intersection points of $\ell_+(a_1)$ and $\ell_-(a_2)$ lie on a single line, which is viewed as the union of two rays. Then we can collectively push these rays into a chosen half-biangle B_p only by smoothly varying the coordinate function θ. By the standard argument involving a smooth cut-off function, we can also modify this "angular" isotopy to be identity near ∂D_p.

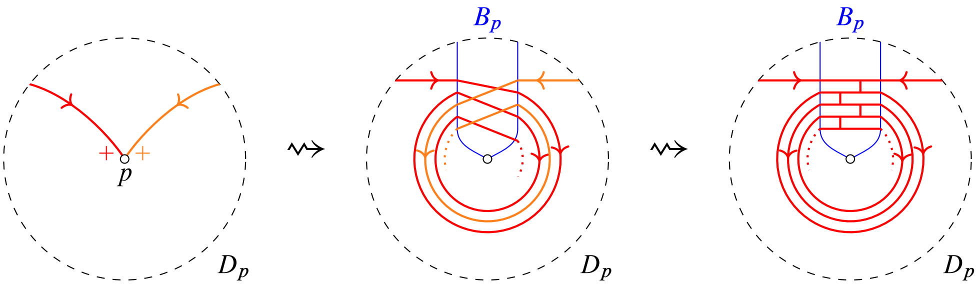


Figure 11: Construction of a spiraling diagram. Replace intersections with H-webs in a periodic manner.

moves, H-moves, and strict isotopies relative to $\widehat{\triangle}$. *Moreover, such a good position is unique up to a sequence of modified H-moves, modified periodic H-moves, loop parallel-moves, boundary H-moves, and strict isotopies relative to* $\widehat{\triangle}$.

Indeed, we can obtain a representative in a good position by successively applying the intersection reduction moves (Figure 8) and then pushing the H-faces into biangles by the H-move (Figure 7). An example of this procedure is illustrated in Figure 13. The main issue here is to ensure that this procedure always terminates in finite steps, which is discussed in Section 6.1 in detail.

While the spiraling diagram itself is suited to discuss its good position, the following *braid representation* will be useful to define the shear coordinates:

Definition 3.11 (braid representation of a spiraling diagram) Let $\mathcal{W}$ be a spiraling diagram in a good position with respect to $\widehat{\triangle}$. Then its *braid representation* $\mathcal{W}_{\mathrm{br}}^{\triangle}$ is obtained from $\mathcal{W}$ by replacing the unbounded essential web $\mathcal{W} \cap B_E$ on each biangle B_E with its braid representation.

The braid representation is closely related to (an unbounded version of) *global picture* [6, Definition 55]. See also Section 6.2.

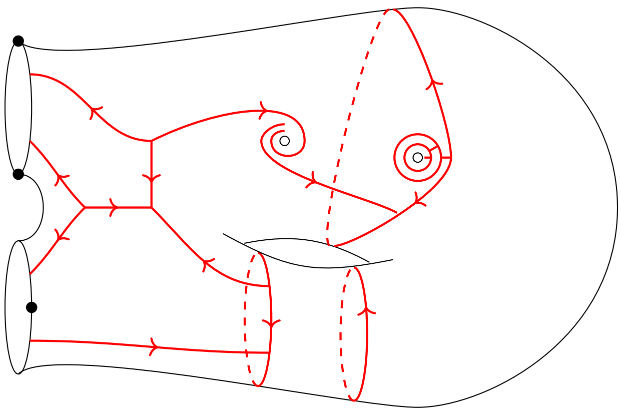


Figure 12: A global example of spiraling diagram arising from the underlying signed nonelliptic web in Figure 3.

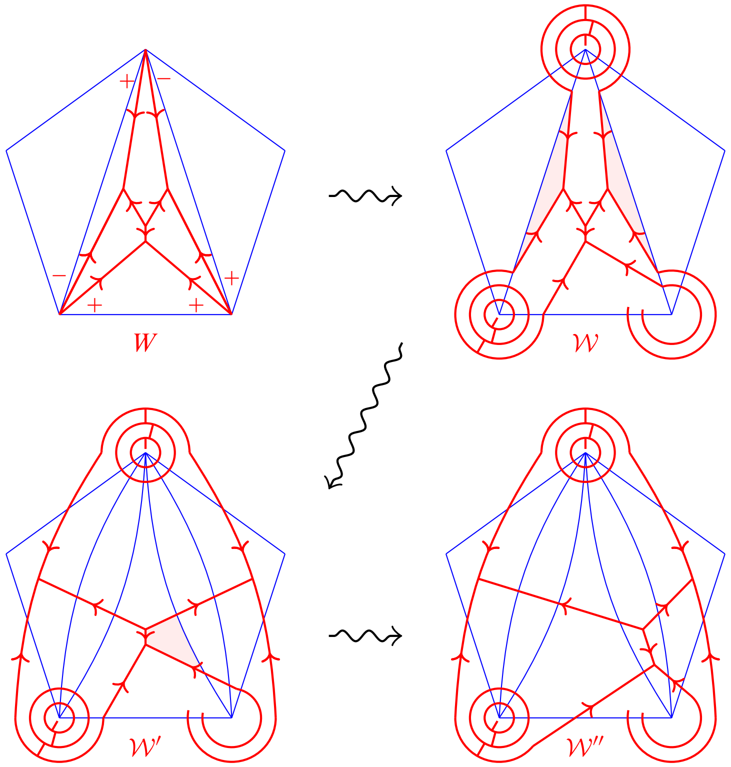


Figure 13: An example of the procedure to place a spiraling diagram in good position.

3.3 Definition of the shear coordinates

Now we define the shear coordinates associated with an ideal triangulation $\triangle$ of Σ without self-folded triangles. Let $\widehat{\triangle}$ be the associated split triangulation.

Given a rational $\mathfrak{sl}_3$-lamination $\hat{L} \in \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$, represent it by an $\mathfrak{sl}_3$-web W together with rational weights on its components and signs at the ends incident to punctures. Let $\mathcal{W}$ be the associated spiraling diagram together with rational weights on the components, placed in good position with respect to $\widehat{\triangle}$. Let $\mathcal{W}^{\triangle}_{\mathrm{br}}$ be its braid representation, together with well-assigned rational weights on its components. The shear coordinates of $\hat{L}$ are going to be defined out of $\mathcal{W}^{\triangle}_{\mathrm{br}}$.

For each $E \in e_{\mathrm{int}}(\triangle)$, let Q_E be the unique quadrilateral containing E as its diagonal, regarded as the union of two triangles, T_L and T_R, and the biangle B_E. By Proposition 3.5, the restriction of $\mathcal{W}^{\triangle}_{\mathrm{br}}$ to each of T_L and T_R has at most one honeycomb web, which is represented by a triangular symbol as in Notation 2.7. We call any strand in the braid representative $\mathcal{W}^{\triangle}_{\mathrm{br}} \cap Q_E$ that is incident to the triangular symbol in T_L (if exists) a T_L*-strand*. Similarly, we define T_R*-strands*. It is possible that an arc is both T_L- and T_R-strand, in which case it connects the two honeycombs. By removing the T_L- and T_R-strands, remaining is a collection of (possibly intersecting) oriented curves, which we call the *curve components*. See Figure 17 below.

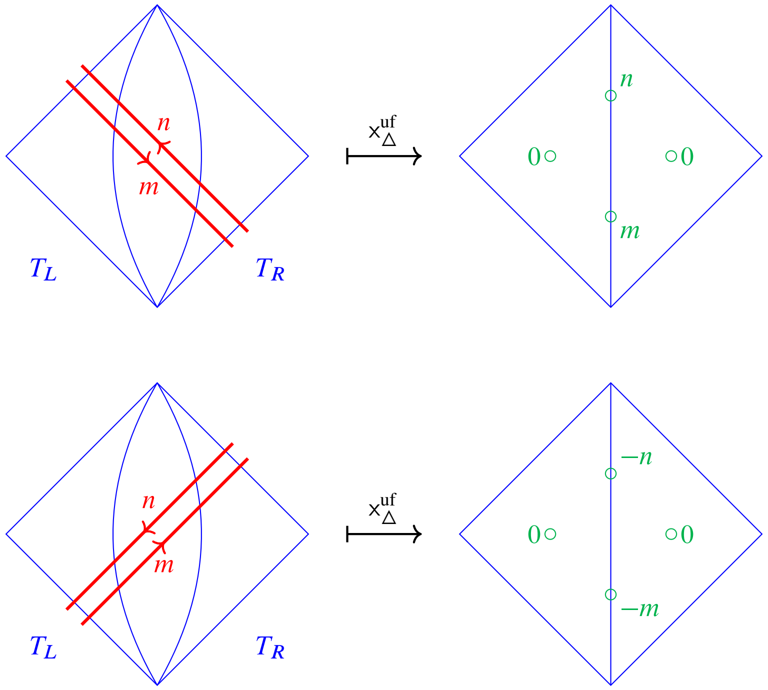

n
m
T_L
T_R
$\mathsf{x}^{\mathrm{uf}}_{\triangle}$
n
0
0
m
n
m
T_L
T_R
$\mathsf{x}^{\mathrm{uf}}_{\triangle}$
$-n$
0
0
$-m$

Figure 14: Contributions from curve components.

Definition 3.12 ($\mathfrak{sl}_3$-shear coordinates) The ($\mathfrak{sl}_3$-)*shear coordinate system*

$$\mathsf{x}^{\mathrm{uf}}_{\triangle}(\hat{L}) = (\mathsf{x}^{\triangle}_i(\hat{L}))_{i \in I_{\mathrm{uf}}(\triangle)} \in \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)}$$

is defined as follows. First, for each $E \in e_{\mathrm{int}}(\triangle)$, the coordinates assigned to the four vertices in the interior of Q_E only depends on the restriction $\mathcal{W}^{\triangle}_{\mathrm{br}} \cap Q_E$.

(1) Each curve component contributes to the edge coordinates according to the rule shown in Figure 14.

(2) The honeycomb on the triangle T_L contributes to $\mathsf{x}^{\mathrm{uf}}_{\triangle}(\hat{L})$ as in Figure 15. Namely, the face coordinate counts the height of the honeycomb web, where a sink (resp. source) is counted positively (resp. negatively). The edge coordinates counts the contributions from T_L-strands, where we have n_1 left-turning ones, n_2 straight-going ones (which are also T_R-strands), and n_3 right-turning ones.

(3) The honeycomb on the triangle T_R and the T_R-strands contribute in the symmetric way with respect to the π rotation of the figure.

Then the shear coordinates are defined to be the weighted sums of these contributions.

Remark 3.13 (1) Notice that the rule shown in Figure 14 is an "oriented version" of the Thurston's shear coordinates (see Section 3.5). Indeed, the sign of contribution is determined by the crossing pattern as in the $\mathfrak{sl}_2$-case, and it contributes to the coordinates *on the right side* of the oriented curve.

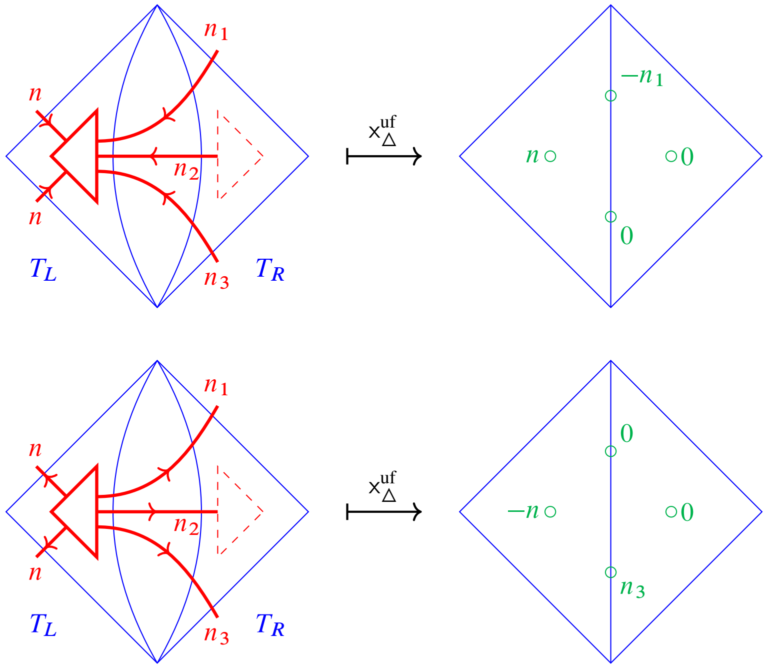


Figure 15: Contributions from the honeycomb of height $n = n_1 + n_2 + n_3$ on the triangle T_L. Observe that the n_2 straight-going T_L-strands do not contribute.

(2) The shear coordinates of the first honeycomb component shown in Figure 15 is the same as the sum of shear coordinates of the three honeycomb components shown in Figure 16.

Proposition 3.14 *The shear coordinate system $\mathsf{x}^{\mathrm{uf}}_{\triangle}(\hat{L}) \in \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)}$ is well defined, and we get a map*

$$\mathsf{x}^{\mathrm{uf}}_{\triangle}\colon \mathcal{L}^{x}_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \to \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)}.$$

Proof It is not hard to see that the operations appearing in Theorem 3.10 that move a spiraling diagram in a good position to another good position do not change the shear coordinates. For example, the modified H-move always involves a pair of oriented curves in the opposite directions in the braid representation,

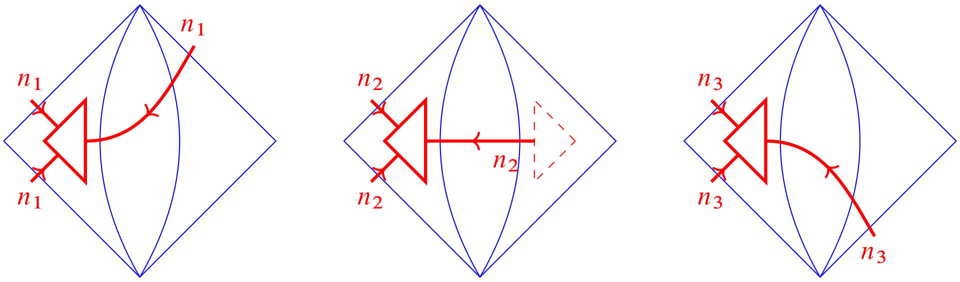


Figure 16: Basic honeycomb components.

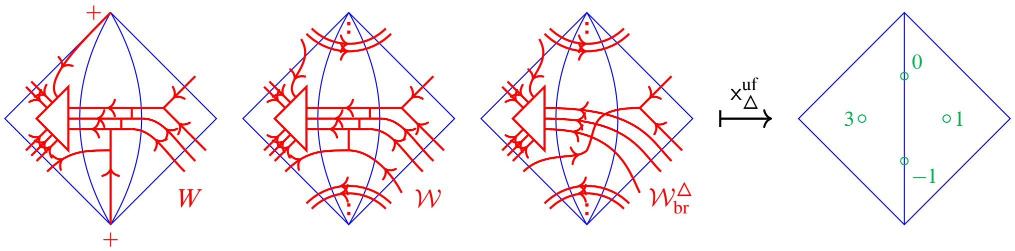
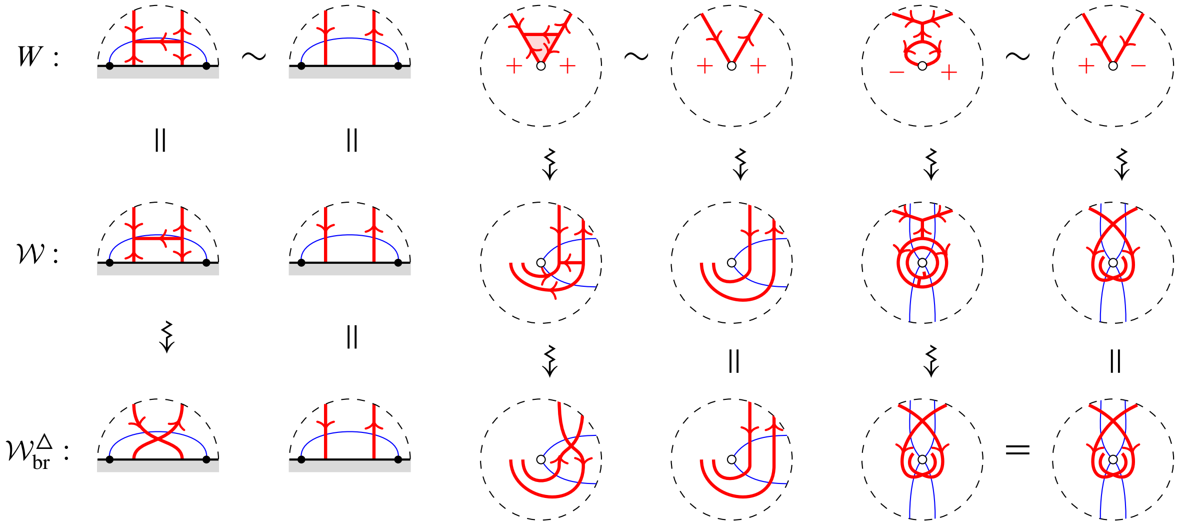

+
+
W
$\mathcal{W}$
$\mathcal{W}^{\triangle}_{\mathrm{br}}$
$\mathsf{x}^{\mathrm{uf}}_{\triangle}$
0
3
1
−1

Figure 17: An example of a signed web W restricted to Q_E, the associated spiraling diagram $\mathcal{W}$, its braid representation $\mathcal{W}^{\triangle}_{\mathrm{br}}$, its shear coordinates are shown order. In $\mathcal{W}^{\triangle}_{\mathrm{br}}$, there are two honeycomb components and infinitely many curve components.

and hence preserves the contribution from the pair. It follows that the shear coordinates are well defined for a given spiraling diagram, not depending on the choice of a good position with respect to $\widehat{\triangle}$.

We need to check that the elementary moves (E1)–(E4) of signed webs do not change the shear coordinates. It is easy to see the invariance for the loop parallel-move (E1). The braid representatives of spiraling diagrams associated with the local signed webs in (2-5)–(2-7) are obtained as follows:



Here the braid representatives are not quite the same in the first two cases, but both have the same shear coordinates. Thus the shear coordinates are invariant under the moves (E2) and (E3). The invariance under the peripheral move (E4) is similarly verified, where the signed web in the left-hand side produces a peripheral component in its spiraling diagram.

The shear coordinates are clearly invariant under operations (2) and (3) in Definition 2.6, and hence do not depend on the choice of a signed $\mathbb{Q}_{>0}$-weighted web representing an unbounded $\mathfrak{sl}_3$-lamination. □

Notation 3.15 We will write $\mathsf{x}^{\triangle}_{T} := \mathsf{x}^{\triangle}_{i(T)}$ for a triangle T of $\triangle$, and $\mathsf{x}^{\triangle}_{E,s} := \mathsf{x}^{\triangle}_{i^s(E)}$ for an oriented edge E of $\triangle$ and $s = 1, 2$. Here recall the notations in Section 2.1.

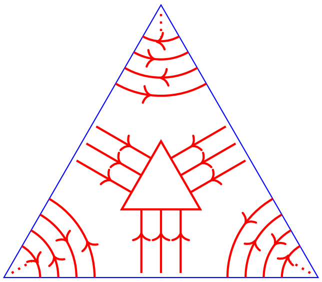


Figure 18: The building block for reconstruction from the shear coordinates when $\mathsf{x}_T = +3$.

3.4 Reconstruction

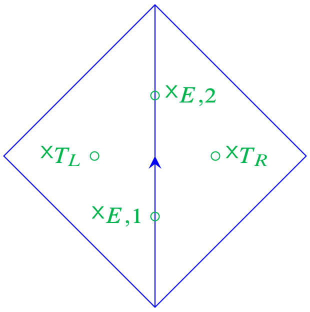
We are going to give an inverse map $\xi_\triangle\colon \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)} \to \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ of the shear coordinate system associated with an ideal triangulation $\triangle$.

Given $(\tilde{\mathsf{x}}_i)_i \in \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)}$, choose a positive integer $u \in \mathbb{Z}_{>0}$ such that $\mathsf{x}_i := u\tilde{\mathsf{x}}_i$ are integral for all $i \in I_{\mathrm{uf}}(\triangle)$. We will use a notation similar to Notation 3.15 for these tuples. On each triangle $T \in t(\triangle)$, first draw a honeycomb web of height $|\mathsf{x}_T|$ of sink type (resp. source type) if $\mathsf{x}_T \geq 0$ (resp. $\mathsf{x}_T < 0$). Moreover, on each corner of T, draw an semi-infinite collection of disjoint corner arcs with alternating orientations such that

- they are disjoint from the honeycomb web (placed on the center of T),
- they accumulate only at the marked points of the triangle, and
- the farthest one from the marked point is oriented clockwise.

See Figure 18. Then we get an unbounded reduced essential web W_T on each triangle T. We are going to glue these local blocks together to form an integral unbounded $\mathfrak{sl}_3$-lamination $\xi_\triangle((\mathsf{x}_i)_i) \in \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Z})$.

Now let us concentrate on a quadrilateral Q_E in the ideal triangulation $\triangle$ which contains two triangles T_L and T_R that share an interior edge E. We fix an orientation of E such that T_L lies on the left; hence we have two edge coordinates $\mathsf{x}_{E,1}$ and $\mathsf{x}_{E,2}$ as well as two face coordinates x_{T_L} and x_{T_R}:



Consider a biangle B_E in the split ideal triangulation $\widehat{\triangle}$ obtained by fattening E, which is bounded by boundary intervals E_L and E_R of T_L and T_R, respectively. For $Z \in \{L, R\}$, let $S_Z = S_Z^+ \sqcup S_Z^-$

denote the set of ends of the web W_{T_Z} on E_Z, where S_Z^+ (resp. S_Z^-) consists of the ends incoming to (resp. outgoing from) the biangle B_E. Then $S = (S_L, S_R)$ defines an asymptotically periodic symmetric strand set (Definition 3.2). Let us define its pinning by the following rule:

- For $Z \in \{L, R\}$, choose orientation-preserving parametrizations
$$\phi_Z^{\pm} \colon \mathbb{R} \to E_Z$$
so that $\phi_Z^{\pm}\big(\frac{1}{2} + \mathbb{Z}\big) = S_Z^{\pm}$, and $\phi_Z^{\pm}(\mathbb{R}_{<0}) \cap S_Z^{\pm}$ consists of all the strands coming from the corner arcs around the initial marked point of E_Z.
- Let $p_Z^{\pm} := \phi_Z^{\pm}(n_Z^{\pm}) \in E_Z$ for $Z \in \{L, R\}$, where $n_Z^{\pm} \in \mathbb{Z}$ are given by
$$n_L^+ := \mathsf{x}_{E,1}, \quad n_L^- := [\mathsf{x}_{T_L}]_+, \quad n_R^+ := \mathsf{x}_{E,2}, \quad n_R^- := [\mathsf{x}_{T_R}]_+, \tag{3-2}$$
where we use the notation $[x]_+ := \max\{0, x\}$.

Then we get a pinned symmetric strand set $\widehat{S}_E := (S; \mathsf{p}_L, \mathsf{p}_R)$ with the pinnings $\mathsf{p}_Z := (p_Z^+, p_Z^-)$ for $Z \in \{L, R\}$. Let $W_{\mathrm{br}}(\widehat{S}_E)$ denote the associated collection of oriented curves in B_E.

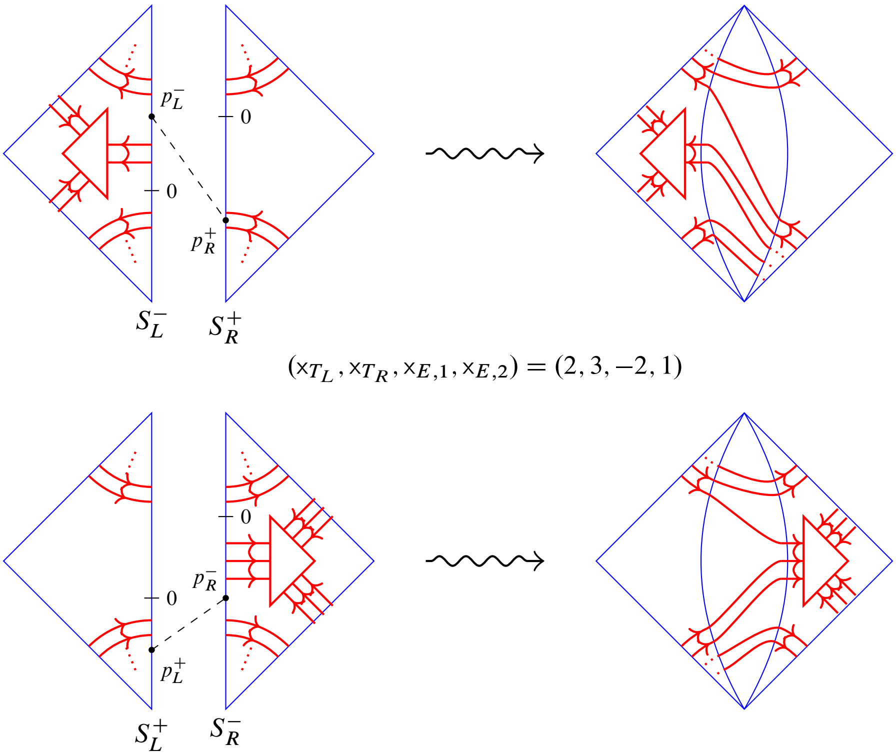
Remark 3.16 The resulting collection $W_{\mathrm{br}}(\widehat{S}_E)$ is invariant under the transformation
$$n_L^+ \mapsto n_L^+ - k, \quad n_L^- \mapsto n_L^- - l, \quad n_R^+ \mapsto n_R^+ + l, \quad n_R^- \mapsto n_R^- + k,$$
for $k, l \in \mathbb{Z}$.

Gluing together the local webs W_T for $T \in t(\triangle)$ and the curves in $W_{\mathrm{br}}(\widehat{S}_E)$ for $E \in e(\triangle)$, we get a (possibly infinite) collection $\mathcal{W}_{\mathrm{br}}^{\triangle}((\mathsf{x}_i)_i)$ of webs on Σ. The following lemma shows that it has correct shear coordinates.

Lemma 3.17 *We have $\mathsf{x}_k^{\triangle}(\mathcal{W}_{\mathrm{br}}^{\triangle}((\mathsf{x}_i)_i)) = \mathsf{x}_k$ for all $k \in I_{\mathrm{uf}}(\triangle)$.*

Proof Let us concentrate on a quadrilateral $Q = T_L \cup B_E \cup T_R$. It is easy to see $\mathsf{x}_{T_Z}^{\triangle}(\widehat{L}) = \mathsf{x}_{T_Z}$ for $Z \in \{L, R\}$. The equalities $\mathsf{x}_{E,1}^{\triangle}(\widehat{L}) = \mathsf{x}_{E,1}$ and $\mathsf{x}_{E,2}^{\triangle}(\widehat{L}) = \mathsf{x}_{E,2}$ can be also verified case-by-case, divided according to the signs of x_{T_L} and x_{T_R}. See Figures 19–21. Here we draw the pictures by separating the gluing procedures $S_L^- \to S_R^+$ and $S_L^+ \to S_R^-$ into two sheets; the result is obtained by overlaying the two diagrams drawn on the right.

For example, let us consider the example shown in Figure 19. In the case $\mathsf{x}_{E,2} \geq 0$ (as in this example), there are $\mathsf{x}_{E,2}$ many lines from south-east to north-west that contribute positively. One can imagine the other cases by varying this example: if we decrease $\mathsf{x}_{E,2}$, then the point p_R^+ moves upward and the gluing pattern is shifted. When $-\mathsf{x}_{T_L} \leq \mathsf{x}_{E,2} < 0$, negative contributions come from the honeycomb in T_L. When $\mathsf{x}_{E,2} < -\mathsf{x}_{T_L}$, there are also lines from south-west to north-east that contribute negatively. Thus we get $\mathsf{x}_{E,2}^{\triangle}(\widehat{L}) = \mathsf{x}_{E,2}$. The check for $\mathsf{x}_{E,1}$ is similar. One can check the other cases from Figures 20 and 21 in a similar manner. □



$(\mathsf{x}_{T_L}, \mathsf{x}_{T_R}, \mathsf{x}_{E,1}, \mathsf{x}_{E,2}) = (2, 3, -2, 1)$

Figure 19: An example for the case $\mathsf{x}_{T_L} \geq 0$ and $\mathsf{x}_{T_R} \geq 0$.

The collection $\mathcal{W}_{\mathrm{br}}^{\triangle}((\mathsf{x}_i)_i)$ is the braid representative of the spiraling diagram associated to an unbounded integral $\mathfrak{sl}_3$-lamination $\xi_\triangle((\mathsf{x}_i)_i)$, which is obtained as follows:

Step 1 First remove the peripheral components around the marked points (both special points and punctures) from $\mathcal{W}_{\mathrm{br}}^{\triangle}((\mathsf{x}_i)_i)$. Then, remaining are finitely many components.

Step 2 Replace each spiraling end around a puncture p with an end incident to p, while encoding the spiraling directions in signs by reversing the rule in Figure 10. Then we get a collection $W_{\mathrm{br}}^{\triangle}((\mathsf{x}_i)_i)$ of signed webs, which we call a *braid representative* of a signed web. It contains at most finitely many intersections of curves only in biangles. Here we can rearrange $W_{\mathrm{br}}^{\triangle}((\mathsf{x}_i)_i)$ so that no pair of curves form a bigon by applying a Reidemeister II-type isotopy if necessary (cf *square removing algorithm* in [6]). See Figure 22. Observe that this operation does not affect the shear coordinates.

Step 3 Replace each intersection of curves in a biangle with an H-web by the rule (3-1). Then we get a signed $\mathfrak{sl}_3$-web W on Σ, which has no elliptic faces. Indeed, we have no 0-gon or 2-gon faces by construction, and possible emergence of 4-gon faces has been eliminated in Step 2.

Then $\xi_\triangle((\mathsf{x}_i)_i) \in \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Z})$ is defined to be the unbounded integral $\mathfrak{sl}_3$-lamination represented by the nonelliptic signed web W (with weight 1 on each component). Set

$$\xi_\triangle((\tilde{\mathsf{x}}_i)_i) := u^{-1} \cdot \xi_\triangle((\mathsf{x}_i)_i) \in \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}).$$

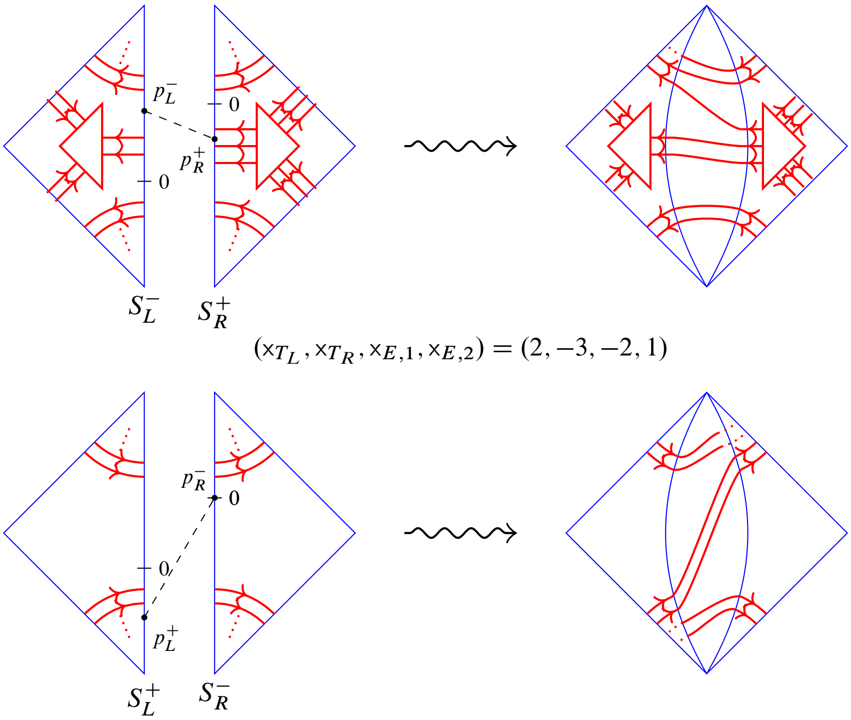

p_L^-
0
p_R^+
0
S_L^-
S_R^+
$(\mathsf{x}_{T_L}, \mathsf{x}_{T_R}, \mathsf{x}_{E,1}, \mathsf{x}_{E,2}) = (2, -3, -2, 1)$
p_R^-
0
0
p_L^+
S_L^+
S_R^-

Figure 20: An example for the case $\mathsf{x}_{T_L} \geq 0$ and $\mathsf{x}_{T_R} \leq 0$. The case $\mathsf{x}_{T_L} \leq 0$ and $\mathsf{x}_{T_R} \geq 0$ follows by symmetry (Remark 3.16).

Thus we get the map $\xi_\triangle \colon \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)} \to \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$, which is clearly $\mathbb{Q}_{>0}$-equivariant. We are going to show that this map indeed gives the inverse map of $\mathsf{x}^{\mathrm{uf}}_\triangle$. The following direction is easier:

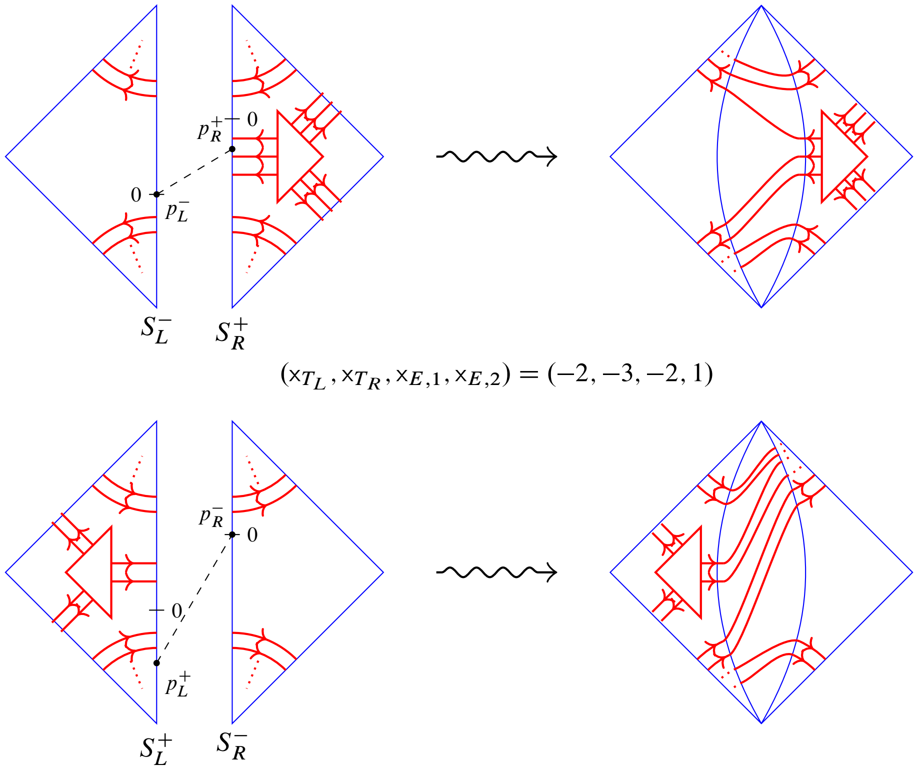
Proposition 3.18 *We have* $\mathsf{x}^{\mathrm{uf}}_\triangle \circ \xi_\triangle = \mathrm{id}_{\mathbb{Q}^{I_{\mathrm{uf}}(\triangle)}}$.

Proof By $\mathbb{Q}_{>0}$-equivariance, it suffices to consider an integral tuple $(\mathsf{x}_i)_i \in \mathbb{Z}^{I_{\mathrm{uf}}(\triangle)}$. Notice that by construction, the collection $\mathcal{W}^\triangle_{\mathrm{br}}((\mathsf{x}_i)_i)$ arising from the gluing construction above is exactly the braid representative of the spiraling diagram associated with the underlying signed web of the $\mathfrak{sl}_3$-lamination $\hat{L} := \xi_\triangle((\mathsf{x}_i)_i) \in \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Z})$. Therefore the shear coordinates $(\mathsf{x}^\triangle_i(\hat{L}))$ can be directly read off from the collection $\mathcal{W}^\triangle_{\mathrm{br}}((\mathsf{x}_i)_i)$. Hence the assertion follows from Lemma 3.17. □

Theorem 3.19 (proof in Section 6.2) *We have* $\xi_\triangle \circ \mathsf{x}^{\mathrm{uf}}_\triangle = \mathrm{id}_{\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})}$. *In particular, the shear coordinates gives a bijection* $\xi_\triangle \colon \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)} \xrightarrow{\sim} \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$.

See Section 6.2 for a proof. The main ingredient of the proof is an unbounded version of the fellow-traveler lemma [6, Lemma 57] with respect to the shear coordinates.

Recall from Section A.3 that the ideal triangulations $\triangle$ correspond to certain seeds in the mutation class $\mathsf{s}(\Sigma, \mathfrak{sl}_3)$. The following theorem states that the associated shear coordinate systems $\mathsf{x}^{\mathrm{uf}}_\triangle$ are related by tropical cluster Poisson transformations:



$$(\mathsf{x}_{T_L}, \mathsf{x}_{T_R}, \mathsf{x}_{E,1}, \mathsf{x}_{E,2}) = (-2, -3, -2, 1)$$

Figure 21: An example for the case $\mathsf{x}_{T_L} \leq 0$ and $\mathsf{x}_{T_R} \leq 0$.

Theorem 3.20 *For any two ideal triangulations $\triangle$ and $\triangle'$ of Σ, the coordinate transformation*

$$\mathsf{x}^{\mathrm{uf}}_{\triangle'} \circ (\mathsf{x}^{\mathrm{uf}}_{\triangle})^{-1} \colon \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)} \to \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)}$$

is a composite of tropical cluster Poisson transformations. In particular, we get an $\mathrm{MC}(\Sigma)$*-equivariant identification* $\mathsf{x}^{\mathrm{uf}}_{\bullet} \colon \mathcal{L}^{x}_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \xrightarrow{\sim} \mathcal{X}^{\mathrm{uf}}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$.

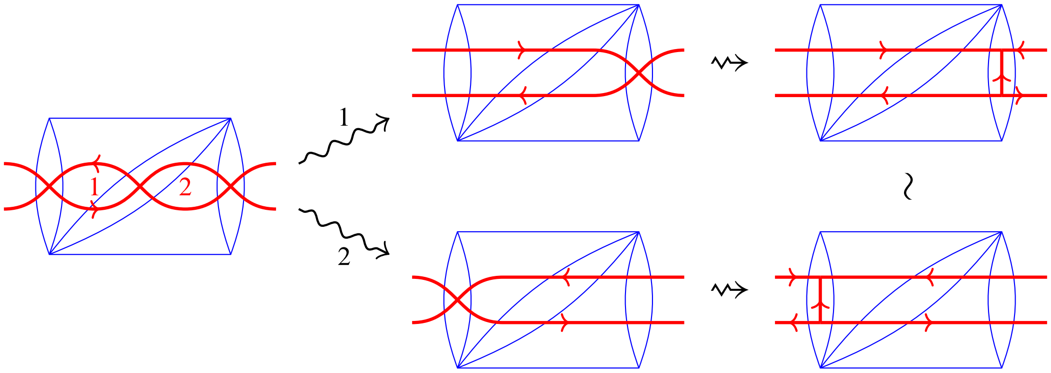


Figure 22: Reidemeister II-type isotopy. We have two ways of applications of this isotopy, which produce equivalent webs.

Since it is classically known that any two ideal triangulations of the same marked surface can be connected by a finite sequence of flips, it suffices to show that a flip corresponds to a composite of tropical cluster Poisson transformations. Although it can be directly checked in a similar way to [7, Section 4], we are going to reduce it to Douglas and Sun's result via the ensemble map and the gluing technique developed in Section 4.

3.5 Relation to the rational unbounded $\mathfrak{sl}_2$-laminations

Recall the space $\mathcal{L}^x_{\mathfrak{sl}_2}(\Sigma, \mathbb{Q})$ of rational unbounded ($\mathfrak{sl}_2$-)laminations from [11]. It consists of the following data:

- A collection of immersed unoriented loops and arcs such that each endpoint lies in $\mathbb{M}_\circ \cup \partial^* \Sigma$, and the other part is embedded in $\operatorname{int} \Sigma$. It is required to have no elliptic faces (the first one in (2-1) or the first and last ones in (2-2)).
- A positive rational weight on each component.
- A sign $\sigma_p \in \{+, 0, -\}$ for each puncture $p \in \mathbb{M}_\circ$ such that $\sigma_p = 0$ if and only if there are no components incident to p.

They are considered modulo removal/creation of peripheral components as in (2-8), and the weighted isotopy as in Definition 2.6(2). Given an ideal triangulation $\triangle$ of Σ, the ($\mathfrak{sl}_2$-)shear coordinate

$$\mathsf{x}_\triangle = (\mathsf{x}^\triangle_E)_{E \in e(\triangle)} \colon \mathcal{L}^x_{\mathfrak{sl}_2}(\Sigma, \mathbb{Q}) \xrightarrow{\sim} \mathbb{Q}^{e(\triangle)}$$

(see [11]) is defined by first constructing a spiraling diagram according to the sign σ_p, and counting contributions with weights from the curves in that diagram, as in Figure 23.

An embedding $\iota_{\mathrm{prin}} \colon \mathcal{L}^x_{\mathfrak{sl}_2}(\Sigma, \mathbb{Q}) \to \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ is defined so that

- each curve γ with weight $u \in \mathbb{Q}_{>0}$ is sent to its parallel copies, γ_1 and γ_2, with the same weight u with the opposite orientations;
- if an arc γ is incident to a puncture p, then the corresponding ends of the oriented curves γ_1 and γ_2 are assigned the sign $\sigma_p \in \{+, -\}$.

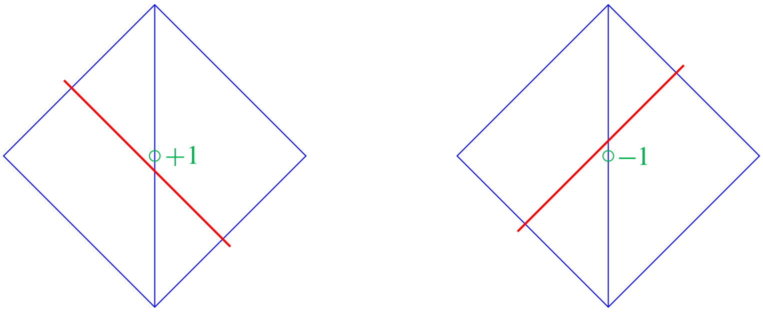


Figure 23: Contributions to the $\mathfrak{sl}_2$-shear coordinates.

One can easily verify that it is indeed well defined. We call ι_{prin} the *principal embedding*, as it is a tropical analogue of the morphism $\mathcal{X}_{SL_2,\Sigma} \to \mathcal{X}_{SL_3,\Sigma}$ induced by the principal embedding $\mathfrak{sl}_2 \to \mathfrak{sl}_3$. The following is a tropical analogue of the statement given in [12, Section 2.5.3]:

Proposition 3.21 *The image $\iota_{\mathrm{prin}}(\mathcal{L}^x_{\mathfrak{sl}_2}(\Sigma, \mathbb{Q}))$ coincides with the fixed point locus of the Dynkin involution $*$ (Definition 2.8). In the shear coordinate system $\mathsf{x}_\triangle$ associated with any ideal triangulation $\triangle$, it is characterized by the equations*

$$\begin{aligned} \mathsf{x}^\triangle_{E,1} &= \mathsf{x}^\triangle_{E,2} \quad \text{for each } E \in e(\triangle), \\ \mathsf{x}^\triangle_T &= 0 \qquad\ \ \text{for each } T \in t(\triangle). \end{aligned}$$

Proof The first assertion follows from the second one, by Proposition 4.13 below. The second assertion is easily verified by comparing the definitions of $\mathfrak{sl}_2$- and $\mathfrak{sl}_3$-shear coordinates. Indeed, we have $\mathsf{x}^\triangle_E(\hat{L}) = \mathsf{x}^\triangle_{E,1}(\iota_{\mathrm{prin}}(\hat{L})) = \mathsf{x}^\triangle_{E,2}(\iota_{\mathrm{prin}}(\hat{L}))$ and $\mathsf{x}^\triangle_T(\iota_{\mathrm{prin}}(\hat{L})) = 0$, where $(\mathsf{x}^\triangle_E)_{E\in e(\triangle)}$ denotes the $\mathfrak{sl}_2$-shear coordinate system. □

4 Rational $\mathcal{P}$-laminations, their gluing and the mutation equivariance

In this section, we introduce the space of *rational $\mathcal{P}$-laminations* by considering some additional data on boundary intervals and define a coordinate system $\mathsf{x}_\triangle$ extending $\mathsf{x}^{\mathrm{uf}}_\triangle$. These additional data allow us to introduce the *gluing map* between these spaces. Under this extended situation, we discuss the relation to Douglas and Sun's tropical $\mathcal{A}$-coordinates [6], and prove that the coordinates $\mathsf{x}_\triangle$ transform correctly under flips.

4.1 Rational unbounded $\mathfrak{sl}_3$-laminations with pinnings

It has been stated that the space $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ of rational unbounded $\mathfrak{sl}_3$-laminations is identified with the unfrozen part $\mathcal{X}^{\mathrm{uf}}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$ of the tropical cluster $\mathcal{X}$-variety. In order to obtain the entire tropical cluster $\mathcal{X}$-variety, we further equip the rational laminations with additional data on boundary intervals. Let $\mathsf{P}^\vee = \mathbb{Z}\varpi_1^\vee \oplus \mathbb{Z}\varpi_2^\vee$ be the coweight lattice of $\mathfrak{sl}_3$, and $\mathsf{P}^\vee_{\mathbb{Q}} := \mathsf{P}^\vee \otimes \mathbb{Q}$. Let us consider the direct sum

$$H_\partial(\mathbb{Q}^T) := \bigoplus_{E \in \mathbb{B}} \mathsf{P}^\vee_{\mathbb{Q}}$$

of the coweight lattices over $\mathbb{Q}$, one for each boundary interval.

Definition 4.1 (rational unbounded $\mathfrak{sl}_3$-laminations with pinnings) We introduce the space

$$\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) := \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \times H_\partial(\mathbb{Q}^T),$$

and call its elements *rational unbounded $\mathfrak{sl}_3$-laminations with pinnings* (or *rational* ($\mathfrak{sl}_3$-)*$\mathcal{P}$-laminations*). The datum in the second factor is written as $\nu = (\nu_E)_{E\in\mathbb{B}}$ with $\nu_E = \nu^+_E\varpi_1^\vee + \nu^-_E\varpi_2^\vee$, $\nu^\pm_E \in \mathbb{Q}$.

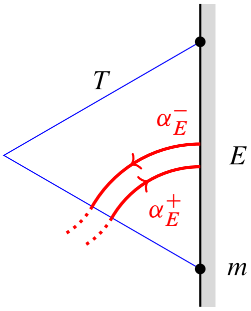


Figure 24: The corner arcs relevant to the boundary shear coordinate.

The data $\nu = (\nu_E)_E$ will be related to the pinning in the sense of Definition 3.4 when we consider their gluings, thus the terminology. We have a natural $\mathbb{Q}_{>0}$-action on $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ given by

$$u.(\hat{L}, \nu) := (u.\hat{L}, (u\nu_E)_E)$$

for $u \in \mathbb{Q}_{>0}$ and $(\hat{L}, \nu = (\nu_E)_E) \in \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$. The Dynkin involution (Definition 2.8) is extended as

$$* \colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \to \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}), \quad (\hat{L}, (\nu_E)_{E \in \mathbb{B}}) \mapsto (\hat{L}^*, (\nu_E^*)_{E \in \mathbb{B}}), \tag{4-1}$$

where $\nu^* = (\nu_E^*)_{E \in \mathbb{B}}$ is obtained from ν by the Dynkin involution on the coweight lattice: $\varpi_s^* := \varpi_{3-s}$ for $s = 1, 2$. There is a projection

$$\pi_{\mathrm{uf}} \colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \to \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$$

forgetting the second factor, which is equivariant under these structures. A rational $\mathcal{P}$-lamination $(\hat{L}, \nu)$ is said to be *integral* if $\hat{L} \in \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Z})$ and $p_E \in \mathsf{P}^\vee$ for all $E \in \mathbb{B}$.

Remark 4.2 The space $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ is introduced as a tropical analogue of the moduli space $\mathcal{P}_{PGL_3,\Sigma}$ of framed PGL_3-local systems with pinnings on Σ [21]. We have a dominant morphism $\mathcal{P}_{PGL_3,\Sigma} \to \mathcal{X}_{PGL_3,\Sigma}$, which is a principal $H_\partial := H^{\mathbb{B}}$-bundle over its image. Here $H \subset PGL_3$ denote the Cartan subgroup. As a tropical analogue, we may naturally consider the bundle

$$0 \to H_\partial(\mathbb{Q}^T) \to \mathcal{P}_{PGL_3,\Sigma}(\mathbb{Q}^T) \to \mathcal{X}_{PGL_3,\Sigma}(\mathbb{Q}^T) \to 0. \tag{4-2}$$

The space $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ is regarded as the total space $\mathcal{P}_{PGL_3,\Sigma}(\mathbb{Q}^T)$ with a fixed trivialization. See also Remark 4.8 below.

Shear coordinates on $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ Given an ideal triangulation $\triangle$ of Σ, we are going to define a shear coordinate system

$$\mathsf{x}_\triangle = (\mathsf{x}_i^\triangle)_{i \in I(\triangle)} \colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \to \mathbb{Q}^{I(\triangle)}$$

which extends $\mathsf{x}_\triangle^{\mathrm{uf}}$ on $\mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$. For $(\hat{L}, \nu) \in \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ and an unfrozen index $i \in I_{\mathrm{uf}}(\triangle)$, let $\mathsf{x}_i^\triangle(\hat{L}, \nu) := \mathsf{x}_i^\triangle(\hat{L})$ be the shear coordinate of the underlying rational unbounded lamination.

We define the frozen coordinate $\mathsf{x}_{E,s}^\triangle(\hat{L}, \nu)$ for $s = 1, 2$ associated to a boundary interval $E \in \mathbb{B}$, as follows. Let W be a nonelliptic signed $\mathbb{Q}_{>0}$-weighted web without peripheral components representing $\hat{L}$, and $\mathcal{W}$ its spiraling diagram in a good position with respect to the split triangulation $\widehat{\triangle}$. By convention, E is endowed with the orientation induced from $\partial\Sigma$. Then $\mathsf{x}^{\triangle}_{E,1}$ (resp. $\mathsf{x}^{\triangle}_{E,2}$) is assigned to the vertex of the $\mathfrak{sl}_3$-triangulation on E closer to the initial (resp. terminal) endpoint. Let $m \in \mathbb{M}_\partial$ be the initial endpoint of E, and $T \in t(\triangle)$ the unique triangle having E as an edge. Let $\alpha^+_E(\widehat{L})$ (resp. $\alpha^-_E(\widehat{L})$) be the total weight of the oriented corner arcs in $\mathcal{W} \cap T$ bounding the special point m in the clockwise (resp. counterclockwise) direction, hence incoming to (resp. outgoing from) the external biangle B_E if we consider the split triangulation $\widehat{\triangle}$. See Figure 24. Then we define

$$\begin{aligned} \mathsf{x}^{\triangle}_{E,1}(\widehat{L},\nu) &:= \nu^+_E - \alpha^+_E(\widehat{L}), \\ \mathsf{x}^{\triangle}_{E,2}(\widehat{L},\nu) &:= \nu^-_E - \alpha^-_E(\widehat{L}) - [\mathsf{x}_T(\widehat{L})]_+. \end{aligned} \tag{4-3}$$

Proposition 4.3 *The shear coordinate system gives a bijection* $\mathsf{x}_\triangle : \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \xrightarrow{\sim} \mathbb{Q}^{I(\triangle)}$.

Proof Given $(\mathsf{x}_i)_{i \in I(\triangle)} \in \mathbb{Q}^{I(\triangle)}$, we can reconstruct the underlying rational unbounded lamination $\widehat{L}$ from the unfrozen part $(\mathsf{x}_i)_{i \in I(\triangle)_{\mathrm{uf}}}$ as in Section 3.4. Then the datum ν is uniquely determined by the relation (4-3). □

The following is immediate from the definition:

Lemma 4.4 *The map* $\pi_{\mathrm{uf}} \colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ *is a* ***cluster projection****. Namely, we have a commutative diagram*

$$\begin{tikzcd} \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \arrow[r, "\mathsf{x}_\triangle"] \arrow[d, "\pi_{\mathrm{uf}}"'] & \mathbb{Q}^{I(\triangle)} \arrow[d] \\ \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \arrow[r, "\mathsf{x}^{\mathrm{uf}}_\triangle"'] & \mathbb{Q}^{I_{\mathrm{uf}}(\triangle)} \end{tikzcd}$$

for any ideal triangulation $\triangle$ of Σ, where the right vertical map is the projection forgetting the frozen coordinates.

4.2 Gluing of laminations

Let Σ be a (possibly disconnected) marked surface, and $E_L, E_R \in B(\Sigma)$ distinct boundary intervals. Then we can form a new marked surface Σ' from Σ by gluing E_L with E_R. As a tropical analogue of the gluing morphism $\mathcal{P}_{PGL_3,\Sigma} \to \mathcal{P}_{PGL_3,\Sigma'}$ [21, Lemma 2.14], we are going to introduce a map

$$q_{E_L,E_R} \colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma',\mathbb{Q})$$

between the corresponding spaces of rational $\mathcal{P}$-laminations. The map q_{E_L,E_R} will be defined to be equivariant with respect to the $\mathbb{Q}_{>0}$-action, and invariant under the action $\alpha_{E_L,E_R} \colon \mathsf{P}^\vee_{\mathbb{Q}} \curvearrowright \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ given by the shift

$$\mu.(\nu_{E_L},\nu_{E_R}) := (\nu_{E_L}+\mu, \nu_{E_R}-\mu^*) \tag{4-4}$$

for $\mu = a\varpi^\vee_1 + b\varpi^\vee_2 \in \mathsf{P}^\vee_{\mathbb{Q}}$, where $\mu^* := b\varpi^\vee_1 + a\varpi^\vee_2$, and keeping other ν_E, $E \neq E_L, E_R$ intact.

Let $(\widehat{L},\nu)\in\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ be an integral $\mathcal{P}$-lamination. Represent the integral unbounded $\mathfrak{sl}_3$-lamination $\widehat{L}$ by a nonelliptic signed web W with weight 1 on every component. Around each special point of E_L and E_R, draw a semi-infinite collection of disjoint corner arcs with alternating orientations that accumulates only at the special point so that they are disjoint from W. Here we choose the orientation of the farthest corner arc from the special point to be clockwise, as in Section 3.4. Insert a biangle B between E_L and E_R, and identify $\Sigma'=\Sigma\cup B$. Notice that the ends of W on E_L and E_R, together with those of the additional corner arcs, defines an asymptotically periodic symmetric strand set $S=(S_L,S_R)$ on B. We equip S with a pinning $\mathsf{p}^{\pm}_Z$ for $Z\in\{L,R\}$ by the following rule:

- Choose continuous parametrizations $\psi^{\pm}_Z\colon\mathbb{R}\to E_Z$ so that $\psi^{\pm}_Z\bigl(\frac{1}{2}+\mathbb{Z}\bigr)=S^{\pm}_Z$, and $\psi^{\pm}_Z(\mathbb{R}_{<0})\cap S^{\pm}_Z$ consists of all the strands coming from the additional corner arcs around the initial marked point of E_Z.
- Then set $p^{\pm}_Z:=\psi^{\pm}_Z(\nu^{\pm}_{E_Z})\in E_Z$.

Then we get a pinned symmetric strand set $\widehat{S}:=(S;\mathsf{p}_L,\mathsf{p}_R)$ on the biangle B. Let $W_{\mathrm{br}}(\widehat{S})$ be the associated collection of oriented curves in B. Gluing the web W with the collection $W_{\mathrm{br}}(\widehat{S})$, we get an infinite collection $\mathcal{W}'_{\mathrm{br}}$ of webs on $\Sigma'=\Sigma\cup B$. The initial (resp. terminal) marked point of E_L is identified with the terminal (resp. initial) marked point of E_R, and regarded as new marked points in Σ'. For each of these new marked points, do the following:

- If it is a special point, then remove the peripheral components around this point from $\mathcal{W}'_{\mathrm{br}}$.
- If it is a puncture, then remove the peripheral components and replace each spiraling end around this point with a signed end, while encoding the spiraling directions in signs by reversing the rule in Figure 10. Then there remain at most finitely many intersections in B.
- Finally, replace each intersection of curves in B with an H-web by the rule (3-1).

Thus we get a nonelliptic signed web W' on Σ', which represents an integral $\mathcal{P}$-lamination

$$\widehat{L}'=q_{E_L,E_R}(\widehat{L})\in\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z}).$$

The construction is clearly invariant for the action of $H_\partial(\mathbb{Q}^T)$ by Remark 3.16, and $\mathbb{Z}_{>0}$-equivariant. Thus it can be extended $\mathbb{Q}_{>0}$-equivariantly.

Definition 4.5 The thus obtained map $q_{E_L,E_R}\colon\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})\to\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma',\mathbb{Q})$ is called the *gluing map* along E_L and E_R.

In view of Remark 3.16, we immediately have:

Lemma 4.6 *The gluing map q_{E_L,E_R} is invariant under the shift action* (4-4) *of* $\mathsf{P}^\vee_{\mathbb{Q}}$.

Any ideal triangulation $\triangle$ of Σ naturally induces a triangulation $\triangle'$ of Σ', where the edges E_L and E_R are identified and give an interior edge E of $\triangle$. The points in $I(\triangle)$ on these edges are identified as $i^s(E_L)=i^{s^*}(E_R)$ for $s=1,2$ with $s^*:=3-s$. The points of $I(\triangle)$ away from the edges E_L and E_R are naturally identified with the corresponding points of $I(\triangle')$.

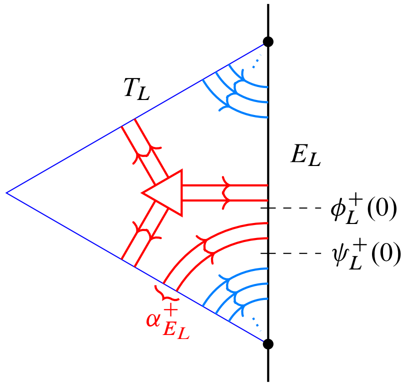


Figure 25: Comparison of two edge parametrizations. A part of the web representing $\widehat{L}$ which will be incoming to the bigon B_E is shown in red, and the additional corner arcs are shown in blue.

Theorem 4.7 *The gluing map q_{E_L,E_R} is the* ***tropicalized amalgamation****. Namely, for any ideal triangulation $\triangle$ of Σ and the induced triangulation $\triangle'$ of Σ', it satisfies*

$$q^*_{E_L,E_R}\mathsf{x}^{\triangle'}_{E,s} = \mathsf{x}^{\triangle}_{E_L,s} + \mathsf{x}^{\triangle}_{E_R,s^*}$$

for $s = 1, 2$. Here E inherits an orientation from E_L (so that from the bottom to the top, when we draw E_L on the left). The other coordinates are kept intact: $q^*_{E_L,E_R}\mathsf{x}^{\triangle'}_i = \mathsf{x}^{\triangle}_i$ for $i \in I(\triangle') \setminus \{i^s(E)\}_{s=1,2}$.

Proof The last statement is clear from the definition. To see the relation between the coordinates on the edges E_L, E_R and E, it suffices to consider an integral lamination $\widehat{L} \in \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Z})$ by $\mathbb{Q}_{>0}$-equivariance. Write $L' := q_{E_L,E_R}(\widehat{L})$ and $\mathsf{x}_i := \mathsf{x}^{\triangle}_i(\widehat{L})$ for $i \in I(\triangle)$. Recall the reconstruction procedure of the integral lamination $\widehat{L}'$ from its shear coordinates, and compare the gluing parameters

$$\begin{aligned} \nu^+_{E_L} &= \mathsf{x}_{E_L,1} + \alpha^+_{E_L}, & \nu^-_{E_R} &= \mathsf{x}_{E_R,2} + [\mathsf{x}_{T_R}]_+ + \alpha^-_{E_R}, \\ \nu^-_{E_L} &= \mathsf{x}_{E_L,2} + [\mathsf{x}_{T_L}]_+ + \alpha^-_{E_L}, & \nu^+_{E_R} &= \mathsf{x}_{E_R,1} + \alpha^+_{E_R}, \end{aligned} \tag{4-5}$$

with the integers appearing in (3-2). By Lemma 4.6, the result of gluing is unchanged under the modification

$$\begin{aligned} \tilde{\nu}^+_{E_L} &:= (\mathsf{x}_{E_L,1} + \mathsf{x}_{E_R,2}) + \alpha^+_{E_L}, & \tilde{\nu}^-_{E_R} &:= [\mathsf{x}_{T_R}]_+ + \alpha^-_{E_R}, \\ \tilde{\nu}^-_{E_L} &:= [\mathsf{x}_{T_L}]_+ + \alpha^-_{E_L}, & \tilde{\nu}^+_{E_R} &:= (\mathsf{x}_{E_L,2} + \mathsf{x}_{E_R,1}) + \alpha^+_{E_R} \end{aligned} \tag{4-6}$$

by the shift action (4-4). On the other hand, since there are "original" corner arcs of $\widehat{L}$ in T_L and T_R before adding infinite collections of corner arcs in the gluing procedure, the parametrizations of edges are related by

$$\phi^{\pm}_Z(n) = \psi^{\pm}_Z(n + \alpha^{\pm}_{E_Z})$$

for $n \in \mathbb{Z}$ and $Z \in \{L, R\}$. See Figure 25. These comparisons on the two gluing constructions show that $\widehat{L}' = q_{E_L,E_R}(\widehat{L})$ if and only if $\mathsf{x}_{E,s}(\widehat{L}') = \mathsf{x}_{E_L,s}(\widehat{L}) + \mathsf{x}_{E_R,s^*}(\widehat{L})$ for $s = 1, 2$. □

Remark 4.8 In view of the gluing construction presented above, the definition of the integral unbounded $\mathfrak{sl}_3$-laminations with pinnings can be modified slightly more geometrically as integral unbounded $\mathfrak{sl}_3$-laminations equipped with infinitely many corner arcs around special points and choices of points $p_E^{\pm} \in E$ for each $E \in \mathbb{B}$, in place of the datum $\nu_E \in \mathsf{P}^\vee$. It gives a right description of the tropical analogue of $\mathcal{P}_{\mathrm{PGL}_3,\Sigma}(\mathbb{Z}^T)$ without fixing a trivialization of the bundle (4-2). We do not pursue an extension of this description to the rational case.

4.3 Extended ensemble map

Recall the geometric ensemble map (2-9). We extend it by

$$\tilde{p}\colon \mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}), \quad L \mapsto (p(L),(\nu_E)_E),$$

where ν_E^+ (resp. ν_E^-) is minus the total weight of the peripheral components with the clockwise (resp. counterclockwise) orientation around the initial marked point of E. We have a commutative diagram

$$\begin{tikzcd}
\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \arrow[r, "\tilde{p}"] \arrow[dr, "p"'] & \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \arrow[d, "\pi_{\mathrm{uf}}"] \\
& \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})
\end{tikzcd}$$

Lemma 4.9 *If Σ has no punctures, then $\tilde{p}\colon \mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ gives a bijection.*

Proof In this case, the only datum that the map p loses is the weights of peripheral components around special points. This can be uniquely recovered from the tuple $(\nu_E)_E$. □

On the integral points, we have $\tilde{p}(\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})) \subset \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$.

Proposition 4.10 *The extended geometric ensemble map $\tilde{p}\colon \mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ coincides with the Goncharov–Shen extension of the ensemble map* (A-6). *Namely, it satisfies*

$$\tilde{p}^*\mathsf{x}_i^{\triangle} = \sum_{j \in I(\triangle)} (\varepsilon_{ij}^{\triangle} + m_{ij})\mathsf{a}_j^{\triangle} \tag{4-7}$$

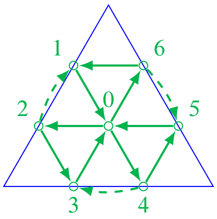
for any ideal triangulation $\triangle$ of Σ and $i \in I(\triangle)$, where

- $(\mathsf{a}_j^{\triangle})_{j\in I(\triangle)}$ *denotes the tropical $\mathcal{A}$-coordinates on $\mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ associated with $\triangle$, which is one-third of the Douglas–Sun coordinates*;
- $\varepsilon^{\triangle} = (\varepsilon_{ij}^{\triangle})_{i,j\in I(\triangle)}$ *denotes the exchange matrix defined in Section A.3*;
- $M = (m_{ij})_{i,j\in I_{\mathrm{f}}(\triangle)}$ *is the half-integral symmetric matrix given in* (A-5).

In particular, by forgetting the pinnings and frozen coordinates, we see that the geometric ensemble map $p\colon \mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^x_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q})$ coincides with the ensemble map (A-2).

Proof In view of the local nature of the definitions of coordinate systems and the exchange matrix, it suffices to consider the case where Σ is a triangle or a quadrilateral. Indeed, for $i = i(T) \in I^{\mathrm{tri}}(\triangle)$, it suffices to focus on the triangle T containing it; for $i = i^s(E) \in I^{\mathrm{edge}}(\triangle) \cap I_{\mathrm{uf}}(\triangle)$ consider the quadrilateral containing the interior edge E as a diagonal; for $i = i^s(E) \in I^{\mathrm{edge}}(\triangle) \cap I_{\mathrm{f}}(\triangle)$ consider the triangle T having the boundary interval E as one of its sides.

Triangle case For the $\mathfrak{sl}_3$-quiver associated with the unique ideal triangulation of a triangle T, label its vertices as:

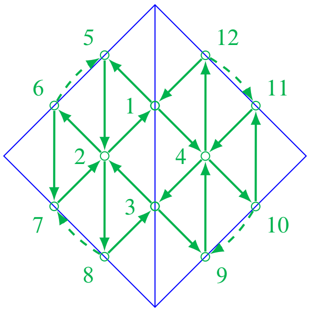


Then the expected relation (4-7) reads as

$$\begin{aligned}
\tilde{p}^*\mathsf{x}_0 &= \mathsf{a}_2 + \mathsf{a}_4 + \mathsf{a}_6 - (\mathsf{a}_1 + \mathsf{a}_3 + \mathsf{a}_5),\\
\tilde{p}^*\mathsf{x}_1 &= \mathsf{a}_0 - \mathsf{a}_1 - \mathsf{a}_6,\\
\tilde{p}^*\mathsf{x}_2 &= \mathsf{a}_1 + \mathsf{a}_3 - \mathsf{a}_2 - \mathsf{a}_0,\\
\tilde{p}^*\mathsf{x}_3 &= \mathsf{a}_0 - \mathsf{a}_3 - \mathsf{a}_2,\\
\tilde{p}^*\mathsf{x}_4 &= \mathsf{a}_3 + \mathsf{a}_5 - \mathsf{a}_4 - \mathsf{a}_0,\\
\tilde{p}^*\mathsf{x}_5 &= \mathsf{a}_0 - \mathsf{a}_5 - \mathsf{a}_4,\\
\tilde{p}^*\mathsf{x}_6 &= \mathsf{a}_5 + \mathsf{a}_1 - \mathsf{a}_6 - \mathsf{a}_0.
\end{aligned}$$

The tropical $\mathcal{A}$-coordinates of essential webs on T are defined as the weighted sum of the coordinates of its components. See [6, Section 4.3]. Therefore it suffices to check the relations for the corner arcs and the sink-/source-honeycombs of height 1, whose coordinates are shown in Figure 26. Then the relations between the two coordinates can be easily verified.

Quadrilateral case For the $\mathfrak{sl}_3$-quiver associated with an ideal triangulation $\triangle$ of a quadrilateral Q, label its vertices as:



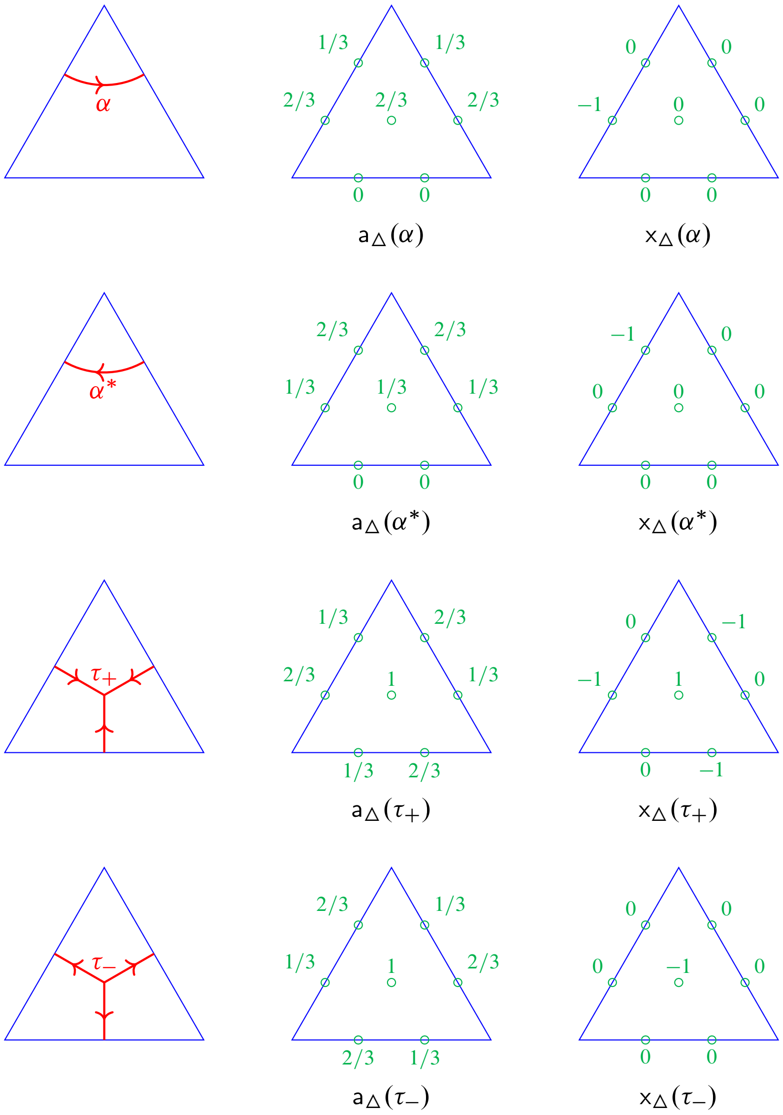

α
1/3
1/3
2/3
2/3
2/3
0
0
$\mathsf{a}_\triangle(\alpha)$
0
0
−1
0
0
0
0
$\mathsf{x}_\triangle(\alpha)$
α^*
2/3
2/3
1/3
1/3
1/3
0
0
$\mathsf{a}_\triangle(\alpha^*)$
−1
0
0
0
0
0
0
$\mathsf{x}_\triangle(\alpha^*)$
τ_+
1/3
2/3
2/3
1
1/3
1/3
2/3
$\mathsf{a}_\triangle(\tau_+)$
0
−1
−1
1
0
0
−1
$\mathsf{x}_\triangle(\tau_+)$
τ_-
2/3
1/3
1/3
1
2/3
2/3
1/3
$\mathsf{a}_\triangle(\tau_-)$
0
0
0
−1
0
0
0
$\mathsf{x}_\triangle(\tau_-)$

Figure 26: Two types of coordinates of component webs on a triangle T. All the webs shown here have weight 1.

The remaining relations to be checked are

$$\begin{aligned} \tilde{p}^*\mathsf{x}_1 &= \mathsf{a}_5 + \mathsf{a}_4 - \mathsf{a}_2 - \mathsf{a}_{12}, \\ \tilde{p}^*\mathsf{x}_3 &= \mathsf{a}_2 + \mathsf{a}_9 - \mathsf{a}_4 - \mathsf{a}_8. \end{aligned} \tag{4-8}$$

The tropical $\mathcal{A}$-coordinate assigned to a vertex $i \in I(\triangle)$ only depends on the restriction of a given web to the triangle which contains i. In particular, we can choose the braid representative with respect to $\widehat{\triangle}$ for the computation, since the biangle part does not matter. Then both $\mathcal{A}$- and $\mathcal{X}$-coordinates are weighted sums of contributions from the components of the braid representative. It is easy to verify that the both sides of the equations in (4-8) vanish for the corner arcs around the marked points Q. For the curve and honeycomb components that contribute to the shear coordinates, the expected relations are easily verified from Figures 27 and 28. Here notice that, for instance, the coordinates of the honeycomb component H_{n_1,n_2,n_3} shown in the top of Figure 15 can be computed as $\mathsf{z}_\triangle(H_{n_1,n_2,n_3}) = n_1\mathsf{z}_\triangle(\tau_+^L) + n_2\mathsf{z}_\triangle(h) + n_3\mathsf{z}_\triangle(\tau_+^R)$ for $\mathsf{z} \in \{\mathsf{a}, \mathsf{x}\}$. Together with this observation, the eight patterns shown in Figures 27 and 28 exhausts all the patterns up to symmetry. □

The following states an extension of Theorem 3.20 with pinnings/frozen variables, as promised before.

Theorem 4.11 *For any two ideal triangulations $\triangle$ and $\triangle'$ of Σ, the coordinate transformation*

$$\mathsf{x}_{\triangle,\triangle'} := \mathsf{x}_{\triangle'} \circ \mathsf{x}_\triangle^{-1} : \mathbb{Q}^{I(\triangle)} \to \mathbb{Q}^{I(\triangle)}$$

is a composite of tropical cluster Poisson transformations. In particular, we get an $MC(\Sigma)$*-equivariant identification* $\mathsf{x}_\bullet : \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \xrightarrow{\sim} \mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$.

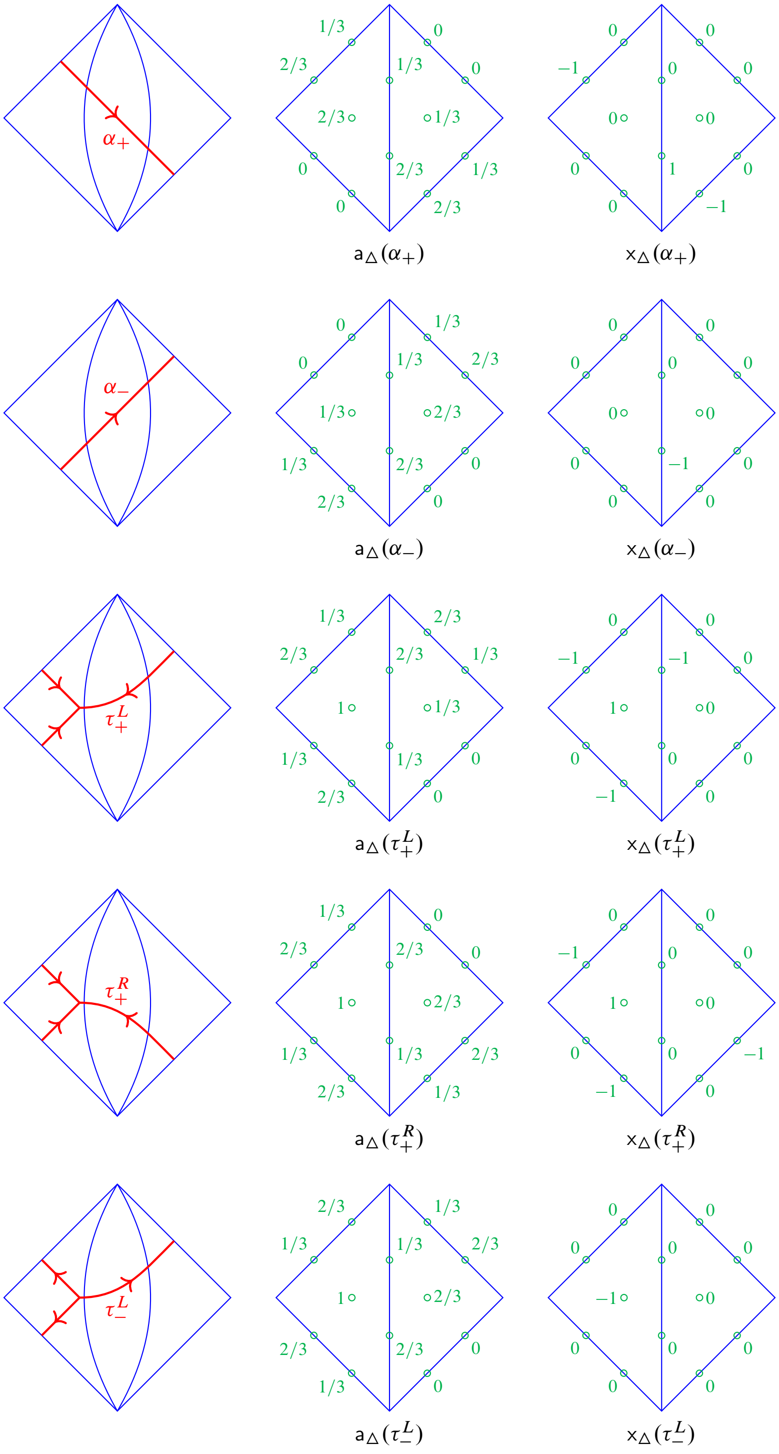
As a corollary, combining with Lemma 4.4, we get a proof of Theorem 3.20.

Proof From Lemma 4.9 and Proposition 4.10, the statement is true when Σ has no puncture (in particular, a quadrilateral). Indeed, the corresponding transformation $\mathsf{a}_{\triangle,\triangle'} := \mathsf{a}_{\triangle'} \circ \mathsf{a}_\triangle^{-1} : \mathbb{Q}^{I(\triangle)} \to \mathbb{Q}^{I(\triangle)}$ is shown to be a composite of tropical cluster $\mathcal{A}$-transformations [7, Proposition 4.2]. Then $\mathsf{x}_{\triangle,\triangle'} = (\tilde{p}^{-1})^* \circ \mathsf{a}_{\triangle,\triangle'} \circ \tilde{p}^*$ is the corresponding composite of tropical cluster $\mathcal{X}$-transformations, since the extended ensemble map commutes with the tropical cluster transformations and is a bijection in this case.

For the general case, it suffices to consider two triangulations, $\triangle$ and $\triangle'$, related by a single flip along an edge $E \in e_{\mathrm{int}}(\triangle)$. Let Q be the unique quadrilateral in $\triangle$ containing E as a diagonal, and $\Sigma' := \Sigma \setminus \mathrm{int}\, Q$ the complement marked surface. It is obvious that the shear coordinates assigned to the vertices outside Q are unchanged. On the other hand, the coordinates assigned to the vertices on Q transform correctly from the argument above under the corresponding coordinate transformation on $\mathcal{L}^p_{\mathfrak{sl}_3}(Q, \mathbb{Q})$. Since Σ is obtained by gluing Q with Σ' and the shear coordinates are obtained by amalgamating those on $\mathcal{L}^p_{\mathfrak{sl}_3}(Q, \mathbb{Q})$ and $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma', \mathbb{Q})$ by Theorem 4.7; the statement follows from the fact that the amalgamations commute with cluster $\mathcal{X}$-transformations [9, Lemma 2.2]. □

Remark 4.12 For an unpunctured surface Σ, the fastest way to introduce the coordinate system $\mathsf{x}_\triangle$ on $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q})$ which transforms correctly under the flips would be to define it via the relation (4-7) in view of Lemma 4.9. Then, however, it becomes rather difficult to obtain the amalgamation formula in Theorem 4.7, since the (tropical) $\mathcal{A}$-coordinates do not behave so simply as the (tropical) $\mathcal{X}$-coordinates under the gluing. Indeed, the following naive diagram does not commute:

$$\begin{tikzcd}
\mathcal{A}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T) \arrow[r] \arrow[d, "\tilde{p}_\Sigma"'] & \mathcal{A}_{\mathfrak{sl}_3,\Sigma'}(\mathbb{Q}^T) \arrow[d, "\tilde{p}_{\Sigma'}"] \\
\mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T) \arrow[r, "q_{E_L,E_R}"'] & \mathcal{X}_{\mathfrak{sl}_3,\Sigma'}(\mathbb{Q}^T).
\end{tikzcd}$$



$\mathsf{a}_\triangle(\alpha_+)$ $\mathsf{x}_\triangle(\alpha_+)$

$\mathsf{a}_\triangle(\alpha_-)$ $\mathsf{x}_\triangle(\alpha_-)$

$\mathsf{a}_\triangle(\tau^L_+)$ $\mathsf{x}_\triangle(\tau^L_+)$

$\mathsf{a}_\triangle(\tau^R_+)$ $\mathsf{x}_\triangle(\tau^R_+)$

$\mathsf{a}_\triangle(\tau^L_-)$ $\mathsf{x}_\triangle(\tau^L_-)$

Figure 27: Two types of coordinates of component webs on a quadrilateral Q. All the webs shown here have weight 1. (Continued in Figure 28.)

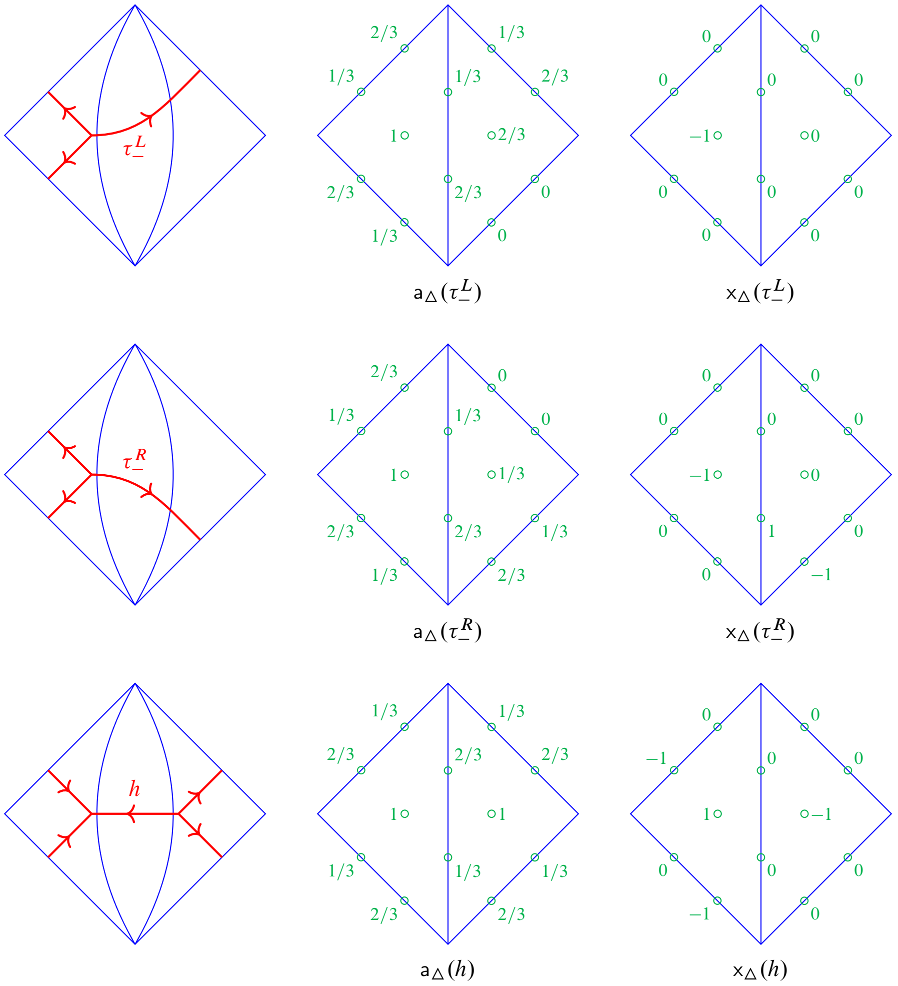


Figure 28: Two types of coordinates of component webs on a quadrilateral Q. All the webs shown here have weight 1. (Continued from Figure 27.)

Here the top right arrow denotes the quotient map given by the equation $\mathsf{a}_i=\mathsf{a}_j$ for any pair $\{i,j\}$ of quiver vertices that are identified under the gluing. Actually, we need to "rescale" some of the $\mathcal{A}$-coordinates for a correct gluing; see [29, Section 6.1] for a more detail. In particular, the sum $\tilde{p}_\Sigma^*\mathsf{x}_i^\triangle+\tilde{p}_\Sigma^*\mathsf{x}_j^\triangle$ does not compute $\tilde{p}_{\Sigma'}^*\mathsf{x}_{\bar{i}}^{\triangle'}$, where the pair $\{i,j\}$ is amalgamated into $\bar{i}$.

4.4 Dynkin involution

Let us discuss the equivariance of the shear coordinates under the Dynkin involution (4-1). The *cluster action* $*_\triangle$ (see the last paragraph of the appendix) of the Dynkin involution in the cluster chart associated to $\triangle$ is given by the mutation sequence

$$\mu_\gamma=\sigma_{e(\triangle)}\circ\mu_{t(\triangle)},$$

where $\sigma_{e(\triangle)}$ denotes the composite of the transpositions of the labels of the two vertices on each edge of $\triangle$, and $\mu_{t(\triangle)}$ is the composite of mutations at the vertex on each triangle of $\triangle$. It induces the tropical cluster $\mathcal{X}$-transformation

$$\begin{aligned} *^x_\triangle : \mathsf{x}_T &\mapsto -\mathsf{x}_T && \text{for } T \in t(\triangle),\\ \mathsf{x}_{E,1} &\mapsto \mathsf{x}_{E,2} + [\mathsf{x}_{T_L}]_+ - [-\mathsf{x}_{T_R}]_+, \\ \mathsf{x}_{E,2} &\mapsto \mathsf{x}_{E,1} + [\mathsf{x}_{T_R}]_+ - [-\mathsf{x}_{T_L}]_+ && \text{for } E \in e(\triangle), \end{aligned}$$

where we use the local labeling as in Section 3.4 for each edge E.

Proposition 4.13 *We have the commutative diagram*

$$\begin{tikzcd} \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \arrow[r, "\mathsf{x}_\triangle"] \arrow[d, "*"'] & \mathbb{Q}^{I(\triangle)} \arrow[d, "*_\triangle"] \\ \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Q}) \arrow[r, "\mathsf{x}_\triangle"'] & \mathbb{Q}^{I(\triangle)}. \end{tikzcd}$$

In particular, the orientation-reversing action of the Dynkin involution coincides with the cluster action.

Proof Mutations commute with amalgamations [9, Lemma 2.2]. Moreover, the permutation term $\sigma_{e(\triangle)}$ also commutes with the amalgamation of edge vertices corresponding to the gluing. Hence $*_\triangle$ commutes with the gluing map. It is also clear from the definitions that the Dynkin involution (4-1) commutes with gluing maps. Therefore it suffices to prove the statement for triangles.

It is easy to verify the equation

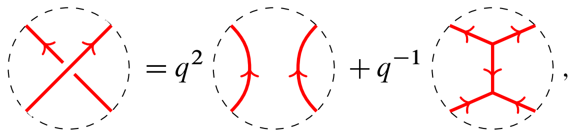
$$*_\triangle \circ \mathsf{x}_\triangle(W) = \mathsf{x}_\triangle(W^*) \tag{4-9}$$

for each component web W shown in Figure 26 by inspection. Consider a disjoint union $W = W_1 \sqcup W_2$ of webs on a triangle T, and suppose that the (4-9) is true for $W = W_1, W_2$. Since sink/source honeycombs cannot coexist, we have $\{\operatorname{sgn} \mathsf{x}_T(W_1), \operatorname{sgn} \mathsf{x}_T(W_2)\} \neq \{+,-\}$. Therefore the coordinate vectors $\mathsf{x}_\triangle(W_1)$ and $\mathsf{x}_\triangle(W_2)$ belong to the same cone on which the tropical cluster transformation $*_\triangle$ is linear. Hence,

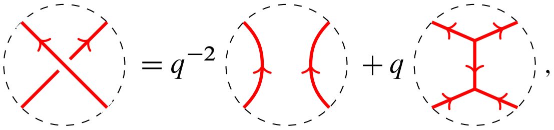
$$\begin{aligned} *_\triangle \circ \mathsf{x}_\triangle(W) &= *_\triangle(\mathsf{x}_\triangle(W_1) + \mathsf{x}_\triangle(W_1)) \\ &= *_\triangle \circ \mathsf{x}_\triangle(W_1) + *_\triangle \circ \mathsf{x}_\triangle(W_2) = \mathsf{x}_\triangle(W_1^*) + \mathsf{x}_\triangle(W_2^*) = \mathsf{x}_\triangle(W^*). \qquad \square \end{aligned}$$

5 A relation to the graphical basis and quantum duality map

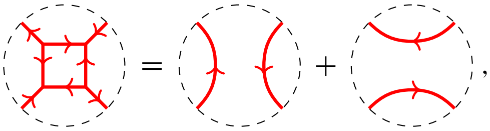
Let Σ be a marked surface without punctures. Recall from [30] the skein algebra $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}$, which is a noncommutative algebra over $\mathbb{Z}_q := \mathbb{Z}[q^{\pm 1/2}]$ consisting of tangled trivalent graphs in Σ with endpoints in $\mathbb{M}$, subject to the $\mathfrak{sl}_3$-skein relations

(5-1) 

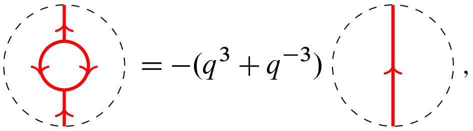
(5-2)



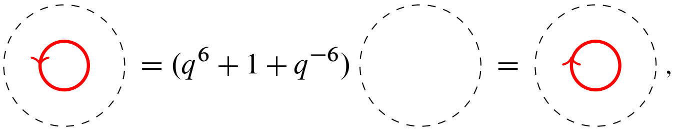
(5-3)



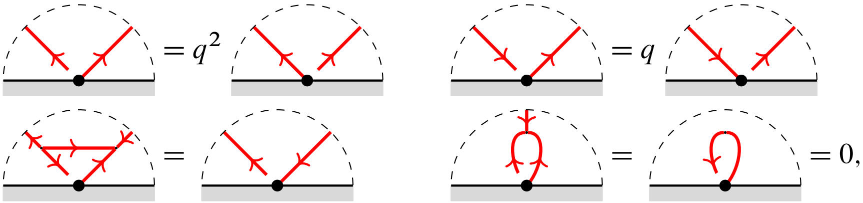
(5-4)



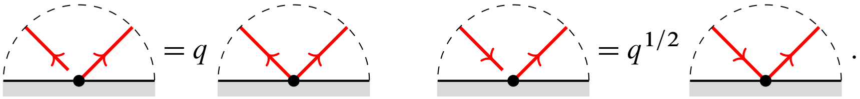
(5-5)



and the boundary skein relations



together with their Dynkin involutions. We included the square-root parameter $q^{1/2}$ so that we can consider the *simultaneous crossing* (or the *Weyl normalization*) as

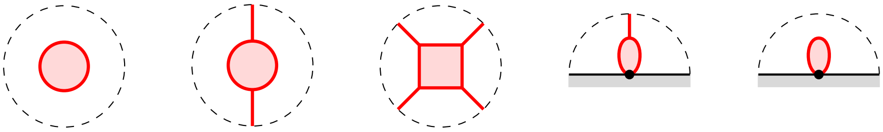


It is proved in [30] that the localized skein algebra $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}]$ along the oriented arcs parallel to boundary intervals is contained in the quantum cluster algebra [3] $\mathscr{A}^q_{\mathfrak{sl}_3,\Sigma}$ associated with a certain choice of compatibility pairs over the mutation class $\mathsf{s}(\mathfrak{sl}_3, \Sigma)$ At least in the classical limit $q = 1$, we have the equalities [29]

$$\mathscr{S}^1_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}] = \mathscr{A}_{\mathfrak{sl}_3,\Sigma} = \mathcal{O}(\mathcal{A}_{\mathfrak{sl}_3,\Sigma}). \tag{5-6}$$

The skein algebra $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}$ has a natural $\mathbb{Z}_q$-basis $\mathsf{BWeb}_{\mathfrak{sl}_3,\Sigma}$ consisting of *nonelliptic flat trivalent graphs*. Here a flat trivalent graph is an immersed oriented uni-trivalent graph on Σ such that each univalent vertex lies in $\mathbb{M}$, and the other part is embedded into int Σ. In particular, it is required to have simultaneous crossings at each special point. It is said to be nonelliptic if it has none of the following *elliptic faces*:

(5-7)



Elements of $\mathsf{BWeb}_{\mathfrak{sl}_3,\Sigma}$ are also called the *basis webs*. We are going to relate the integral $\mathfrak{sl}_3$-laminations with pinnings to the basis webs.

Definition 5.1 (negative $\mathbb{M}$-shifting of webs (cf "moving left" in [36, Figure 2])) Given a web W on Σ in the sense of Section 2.2, let $W^{\mathbb{M}} \in \mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}$ be the flat trivalent graph obtained by shifting the endpoints of W to the nearest special point in the negative direction along the boundary (with respect to the orientation induced from Σ), and taking the simultaneous crossing. See Figure 30.

For an integral $\mathfrak{sl}_3$-lamination with pinnings $(\hat{L},\nu) \in \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$, represent $\hat{L}$ by a nonelliptic $\mathfrak{sl}_3$-web W only with components with weight one, and define

$$\mathbb{I}^q_{\mathcal{X}}(\hat{L}) := \Big[W^{\mathbb{M}} \cdot \prod_{E\in\mathbb{B}} (e^+_E)^{\nu^+_E}(e^-_E)^{\nu^-_E}\Big] \in \mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}].$$

Here $\nu_E = \nu^+_E\varpi^\vee_1 + \nu^-_E\varpi^\vee_2 \in \mathsf{P}^\vee$ for each $E \in \mathbb{B}$, and the symbol $[-]$ stands for the Weyl normalization. Then $\mathbb{I}^q_{\mathcal{X}}(\hat{L})$ does not depend on the choice of the representative W, since the loop parallel-move is also realized in the skein algebra (by using the Reidemeister II move twice), and the boundary H-move exactly corresponds to the third boundary skein relation. Moreover, it is a basis web since the two notions of elliptic faces correspond to each other via the shift of endpoints.

Note that $\mathbb{I}^q_{\mathcal{X}}(\hat{L}) \in \mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}$ if and only if $\nu_E \in \mathsf{P}^\vee_+ := \mathbb{Z}_+\varpi^\vee_1 + \mathbb{Z}_+\varpi^\vee_2$ for all $E \in \mathbb{B}$. In this case, we say that $(\hat{L},(\nu_E)) \in \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ is *dominant*. Let $\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})_+ \subset \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ denote the subspace of dominant integral $\mathfrak{sl}_3$-laminations. From the above discussion, we get:

Theorem 5.2 *Assume that Σ has no punctures. Then we have an* $MC(\Sigma)\times\mathrm{Out}(SL_3)$*-equivariant bijection*

$$\mathbb{I}^q_{\mathcal{X}} \colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})_+ \xrightarrow{\sim} \mathsf{BWeb}_{\mathfrak{sl}_3,\Sigma} \subset \mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}.$$

Moreover, it is extended to a map $\mathbb{I}^q_{\mathcal{X}} \colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z}) \hookrightarrow \mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}]$, whose image again gives a $\mathbb{Z}_q$-basis.

The latter correspondence should be a basic ingredient for a construction of Fock and Goncharov's *quantum duality map* [13] (see [41, Conjecture 4.14] for a finer formulation as well as [5]), which requires a basis of the quantum upper cluster algebra parametrized by the tropical set $\mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T) = \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$ with certain positivity properties. Let us interpret Theorem 5.2 in this context.

Langlands dual coordinates It turns out that it is more convenient to use a slight modification[6] of frozen shear coordinates to make the correspondence suited to the Fock–Goncharov conjecture. For an ideal triangulation $\triangle$ of Σ, we define the *Langlands dual coordinates*

$$\check{\mathsf{x}}_\triangle = (\check{\mathsf{x}}^\triangle_i)_{i\in I(\triangle)} \colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \xrightarrow{\sim} \mathbb{Q}^{I(\triangle)}$$

[6] In the language of Goncharov and Shen [21], it amounts to take the decoration at the *terminal* endpoint of a boundary interval rather than its *initial* endpoint along the boundary orientation to make a pinning.

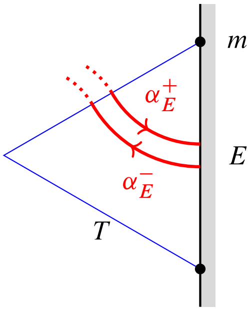


Figure 29: The corner arcs relevant to the Langlands dual coordinate.

as follows. For $i \in I_{\mathrm{uf}}(\triangle)$, let $\check{\mathsf{x}}_i^\triangle := \mathsf{x}_i^\triangle$. For $E \in \mathbb{B}$, we define the frozen coordinates on E by

$$\check{\mathsf{x}}_{E,1}^\triangle(\hat{L},\nu) := \nu_E^+ + \check{\alpha}_E^+(\hat{L}) + [\mathsf{x}_T(\hat{L})]_+,$$
$$\check{\mathsf{x}}_{E,2}^\triangle(\hat{L},\nu) := \nu_E^- + \check{\alpha}_E^-(\hat{L}).$$

Here T is the unique triangle having E as an edge; $\check{\alpha}_E^+(\hat{L})$ (resp. $\check{\alpha}_E^-(\hat{L})$) is the total weight of the oriented corner arcs in $\mathcal{W} \cap T$ bounding the *terminal* endpoint of E in the counterclockwise (resp. clockwise) direction. Compare with (4-3). The map $\check{\mathsf{x}}_\triangle$ gives a bijection, which can be verified similarly to the proof of Proposition 4.3.

We define the *Langlands dual ensemble map*

$$\check{p}\colon \mathcal{L}^a_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \to \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Q}) \tag{5-8}$$

by forgetting the peripheral components, and defining the pinning $\nu_E^+ \in \mathbb{Q}$ (resp. $\nu_E^- \in \mathbb{Q}$) to be the weight of the peripheral component around the *terminal* endpoint of E in the counterclockwise (resp. clockwise) direction. The name "Langlands dual" is inspired by the following property:

Proposition 5.3 *The Langlands dual ensemble map* (5-8) *satisfies*

$$\check{p}^*\mathsf{x}_i^\triangle = \sum_{j \in I(\triangle)} (\varepsilon_{ij}^\triangle - m_{ij})\mathsf{a}_j^\triangle$$

for any ideal triangulation.

Compare with (A-2), and observe that the presentation matrix is changed to the Langlands dual

$$-(\varepsilon^\triangle + M)^\top = \varepsilon^\triangle - M.$$

The verification of Proposition 5.3 is similar to Proposition 4.10, which is left to the reader.

For each $v \in \mathbb{E}\mathrm{xch}_{\mathfrak{sl}_3,\Sigma}$ and $k \in I$, the *elementary lamination* is the tropical point $\ell_k^{(v)} \in \mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T)$ characterized by $\check{\mathsf{x}}_i^{(v)}(\ell_k^{(v)}) = \delta_{i,k}$. We have the cone

$$\mathcal{C}^+_{(v)} := \mathrm{span}_{\mathbb{R}_+}\{\ell_k^{(v)} \mid k \in I\} = \{\ell \in \mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{R}^T) \mid \check{\mathsf{x}}_k^{(v)}(\ell) \geq 0 \text{ for all } k \in I\}$$

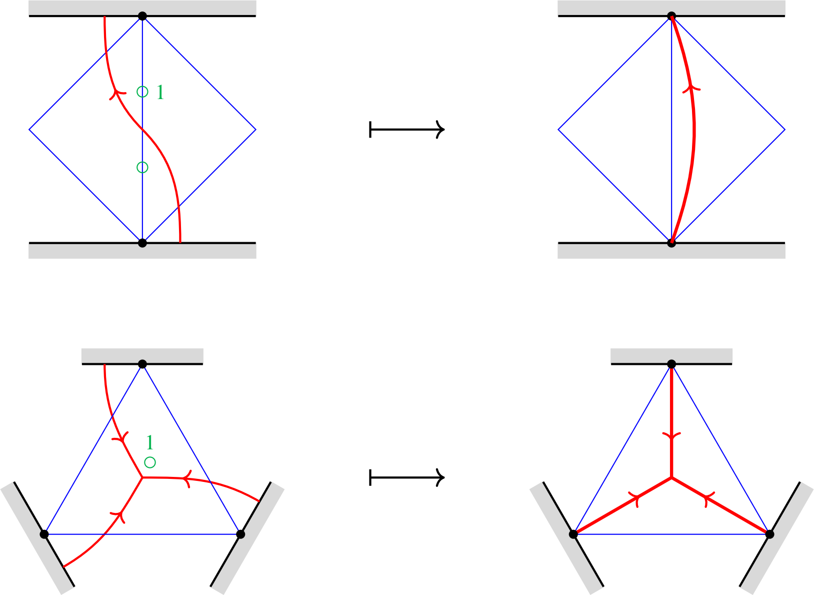


Figure 30: Negative $\mathbb{M}$-shifting of elementary laminations associated with a triangulation. Here exactly one of the Langlands dual coordinates $\check{\mathsf{x}}_i^{\Delta}$ is $+1$, while the others are zero (including the frozen ones).

and its integral points $\mathcal{C}^+_{(v)}(\mathbb{Z}) := \mathcal{C}^+_{(v)} \cap \mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T)$. The following gives a partial verification of a condition for the quantum duality map:

Lemma 5.4 *For any elementary lamination $\ell^{(v)}_k$ associated with a labeled $\mathfrak{sl}_3$-triangulation $v = (\Delta, \ell)$ in $\mathbb{E}\mathrm{xch}_{\mathfrak{sl}_3,\Sigma}$, the element $\mathbb{I}^q_{\mathcal{X}}(\ell^{(v)}_k)$ coincides with the quantum cluster variable $A^{(v)}_k \in \mathscr{A}^q_{\mathfrak{sl}_3,\Sigma}$. In particular, any point $\ell = \sum_k x_k \ell^{(v)}_k \in \mathcal{C}^+_{(v)}(\mathbb{Z})$ gives a quantum cluster monomial $\left[\prod_k (A^{(v)}_k)^{x_k}\right]$.*

Proof Via the isomorphism

$$\check{\mathsf{x}}_{\Delta}^{-1}\colon \mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T) \cong \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Z}),$$

the elementary laminations $\ell^{(v)}_k$ for unfrozen $k \in I(\Delta)_{\mathrm{uf}}$ correspond to the integral $\mathfrak{sl}_3$-laminations as shown in the left of Figure 30. The elementary laminations $\ell^{(v)}_k$ for frozen $k = i^s(E) \in I(\Delta)_{\mathrm{f}}$ with $E \in \mathbb{B}$ and $s \in \{1,2\}$ correspond to the pinning data $\nu_E = \varpi_s^\vee$. Then via the quantum duality map

$$\mathbb{I}^q_{\mathcal{X}}\colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Z})_+ \xrightarrow{\sim} \mathsf{BWeb}_{\mathfrak{sl}_3,\Sigma} \subset \mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}$$

these laminations are sent to the *elementary webs* associated with Δ in the sense of [30]. They correspond to the quantum cluster variables [30, Section 5]. □

Remark 5.5 By the equivariance of the map $\mathbb{I}^q_{\mathcal{X}}$ under the Dynkin involution, the above lemma can be immediately generalized for *decorated triangulations* (see [30, Section 1]).

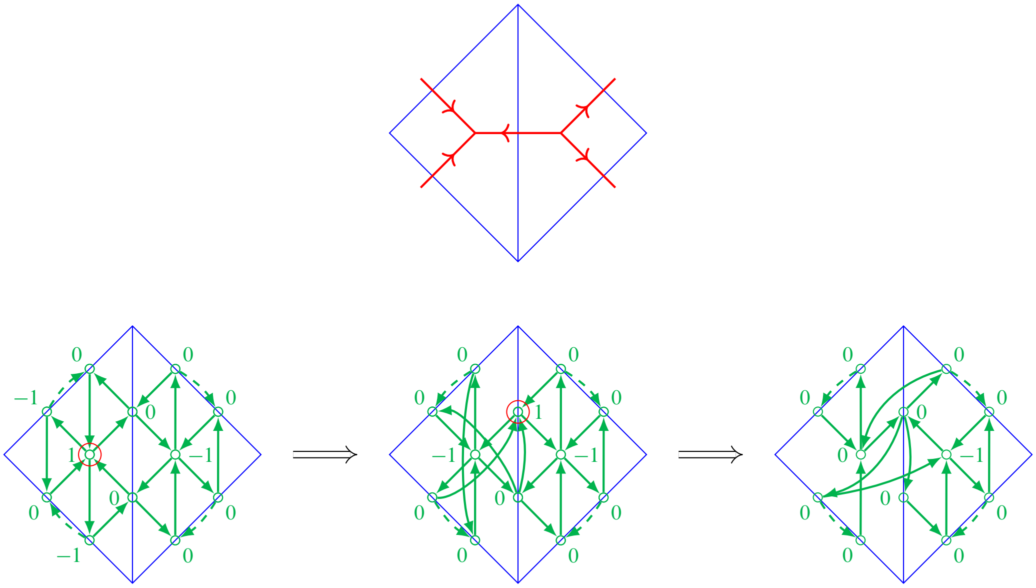


Figure 31: An elementary lamination of H-shape. Its shear coordinates associated with a triangulation is shown in the bottom left, and their transformations under the mutation sequence shown in red circles continue to the right.

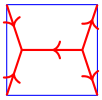
Remark 5.6 When Σ is not a k-gon with $k = 3, 4, 5$, the mutation class $\mathsf{s}(\mathfrak{sl}_3, \Sigma)$ is of infinite-mutation type. In this case, the union $\bigcup_{v\in\mathbb{E}\mathrm{xch}_{\mathfrak{sl}_3,\Sigma}} \mathcal{C}^+_{(v)}$ is not dense in $\mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{R}^{\mathrm{trop}})$ [44, Theorem 2.27]. Therefore Lemma 5.4 is far from characterizing the map $\mathbb{I}^q_{\mathcal{X}}$.

For the simplest cases that Σ is a triangle or a quadrilateral (where the mutation class $\mathsf{s}(\mathfrak{sl}_3, \Sigma)$ is finite types A_1 and D_4, respectively), we actually get a quantum duality map:

Proposition 5.7 *When Σ is a triangle or a quadrilateral, the image $\mathbb{I}^q_{\mathcal{X}}(\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma, \mathbb{Z})) \subset \mathcal{O}_q(\mathcal{A}_{\mathfrak{sl}_3,\Sigma})$ gives a $\mathbb{Z}_q$-basis consisting of quantum cluster monomials. In particular, it has positive structure constants.*

Proof For these cases, it is easy to see that $\mathscr{S}^q_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}] = \mathscr{A}^q_{\mathfrak{sl}_3,\Sigma} = \mathcal{O}_q(\mathcal{A}_{\mathfrak{sl}_3,\Sigma})$ [30, Corollary 6.1]. Moreover, the tropical set $\mathcal{X}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Z}^T)$ is covered by finitely many cones $\mathcal{C}^+_{(v)}(\mathbb{Z}^T)$ for $v \in \mathbb{E}\mathrm{xch}_{\mathfrak{sl}_3,\Sigma}$.

For the triangle case, we have only two clusters (up to permutations), and hence Lemma 5.4 with Remark 5.5 already gives the desired statement. For the quadrilateral case (type D_4), we have 16 unfrozen variables and 8 frozen variables. For instance, see [30, Appendix A and Corollary 6.1]. Up to symmetry, we have already seen in the proof of Lemma 5.4 (see Figure 30) that all of them are the images of some elementary laminations under the map $\mathbb{I}^q_{\mathcal{X}}$, except for the one represented by the elementary web



This one also comes from an elementary lamination, as seen from Figure 31. □

Conjecture 5.8 The basis $\mathbb{I}^q_{\mathcal{X}}(\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z}))$ is *parametrized by tropical points* in the sense of [41, Definition 4.13]. Namely, for any integral $\mathfrak{sl}_3$-lamination $\hat{L} \in \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})$, the quantum Laurent expression of $\mathbb{I}^q_{\mathcal{X}}(\hat{L}) \in \mathscr{A}^q_{\mathfrak{sl}_3,\Sigma}$ in the quantum cluster $\{A_i\}_{i\in I}$ associated with a vertex $\omega \in \mathbb{E}\mathrm{xch}_{\mathfrak{sl}_3,\Sigma}$ has the leading term $\big[\prod_{i\in I} A_i^{\check{\mathsf{x}}_i(\hat{L})}\big]$ with respect to the dominance order [41, Definition 4.6], where $\check{\mathsf{x}}^{(\omega)} = (\check{\mathsf{x}}_i)_{i\in I}$ is the Langlands dual shear coordinate system associated with ω.

Classical limit Recall that the set $\mathsf{BWeb}_{\mathfrak{sl}_3,\Sigma}$ also gives a $\mathbb{Z}$-basis of the classical (commutative) skein algebra $\mathscr{S}^1_{\mathfrak{sl}_3,\Sigma}$. Then Theorem 5.2 tells us that the map $\mathbb{I}^q_{\mathcal{X}}$ induces a bijection

$$\mathbb{I}_{\mathcal{X}}\colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z})_+ \xrightarrow{\sim} \mathsf{BWeb}_{\mathfrak{sl}_3,\Sigma} \subset \mathscr{S}^1_{\mathfrak{sl}_3,\Sigma},$$

which is also extended to a map $\mathbb{I}_{\mathcal{X}}\colon \mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z}) \hookrightarrow \mathscr{S}^1_{\mathfrak{sl}_3,\Sigma}[\partial^{-1}]$. Then by (5-6), we get the following:

Corollary 5.9 *The image $\mathbb{I}_{\mathcal{X}}(\mathcal{L}^p_{\mathfrak{sl}_3}(\Sigma,\mathbb{Z}))$ gives a $\mathbb{Z}$-basis of the cluster algebra $\mathscr{A}_{\mathfrak{sl}_3,\Sigma}$.*

6 Proofs of Theorems 3.10 and 3.19

6.1 Proof of Theorem 3.10

General position Recall that an *ideal arc* in $(\Sigma,\mathbb{M})$ is an immersed arc γ in Σ with endpoints in $\mathbb{M}$ which has no self-intersection except possibly at its endpoints, and not isotopic to one point. In particular γ is one-sided differentiable at each endpoint p, hence there exists a small coordinate neighborhood D_p of p such that $D_p \cap \gamma$ consists of (at most two) rays incident to p.

We say that two immersed arcs or webs in Σ are in *general position* with each other if their intersections are finite, transverse and avoiding the trivalent vertices. Moreover, we say that the spiraling diagram $\mathcal{W}$ (Definition 3.8) associated with a nonelliptic signed web is in *general position* with an ideal arc if their intersection points do not accumulate in $\operatorname{int}\Sigma$, transverse and avoiding the trivalent vertices. We may always assume the general position by the concrete construction of a spiraling diagram as logarithmic spirals near punctures.

Relative intersection number Let γ and γ' be two ideal arcs isotopic to each other with common endpoints $p_1, p_2 \in \mathbb{M}$, and $\mathcal{W}$ a spiraling diagram. Assume that these three are in a general position with each other. Then the ideal arcs γ and γ' bounds a region $B(\gamma,\gamma')$, which is a union of finitely many biangles (or such a region minus small biangles; see γ and γ'_2 in Figure 35).

By the construction of the spiraling diagram, there exists a small disk neighborhood $p_i \in D_i$ for $i = 1, 2$ such that $\rho_i := \gamma \cap D_i$ and $\rho'_i := \gamma' \cap D_i$ are rays incident to p_i, and $\mathcal{W} \cap D_i$ is a logarithmic spiral. The rays ρ_i and ρ'_i separate D_i into two sectors, and exactly one of them corresponds to the region bounded by γ and γ'. Then we can find a circular segment in this sector which does not intersect with $\mathcal{W}$, and the restriction of $\mathcal{W}$ to the circular sector separated by this segment is a periodic ladder-web. We call this circular sector $S(p_i)$ a *cut-off sector* at p_i. See Figure 32. Then $\mathcal{W}_{\mathrm{reg}} := \mathcal{W} \cap (B \setminus S(p_1) \cup S(p_2))$ is a finite web.

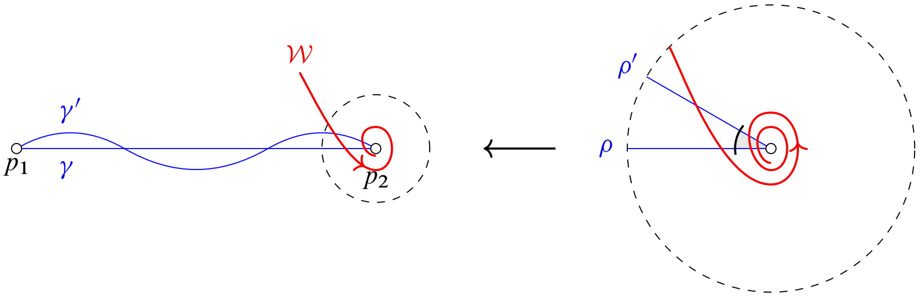


Figure 32: Two isotopic ideal arcs and a spiraling diagram. A cut-off sector is shown in gray in the right.

Definition 6.1 (relative intersection number) Let γ, γ' and $\mathcal{W}$ be as above, and choose cut-off sectors $S(p_1)$ and $S(p_2)$ at the common endpoints $p_1, p_2 \in \mathbb{M}$. Then we define the *relative intersection number* of $\mathcal{W}$ with (γ, γ') to be

$$i(\mathcal{W}; \gamma, \gamma') := i(\mathcal{W}_{\mathrm{reg}}, \gamma) - i(\mathcal{W}_{\mathrm{reg}}, \gamma').$$

Here $i(-,-)$ denotes the usual geometric intersection number of two webs.

Notice that it is independent of the choice of the cut-off sectors since a periodic ladder-web has an equal number of intersections with γ and γ' in each of its period. Clearly, we have $i(\mathcal{W}; \gamma', \gamma) = -i(\mathcal{W}; \gamma, \gamma')$.

Lemma 6.2 *Let γ_1, γ_2 and γ_3 be three ideal arcs isotopic to each other with common endpoints, and $\mathcal{W}$ a spiraling diagram. Assume that they are in general position with each other. Then we have*

$$i(\mathcal{W}; \gamma_1, \gamma_3) = i(\mathcal{W}; \gamma_1, \gamma_2) + i(\mathcal{W}; \gamma_2, \gamma_3).$$

Proof Immediately verified by choosing a common cut-off sector. □

Definition 6.3 We say that an ideal arc γ is in *minimal position* with a spiraling diagram $\mathcal{W}$ if it satisfies $i(\mathcal{W}; \gamma', \gamma) \geq 0$ for any ideal arc γ' isotopic to γ with common endpoints, and in general position with $\mathcal{W}$.

See Figure 33 for an example of an ideal arc not in a minimal position.

Realization of a minimal position We are going to prove:

Proposition 6.4 (unbounded version of [19, Corollary 12]) *Let $\mathcal{W}$ be the spiraling diagram associated with a nonelliptic signed web, and γ an ideal arc in a general position with $\mathcal{W}$. Then we can isotope $\mathcal{W}$*

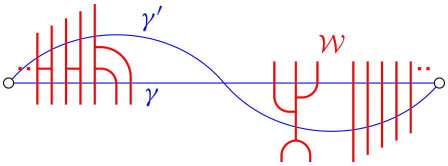


Figure 33: A spiraling diagram $\mathcal{W}$ that is not in minimal position with an ideal arc γ. Indeed, $i(\mathcal{W}; \gamma', \gamma) = -4$.

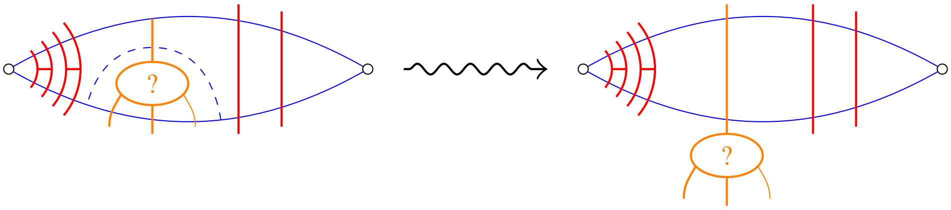


Figure 34: The restriction of a spiraling diagram $\mathcal{W}$ to a biangle bounded by two ideal arcs. Its loose part is shown in orange, which can be pushed out through a sequence of intersection reduction moves and H-moves.

into a spiraling diagram $\mathcal{W}'$ in minimal position with γ via a finite sequence of intersection reduction moves, H-moves, and an isotopy relative to γ.

To prove this, the following lemma is useful:

Lemma 6.5 (unbounded version of [19, Lemma 15]) *Let B be a biangle in Σ bounded by two immersed arcs α and α', and $\mathcal{W}$ a spiraling diagram in a general position. If some of the endpoints of α and α' are punctures, then choose any cut-off sectors and consider $\mathcal{W}_{\mathrm{reg}}$ as above. Otherwise, set $\mathcal{W}_{\mathrm{reg}} := \mathcal{W}$. Then $\mathcal{W}_{\mathrm{reg}}$ can be isotoped through a finite number of intersection reduction moves and H-moves so that $\mathcal{W}_{\mathrm{reg}} \cap B$ consists of disjoint parallel arcs connecting α and α'. This can be done by preserving the cut-off sectors, and the resulting web does not depend on the choice of cut-offs.*

Proof Since $\mathcal{W}_{\mathrm{reg}}$ is finite, the statement follows from [19, Lemma 15]. □

Notice that each of the H-moves and the intersection reduction moves are accompanied with a small biangle (shown by dashed lines in Figures 7 and 8) that cuts out a part of the web which we push out. Therefore the finite sequence of these moves in Lemma 6.5 is accompanied with a finite collection $\{B^{(j)}\}_{j\in J}$ of biangles that is partially ordered for the inclusion according to the order of moves, which we call the *tightening biangles*. Let us denote by $W^{(j)}$ the part of $\mathcal{W}$ cut out by the tightening biangle $B^{(j)}$, which we call the *loose part* of $\mathcal{W}$. See Figure 34.

The following lemma ensures that the intersection reduction procedures of a spiraling diagram associated with a nonelliptic signed web always terminate in finite steps.

Lemma 6.6 *For any spiraling diagram $\mathcal{W}$ associated with a nonelliptic signed web W and an ideal arc γ in general position, the relative intersection number $i(\mathcal{W};\gamma,\gamma')$ is bounded from above when γ' runs over the ideal arcs homotopic to γ and in general position with γ and $\mathcal{W}$.*

Proof If W has punctured H-faces, then applying appropriate puncture H-moves, we obtain another signed web W' which is puncture-reduced. The corresponding spiraling diagrams $\mathcal{W}$ and $\mathcal{W}'$ differ only by some finitely many H-shaped parts in the spiraling part, and hence $i(\mathcal{W};\gamma,\gamma') = i(\mathcal{W}';\gamma,\gamma')$. Therefore it suffices to consider the case where the signed web W giving rise to $\mathcal{W}$ is puncture-reduced.

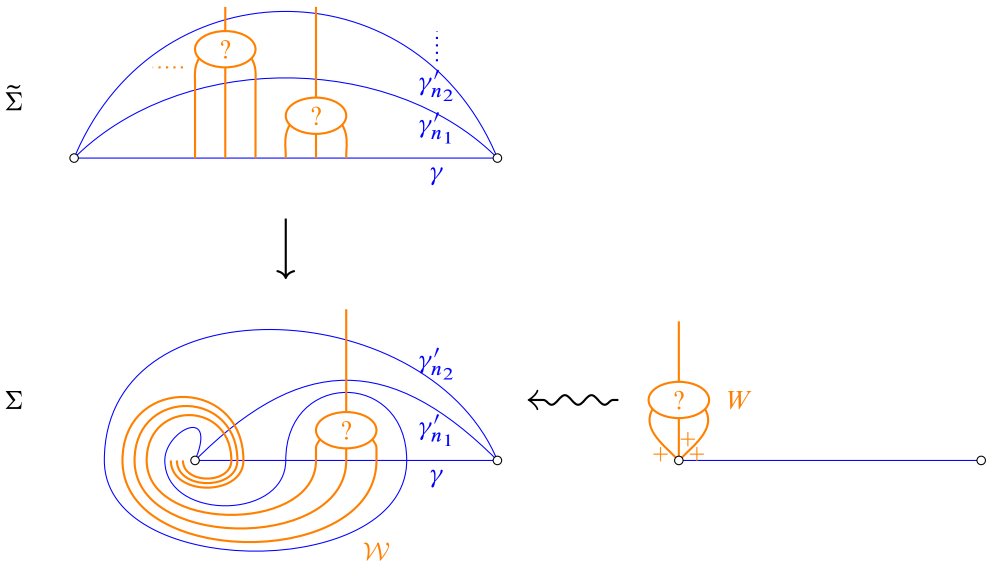


Figure 35: The situation that the relative intersection numbers $i(\mathcal{W};\gamma,\gamma'_n)$ diverge. The top left shows a covering of Σ around the puncture. The infinite sequence of portions are projected to the same portion of $\mathcal{W}$.

We prove the assertion by contradiction: suppose that there exists a sequence $\gamma'_n \simeq \gamma$ of ideal arcs satisfying the condition and $i(\mathcal{W};\gamma,\gamma'_n) \geq n$ for all $n \in \mathbb{Z}_{\geq 0}$. Let $\{B_n^{(j)}\}_{j \in J_n}$ be the collection of tightening biangles for the pair (γ,γ'_n), and $W_n^{(j)} \subset \mathcal{W}$ the corresponding loose part.

(a) Since we are interested in a sequence γ'_n such that $i(\mathcal{W};\gamma,\gamma'_n)$ diverges, we may assume that all of the tightening biangles $B_n^{(j)}$ are stuck to γ rather than γ'_n. Otherwise, a biangle stuck to γ'_n contributes negatively to $i(\mathcal{W};\gamma,\gamma'_n)$. Then we may isotope γ'_n to avoid this biangle without decreasing $i(\mathcal{W};\gamma,\gamma'_n)$.

(b) Shrinking each tightening biangle (without changing the intersection number of its boundary with $\mathcal{W}$) if necessary, we may assume that either $B_n^{(j)} \cap B_m^{(\ell)} = \varnothing$, $B_n^{(j)} \subset B_m^{(\ell)}$ or $B_m^{(\ell)} \subset B_n^{(j)}$ holds for any pair in this collection. Also we can ensure that each tightening biangle does not intersect with the cut-off sectors at punctures.

Let us consider the compact interval $K = \gamma \setminus (\text{cut-off sectors})$. From the assumption of general position, the intersection of $\mathcal{W}$ with K is finite. The intersections $I_n^{(j)} := \gamma \cap \operatorname{int} B_n^{(j)}$ give open intervals in K. Observe that the union $\bigcup_{n \geq 0,\ j \in J_n} I_n^{(j)}$ has finitely many path-connected components, since each such component contains a distinct point in $\mathcal{W} \cap K$, which is finite. Therefore we see that there exist subsequences n_k and $j_k \in J_{n_k}$ such that $B_{n_k}^{(j_k)} \subset B_{n_{k+1}}^{(j_{k+1})}$.

Such a nested situation is illustrated in Figure 35. Indeed, the situation says that distinct reduction moves are applied infinitely many times, while the original signed web W is finite. It means that there is a portion P of the signed web that is referred infinitely many times. Therefore the nested biangles $B_{n_k}^{(j_k)}$ (or the

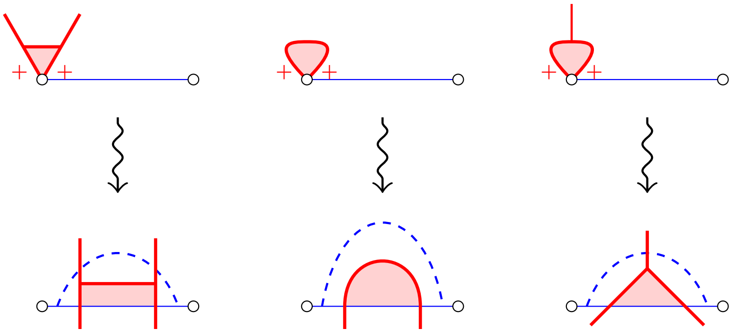


Figure 36: The correspondence between the puncture-faces (top) and the faces stuck to γ (bottom).

arcs γ'_{n_k}) must be winding around one of the punctures p_1 or p_2, while the portion $\mathcal{P}$ in $\mathcal{W}$ corresponding to P is spiraling around the same puncture as in the bottom left of the figure. Notice that such a spiraling diagram $\mathcal{W}$ arises from the signed web W shown in the bottom right.

Moreover, observe the correspondence shown in Figure 36 between the faces stuck to γ and the puncture-faces. Therefore, the sequence of loose parts $W_{n_k}^{(j_k)}$ must come from these puncture-faces in the signed web W, which contradicts to either the puncture-reduced assumption, nonelliptic condition, or the no bad ends condition. Thus the assertion is proved. □

Proof of Proposition 6.4 Suppose that $\mathcal{W}$ is not in minimal position with γ. Then there exists an ideal arc $\gamma_0 \simeq \gamma$ such that $i(\mathcal{W};\gamma,\gamma_0) > 0$ and in general position with γ and $\mathcal{W}$. Choose γ_0 so that $i(\mathcal{W};\gamma,\gamma_0)$ is maximal, whose existence is ensured by Lemma 6.6. Then for any other ideal arc γ' isotopic to γ, we have

$$i(\mathcal{W};\gamma',\gamma_0) = i(\mathcal{W};\gamma',\gamma) + i(\mathcal{W};\gamma,\gamma_0) = -i(\mathcal{W};\gamma,\gamma') + i(\mathcal{W};\gamma,\gamma_0) \geq 0$$

by Lemma 6.2 and the maximality of γ_0. It implies that γ_0 is in minimal position with $\mathcal{W}$, as desired. □

Corollary 6.7 (cf [19, Corollary 12 and Proposition 13]) *Any spiraling diagram $\mathcal{W}$ associated with a signed web on Σ can be isotoped through a finite number of intersection reduction moves and H-moves so that it is in minimal position simultaneously with any disjoint finite collection $\{\gamma_i\}_{i=1}^N$ of ideal arcs. Such a minimal position with $\{\gamma_i\}_{i=1}^N$ is unique up to isotopy relative to these arcs, H-moves, periodic H-moves and parallel moves.*

Proof As in the discussion above, we isotope the arcs instead of the spiraling diagram. Let $\{\gamma_i\}_i$ be the original collection of ideal arcs, and $\{\gamma'_i\}_i$ the collection of modified arcs such that $i(\mathcal{W};\gamma_i,\gamma'_i)$ is maximal. Let B_i be the biangle bounded by γ_i and γ'_i. We claim that we can slightly modify B_i as in (b) above so that it does not cross γ'_j for any $i \neq j$. Indeed, suppose B_i crosses γ'_j. If we can shrink B_i without changing the intersection with $\mathcal{W}$, do so. Otherwise, it implies that γ_i and γ'_j bound together at least one biangle $B' \subset B_j$, for which we can apply a reduction move (see Figure 37). It contradicts to the maximality of $i(\mathcal{W};\gamma_j,\gamma'_j)$.

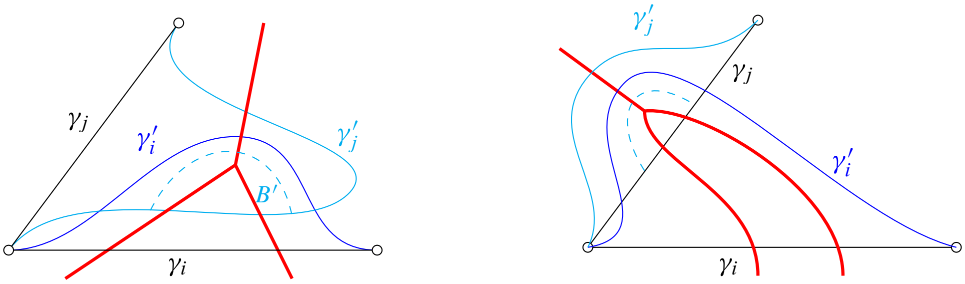


Figure 37: Situations where B_i essentially crosses γ'_j (left) and B_i crosses γ_j essentially (right). Both pictures show the case where B_i intersects B_j only once.

Hence, the biangle B_i is either disjoint from B_j or intersect with B_j only through γ_j. In the former case, the reduction moves are independently applied. In the latter case, some of the reduction moves are common for γ_i and γ_j but still the minimal positions can be simultaneously realized. Thus we get the first statement.

The second one is proved by induction on the number N of arcs, just in the same way as the proof of [19, Proposition 13]. □

Proof of Theorem 3.10: realization of a good position By Corollary 6.7, we can place any spiraling diagram $\mathcal{W}$ in a minimal position with the ideal arcs in the split triangulation $\widehat{\triangle}$. Then by applying a finite number of H-moves and periodic H-moves, we can push all the *ladders* as in Figure 38 into biangles (the "tidying up" operation in [19]). Assume that these moves can be no longer applied to $\mathcal{W}$. We are going to prove that this position (the "joy-sparking" position in [19]) is a good position with respect to $\widehat{\triangle}$.

For each $E \in e(\triangle)$, the intersection $\mathcal{W} \cap B_E$ is an unbounded essential web by Lemma 6.5, since it is in minimal position with the ideal arcs bounding B_E. For each $T \in t(\triangle)$, we see that the only components of $\mathcal{W} \cap T$ which do not touch all sides of T are corner arcs by Lemma 6.5. Indeed, such a component can be viewed as a web in a biangle obtained from T by collapsing one edge that is not touched, and the ladders in the periodic part have been pushed into the biangles neighboring to T. Let W' be the web

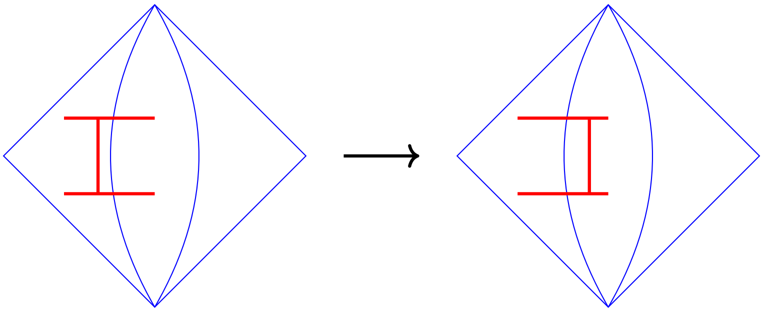


Figure 38: Pushing a ladder into a biangle.

obtained from $\mathcal{W} \cap T$ by removing these corner arcs, which must be finite. Then we see that W' must be a honeycomb in the same as in the last part in the proof of [19, Theorem 19]. Hence $\mathcal{W} \cap T$ is an unbounded rung-less essential web. The uniqueness statement follows from that of Corollary 6.7. Thus Theorem 3.10 is proved.

6.2 Proof of Theorem 3.19

We are going to prove Theorem 3.19 by following the strategy for the proof of [6, Theorem 47]. We remark here that another proof of the latter statement is given in [19, Section 14] based on the graded skein algebras.

The main issue here is that we have fixed the periodic pattern of corner arcs in the reconstruction procedure. Hence the resulting spiraling diagram may differ from the original one by a periodic permutation of corner arcs ("periodic local parallel-moves") on each triangle. Our claim is that these local adjustments glue together to give a global parallel-move, thus we get equivalent $\mathfrak{sl}_3$-laminations. See Figure 40 for a typical example.

By the $\mathbb{Q}_{>0}$-equivariance, it suffices to consider integral unbounded $\mathfrak{sl}_3$-laminations, which are represented by signed nonelliptic webs. Therefore it suffices to prove the following statement:

Proposition 6.8 *If two signed nonelliptic webs W_1 and W_2 have the same shear coordinates $(\mathsf{x}_i)_{i \in I_{\mathrm{uf}}(\triangle)}$ with respect to an ideal triangulation $\triangle$, then W_1 and W_2 are equivalent as unbounded $\mathfrak{sl}_3$-laminations.*

In what follows, the index $\nu \in \{1, 2\}$ will always given to the objects associated to the web W_ν. For a discrete subset $A \subset \mathbb{R}$ (eg $A = \mathbb{Z}$), we call a subset $I \subset A$ of the form $I = [a, b] \cap A$ for a (possibly unbounded) interval $[a, b] \subset \mathbb{R}$ an *interval* in A.

Global pictures Let W_1 and W_2 be as in Proposition 6.8. For $\nu = 1, 2$, we may assume that the associated spiraling diagram $\mathcal{W}_\nu$ is placed in a good position with respect to the split triangulation $\widehat{\triangle}$ by Theorem 3.10. Then its braid representative $\mathcal{W}^{\triangle}_{\nu,\mathrm{br}}$ has at most one honeycomb component on each triangle. Let Σ° be the holed surface, which is a compact surface obtained by removing a small open disk D_T in each $T \in t(\triangle)$ from Σ. We may isotope the unique honeycomb component of $\mathcal{W}^{\triangle}_{\nu,\mathrm{br}}$ into the disk D_T, so that $\langle \mathcal{W}_\nu \rangle := \mathcal{W}^{\triangle}_{\nu,\mathrm{br}} \cap \Sigma^\circ$ is a collection of oriented curves, whose ends either lie on $\partial\Sigma^\circ$ or spiral around punctures. Following [6], we call $\langle \mathcal{W}_\nu \rangle$ the *global picture* associated with $\mathcal{W}^{\triangle}_{\nu,\mathrm{br}}$. It is obvious to reconstruct the braid representative from its global picture. We call each oriented curve in $\langle \mathcal{W}_\nu \rangle$ a *traveler.*

Recall from Steps 1 and 2 in the reconstruction procedure (Section 3.4) that we can construct a braid representative $W^{\triangle}_{\nu,\mathrm{br}}$ of signed web by replacing the spiraling ends with signed ends. We similarly define its global picture by $\langle W_\nu \rangle := W^{\triangle}_{\nu,\mathrm{br}} \cap \Sigma^\circ$. For the scheme of our proof, see Figure 39. Our strategy is as follows:

$$\begin{tikzcd}
\boxed{\text{signed web } W_\nu} \arrow[r] \arrow[rrd, dashed, bend left, "\mathsf{x}_\triangle"] & \boxed{\text{spiraling diagram } \mathcal{W}_\nu} \arrow[d, leftrightarrow, "1:1"] & \\
\boxed{\text{global picture } \langle W_\nu \rangle} \arrow[u, "\text{Step 3}"'] \arrow[r, leftrightarrow, "1:1", "\text{Steps 1 \& 2}"'] & \boxed{\text{global picture } \langle \mathcal{W}_\nu \rangle} \arrow[r, leftrightarrow, "1:1", "\text{Lemma 3.17}"'] & \mathbb{Z}^{I_{\mathrm{uf}}(\triangle)} \arrow[llu, dashed, bend left, "\xi_\triangle"]
\end{tikzcd}$$

Figure 39: The scheme for a proof of Proposition 6.8. It is obvious that the three objects $\mathcal{W}_\nu$, $\langle \mathcal{W}_\nu \rangle$ and $\langle W_\nu \rangle$ are in one-to-one correspondences (up to strict isotopies), when one fixes a triangulation $\triangle$. It will be proved that we get the identity (up to equivalence of signed webs) after going through the square.

(1) Starting from the assumption in Proposition 6.8, we are going to make a correspondence between the topological data of global pictures $\langle W_1 \rangle$ and $\langle W_2 \rangle$ (namely, their travelers and intersection points among them) by an unbounded version of the "fellow-traveler lemma" [6, Lemma 57].

(2) From such a correspondence, we can describe a sequence of elementary moves relation W_1 and W_2 by just following the argument of Douglas and Sun [6, Section 7.4] for the bounded case.

Unbounded fellow-traveler lemma For each traveler γ in $\langle \mathcal{W}_\nu \rangle$, fix a basepoint $x_0 \in \gamma$ so that it does not lie on any edge of $\triangle$. Associated to such a based traveler (γ, x_0) is the *route* $(E_i)_{i \in I}$, where $I \subset \mathbb{Z}$ is an interval and E_i is the i^{th} edge of $\triangle$ crossed by γ listed in order according to the orientation of γ: the 0^{th} edge is the first one encountered by γ after passing x_0. We also define the *turning pattern* $(\tau_i)_{i \in I} \subset \{L, S, R\}^I$ of the based traveler (γ, x_0) as follows:

$$\tau_i := \begin{cases} L & \text{if } E_{i+1} \text{ follows } E_i \text{ in the counterclockwise direction at their common endpoints,} \\ S & \text{if } \gamma \text{ ends at the boundary of } D_T \text{ right after passing } E_i, \\ R & \text{if } E_{i+1} \text{ follows } E_i \text{ in the clockwise direction at their common endpoints.} \end{cases}$$

The following is immediately verified:

Lemma 6.9 *The topological types of the travelers γ are distinguished by the periodicity of the data $(E_i, \tau_i)_{i \in I}$, as follows:*

- *γ is a bounded arc both of whose ends lie on $\partial \Sigma^\circ$ if $I \subset \mathbb{Z}$ is bounded;*
- *γ is a loop if $I = \mathbb{Z}$ and the route is totally periodic (namely, $E_{i+k} = E_i$ for some $k \in \mathbb{Z}$). Moreover, it is peripheral if the turning pattern $(\tau_i)_{i \in I}$ is constant;*
- *γ has an end spiraling to a puncture p, say in the forward direction, if $I \subset \mathbb{Z}$ is unbounded from above, the route $(E_i)_i$ is not totally periodic but eventually periodic, and the turning pattern $(\tau_i)_i$ is eventually constant in the forward direction.*

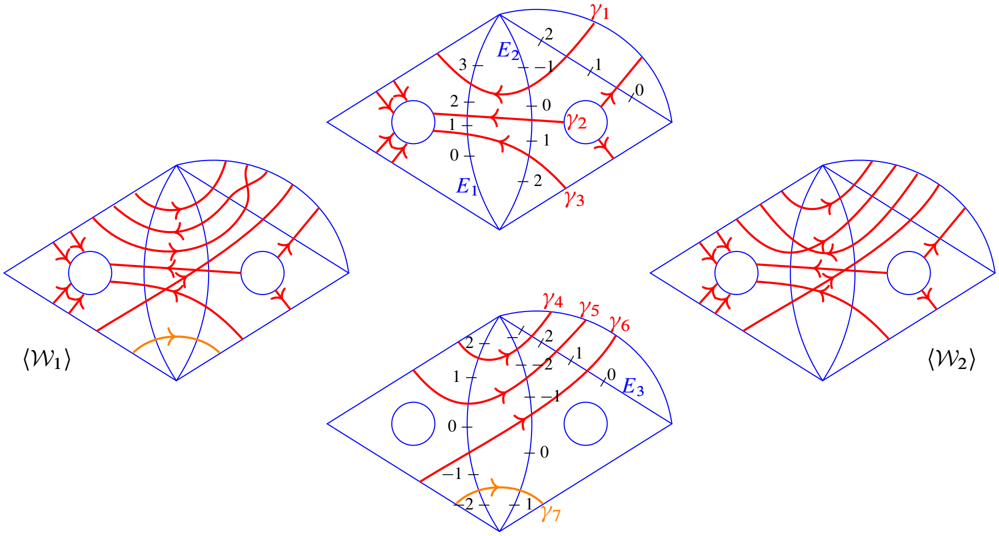


Figure 40: Example of local picture of a pair $(\langle \mathcal{W}_1 \rangle, \langle \mathcal{W}_2 \rangle)$ having the same shear coordinates. Here the top (resp. bottom) picture shows the collection of oriented curves going through the central biangle from the right to the left (resp. from the left to the right), which is common for $\langle \mathcal{W}_1 \rangle$ and $\langle \mathcal{W}_2 \rangle$ except for γ_7.

We say that two travelers $\gamma^{(1)}$ in $\langle \mathcal{W}_1 \rangle$ and $\gamma^{(2)}$ in $\langle \mathcal{W}_2 \rangle$ are *fellow-travelers* if their data $(E_i^{(1)}, \tau_i^{(1)})_{i \in I_1}$ and $(E_i^{(2)}, \tau_i^{(2)})_{i \in I_2}$ are the same, in the sense that there exists an order-preserving bijection $f \colon I_1 \to I_2$ such that $E_{f(i)}^{(2)} = E_i^{(1)}$ and $\tau_{f(i)}^{(2)} = \tau_i^{(1)}$ for all $i \in I_1$. Notice that the notion of fellow-traveler does not depend on the choice of basepoints, and that two fellow-travelers have the same topological type by Lemma 6.9.

Lemma 6.10 (unbounded fellow-traveler lemma, cf [6, Lemma 57]) *Under the assumption of Proposition 6.8, there exists a bijection*

$$\varphi \colon \{\textit{nonperipheral travelers in } \langle \mathcal{W}_1 \rangle\} \xrightarrow{\sim} \{\textit{nonperipheral travelers in } \langle \mathcal{W}_2 \rangle\}$$

such that γ and $\varphi(\gamma)$ are fellow-travelers.

Traveler identifier In order to prove Lemma 6.10, let us introduce another data that identifies the traveler and can be characterized by the shear coordinates. Let us consider two triangles $T_L, T_R \in t(\triangle)$ that shares a biangle B_E. For $Z \in \{L, R\}$, let E_Z denote the edge of $\widehat{\triangle}$ shared by T_Z and B_E. Let $S_{E_Z}^{+,(\nu)}$ (resp. $S_{E_Z}^{-,(\nu)}$) denote the set of strands on E_Z incoming to (resp. outgoing from) the biangle B_E, which are given by the intersections of travelers in $\langle \mathcal{W}_\nu \rangle$ and E_Z for $\nu = 1, 2$. We endow E_Z with the orientation induced from the triangle T_Z.

Choose two orientation-preserving parametrizations of E_Z in the same way as in Section 3.4. Namely, choose $\phi_{E_Z}^{\pm,(\nu)} \colon \mathbb{R} \to E_Z$ so that the inverse image of $S_{E_Z}^{\pm,(\nu)}$ is an interval $I_{E_Z}^{\pm,(\nu)} \subset \frac{1}{2} + \mathbb{Z}$, and $\phi_{E_Z}^{\pm,(\nu)}(\mathbb{R}_{<0}) \cap S_{E_Z}^{\pm,(\nu)}$ consists of all the strands coming from the corner arcs around the initial marked point of E_Z. Let $f^{\pm,(\nu)}_{E_Z}\colon E_Z \to \mathbb{R}$ be the inverse map of $\phi^{\pm,(\nu)}_{E_Z}$. For a traveler $\gamma^{(\nu)}$ in $\langle \mathcal{W}_\nu \rangle$ that intersects with the edge E_Z at a point x, its *traveler identifier* at E_Z is the pair $(k,\epsilon) \in \bigl(\frac{1}{2}+\mathbb{Z}\bigr) \times \{\pm 1\}$ given by

$$(k,\epsilon) := \begin{cases} (f^{+,(\nu)}_{E_Z}(x), +) & \text{if } \gamma \text{ enters } B_E \text{ from } E_Z, \\ (f^{-,(\nu)}_{E_Z}(x), -) & \text{if } \gamma \text{ exits } B_E \text{ from } E_Z. \end{cases}$$

Then we write $\gamma^{(\nu)} = \gamma^{(\nu)}_{E_Z}(k,\epsilon)$.

Example 6.11 In the example shown in Figure 40, we have

$$\begin{aligned}
\gamma_1 &= \gamma^{(\nu)}_{E_1}(5/2,-) = \gamma^{(\nu)}_{E_2}(-1/2,+) = \gamma^{(\nu)}_{E_3}(3/2,-),\\
\gamma_2 &= \gamma^{(\nu)}_{E_1}(3/2,-) = \gamma^{(\nu)}_{E_2}(1/2,+),\\
\gamma_3 &= \gamma^{(\nu)}_{E_1}(1/2,-) = \gamma^{(\nu)}_{E_2}(3/2,+),\\
\gamma_4 &= \gamma^{(\nu)}_{E_1}(3/2,+) = \gamma^{(\nu)}_{E_2}(-5/2,-) = \gamma^{(\nu)}_{E_3}(5/2,+),\\
\gamma_5 &= \gamma^{(\nu)}_{E_1}(1/2,+) = \gamma^{(\nu)}_{E_2}(-3/2,-) = \gamma^{(\nu)}_{E_3}(3/2,+),\\
\gamma_6 &= \gamma^{(\nu)}_{E_1}(-1/2,+) = \gamma^{(\nu)}_{E_2}(-1/2,-) = \gamma^{(\nu)}_{E_3}(1/2,+),\\
\gamma_7 &= \gamma^{(\nu)}_{E_1}(-3/2,+) = \gamma^{(\nu)}_{E_2}(1/2,+).
\end{aligned}$$

Lemma 6.12 *Let $\gamma^{(\nu)}$ be a traveler in $\langle \mathcal{W}_\nu \rangle$ that passes through B_E from E_L to E_R. Let $(k_L,+)$ and $(k_R,-)$ be its traveler identifier at E_L and E_R, respectively. Then we have*

$$k_L + k_R = \mathsf{x}_{E,1}(W_\nu) + [\mathsf{x}_{T_R}(W_\nu)]_+.$$

If $\gamma^{(\nu)}$ passes through B_E from E_R to E_L, then its traveler identifiers $(k_R,+)$ and $(k_L,+)$ satisfy $k_L + k_R = \mathsf{x}_{E,2}(W_\nu) + [\mathsf{x}_{T_L}(W_\nu)]_+$.

Proof Just observe that our choice of parametrizations $\phi^{\pm,(\nu)}_{E_Z}$ is the same as in the reconstruction procedure (Section 3.4), except for the difference that we do not necessarily have an infinite number of corner arcs here. Then the assertion is obtained from the gluing rule (3-2). □

Lemma 6.13 *The traveler identifiers characterizes the traveler and its topological type. Namely,*

(1) *the traveler identifier determines the data $(E_i,\tau_i)_{i\in I}$ for each traveler;*

(2) *if $\gamma^{(1)}_E(k,\epsilon) = \gamma^{(1)}_{E'}(k',\epsilon')$ for two edges E and E' of the split triangulation $\widehat{\triangle}$, then*

$$\gamma^{(2)}_E(k,\epsilon) = \gamma^{(2)}_{E'}(k',\epsilon').$$

Proof (1) The initial edge E_0 is determined from the basepoint x_0. Assume that we have determined the data E_i for $0 \le i \le k$ and τ_j for $0 \le j \le k-1$. Let $E := E_k$. Then E_{k-1} and τ_{k-1} tell us from which direction our traveler passes through the biangle B_E. Assume it is from T_L to T_R, without loss of

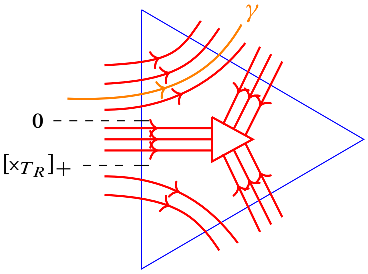


Figure 41: The turning pattern determined by the value of k_R.

generality. Then by Lemma 6.12, we have $k_R = \mathsf{x}_{E,1}(W_\nu) + [\mathsf{x}_{T_R}(W_\nu)]_+ - k_L$. Then by the choice of the parametrization $\phi_{E_R}^{-,(\nu)}$, we see that

$$\tau_k = \begin{cases} L & \text{if } k_R < 0, \\ S & \text{if } 0 < k_R < [\mathsf{x}_{T_R}(W_\nu)]_+, \\ R & \text{if } k_R > [\mathsf{x}_{T_R}(W_\nu)]_+. \end{cases}$$

See Figure 41. Moreover, the pattern τ_k tells us the next edge E_{k+1} or its absence.

(2) Recall that the shear coordinates of W_1 and W_2 are the same. Since the reconstruction given in (1) is characterized by the shear coordinates, the assertion follows. □

Proof of Lemma 6.10 Define the bijection φ by

$$\varphi\colon \gamma_E^{(1)}(k,\epsilon) \mapsto \gamma_E^{(2)}(k,\epsilon). \tag{6-1}$$

It is well defined by Lemma 6.13(2), and preserves the topological types of travelers by Lemma 6.13(1). □

Remark 6.14 From the proof of Lemma 6.12, a traveler $\gamma = \gamma_E^{(1)}(k,\epsilon)$ with $k \in I_E^{\epsilon,(1)} \setminus I_E^{\epsilon,(2)}$ must be peripheral. For example, if $\gamma = \gamma_{E_L}^{(1)}(k_L,+)$ and $k_L < \min I_{E_L}^{\epsilon,(2)}$ is a lower excess, then it must have $\tau_k = R$, since otherwise it has a nontrivial contribution to the edge coordinates. It follows that such a traveler also has an identifier of lower excess in the next biangle, concluding $\tau_k = R$ for all $k \in \mathbb{Z}$ inductively for both directions. See γ_7 in Figure 40 for an example.

Correspondence between the global pictures $\langle W_1 \rangle$ and $\langle W_2 \rangle$ Let W_1 and W_2 be as in Proposition 6.8. Then by the unbounded fellow-traveler Lemma, we have a bijective correspondence φ between the travelers in $\langle \mathcal{W}_1 \rangle$ and $\langle \mathcal{W}_2 \rangle$. Let us consider the global pictures $\langle W_1 \rangle$ and $\langle W_2 \rangle$, and call each oriented curve in $\langle W_\nu \rangle$ a traveler again. Since the bijection φ preserves the spiraling types of travelers in $\langle \mathcal{W}_\nu \rangle$, it induces a bijection

$$\varphi\colon \{\text{travelers in } \langle W_1 \rangle\} \xrightarrow{\sim} \{\text{travelers in } \langle W_2 \rangle\}. \tag{6-2}$$

Here we make the intersection of each traveler in $\langle W_\nu \rangle$ with each edge of $\widehat{\Delta}$ minimal, by applying the same isotopy for each pair $(\gamma, \varphi(\gamma))$ of travelers. Notice that each traveler in $\langle W_\nu \rangle$ is either a closed loop or a compact arc, and their intersections are finite. Therefore we can proceed by applying Douglas and Sun's argument [6, Section 7.4] for the rest of discussion.

Recall the notion of a *shared route* of two ordered travelers (γ, γ') from [6, Definition 59]. Roughly speaking, it is a maximal interval shared by the routes of two travelers with opposite orientations. The definition is extended for the travelers in $\langle W_\nu \rangle$ in a straightforward way. A shared route is either *crossing* or *noncrossing*. A noncrossing shared route is said to be *left-oriented* if one traveler is always seen on the left from the other traveler. A crossing shared route is said to be *left-oriented* if the same situation occurs near its source-end [6, Definition 61].

By applying the boundary and puncture H-moves if necessary, we may assume that these webs are reduced. Then we see that each shared route has at most one intersection point (see [6, Lemma 60]). Indeed, two intersecting travelers cannot have a common endpoint at a puncture, since such a situation would come from a puncture H-face. Hence the situation regarding the crossing shared routes is exactly the same as in the bounded case. From these observations, together with the bijection (6-2), we get:

Lemma 6.15 (cf [6, Corollary 64]) *For $\nu = 1, 2$, let $P_{\langle W_\nu \rangle}$ denote the set of intersections of travelers in $\langle W_\nu \rangle$. Then we have a bijection*

$$\varphi_{\mathrm{int}}\colon P_{\langle W_1 \rangle} \xrightarrow{\sim} P_{\langle W_2 \rangle}$$

such that the unique intersection point p of a left-oriented shared route of two travelers (γ, γ') in $\langle W_1 \rangle$ is sent to the unique intersection point $\varphi_{\mathrm{int}}(p)$ of the corresponding shared route of $(\varphi(\gamma), \varphi(\gamma'))$ in $\langle W_2 \rangle$.

Proof of Proposition 6.8: a sequence of elementary moves relating W_1 and W_2 As in the previous paragraph, we may assume that W_1 and W_2 are reduced by applying the boundary/puncture H-moves. Moreover by applying the loop parallel-moves and the arc parallel-moves (Lemma 2.4), we may assume that both W_1 and W_2 are *left-oriented* in the sense that for each pair of parallel loop or arc components with opposite orientations, one is always seen on the left from the other. It includes the *closed-left-oriented* condition [6, Definition 62]. Now we are going to see that the intersection points $p \in P_{\langle W_1 \rangle}$ and $\varphi_{\mathrm{int}}(p) \in P_{\langle W_2 \rangle}$ can be adjusted to a common position by a sequence of modified H-moves; see Figure 42. The techniques developed in [6, Section 7.8] can be directly applied to our situation without any essential modification, since the situation around a crossing shared route is exactly the same as in the bounded case, and the sets $P_{\langle W_\nu \rangle}$ are finite. Then we get:

Lemma 6.16 (cf [6, Lemma 66]) *There are sequences of modified H-moves applicable to the webs W_1 and W_2 respectively, after which the bijection φ_{int} satisfies the property that for each intersection point p in $\langle W_1 \rangle$, the two points p and $\varphi_{\mathrm{int}}(p)$ lie in the same shared-route-biangle* [6, Definition 65].

Apply the sequence of modified H-moves to W_1 and W_2 prescribed above. We claim that the two signed webs W_1 and W_2 are now isotopic.

In the same way as in the proof of [6, Lemma 67], we see that the finite sequences of oriented strands on each edge of the split triangulation $\widehat{\Delta}$ are the same for $\langle W_1 \rangle$ and $\langle W_2 \rangle$. We have a correspondence (6-2) that relates the travelers in $\langle W_1 \rangle$ and $\langle W_2 \rangle$, in particular the ends incident to punctures and their signs.

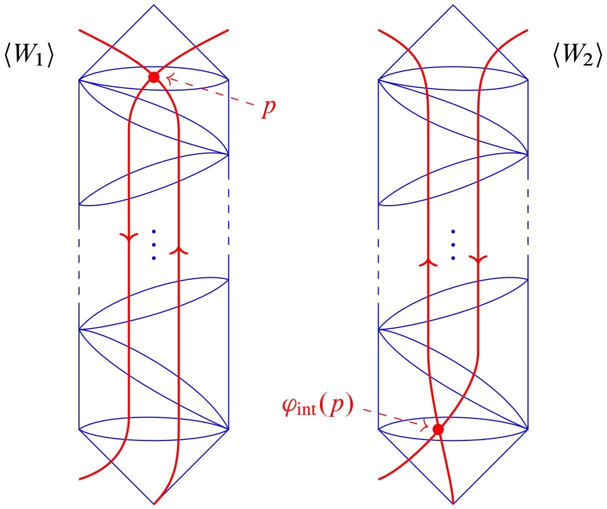


Figure 42: Adjustment of intersection points. Here only the difference from the situation in [6] is that some of the travelers can end at a puncture.

The travelers can intersect with each other inside biangles, whose pattern is uniquely determined by the sequence of oriented strands on the side edges. Thus $\langle W_1 \rangle$ and $\langle W_2 \rangle$ restricts to the same collection of oriented curves (with signed ends at punctures) in each triangle and biangle in $\widehat{\triangle}$. Since we can uniquely recover the honeycombs from these diagrams, we get $W_1 = W_2$ up to isotopy. Thus Proposition 6.8 is proved.

Proof of Theorem 3.19 Let us consider an integral unbounded $\mathfrak{sl}_3$-lamination, which is represented by a signed nonelliptic web W_1. Let $W_2 := \xi_\triangle \circ \mathsf{x}^{\mathrm{uf}}_\triangle(W_1)$ be the signed nonelliptic web obtained from the reconstruction. By Proposition 3.18, we have $\mathsf{x}^{\mathrm{uf}}_\triangle(W_1) = \mathsf{x}^{\mathrm{uf}}_\triangle(W_2)$. Then Proposition 6.8 tells us that W_1 and W_2 determine an equivalent $\mathfrak{sl}_3$-lamination. Combining with the $\mathbb{Q}_{>0}$-equivariance, we get the desired assertion. □

Appendix Cluster varieties associated with the pair $(\mathfrak{sl}_3, \Sigma)$

Here we briefly recall the general theory of cluster varieties [13], and the construction of the seed pattern $\boldsymbol{s}(\mathfrak{sl}_3, \Sigma)$ that encodes the cluster structures of the spaces of $\mathfrak{sl}_3$-laminations in consideration.

A.1 Seeds, mutations and the labeled exchange graph

Fix a finite set $I = \{1, \dots, N\}$ of indices, and let $\mathcal{F}_A$ and $\mathcal{F}_X$ be fields both isomorphic to the field $\mathbb{Q}(z_1, \dots, z_N)$ of rational functions on N variables. We also fix a subset $I_{\mathrm{uf}} \subset I$ ("unfrozen") and let $I_{\mathrm{f}} := I \setminus I_{\mathrm{uf}}$ ("frozen"). A (*labeled*, *skew-symmetric*) *seed* in $(\mathcal{F}_A, \mathcal{F}_X)$ is a triple $(\varepsilon, \boldsymbol{A}, \boldsymbol{X})$, where

- $\varepsilon = (\varepsilon_{ij})_{i,j \in I}$ is a skew-symmetric matrix (*exchange matrix*) with values in $\frac{1}{2}\mathbb{Z}$ such that $\varepsilon_{ij} \in \mathbb{Z}$ unless $(i, j) \in I_{\mathrm{f}} \times I_{\mathrm{f}}$;

- $\boldsymbol{A} = (A_i)_{i \in I}$ and $\boldsymbol{X} = (X_i)_{i \in I}$ are tuples of algebraically independent elements (*cluster $\mathcal{A}$- and $\mathcal{X}$-variables*) in $\mathcal{F}_A$ and $\mathcal{F}_X$, respectively.

The exchange matrix ε can be encoded in a quiver with vertices parametrized by the set I and $|\varepsilon_{ij}|$ arrows from i to j (resp. j to i) if $\varepsilon_{ij} > 0$ (resp. $\varepsilon_{ji} > 0$). In figures, we draw n dashed arrows from i to j if $\varepsilon_{ij} = n/2$ for $n \in \mathbb{Z}$, where a pair of dashed arrows is replaced with a solid arrow.

For an unfrozen index $k \in I_{\mathrm{uf}}$, the *mutation* directed to k produces a new seed $(\varepsilon', \boldsymbol{A}', \boldsymbol{X}') = \mu_k(\varepsilon, \boldsymbol{A}, \boldsymbol{X})$ according to an explicit formula [17]. See, for instance, [23, (2.1), (2.3) and (2.4)] for a formula which fits in with our convention. A permutation $\sigma \in \mathfrak{S}_{I_{\mathrm{uf}}} \times \mathfrak{S}_{I_{\mathrm{f}}}$ induces a transformation $\sigma : (\varepsilon, \boldsymbol{A}, \boldsymbol{X}) \to (\varepsilon', \boldsymbol{A}', \boldsymbol{X}')$ by the rule

$$\text{(A-1)} \qquad \varepsilon'_{ij} := \varepsilon_{\sigma^{-1}(i),\sigma^{-1}(j)}, \quad A'_i := A_{\sigma^{-1}(i)}, \quad X'_i := X_{\sigma^{-1}(i)}.$$

We say that two seeds in $(\mathcal{F}_A, \mathcal{F}_X)$ are *mutation-equivalent* if they are transformed to each other by a finite sequence of mutations and permutations. The equivalence class is usually called a *mutation class*.

The relations among the seeds in a given mutation class s can be encoded in the (*labeled*) *exchange graph* $\mathbb{E}\mathrm{xch}_{\mathsf{s}}$. It is a graph with vertices v corresponding to the seeds $\mathsf{s}^{(v)}$ in s, together with labeled edges of the following two types:

- edges of the form $v \overset{k}{\text{---}} v'$ whenever the seeds $\mathsf{s}^{(v)}$ and $\mathsf{s}^{(v')}$ are related by the mutation μ_k for $k \in I_{\mathrm{uf}}$;
- edges of the form $v \overset{\sigma}{\text{---}} v'$ whenever the seeds $\mathsf{s}^{(v)}$ and $\mathsf{s}^{(v')}$ are related by the transposition $\sigma = (j\ k)$ for $(j,k) \in I_{\mathrm{uf}} \times I_{\mathrm{uf}}$ or $I_{\mathrm{f}} \times I_{\mathrm{f}}$.

When no confusion can occur, we simply denote a vertex of the labeled exchange graph by $v \in \mathbb{E}\mathrm{xch}_{\mathsf{s}}$ instead of $v \in V(\mathbb{E}\mathrm{xch}_{\mathsf{s}})$. When we write $\mathsf{s}^{(v)} = (\varepsilon^{(v)}, \boldsymbol{A}^{(v)}, \boldsymbol{X}^{(v)})$, it is known that $(\varepsilon^{(v)}, \boldsymbol{A}^{(v)}) = (\varepsilon^{(v')}, \boldsymbol{A}^{(v')})$ if and only if $(\varepsilon^{(v)}, \boldsymbol{X}^{(v)}) = (\varepsilon^{(v')}, \boldsymbol{X}^{(v')})$ for two vertices v and v' (the *synchronicity phenomenon* [38]). We call $(\varepsilon^{(v)}, \boldsymbol{A}^{(v)})$ and $(\varepsilon^{(v)}, \boldsymbol{X}^{(v)})$ an *$\mathcal{A}$-seed* and an *$\mathcal{X}$-seed*, respectively. We also remark that the labeled exchange graph depends only on the mutation class of the underlying exchange matrices. Indeed, it is unchanged if we transform the cluster variables simultaneously by an automorphism of the ambient field.

Remark A.1 In geometric applications, $\mathcal{A}$- and $\mathcal{X}$-seeds are constructed in the field of rational functions on a space of interest. For $\mathcal{Z} \in \{\mathcal{A}, \mathcal{X}\}$, a *cluster $\mathcal{Z}$-atlas* on a variety (scheme, stack) V is a collection of $\mathcal{Z}$-seeds in the field $\mathcal{K}(V)$ of rational functions which are mutation-equivalent to each other. A cluster atlas can be uniquely extended to a *cluster $\mathcal{Z}$-structure*, which is a maximal collection of $\mathcal{Z}$-seeds in $\mathcal{K}(V)$, thus forming a mutation class s. See Remark A.3 below.

A.2 Cluster varieties

The cluster varieties associated with a mutation class s are constructed by patching algebraic tori parametrized by the vertices of the labeled exchange graph.

Notation A.2 A multiplicative algebraic group is denoted by $\mathbb{G}_m = \operatorname{Spec} \mathbb{Z}[u, u^{-1}]$. For a lattice Λ (ie a free abelian group of finite rank), let $\mathbb{T}_\Lambda := \operatorname{Hom}(\Lambda, \mathbb{G}_m)$ denote the associated algebraic torus. For a (split) algebraic torus $T \cong (\mathbb{G}_m)^N$, let

$$X^*(T) := \operatorname{Hom}(T, \mathbb{G}_m) \quad \text{and} \quad X_*(T) := \operatorname{Hom}(\mathbb{G}_m, T)$$

denote the lattices of characters and cocharacters, respectively. These lattices are dual to each other by via the canonical pairing $X_*(T) \otimes X^*(T) \to \operatorname{Hom}(\mathbb{G}_m, \mathbb{G}_m) \cong \mathbb{Z}$. The contravariant functors $\mathbb{T}_\bullet \colon \Lambda \mapsto \mathbb{T}_\Lambda$ and $X^* \colon T \mapsto X^*(T)$ are inverses to each other: $\Lambda = X^*(\mathbb{T}_\Lambda)$, $T = \mathbb{T}_{X^*(T)}$. A vector $\lambda \in \Lambda$ gives rise to a character $\chi_\lambda \colon \mathbb{T}_\Lambda \to \mathbb{G}_m$.

For $v \in \mathbb{E}\mathrm{xch}_{\mathfrak{s}}$, consider a lattice $N^{(v)} = \bigoplus_{i \in I} \mathbb{Z} e_i^{(v)}$ with a fixed basis and its dual $M^{(v)} = \bigoplus_{i \in I} \mathbb{Z} f_i^{(v)}$. Let $\mathcal{X}_{(v)} := \mathbb{T}_{N^{(v)}}$ and $\mathcal{A}_{(v)} := \mathbb{T}_{M^{(v)}}$ denote the associated algebraic tori of dimension $|I|$. The characters $X_i^{(v)} := \chi_{e_i^{(v)}} \colon \mathcal{X}_{(v)} \to \mathbb{G}_m$ and $A_i^{(v)} := \chi_{f_i^{(v)}} \colon \mathcal{A}_{(v)} \to \mathbb{G}_m$ are called the *cluster coordinates*. The exchange matrix $\varepsilon^{(v)}$ defines a $\frac{1}{2}\mathbb{Z}$-valued bilinear form on $N^{(v)}$ by $(e_i^{(v)}, e_j^{(v)}) := \varepsilon_{ij}^{(v)}$, which induces Poisson and K_2-structures on $\mathcal{X}_{(v)}$ and $\mathcal{A}_{(v)}$, respectively. The mutation rule turns into birational maps $\mu_k^x \colon \mathcal{X}_{(v)} \to \mathcal{X}_{(v')}$ and $\mu_k^a \colon \mathcal{A}_{(v)} \to \mathcal{A}_{(v')}$, called the *cluster transformations* [13, (13) and (14)]. Then the *cluster $\mathcal{X}$- and $\mathcal{A}$-varieties* are the schemes defined as

$$\mathcal{X}_{\mathfrak{s}} := \bigcup_{v \in \mathbb{E}\mathrm{xch}_{\mathfrak{s}}} \mathcal{X}_{(v)}, \quad \mathcal{A}_{\mathfrak{s}} := \bigcup_{v \in \mathbb{E}\mathrm{xch}_{\mathfrak{s}}} \mathcal{A}_{(v)}.$$

Here for $(z, \mathbb{Z}) \in \{(a, \mathcal{A}), (x, \mathcal{X})\}$, (open subsets of) tori $\mathcal{Z}_{(v)}$ and $\mathcal{Z}_{(v')}$ are identified via the cluster transformation μ_k^z if there is an edge of the form $v \overset{k}{-\!\!-} v'$, or via the coordinate permutation (A-1) if there is an edge of the form $v \overset{\sigma}{-\!\!-} v'$. As a slight variant, let $\mathcal{X}_{(v)}^{\mathrm{uf}} := \mathbb{T}_{N_{\mathrm{uf}}^{(v)}}$, and $\mathcal{X}_{\mathfrak{s}}^{\mathrm{uf}} := \bigcup_{v \in \mathbb{E}\mathrm{xch}_{\mathfrak{s}}} \mathcal{X}_{(v)}^{\mathrm{uf}}$ the cluster $\mathcal{X}$-variety without frozen coordinates. Since the cluster transformation of unfrozen $\mathcal{X}$-coordinates does not refer the frozen ones, we have a natural projection $\mathcal{X}_{\mathfrak{s}} \to \mathcal{X}_{\mathfrak{s}}^{\mathrm{uf}}$. We remark that the cluster varieties are constructed only from the mutation class of the underlying exchange matrices.

For $(Z, \mathcal{Z}) \in \{(A, \mathcal{A}), (X, \mathcal{X})\}$, each pair $(\varepsilon^{(v)}, (Z_i^{(v)})_{i \in I})$ of the exchange matrix and the cluster $\mathcal{Z}$-coordinates defines a $\mathcal{Z}$-seed in the field $\mathcal{F}_Z := \mathcal{K}(\mathcal{Z}_{\mathfrak{s}})$ of rational functions in the sense of the previous section. The rings $\mathcal{O}(\mathcal{A}_{\mathfrak{s}}) \subset \mathcal{F}_A$ and $\mathcal{O}(\mathcal{X}_{\mathfrak{s}}) \subset \mathcal{F}_X$ of regular functions are called the *upper cluster algebra* and the *cluster Poisson algebra*, respectively. The *cluster algebra* [16] is the subring $\mathscr{A}_{\mathfrak{s}} \subset \mathcal{O}(\mathcal{A}_{\mathfrak{s}})$ generated by all the cluster coordinates $A_i^{(v)}$, $i \in I$, $v \in \mathbb{E}\mathrm{xch}_{\mathfrak{s}}$.

Ensemble maps and their extensions The cluster varieties $\mathcal{X}_{\mathfrak{s}}$ and $\mathcal{A}_{\mathfrak{s}}$ are coupled as a *cluster ensemble*. For $v \in \mathbb{E}\mathrm{xch}_{\mathfrak{s}}$, let $N_{\mathrm{uf}}^{(v)} \subset N^{(v)}$ denote the sublattice spanned by $e_i^{(v)}$ for $i \in I_{\mathrm{uf}}$. Then by the assumption on the exchange matrix, we have the linear map

$$p_{(v)}^* \colon N_{\mathrm{uf}}^{(v)} \to M^{(v)}, \quad e_i^{(v)} \mapsto \sum_{j \in I} \varepsilon_{ij}^{(v)} f_j^{(v)}. \tag{A-2}$$

Moreover, it can be verified that the maps between tori induced by (A-2) commute with cluster transformations, and combine to give a morphism $p \colon \mathcal{A}_{\mathfrak{s}} \to \mathcal{X}_{\mathfrak{s}}^{\mathrm{uf}}$. We call this map the *ensemble map*, and the triple $(\mathcal{A}_{\mathsf{s}}, \mathcal{X}_{\mathsf{s}}, p)$ the *cluster ensemble* associated with s. If we pick up a suitable extension $\tilde{p}^*_{(v)} \colon N^{(v)} \to M^{(v)}$ of the map (A-2) (see [22, (A.2)] for the required condition), then it still commutes with cluster transformations and hence we get an *extended ensemble map* $\tilde{p} \colon \mathcal{A}_{\mathsf{s}} \to \mathcal{X}_{\mathsf{s}}$. It is shown in [21, Section 13.3] that such a choice exactly corresponds to a choice of compatibility pairs [3] defining a quantum cluster algebra.

Tropicalizations The positive structures on the cluster varieties allow us to consider their semifield-valued points. For $\mathbb{A} = \mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$, let $\mathbb{A}^T := (\mathbb{A}, \max, +)$ denote the corresponding *tropical semifield* (or the *max-plus semifield*). For an algebraic torus H, let $H(\mathbb{A}^T) := X_*(H) \otimes_{\mathbb{Z}} (\mathbb{A}, +)$. A positive rational map $f \colon H \to H'$ between algebraic tori naturally induces a piecewise-linear (PL for short) map $f^T \colon H(\mathbb{A}^T) \to H'(\mathbb{A}^T)$. We call f^T the *tropicalized map*. In particular we have the tropicalized cluster transformations $\mu_k^T \colon \mathcal{Z}_{(v)}(\mathbb{A}^T) \to \mathcal{Z}_{(v')}(\mathbb{A}^T)$ for $(\mathsf{z}, \mathcal{Z}) \in \{(\mathsf{a}, \mathcal{A}), (\mathsf{x}, \mathcal{X})\}$, explicitly given as

$$(\mu_k^T)^* \mathsf{x}_i^{(v')} = \begin{cases} -\mathsf{x}_k^{(v)} & \text{if } i = k, \\ \mathsf{x}_i^{(v)} - \varepsilon_{ik}^{(v)}[-\operatorname{sgn}(\varepsilon_{ik}^{(v)})\mathsf{x}_k^{(v)}]_+ & \text{if } i \neq k, \end{cases} \tag{A-3}$$

$$(\mu_k^T)^* \mathsf{a}_i^{(v')} = \begin{cases} -\mathsf{a}_k^{(v)} + \max\{\sum_{j \in I} [\varepsilon_{kj}^{(v)}]_+ \mathsf{a}_j^{(v)}, \sum_{j \in I} [-\varepsilon_{kj}^{(v)}]_+ \mathsf{a}_j^{(v)}\} & \text{if } i = k, \\ \mathsf{a}_i^{(v)} & \text{if } i \neq k. \end{cases} \tag{A-4}$$

Here $\mathsf{x}_i^{(v)}$ and $\mathsf{a}_i^{(v)}$ are the coordinate functions induced by the basis vectors $e_i^{(v)}$ and $f_i^{(v)}$ respectively, and $[u]_+ := \max\{0, u\}$ for $u \in \mathbb{A}$. We can use them to define the *tropical cluster varieties*

$$\mathcal{X}_{\mathsf{s}}(\mathbb{A}^T) := \bigcup_{v \in \mathbb{E}\mathrm{xch}_{\mathsf{s}}} \mathcal{X}_{(v)}(\mathbb{A}^T), \quad \mathcal{A}_{\mathsf{s}}(\mathbb{A}^T) := \bigcup_{v \in \mathbb{E}\mathrm{xch}_{\mathsf{s}}} \mathcal{A}_{(v)}(\mathbb{A}^T),$$

which are naturally equipped with canonical PL structures. Since the PL maps are equivariant for the scaling action of $\mathbb{A}_{>0}$, the tropical cluster varieties inherit this $\mathbb{A}_{>0}$-action. We also consider the tropical $\mathcal{X}$-varieties $\mathcal{X}_{\mathsf{s}}^{\mathrm{uf}}(\mathbb{A}^T)$ without frozen coordinates. In the body of this paper, the main objects of study are the spaces $\mathcal{X}_{\mathsf{s}}^{\mathrm{uf}}(\mathbb{Q}^T)$ and $\mathcal{X}_{\mathsf{s}}(\mathbb{Q}^T)$ associated with a particular mutation class s.

Cluster modular group The cluster ensemble is naturally equipped with a discrete symmetry group. Let $\mathrm{Mat}_{\mathsf{s}}$ denote the mutation class of exchange matrices underlying the mutation class s. Then we have a map

$$\varepsilon^\bullet \colon V(\mathbb{E}\mathrm{xch}_{\mathsf{s}}) \to \mathrm{Mat}_{\mathsf{s}}, \quad v \mapsto \varepsilon^{(v)}.$$

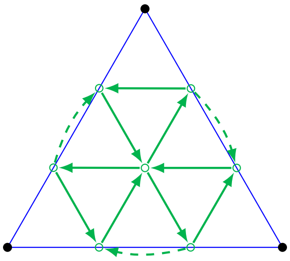
Then the *cluster modular group* $\Gamma_{\mathsf{s}} \subset \mathrm{Aut}(\mathbb{E}\mathrm{xch}_{\mathsf{s}})$ consists of graph automorphism ϕ which preserves the fibers of the map $\varepsilon^\bullet$ and the labels on the edges. An element of the cluster modular group is called a *mutation loop*. The cluster modular group acts on the cluster varieties $\mathcal{A}_{\mathsf{s}}$ and $\mathcal{X}_{\mathsf{s}}$ so that $\phi^* Z_i^{(v)} = Z_i^{(\phi^{-1}(v))}$ for all $\phi \in \Gamma_{\boldsymbol{s}}$, $v \in \mathbb{E}\mathrm{xch}_{\boldsymbol{s}}$ and $i \in I$, where $(Z, \mathcal{Z}) \in \{(A, \mathcal{A}), (X, \mathcal{X})\}$. These actions commute with the ensemble map.

Since the actions are by positive rational maps, they induce actions of Γ_{s} on $\mathcal{A}_{\mathsf{s}}(\mathbb{A}^T)$ and $\mathcal{X}_{\mathsf{s}}(\mathbb{A}^T)$ by PL automorphisms, which commute with the (extended) ensemble map. Moreover, these actions commute with the rescaling action of $\mathbb{A}_{>0}$.

A.3 The cluster ensemble associated with the pair $(\mathfrak{sl}_3, \Sigma)$

Here we quickly recall the cluster structures on the moduli spaces $\mathcal{A}_{SL_3,\Sigma}$, $\mathcal{X}_{PGL_3,\Sigma}$ and $P_{PGL_3,\Sigma}$ constructed in [10; 21]. We are going to recall the *Fock–Goncharov atlas* associated with ideal triangulations of Σ and their mutation-equivalences, since it is typical difficult to describe the entire cluster structure.

Let $\triangle$ be an ideal triangulation of Σ. Then we construct a quiver $Q_\triangle$ with the vertex set $I(\triangle)$ by drawing the quiver



on each triangle, and glue them by the *amalgamation* construction [9]. In our case, this just means that we glue the quivers on adjacent triangles by identifying the two vertices on the shared edge and eliminate the pair of opposite dashed arrows. The vertices on the boundary intervals of Σ are declared to be frozen, forming the subset $I_f(\triangle) \subset I(\triangle)$ as in Section 2.1. Let $\varepsilon^\triangle = (\varepsilon^\triangle_{ij})_{i,j \in I(\triangle)}$ be the corresponding exchange matrix.

These quivers $Q^\triangle$ (or the exchange matrices $\varepsilon^\triangle$) associated with ideal triangulations of Σ are mutation-equivalent to each other. Indeed, the quivers $Q^\triangle$ and $Q^{\triangle'}$ associated with two triangulations $\triangle$ and $\triangle'$ connected by a single flip $f_E \colon \triangle \to \triangle'$ are transformed to each other via one of the mutation sequences shown in Figure 43. Then the assertion follows from the classical fact that any two ideal triangulations of the same marked surface can be transformed to each other by a finite sequence of flips.

Remark A.3 For each ideal triangulation $\triangle$, we can associate an $\mathcal{A}$-seed $(\varepsilon^\triangle, \boldsymbol{A}^\triangle)$ (resp. $\mathcal{X}$-seed $(\varepsilon^\triangle, \boldsymbol{X}^\triangle)$) in the field of rational functions on the moduli space $\mathcal{A}_{SL_3,\Sigma}$ (resp. $\mathcal{P}_{PGL_3,\Sigma}$). Forgetting the frozen part in the latter, we get an $\mathcal{X}$-seed for the moduli space $\mathcal{X}_{PGL_3,\Sigma}$. See [10, Section 9] or [21, Section 3] for construction. These birational coordinate systems define cluster atlases on these moduli spaces in the sense of Remark A.1.

Then there exists a unique mutation class $\mathsf{s}(\mathfrak{sl}_3, \Sigma)$ containing the seeds associated with any ideal triangulations $\triangle$. More precisely, a *labeled* $\mathfrak{sl}_3$*-triangulation* $(\triangle, \ell)$, namely an ideal triangulation $\triangle$ together with a bijection $\ell \colon I(\triangle) \to \{1, \dots, N\}$, give rise to vertices of the labeled exchange graph $\mathbb{E}\mathrm{xch}_{\mathsf{s}(\mathfrak{sl}_3,\Sigma)}$. Figure 43 describes a subgraph containing $(\triangle, \ell)$ and $(\triangle', \ell')$, where the labels ℓ and ℓ' are consistently chosen. Let us simply denote the objects related to $\mathsf{s}(\mathfrak{sl}_3, \Sigma)$ by

$$\mathcal{A}_{\mathfrak{sl}_3,\Sigma} := \mathcal{A}_{\mathsf{s}(\mathfrak{sl}_3,\Sigma)}, \quad \mathcal{X}_{\mathfrak{sl}_3,\Sigma} := \mathcal{X}_{\mathsf{s}(\mathfrak{sl}_3,\Sigma)}, \quad \mathbb{E}\mathrm{xch}_{\mathfrak{sl}_3,\Sigma} := \mathbb{E}\mathrm{xch}_{\mathsf{s}(\mathfrak{sl}_3,\Sigma)}, \quad \Gamma_{\mathfrak{sl}_3,\Sigma} := \Gamma_{\mathsf{s}(\mathfrak{sl}_3,\Sigma)},$$

and so on.

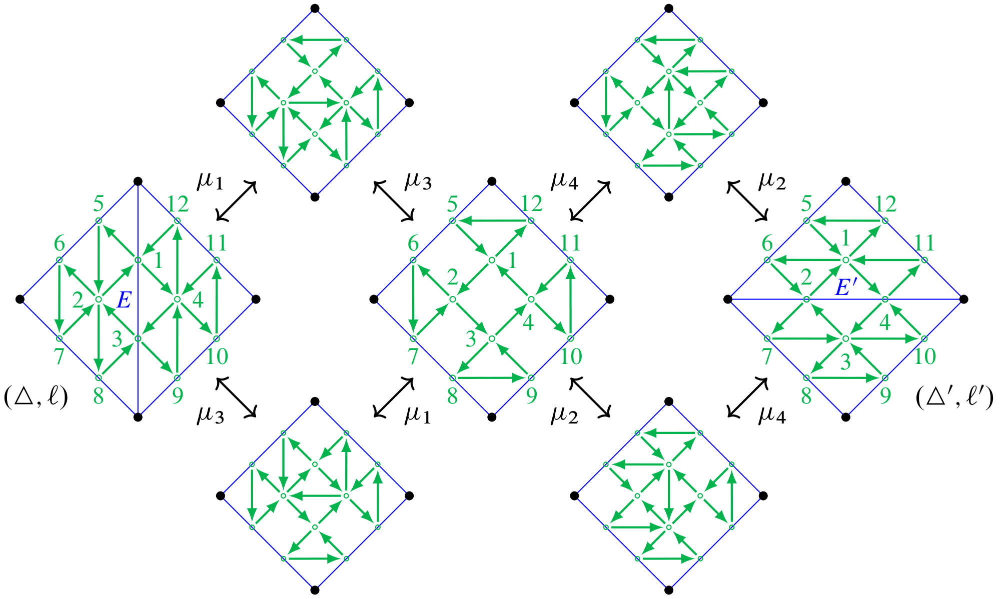


Figure 43: Some of the sequences of mutations that realize the flip $f_E\colon \triangle\to\triangle'$. Here we partially fix labelings ℓ and ℓ' of vertices in $I(\triangle)$ and $I(\triangle')$, respectively.

The following can be verified from (A-3) by a direct computation:

Lemma A.4 *For two labeled $\mathfrak{sl}_3$-triangulations $v=(\triangle,\ell)$, $v'=(\triangle',\ell')\in\mathbb{E}\mathrm{xch}_{\mathfrak{sl}_3,\Sigma}$ as in Figure 43, the (max-plus) tropical coordinates $\mathsf{x}_i:=\mathsf{x}_i^{(v)}$ and $\mathsf{x}'_i:=\mathsf{x}_i^{(v')}$ for $i\in\{1,\dots,12\}$ are related as follows:*

$$\begin{aligned}
\mathsf{x}'_1&=\mathsf{x}_2+[\mathsf{x}_3,\mathsf{x}_4,\mathsf{x}_1]_+-[\mathsf{x}_1,\mathsf{x}_2,\mathsf{x}_3]_+, & \mathsf{x}'_2&=-\mathsf{x}_1-\mathsf{x}_2+[\mathsf{x}_1]_+-[\mathsf{x}_3]_+,\\
\mathsf{x}'_3&=\mathsf{x}_4+[\mathsf{x}_1,\mathsf{x}_2,\mathsf{x}_3]_+-[\mathsf{x}_3,\mathsf{x}_4,\mathsf{x}_1]_+, & \mathsf{x}'_4&=-\mathsf{x}_3-\mathsf{x}_4+[\mathsf{x}_3]_+-[\mathsf{x}_1]_+,\\
\mathsf{x}'_5&=\mathsf{x}_5+[\mathsf{x}_1]_+, & \mathsf{x}'_6&=\mathsf{x}_6+[\mathsf{x}_1,\mathsf{x}_2,\mathsf{x}_3]_+-[\mathsf{x}_1]_+,\\
\mathsf{x}'_7&=\mathsf{x}_7+\mathsf{x}_1+\mathsf{x}_2+[\mathsf{x}_3]_+-[\mathsf{x}_1,\mathsf{x}_2,\mathsf{x}_3]_+, & \mathsf{x}'_8&=\mathsf{x}_8-[-\mathsf{x}_3]_+,\\
\mathsf{x}'_9&=\mathsf{x}_9+[\mathsf{x}_3]_+, & \mathsf{x}'_{10}&=\mathsf{x}_{10}+[\mathsf{x}_3,\mathsf{x}_4,\mathsf{x}_1]_+-[\mathsf{x}_3]_+,\\
\mathsf{x}'_{11}&=\mathsf{x}_{11}+\mathsf{x}_3+\mathsf{x}_4+[\mathsf{x}_1]_+-[\mathsf{x}_3,\mathsf{x}_4,\mathsf{x}_1]_+, & \mathsf{x}'_{12}&=\mathsf{x}_{12}-[-\mathsf{x}_1]_+.
\end{aligned}$$

Here $[x]_+:=\max\{0,x\}$ and $[x,y,z]_+:=\max\{0,x,x+y,x+y+z\}$.

Goncharov–Shen extension of the ensemble map Following [21], we choose the following extension of the ensemble map. Let

$$C(\mathfrak{sl}_3)=(C_{st})_{s,t\in\{1,2\}}=\begin{pmatrix}2&-1\\-1&2\end{pmatrix}$$

denote the Cartan matrix of the Lie algebra $\mathfrak{sl}_3$. For an ideal triangulation $\triangle$, let $\tilde{\varepsilon}^{\triangle}=(\tilde{\varepsilon}^{\triangle}_{ij})_{i,j\in I(\triangle)}$ be the matrix given by $\tilde{\varepsilon}^{\triangle}_{ij}:=\varepsilon^{\triangle}_{ij}+m_{ij}$, where

$$m_{ij}:=\begin{cases}-\frac{1}{2}C_{st} & \text{if } i=i^s(E) \text{ and } j=i^t(E) \text{ lie on a common boundary interval } E\in\mathbb{B},\\ 0 & \text{otherwise.}\end{cases}\tag{A-5}$$

Then we define $\tilde{p}^*_{\triangle}\colon N^{\triangle} \to M^{\triangle}$ by $e_i^{\triangle} \mapsto \sum_{i,j\in I(\triangle)} \tilde{\varepsilon}_{ij}^{\triangle} f_j^{\triangle}$ inducing a morphism

$$\tilde{p}_{\mathrm{GS}}\colon \mathcal{A}_{\mathfrak{sl}_3,\Sigma} \to \mathcal{X}_{\mathfrak{sl}_3,\Sigma}, \tag{A-6}$$

which we call the *Goncharov–Shen extension of the ensemble map*. This choice naturally comes from the geometry of the moduli spaces of local systems on Σ, so $\tilde{p}_{\mathrm{GS}}$ agrees with the map $p\colon \mathcal{A}^{\times}_{\mathrm{SL}_3,\Sigma} \to \mathcal{P}_{\mathrm{PGL}_3,\Sigma}$ [21, Proposition 9.4].

Cluster modular group Although the entire structure of the cluster modular group $\Gamma_{\mathfrak{sl}_3,\Sigma}$ is yet unknown, it is known to include the subgroup $(\mathrm{MC}(\Sigma) \times \mathrm{Out}(\mathrm{SL}_3)) \ltimes W(\mathfrak{sl}_3)^{\mathbb{M}_\circ} \subset \Gamma_{\mathfrak{sl}_3,\Sigma}$ [20]. Here $\mathrm{MC}(\Sigma)$ denotes the mapping class group of the marked surface Σ, $\mathrm{Out}(\mathrm{SL}_3)) = \mathrm{Aut}(\mathrm{SL}_3)/\mathrm{Inn}(\mathrm{SL}_3)$ is the outer automorphism group of SL_3, and $W(\mathfrak{sl}_3)$ is the Weyl group of the Lie algebra $\mathfrak{sl}_3$. The group $\mathrm{Out}(\mathrm{SL}_3)$ has order 2, and generated by the *Dynkin involution* $*\colon G \to G$, $g \mapsto (g^{-1})^{\mathsf{T}}$. For each element ϕ in this subgroup, let us call the induced PL action $\phi\colon \mathcal{Z}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T) \to \mathcal{Z}_{\mathfrak{sl}_3,\Sigma}(\mathbb{Q}^T)$ the *cluster action*, in comparison to the geometric action we introduce in the body of this paper in terms of signed $\mathfrak{sl}_3$-webs.

References

[1] **D G L Allegretti**, *Quantization of canonical bases and the quantum symplectic double*, Manuscripta Math. 167 (2022) 613–651 MR Zbl

[2] **D G L Allegretti**, **H K Kim**, *A duality map for quantum cluster varieties from surfaces*, Adv. Math. 306 (2017) 1164–1208 MR Zbl

[3] **A Berenstein**, **A Zelevinsky**, *Quantum cluster algebras*, Adv. Math. 195 (2005) 405–455 MR Zbl

[4] **S Y Cho**, **H Kim**, **H K Kim**, **D Oh**, *Laurent positivity of quantized canonical bases for quantum cluster varieties from surfaces*, Comm. Math. Phys. 373 (2020) 655–705 MR Zbl

[5] **B Davison**, **T Mandel**, *Strong positivity for quantum theta bases of quantum cluster algebras*, Invent. Math. 226 (2021) 725–843 MR Zbl

[6] **D C Douglas**, **Z Sun**, *Tropical Fock–Goncharov coordinates for* SL_3*–webs on surfaces, I: Construction*, Forum Math. Sigma 12 (2024) art. id. e5 MR Zbl

[7] **D C Douglas**, **Z Sun**, *Tropical Fock–Goncharov coordinates for* SL_3*-webs on surfaces, II: Naturality*, Algebr. Comb. 8 (2025) 101–156 MR Zbl

[8] **V V Fock**, **L O Chekhov**, *Quantum Teichmüller spaces*, Teoret. Mat. Fiz. 120 (1999) 511–528 MR Zbl In Russian; translated in Theoret. Math. Phys. 120 (1999) 1245–1259

[9] **V V Fock**, **A B Goncharov**, *Cluster* $\mathscr{X}$*-varieties, amalgamation, and Poisson–Lie groups*, from "Algebraic geometry and number theory", Progr. Math. 253, Birkhäuser, Boston, MA (2006) 27–68 MR Zbl

[10] **V Fock**, **A Goncharov**, *Moduli spaces of local systems and higher Teichmüller theory*, Publ. Math. Inst. Hautes Études Sci. 103 (2006) 1–211 MR Zbl

[11] **V V Fock**, **A B Goncharov**, *Dual Teichmüller and lamination spaces*, from "Handbook of Teichmüller theory, I", IRMA Lect. Math. Theor. Phys. 11, Eur. Math. Soc., Zürich (2007) 647–684 MR Zbl

[12] **V V Fock**, **A B Goncharov**, *Moduli spaces of convex projective structures on surfaces*, Adv. Math. 208 (2007) 249–273 MR Zbl

[13] **V V Fock**, **A B Goncharov**, *Cluster ensembles, quantization and the dilogarithm*, Ann. Sci. École Norm. Sup. 42 (2009) 865–930 MR Zbl

[14] **V V Fock**, **A B Goncharov**, *Cluster Poisson varieties at infinity*, Selecta Math. 22 (2016) 2569–2589 MR Zbl

[15] **S Fomin**, **M Shapiro**, **D Thurston**, *Cluster algebras and triangulated surfaces, I: Cluster complexes*, Acta Math. 201 (2008) 83–146 MR Zbl

[16] **S Fomin**, **A Zelevinsky**, *Cluster algebras, I: Foundations*, J. Amer. Math. Soc. 15 (2002) 497–529 MR Zbl

[17] **S Fomin**, **A Zelevinsky**, *Cluster algebras, IV: Coefficients*, Compos. Math. 143 (2007) 112–164 MR Zbl

[18] **B Fontaine**, **J Kamnitzer**, **G Kuperberg**, *Buildings, spiders, and geometric Satake*, Compos. Math. 149 (2013) 1871–1912 MR Zbl

[19] **C Frohman**, **A S Sikora**, *$SU(3)$-skein algebras and webs on surfaces*, Math. Z. 300 (2022) 33–56 MR Zbl

[20] **A Goncharov**, **L Shen**, *Donaldson–Thomas transformations of moduli spaces of G-local systems*, Adv. Math. 327 (2018) 225–348 MR Zbl

[21] **A Goncharov**, **L Shen**, *Quantum geometry of moduli spaces of local systems and representation theory*, preprint (2019) arXiv 1904.10491

[22] **M Gross**, **P Hacking**, **S Keel**, **M Kontsevich**, *Canonical bases for cluster algebras*, J. Amer. Math. Soc. 31 (2018) 497–608 MR Zbl

[23] **R Inoue**, **T Ishibashi**, **H Oya**, *Cluster realizations of Weyl groups and higher Teichmüller theory*, Selecta Math. 27 (2021) art. id. 37 MR Zbl

[24] **T Ishibashi**, *On a Nielsen–Thurston classification theory for cluster modular groups*, Ann. Inst. Fourier (Grenoble) 69 (2019) 515–560 MR Zbl

[25] **T Ishibashi**, **S Kano**, *Sign stability of mapping classes on marked surfaces*, *I*: *Empty boundary case*, preprint (2020) arXiv 2010.05214

[26] **T Ishibashi**, **S Kano**, *Sign stability of mapping classes on marked surfaces*, *II*: *General case via reductions*, preprint (2020) arXiv 2011.14320

[27] **T Ishibashi**, **S Kano**, *Algebraic entropy of sign-stable mutation loops*, Geom. Dedicata 214 (2021) 79–118 MR Zbl

[28] **T Ishibashi**, **S Kano**, *Unbounded $\mathfrak{sl}_3$-laminations around punctures*, preprint (2024) arXiv 2404.18236

[29] **T Ishibashi**, **H Oya**, **L Shen**, *$\mathscr{A} = \mathfrak{U}$ for cluster algebras from moduli spaces of G–local systems*, Adv. Math. 431 (2023) art. id. 109256 MR Zbl

[30] **T Ishibashi**, **W Yuasa**, *Skein and cluster algebras of unpunctured surfaces for $\mathfrak{sl}_3$*, Math. Z. 303 (2023) art. id. 72 MR Zbl

[31] **S Kano**, *Train track combinatorics and cluster algebras*, preprint (2023) arXiv 2303.03190

[32] **H K Kim**, SL_3*-laminations as bases for* PGL_3 *cluster varieties for surfaces* (2020) arXiv 2011.14765 To appear in Mem. Amer. Math. Soc.

[33] **G Kuperberg**, *Spiders for rank* 2 *Lie algebras*, Comm. Math. Phys. 180 (1996) 109–151 MR Zbl

[34] **I Le**, *Higher laminations and affine buildings*, Geom. Topol. 20 (2016) 1673–1735 MR Zbl

[35] **I Le**, *Cluster structures on higher Teichmüller spaces for classical groups*, Forum Math. Sigma 7 (2019) art. id. e13 MR Zbl

[36] **T T Q Lê**, **T Yu**, *Quantum traces and embeddings of stated skein algebras into quantum tori*, Selecta Math. 28 (2022) art. id. 66 MR Zbl

[37] **T Mandel**, **F Qin**, *Bracelets bases are theta bases*, preprint (2023) arXiv 2301.11101

[38] **T Nakanishi**, *Synchronicity phenomenon in cluster patterns*, J. Lond. Math. Soc. 103 (2021) 1120–1152 MR Zbl

[39] **A Papadopoulos**, **R C Penner**, *The Weil–Petersson symplectic structure at Thurston's boundary*, Trans. Amer. Math. Soc. 335 (1993) 891–904 MR Zbl

[40] **R C Penner**, *Decorated Teichmüller theory*, Eur. Math. Soc., Zürich (2012) MR Zbl

[41] **F Qin**, *Cluster algebras and their bases*, from "Representations of algebras and related structures", Eur. Math. Soc., Berlin (2023) 335–369 MR Zbl

[42] **D P Thurston**, *Positive basis for surface skein algebras*, Proc. Natl. Acad. Sci. USA 111 (2014) 9725–9732 MR Zbl

[43] **W P Thurston**, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. 19 (1988) 417–431 MR Zbl

[44] **T Yurikusa**, *Acyclic cluster algebras with dense g-vector fans*, from "McKay correspondence, mutation and related topics", Adv. Stud. Pure Math. 88, Math. Soc. Japan, Tokyo (2023) 437–459 MR Zbl

Mathematical Institute, Tohoku University
Sendai, Japan

Mathematical Science Center for Co-creative Society, Tohoku University
Sendai, Japan

tsukasa.ishibashi.a6@tohoku.ac.jp, s.kano@tohoku.ac.jp

Received: 24 September 2022 Revised: 24 January 2024

Algebraic & Geometric Topology **25**:3 (2025) 1501–1522
DOI: 10.2140/agt.2025.25.1501
Published: 20 June 2025

Bridge trisections and Seifert solids

JASON JOSEPH
JEFFREY MEIER
MAGGIE MILLER
ALEXANDER ZUPAN

We adapt Seifert's algorithm for classical knots and links to the setting of triplane diagrams for bridge trisected surfaces in the 4-sphere. Our approach allows for the construction of a Seifert solid that is described by a Heegaard diagram. The Seifert solids produced can be assumed to have exteriors that can be built without 3-handles; in contrast, we give examples of Seifert solids (not coming from our construction) whose exteriors require arbitrarily many 3-handles. We conclude with two classification results. The first shows that surfaces admitting doubly standard shadow diagrams are unknotted. The second says that a b-bridge trisection in which some sector contains at least $b-1$ patches is completely decomposable, thus the corresponding surface is unknotted. This settles affirmatively a conjecture of the second and fourth authors.

57K45; 57K10

1 Introduction

One of the most important avenues available for study in knotted surfaces in 4-space is the analysis of the 3-dimensional Seifert solids bounded by such surfaces. There are many situations in which information about such a Seifert solid gives rise to useful information about the corresponding knotted surface. Examples, ranging from classical to modern, include Gordon's proof that 2-knots are not determined by their complements [7], Cochran's characterization of fibered homotopy-ribbon 2-knots [3], and recent work of Dai and Miller analyzing the relevance of homology cobordism invariants of Seifert solids [4].

Here we show how topological information about a knotted surface can be recovered from a bridge trisection of the surface, which allows for the diagrammatic study of knotted surfaces and their Seifert solids. A *bridge trisection* of a surface $\mathcal{S}$ in S^4 is a certain decomposition of $(S^4, \mathcal{S})$ into three trivial disk systems $(B^4_1, \mathcal{D}_1)$, $(B^4_2, \mathcal{D}_2)$, and $(B^4_3, \mathcal{D}_3)$ that can be encoded diagrammatically either as a triple of tangles called a *triplane diagram* or as a corresponding *shadow diagram.*

In Section 3, we give a version of Seifert's algorithm for bridge-trisected surfaces, showing how a triplane diagram can be used to produce a 3-manifold bounded by a connected surface $\mathcal{S}$ with normal Euler number zero.

Theorem 3.4 *If $\mathcal{S}$ is connected and $e(\mathcal{S}) = 0$, then there is a procedure to produce a Seifert solid for $\mathcal{S}$ that takes as input a triplane diagram for $\mathcal{S}$.*

In Section 3.2, we give an explicit procedure for constructing a Heegaard diagram for such a 3-manifold when $\mathcal{S} \cong S^2$. As a corollary of the work in building Seifert solids, we recover a combinatorial proof of the existence of Seifert solids. Although the literature already contains a method for producing a Heegaard diagram for a Seifert solid — namely, the work of Carter and Saito [2] — the procedure described here is a bit more practical. In [2, Section 3], the authors employ their methods to take a broken surface diagram and produce a genus 21 Heegaard diagram for a punctured $L(3,1) \# \big(\#^3(S^1 \times S^2)\big)$ bounded by the 2-twist spun trefoil, noting that this solid is nonminimal, since the same 2-knot also bounds a punctured $L(3,1)$. In contrast, in Section 3.3 we use our procedure to find genus three Heegaard diagrams for Seifert solids bounded by the spun trefoil and 1-twist spun trefoil, where these solids are minimal. For the 2-twist spun trefoil, the procedure yields a genus four Heegaard diagram for a Seifert solid (calculations omitted here). The 2-dimensional data contained in a triplane diagram can often be easier to manipulate and simplify than the data in a broken surface diagram; as such, both the solids and their Heegaard diagrams produced by Theorem 3.4 are likely to be less complicated.

We also show that certain bridge trisected surfaces are unknotted.

Theorem 3.3 *If a surface $\mathcal{S}$ has a doubly standard shadow diagram, then $\mathcal{S}$ is unknotted.*

In practice, Theorem 3.3 offers a new and effective method to show unknottedness for bridge trisected surfaces. The doubly standard criterion has considerable potential to aid in the tabulation of low-complexity knotted surfaces, since verifying that a shadow diagram is doubly standard can be much easier than proving unknottedness via other methods.

One of the key features of trisection theory is that it provides a vehicle to adapt 3-dimensional ideas to dimension four, and in Section 4, we prove another result that fits into this line of research. It is well-known that the complement of every *canonical* Seifert surface (ie one obtained from Seifert's algorithm) is a handlebody. Thus, it is natural to attempt to extend this notion to dimension four. In this vein, we call a Seifert solid *canonical* if it is obtained from the procedure presented in Section 3, and we call a Seifert solid *spinal* if its exterior in S^4 can be built without 3-handles. We prove the following two results relating (and distinguishing) these concepts:

Theorem 4.1 *If a surface-knot $\mathcal{S}$ admits a Seifert solid, then it admits a canonical Seifert solid that is spinal.*

In fact, modulo some additional, easily satisfied connectivity conditions, every canonical Seifert solid is spinal. The next result shows that some Seifert solids (in contrast to canonical Seifert solids and many standard examples) are "far" from being spinal.

Theorem 4.2 *Given any $n \in \mathbb{N}$, there exists a 2-knot $\mathcal{K}$ that bounds a Seifert solid Y homeomorphic to $(S^1 \times S^2)^\circ$ such that $S^4 \setminus \nu(Y)$ requires at least n 4-dimensional 3-handles.*

Finally, in Section 5 we prove the following standardness result, affirmatively settling Conjecture 4.3 of Meier and Zupan [15].

Theorem 5.2 *Let $\mathfrak{T}$ be a $(b;\boldsymbol{c})$-bridge trisection with $c_i = b-1$ for some $i \in \mathbb{Z}_3$. Then $\mathfrak{T}$ is completely decomposable, and the underlying surface-link is either the unlink of $\min\{c_i\}$ 2-spheres or the unlink of $\min\{c_i\}$ 2-spheres and one projective plane, depending on whether $|c_{i-1} - c_{i+1}| = 1$ or 0.*

The proof relies on theorems of Scharlemann [19] and Bleiler and Scharlemann [1] regarding planar surfaces in 3-manifolds. The methods of the proof are somewhat unrelated to the methods used in the preceding sections and may be of independent interest. The second and fourth authors previously handled this case when $c_i = b$ for some $i \in \mathbb{Z}_3$ [15, Proposition 4.1]. Theorem 5.2 can be seen as the analog for bridge trisections of Theorem 1.2 of Meier, Schirmer and Zupan [13], which establishes a similar standardness result for trisections of closed manifolds; as such, our theorem fills an important gap in the trisections literature and provides yet another avenue to verify that a surface in S^4 is unknotted.

Acknowledgements

This paper began following discussions at the workshop *Unifying 4-dimensional knot theory*, which was hosted by the Banff International Research Station in November 2019, and the authors would like to thank BIRS for providing an ideal space for sparking collaboration. We are grateful to Masahico Saito for sharing his interest in adapting Seifert's algorithm to bridge trisections and motivating this paper. The authors would like to thank Román Aranda, Scott Carter, and Peter Lambert-Cole for helpful conversations. The authors are grateful to the referee for their careful reading and useful comments. Joseph was supported by MPIM during part of this project, as well as NSF grants DMS-1664567 and DMS-1745670. Meier was supported by NSF grants DMS-1933019 and DMS-2006029. Miller was supported by MPIM during part of this project, as well as NSF grants DGE-1656466 (at Princeton) and DMS-2001675 (at MIT) and a research fellowship from the Clay Mathematics Institute (at Stanford). Zupan was supported by MPIM during part of this project, by NSF grants DMS-1664578 and DMS-2005518, and by a Simons Fellowship.

2 Preliminaries

We work in the smooth category. This section includes an abbreviated introduction to the concepts relevant to this paper, but the interested reader is encouraged to consult [5] for further information about 4-manifold trisections and [9, Section 2; 15] for more detailed discussions of bridge trisections. We limit our work here to surfaces in S^4, but there is also a theory of bridge trisections in arbitrary 4-manifolds; see [16].

2.1 Bridge trisections

Let $\mathcal{S}$ be an embedded surface in S^4, let b be a positive integer, and let $\boldsymbol{c} = (c_1, c_2, c_3)$ be a triple of positive integers. A $(b; \boldsymbol{c})$*-bridge trisection* of $(S^4, \mathcal{S})$ is a decomposition

$$(S^4, \mathcal{S}) = (X_1, \mathcal{D}_1) \cup (X_2, \mathcal{D}_2) \cup (X_3, \mathcal{D}_3)$$

such that

(1) each $\mathcal{D}_i$ is a collection of c_i boundary-parallel disks in the 4-ball X_i,

(2) each intersection $\mathcal{T}_i = \mathcal{D}_{i-1} \cap \mathcal{D}_i$ is a boundary-parallel tangle in the 3-ball $H_i = X_{i-1} \cap X_i$ (with indices considered mod 3),

(3) the triple intersection $\mathcal{D}_1 \cap \mathcal{D}_2 \cap \mathcal{D}_3$ is a collection of b points in the 2-sphere $\Sigma = X_1 \cap X_2 \cap X_3$.

In [15], it was proved that every surface $\mathcal{S}$ admits a $(b; \boldsymbol{c})$-bridge trisection for some $(b; \boldsymbol{c})$. We choose orientations so that $\partial(X_i, \mathcal{D}_i) = (H_i, \mathcal{T}_i) \cup (\overline{H}_{i+1}, \overline{\mathcal{T}}_{i+1})$. When we wish to be succinct, we use $\mathfrak{T}$ to represent a bridge trisection, with components labeled as above.

2.2 Diagrams for bridge trisections

The existence of bridge trisections gives rise to a new diagrammatic theory for surfaces in S^4, using an object called a *triplane diagram*, a triple $\mathbb{D} = (\mathbb{D}_1, \mathbb{D}_2, \mathbb{D}_3)$ of trivial planar diagrams with the additional condition that each $\mathbb{D}_i \cup \overline{\mathbb{D}}_{i+1}$ is a classical diagram for an unlink. In [15], it was shown that every triplane diagram determines a bridge trisection $\mathfrak{T}$. Conversely, given a bridge trisection $\mathfrak{T}$ of $(S^4, \mathcal{S})$, we can choose a triple of disks $E_i \subset H_i$ with common boundary and project the tangles $\mathcal{T}_i$ onto E_i to obtain a triplane diagram. Of course, the choices of disks and projections are not unique, but any two triplane diagrams corresponding the same bridge trisection $\mathfrak{T}$ are related by a finite collection of *interior Reidemeister moves* and *mutual braid transpositions*, while any two bridge trisections $\mathfrak{T}$ and $\mathfrak{T}'$ for the same surface $\mathcal{S}$ are related by *perturbation* and *deperturbation* moves.

In addition, bridge trisections yield another type of diagram: each trivial tangle $\mathcal{T}_i$ can be isotoped rel-boundary into the surface Σ, yielding a triple (A, B, C) of pairwise disjoint collections of arcs called a *shadow diagram*, which has the property that $\partial A = \partial B = \partial C$, and the pairwise unions of any two of the tangles $\mathcal{T}_A$, $\mathcal{T}_B$, and $\mathcal{T}_C$ determined by the arcs are unlinks. As with triplane diagrams, any shadow diagram determines a bridge trisection. Further details about shadow diagrams can be found in [14].

Here we consider special types of shadow diagrams. We say that a pair of collections of arcs in a shadow diagram is *standard* if their union is embedded. Any bridge trisection admits a shadow diagram (A, B, C) in which one of the pairs is standard. If two or three pairs of shadows in a shadow diagram (A, B, C) are standard, then we say that (A, B, C) is *doubly standard* or *triply standard*, respectively. Theorem 3.3 says that doubly standard (and thus triply standard) diagrams always describe unknotted surfaces.

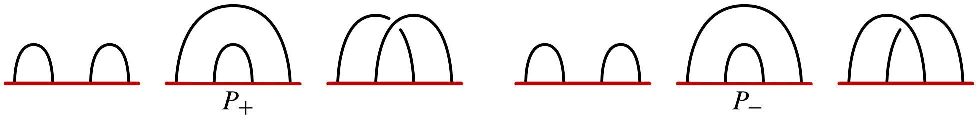


Figure 1: Triplane diagrams for P_+ and P_-.

2.3 Unknotted surfaces

In this subsection, we review standard notions of unknottedness for surfaces in S^4. A closed connected orientable surface $\mathcal{S}$ in S^4 is *unknotted* if it bounds an embedded 3-dimensional handlebody $H \subset S^4$. For nonorientable surfaces, the definition is slightly more involved. We define the two unknotted projective planes, $P_\pm$, to be the two standard projective planes in S^4, pictured via their triplane diagrams in Figure 1, where $e(P_\pm) = \pm 2$.

In general, for a nonorientable surface $\mathcal{S}$, we say that $\mathcal{S}$ is *unknotted* if $\mathcal{S}$ is isotopic to a connected sum of some number of copies of P_+ and P_-. See [9, Remark 2.6] for a detailed discussion of the orientation conventions used here.

3 Seifert solids

Classical results of Gluck [6] (resp. Gordon–Litherland [8]) assert that every orientable surface $\mathcal{S}$ (resp. surface $\mathcal{S}$ with $e(\mathcal{S}) = 0$) in S^4 bounds an embedded 3-manifold, called a *Seifert solid* in the orientable case. In the setting of broken surface diagrams, Carter and Saito provided a procedure that in many respects mimics Seifert's algorithm for classical knots [2]. In this section, we describe an extension of Seifert's algorithm that takes an oriented triplane diagram $\mathbb{D}$ and produces a Seifert solid whose intersection with ∂X_i agrees with the classical Seifert algorithm performed on the oriented unlink diagram $\mathbb{D}_i \cup \overline{\mathbb{D}}_{i+1}$. We also obtain alternative proofs of the theorems of Gluck and Gordon–Litherland mentioned above.

3.1 Existence of Seifert solids

Given a spanning surface F for an unlink U, we define the *cap-off* $\mathcal{F}$ of F to be the closed surface $\mathcal{F} \subset S^4$ obtained by gluing a collection of trivial disks in B^4_- to F along U. (There is a unique such choice of disks up to isotopy rel-boundary in B^4_- by eg [10] or [12].) Let $F_+ \subset S^3$ denote the Möbius band bounded by the unknot so that F_+ contains a positive half-twist and has boundary slope $+2$, and let $F_- \subset S^3$ denote the Möbius band bounded by the unknot with a negative half-twist and boundary slope -2. For $n > 0$, let F_n be the connected surface obtained by attaching $n-1$ trivial bands to the split union of n copies of F_+; that is, F_n is obtained by taking the boundary connected sum of n copies of F_+. For $n < 0$, let F_n be obtain by taking the boundary connected sum of $(-n)$ copies of F_-. Finally, let F_0 be the disk bounded by the unknot in S^3. Additionally, let $\mathcal{F}_n$ be the cap-off of F_n. In Figure 1, the negative Möbius band is shown to cap off into B^4_+ to obtain P_+; see also [9, Figure 2]. Here, we are capping off into B^4_-, so that by definition the cap-off $\mathcal{F}_{-1}$ of the negative Möbius band F_- is P_-. In contrast, the cap-off $\mathcal{F}_1$ of the positive Möbius band F_+ is P_+. (Recall that P_+ and P_- denote the two unknotted projective planes in S^4; see Section 2.3.) It follows that

$$\mathcal{F}_n = \begin{cases} \text{a connected sum of } n \text{ copies of } P_+ & \text{if } n > 0, \\ \text{a connected sum of } -n \text{ copies of } P_- & \text{if } n < 0, \\ \text{an unknotted 2-sphere} & \text{if } n = 0. \end{cases}$$

The intent of the cap-off notation is to emphasize the way in which $\mathcal{F}_n$ can be obtained from a specific surface in S^3, which will be useful in the rest of this section — especially given the following lemma:

Lemma 3.1 *Every incompressible spanning surface F for the unknot is isotopic to F_n for some $n \in \mathbb{Z}$.*

Proof First, we argue that F_n is incompressible for all n. This follows from [20], but we include a proof here. Certainly, F_0 and $F_{\pm 1}$ are incompressible, since a compression increases Euler characteristic by two. Suppose now that F_n is compressible for some $n > 1$, and let F'_n be the component of the surface obtained by compressing F_n so that $\partial F'_n = \partial F_n$. In addition, let $\mathcal{F}'_n \subset S^4$ be the cap-off of F'_n. Then the embedded surface $\mathcal{F}_n$ can be obtained from $\mathcal{F}'_n$ by a 1-handle attachment, and thus $e(\mathcal{F}'_n) = e(\mathcal{F}_n) = 2n$. However, since the nonorientable genus of $\mathcal{F}'_n$ is strictly less than n, this contradicts the Whitney–Massey theorem (see discussion in [9]). We conclude that F_n is incompressible.

On the other hand, suppose that F is an arbitrary incompressible spanning surface for the unknot U. The exterior of U is a solid torus V, and every simple closed curve $c \subset \partial V$ is homotopic to a (p,q)-curve, where a $(0,1)$-curve is the boundary of a meridian disk of V and a $(1,0)$-curve is the boundary of a meridian disk of $N(U)$. The boundary of F is a $(2k,1)$-curve for some integer k. (The spanning surface F intersects the disk bounded by U in some number of arcs, the endpoints of which correspond to the intersections of the (p,q)-curve with the $(0,1)$-curve.) If F is orientable, then it is well-known that F is isotopic to the meridian disk F_0.

Suppose that F is nonorientable. By [20, Corollary 12], the nonorientable genus of F is equal to $|k|$. Assuming that ∂F and ∂F_0 meet efficiently, isotope F so that it intersects F_0 minimally. By standard cut-and-paste arguments, an arc of $F \cap F_0$ which is outermost in F_0 gives rise to a boundary-compressing disk Δ for F. Since ∂F and ∂F_0 meet efficiently, the result F' of boundary-compressing F along Δ has a single boundary component and nonorientable genus $k-1$. Reversing the process, we see that F can be obtained from F' by attaching a boundary-parallel band to F' along opposite sides of $\partial F'$. Note that $\partial V \setminus \partial F'$ is an annulus and the band is determined by a spanning arc. Working rel-boundary, all choices of spanning arcs are related by Dehn twists about $\partial F'$, and so it follows that up to isotopy, there is a unique band taking F' to F.

Finally, we claim that F is isotopic to F_k, and we prove this fact by inducting on k. If $k = \pm 1$, then F has genus one and is obtained from the disk $F' = F_0$ by a single boundary tubing. By the above argument, there is precisely one way to do this, and thus $F = F_{\pm 1}$. Now suppose that $k > 1$ and the claim holds for $j = k-1$. As above, isotope F to meet F_0 minimally, and since $k > 1$, there are at least two arcs a_0 and a_1 of $F \cap F_0$ that are outermost in F_0. Let ℓ be a $(0,1)$-curve that meets ∂F in a single point contained in a_0. Then a_1 gives rise to a boundary-compressing disk Δ_1 and the result F' of boundary-compressing F along Δ_1 also satisfies $|\partial F' \cap \ell| = 1$, since the modification was carried out away from the arc a_0. We conclude that F' has genus $k-1$ and boundary slope $(2(k-1), 1)$. By induction $F' = F_{k-1}$, and since there is a unique way to obtain F from F' by boundary-tubing, it follows that $F = F_k$. The case $k < -1$ follows symmetrically. □

In the next proposition, we use Lemma 3.1 to understand the cap-off of any spanning surface F for an unlink in S^3:

Proposition 3.2 *Let F be a spanning surface for an unlink U in S^3.*

1. *If every component of ∂F has slope 0, then the cap-off $\mathcal{F}$ bounds a (possibly nonorientable, possibly disconnected) handlebody $V \subset B^4$ such that $V \cap \partial B^4 = F$.*
2. *The normal Euler number $e(\mathcal{F})$ is equal to the sum of the slopes of the boundary components of F.*
3. *The cap-off $\mathcal{F}$ is a split union of unknotted surfaces in S^4.*

Proof Suppose F and F' are two spanning surfaces for an unlink U in S^3 such that F' is isotopic relative to U to the surface obtained by surgering F along a compressing disk D for F. Then there is a compression body $C \subset S^3 \times [0,1]$ such that

- $C \cap (S^3 \times \{1\}) = F \times \{1\}$,
- $C \cap (S^3 \times \{0\}) = F' \times \{0\}$, and
- $\partial C = (F \times \{1\}) \cup (\overline{F' \times \{0\}}) \cup (U \times [0,1])$,

and C has a single critical point (of index 1) with respect to the Morse function $S^3 \times [0,1] \to [0,1]$, which we assume lies in $S^3 \times \{\frac{1}{2}\}$. Note that C is a product cobordism above and below $S^3 \times \{\frac{1}{2}\}$.

Any spanning surface F for U can be reduced to F', a union of 2-spheres and incompressible spanning surfaces for components of U via a sequence of compressions and isotopies. If each component of ∂F has slope 0, then F' is a collection of disks and spheres. Applying the compression body construction described above for each compression taking F to F' and stacking the results, we get a compression body C cobounded by F and F'. Since F' is a collection of disks and spheres, there is a handlebody with boundary $\mathcal{F} = F \cup \mathcal{D}$, where $\mathcal{D} = \overline{F'} \cup (U \times [0,1])$ is a collection of properly embedded disks in B^4: simply cap-off the sphere components of C with 3-balls whose interiors are pushed sufficiently deep into B^4. This handlebody is nonorientable (resp. disconnected) if and only if F is. This establishes (1).

Let F be any spanning surface for an unlink $U = \bigsqcup_{i=1}^n U_i$. Let $B = \bigsqcup_{i=1}^n B_i$ be a collection of disjoint 3-balls with $U_i \subset \operatorname{Int}(B_i)$. Let $F' = \bigsqcup_{i=1}^n F_i$ be a split union of incompressible spanning surfaces for the components of U, with $F_i \subset \operatorname{Int}(B_i)$, so that the slopes of F and F' agree at each component of U. Let F'' be the result of surgering F' along a collection of arcs so that F'' and F have the same homeomorphism type relative to U; moreover, assume that every arc of the collection intersects each component of ∂B in at most one point. It follows that F'' decomposes as a split union of connected sums of surfaces, each summand of which is either a torus or an incompressible spanning surface for an unknot. Therefore, the cap-off $\mathcal{F}''$ is the split union of connected sums of surfaces, each summand of which is an unknotted surface in S^4. Livingston showed that F and F'' are isotopic rel-boundary in B^4 [12]. It follows that the cap-off $\mathcal{F}$ will isotopic to the cap-off $\mathcal{F}''$, which completes the proof of (3). Since (2) holds for $\mathcal{F}_1$ and $\mathcal{F}_{-1}$, and since the normal Euler number is additive under connected sum, (2) follows, as well. □

Recall that a shadow diagram is doubly standard if two of the pairings of arcs yield embedded curves. We can use Proposition 3.2 to obtain the following classification result for doubly standard diagrams:

Theorem 3.3 *If $\mathcal{S}$ has a doubly standard shadow diagram, then $\mathcal{S}$ is unknotted.*

Note that Theorem 3.3 also applies to surfaces with triply standard shadow diagrams, as a special class of doubly standard shadow diagrams.

Proof Suppose $\mathcal{S}$ has a shadow diagram (A, B, C) such that the pairings (A, B) and (B, C) are standard. Consider the standard Heegaard splitting $\partial X_3 = S^3 = H_+ \cup_\Sigma H_-$, and let $\Sigma_\pm$ be a parallel copy of Σ pushed slightly into $H_\pm$. Note that $A \cup B$ may have nested components (so that components of $A \cup B$ don't necessarily bound a collection of disjoint disks). After a sequence of arc slides, however, performed only on the arcs in A, we obtain arcs A' such that the embedded curves $A' \cup B$ bound a pairwise disjoint collection of disks. We perform a similar procedure with $B \cup C$ to obtain $B \cup C'$. Now embed parallel copies $A'_+ \cup B_+$ of the curves $A' \cup B$ in Σ_+ so that they bound a pairwise disjoint collection D_+ of disks in Σ_+, and embed parallel copies $B_- \cup C'_-$ of the curves $B \cup C'$ in Σ_- so that they bound a pairwise disjoint collection D_- of disks in Σ_-. In H_+, there is an isotopy of B_+ to $B \subset \Sigma$ taking the disks D_+ to disks $D_1 \subset H_+$ such that $D_1 \cap \Sigma = B$. The tangle $\mathcal{T}_1 = \mathcal{S} \cap (H_+)$ is the image of A'_+ under this isotopy. Similarly, in H_- there is an isotopy of B_- to B taking the disks D_- to disks $D_2 \subset H_-$ such that $D_2 \cap \Sigma = B$. The tangle $\mathcal{T}_3 = \mathcal{S} \cap H_-$ is the image of C'_- under this isotopy. See Figure 2.

By construction $D_1 \cap D_2 = B$, so that $F = D_1 \cup D_2$ is a spanning surface for the unlink $\mathcal{T}_1 \cup \mathcal{T}_3$. Note further that D_1 is a trivial disk system for $\mathcal{T}_1 \cup B$, and D_2 is a trivial disk system for $B \cup \mathcal{T}_3$; hence, $\mathcal{S}$ is

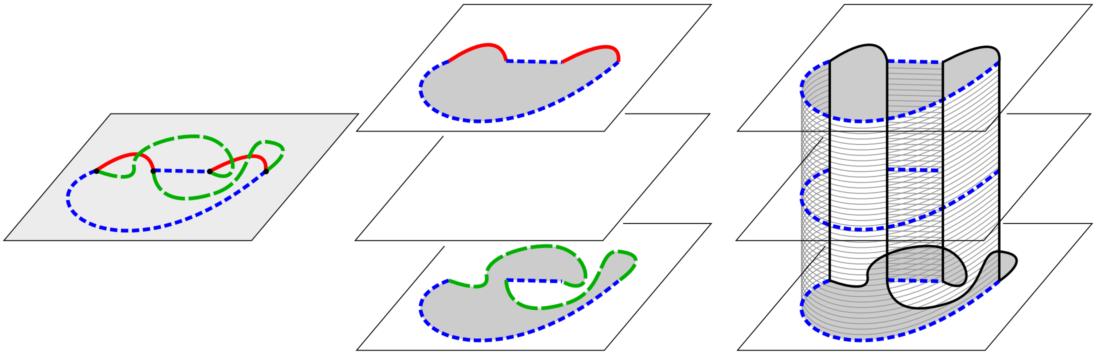


Figure 2: Left: a doubly standard shadow diagram (A, B, C); the pairings (A, B) and (B, C) are standard. Middle: disks in Σ_+ and Σ_- bounded by parallel copies of $A \cup B$ and $B \cup C$, respectively. Right: a spanning surface F for $\mathcal{T}_1 \cup \mathcal{T}_2$ in $\partial X_3 = S^3$.

the union of D_1, D_2, and D_3, where D_3 is a trivial disk system for $\mathcal{T}_1 \cup \mathcal{T}_3$ pushed into B^4. However, since $F = D_1 \cup D_2 \subset S^3$, it follows that $\mathcal{S}$ is also isotopic to the cap-off $\mathcal{F}$ of F, which is unknotted by Proposition 3.2. □

We are now ready to prove our main result:

Theorem 3.4 *If $\mathcal{S}$ is connected and $e(\mathcal{S}) = 0$, then there is a procedure to produce a Seifert solid for $\mathcal{S}$ that takes as input a triplane diagram for $\mathcal{S}$.*

Proof The proof follows from the proofs of Propositions 3.5 and 3.6 below. □

In Section 3.2, we show that there is a procedure to produce a Heegaard splitting for the Seifert solid when $\mathcal{S}$ is a 2-knot.

In addition to providing the proof of the above theorem, the next two propositions provide alternative proofs of the results in [6; 8] mentioned above.

Proposition 3.5 *Every orientable surface-link $\mathcal{S}$ bounds a Seifert solid in S^4.*

Proof Let $\mathbb{D}$ be a triplane diagram for $\mathcal{S}$, with induced orientation on the bridge points $\boldsymbol{x}$. Perform mutual braid transpositions so that the bridge points alternate sign (orientation). Then there are b pairwise disjoint arcs ε contained in the equator e connecting bridge points of opposite signs, so that $\mathbb{D}_i \cup \varepsilon$ is an oriented link diagram. Let F_i be the Seifert surface obtained by performing Seifert's procedure on the diagram $\mathbb{D}_i \cup \varepsilon$, and let $\widehat{F}_i = F_i \cup \overline{F}_{i+1}$ be the spanning surface obtained by gluing F_i to $\overline{F}_{i+1}$ along ε. By Proposition 3.2, there exists a handlebody $V_i \subset X_i$ such that $\partial V_i = \widehat{F}_i \cup \mathcal{D}_i$ and $V_i \cap \partial X_i = \widehat{F}_i$. Finally, $Y = V_1 \cup V_2 \cup V_3$ is an embedded 3-manifold whose boundary is $\mathcal{D}_1 \cup \mathcal{D}_2 \cup D_3 = S$, and so Y is a Seifert solid for $\mathcal{S}$. □

Proposition 3.6 *If $\mathcal{S}$ is connected and $e(\mathcal{S}) = 0$, then $\mathcal{S}$ bounds a spanning solid in S^4.*

Proof Consider a bridge trisection $\mathfrak{T}$ of $\mathcal{S}$, with $U_i = \partial \mathcal{D}_i$ and $\tau = \mathcal{T}_1 \cup \mathcal{T}_2 \cup \mathcal{T}_3$. By taking, for example, a triplane diagram $\mathbb{D}$ and compatible checkerboard surfaces in $\mathbb{D}_i$, we can produce spanning surfaces $\widehat{F}_i$ for U_i such that $\widehat{F}_i \cap H_i = \widehat{F}_{i-1} \cap H_i$. Let F_i denote $\widehat{F}_i \cap H_i$. For each component J of $U_i = \partial \widehat{F}_i$, let $\iota_{\widehat{F}}(J)$ denote the induced boundary slope on the curve J by the surface $\widehat{F}_i$. Then by Proposition 3.2, we have

$$\sum_{J \subset U_1 \cup U_2 \cup U_3} \iota_{\widehat{F}}(J) = 0.$$

Choose a triple of spanning surfaces $\widehat{F}_i$ such that $\sum |\iota_{\widehat{F}}(J)|$ is minimal over all possible choices. We claim that $\sum |\iota_{\widehat{F}}(J)| = 0$. If not, then there exist boundary curves J_+ and J_- such that $\iota_{\widehat{F}}(J_+) > 0$ and $\iota_{\widehat{F}}(J_-) < 0$. Noting that the surface $\mathcal{S}$ contains all curves $J \subset U_i \subset \tau$, push each curve $J \subset U_i$ slightly off of τ into the corresponding disk component of $\mathcal{D}_i$, so that the collection of curves J is embedded in $\mathcal{S}$ and disjoint from τ. Choose a path $\gamma \subset \mathcal{S}$ from J_+ to J_-, avoiding the bridge points, noting that $|\gamma \cap \tau| > 0$.

At each point of $\gamma \cap \tau$, modify the corresponding component of F_i by taking the boundary connected sum of F_i with a trivial Möbius band to obtain new surfaces $\widehat{F}_i'$ and F_i', so that the corresponding boundary curves satisfy $\iota_{\widehat{F}'}(J_+') = \iota_{\widehat{F}}(J_+) - 2$, $\iota_{\widehat{F}'}(J_-') = \iota_{\widehat{F}}(J_-) + 2$, and $\iota_{\widehat{F}'}(J') = \iota_{\widehat{F}}(J)$ for all other curves J'. It follows that $\sum |\iota_{\widehat{F}'}(J')| < \sum |\iota_{\widehat{F}}(J)|$, contradicting our assumption of minimality. (Note that $\iota_{\widehat{F}}(J)$ is always even, since it represents the number of intersection points between the boundary curves of spanning surfaces; see the proof of Lemma 3.1.)

We conclude that $\iota_{\widehat{F}}(J) = 0$ for all curves J, and thus by Proposition 3.2, each spanning surface $\widehat{F}_i$ cobounds a (possibly) nonorientable handlebody $V_i \subset X_i$ with the disks $\mathcal{D}_i$. It follows that $V_1 \cup V_2 \cup V_3$ is a spanning solid for $\mathcal{S}$ in S^4. □

3.2 Procedure to find a Heegaard diagram for a Seifert solid

In this subsection, we describe a procedure for finding a Heegaard diagram for the Seifert solid coming from a bridge trisection $\mathfrak{T}$ of a 2-knot $\mathcal{S}$. We use labels consistent with those appearing above in the proof of Proposition 3.5. The process is illustrated in Figures 3–6.

Step 1 Given a triplane diagram $\mathbb{D}$ for $\mathcal{S}$, perform interior Reidemeister moves and mutual braid transpositions so that the induced Seifert surfaces satisfy the following conditions:

(a) Each of F_1, F_2, and $\widehat{F}_1$ is a collection of disks.

(b) Surfaces $\widehat{F}_2$ and $\widehat{F}_3$ are connected.

(c) $g(\widehat{F}_2) = g(F_3)$.

See Figure 3. Note that attaining condition (a) is possible since any triplane diagram can be converted to one in which two of the tangles have no crossings. Condition (b) can be attained by performing interior Reidemeister moves on the diagram $\mathbb{D}_3$. Attaining condition (c) is possible since we can arrange so that F_2 is a collection of b bridge disks, in which case $\widehat{F}_2$ deformation retracts onto F_3 (although in general, we need not assume that F_2 has b components, as shown below).

Step 2 Following the proof of Proposition 3.2, the surfaces $\widehat{F}_2$ and $\widehat{F}_3$ compress completely to disks in S^3. Let α be a complete collection of pairwise disjoint compressing curves in $\widehat{F}_3$, and let β be a complete collection of pairwise disjoint compressing curves in $\widehat{F}_2$. See Figure 4 (top row).

Step 3 If necessary, slide the curves β over the components of $\partial\mathcal{D}_2$ to obtain a collection of curves $\beta' \subset F_3$. Note that since $g(F_3) = g(\widehat{F}_2)$, as curves in $\mathcal{F}_2 = \widehat{F}_2 \cup \mathcal{D}_2$, the collection β can be isotoped to be contained in F_3, and any isotopy of a curve over a disk component of $\mathcal{D}_2$ can be realized as a slide over $\partial\mathcal{D}_2$. Thus, such a sequence of slides exists. See Figure 4 (middle row).

Step 4 Let $P = \mathcal{D}_1 \cup \mathcal{D}_2$, so that P is a planar surface with c_3 boundary components, let Q be the surface obtained by gluing P to $\widehat{F}_3$ along their boundaries, and let α^* be a choice of $c_3 - 1$ boundary components of P and some minimal number of curves in α such that α^* forms a cut system for Q.

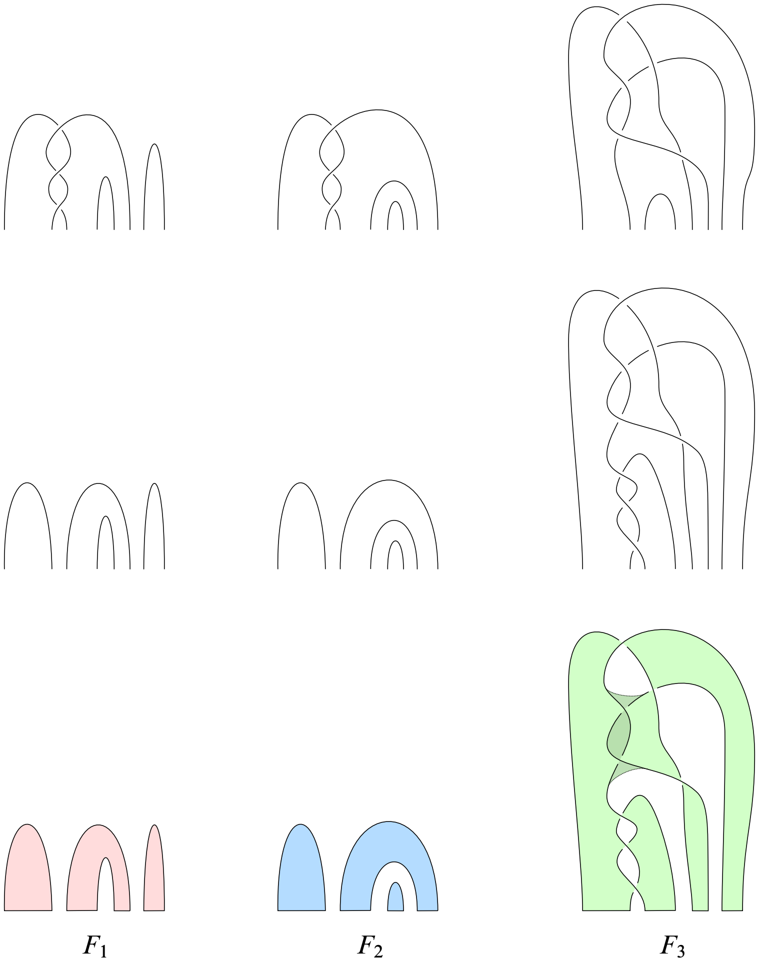


Figure 3: To perform the Seifert solid procedure on a triplane diagram, we first perform mutual braid transposition until the tangle diagrams in V_1 and V_2 have no crossings. Then we perform the usual Seifert's procedure for knot diagrams to obtain surfaces F_1, F_2, and F_3 that agree in the bridge sphere Σ, with F_1, F_2, and $\widehat{F}_1$ all collections of disks and $g(\widehat{F}_2) = g(F_3)$.

Step 5 Let β^* be the union of β' and a collection of curves in Q obtained by the following instructions: For each component of J of $\partial\mathcal{D}_1$, suppose that J meets d disk components of F_2. Choose $d-1$ of these components, isotope them off of F_2 in $\mathcal{F}_2 = F_2 \cup F_3 \cup \mathcal{D}_2$, and add these $d-1$ curves to β^*. Discard any superfluous curves of β' so that β^* is a cut system for Q.

Proposition 3.7 *Using the procedure described above*, $\mathcal{S}$ *bounds a punctured copy of the* 3*-manifold determined by the Heegaard diagram* $(Q; \alpha^*, \beta^*)$.

Proof Suppose that $\mathbb{D}$ is a triplane diagram satisfying conditions (a), (b), and (c) given in Step 1 above. Following the proofs of Propositions 3.2 and 3.5, for each i, the surface $\widehat{F}_i \cup \mathcal{D}_i$ bounds a handlebody V_i, where V_1 is a collection of 3-balls, say $B_1, \dots, B_n$, and V_2 and V_3 are connected. Moreover, α contains a

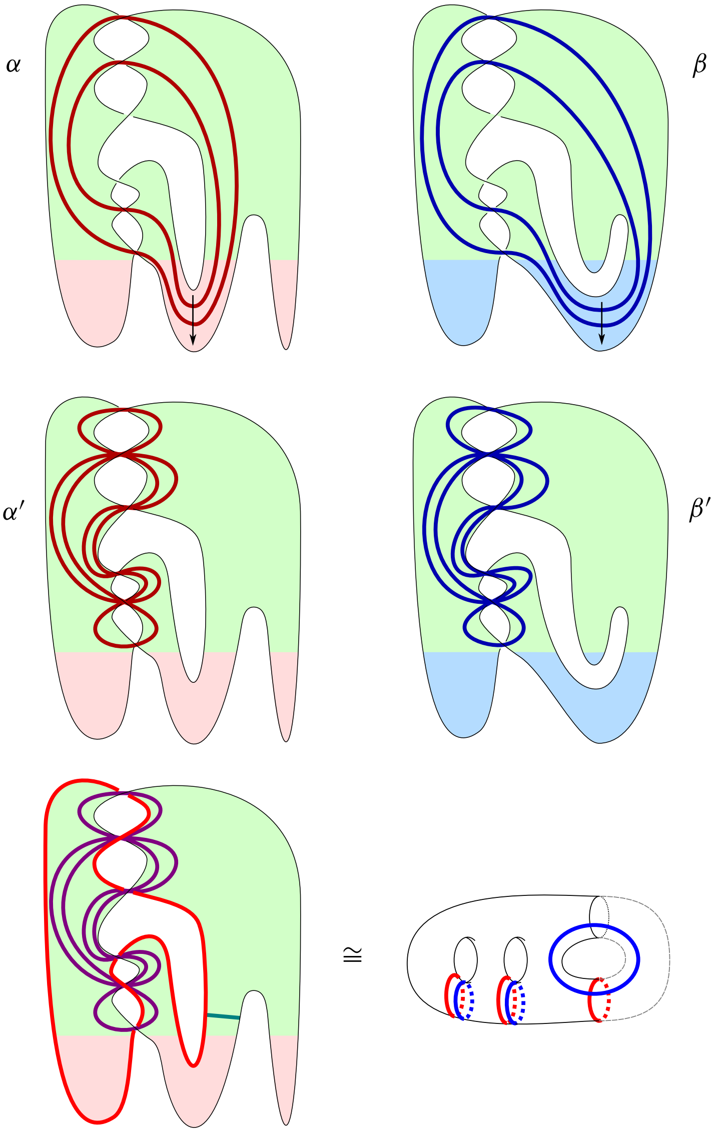


Figure 4: Top: we find complete sets of compressing curves α and β for $\widehat{F}_3$ and $\widehat{F}_2$, respectively. Middle: we slide α and β (with slides indicated in top row) over $\partial\widehat{F}_3$ and $\partial\widehat{F}_2$ to obtain curve systems α' and β' that are each completely within F_3. Bottom: We obtain α^* (red and purple curves) by adding boundary curves as in Step (4) of Section 3.2. We obtain β^* by adding arcs as in Step (5). Then $(Q;\alpha^*,\beta^*)$ is a Heegaard diagram for a (closure of a) Seifert solid for the 2-knot described by the initial triplane diagram.

cut system for V_3 and β contains a cut system for V_2. Since β' is homotopic to β in ∂V_2, it follows that β' also contains a cut system for V_2. Thus, the Seifert solid bounded by $\mathcal{S}$ is equal to $V_2 \cup V_3 \cup B_1 \cup \cdots \cup B_n$. Let Y be the closed 3-manifold obtained by capping off the boundary $\mathcal{S}$ of this Seifert solid with an abstract 3-ball B_0. We will show that $(Q;\alpha^*,\beta^*)$ is a Heegaard diagram for Y.

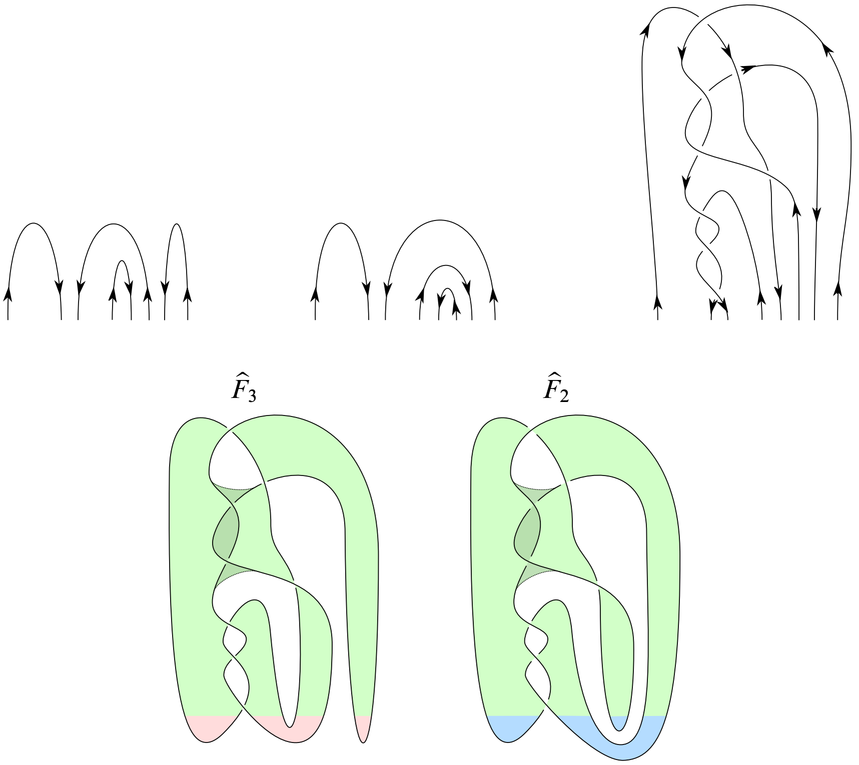


Figure 5: We start performing the Seifert solid procedure (Section 3.2) on the triplane diagram in the top row.

To this end, consider $W = V_3 \cup B_0$ and $W' = V_2 \cup B_1 \cup \cdots \cup B_n$. Considering that $\partial V_2 = F_2 \cup F_3 \cup \mathcal{D}_2$ and $\partial(B_1 \cup \cdots \cup B_n) = F_1 \cup F_2 \cup \mathcal{D}_1$, we have that

$$\partial W' = F_3 \cup F_1 \cup \mathcal{D}_2 \cup \mathcal{D}_1 = \widehat{F}_3 \cup P = Q.$$

Additionally, the 3-balls B_i are attached to V_2 along F_2, which is a collection of disks by condition (a). It follows that the curves $\beta' \cup \partial F_2$ bound compressing disks in W' cutting W' into a collection of 3-balls, so W' is a handlebody. In addition, choosing all but one curve of ∂F_2 for each component B_i and a subset of β' as in Step 5 above yields a cut system β^* for W'.

Turning our attention to W, we have $\partial V_3 = \widehat{F}_3 \cup \mathcal{D}_3$ and $\partial B_0 = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$, so $\partial W = \widehat{F}_3 \cup \mathcal{D}_1 \cup \mathcal{D}_2 = Q$, and in addition, the curves α and $\partial \mathcal{D}_3$ bound disks cutting W into 3-balls. Choosing α^* to contain all but one curve of $\partial \mathcal{D}_3$ and a subset of α as in Step 4, the curves in α^* bound disks cutting W into a single 3-ball, so α^* is a cut system for W. We conclude that $(Q; \alpha^*, \beta^*)$ is a Heegaard diagram for Y, as desired. □

Remark 3.8 It may be the case that the surface F_3 compresses in H_3, in which case α and β could have one or more curves in F_3 in common. Following the procedure with such α and β produces one or more extra $S^1 \times S^2$ summands for the 3-manifold Y, and a simpler Seifert solid can be obtained by first compressing F_3 maximally in H_3.

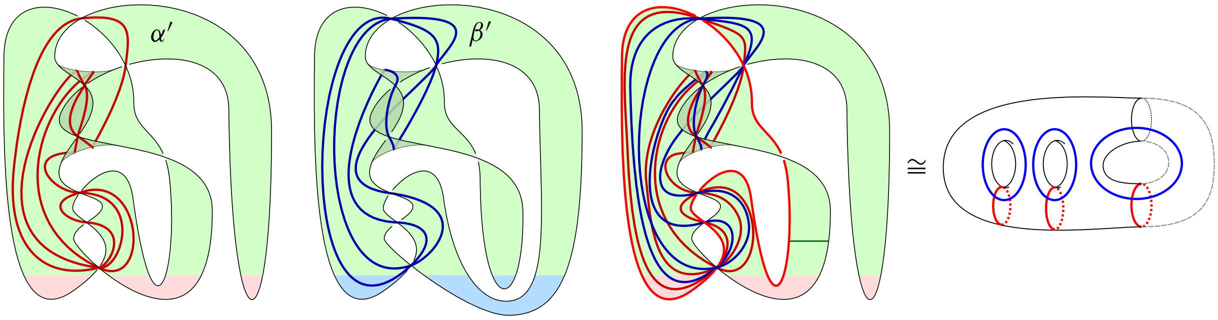


Figure 6: Left: the curves α' in F_3. Center left: the curves β' in F_2. Center right: we add some boundary curves of F_2 to α to obtain α^* and some arc to β' to obtain β^*. Right: we simplify the resulting Heegaard diagram $(\Sigma; \alpha^*, \beta^*)$ to see that it is a diagram of S^3. Thus, the initial 2-knot bounds a copy of B^3 in S^4, so is unknotted.

Remark 3.9 The procedure above can be generalized: We can relax conditions (a), (b), and (c) from Step 1; the only assumption necessary to ensure that $V_1 \cup V_2$ is a handlebody is that their intersection F_2 is a collection of disks. However, the weaker conditions make it somewhat more difficult to draw the diagram, since we are no longer guaranteed the existence of the slides of Step 3 — it may be the case that β curves necessarily intersect the disks $\mathcal{D}_1$ and $\mathcal{D}_2$.

Remark 3.10 The observant reader might notice that we call our process the Seifert solid *procedure*, rather than *algorithm*. An algorithm gives an output completely determined from the input, independent of further choices. A procedure may require additional choices for the output to be determined. In the procedure we give in this section to find a description of a Seifert solid for a 2-knot, we are forced to choose compressing circles for surfaces in S^3. These circles are generally not unique (and in fact, different choices can determine different Seifert solids), so we do not refer to this procedure as an algorithm.

3.3 Some examples

In this subsection, we carry out the procedure described above for a couple of specific examples. The first is the spun trefoil. In Figure 3, we see a triplane diagram for the spun trefoil coming from [15], followed by the result of performing triplane moves so that the induced Seifert surfaces F_i satisfy conditions (a), (b), and (c) from Step 1 above.

In the top row of Figure 4, we find the compressing curves α on $\widehat{F}_3$ and β on $\widehat{F}_2$. Note that in this case $\mathcal{D}_3$ contains two disks, so that $P = \mathcal{D}_1 \cup \mathcal{D}_2$ is an annulus, and $Q = \widehat{F}_3 \cup P$ can be obtained by identifying the two boundary components of $\widehat{F}_3$. Under this identification, the identified boundary components constitute the third curve in the cut system α^*. In the second row at left, we slide the two curves of α over the third curve of α^* in Q. In the second row at right, we slide the two curves of β over a boundary component as shown to get the curves $\beta' \subset F_3$ (which are identical to the image of α under the slides described above). Finally, the third curve of β^* consists of the teal arc depicted in F_3 and a spanning arc in the annulus A, or equivalently, we can identify the endpoints of the teal arc. In the lower row, we see the diagram for the Seifert solid, the standard (once-stabilized) Heegaard diagram for $\#^2(S^1 \times S^2)$.

Remark 3.11 These diagrams and arguments easily generalize to produce the Seifert solid $\#^{p-1}(S^1 \times S^2)$ for the spun $(p,2)$-torus knot. Miyazaki proved that the degree of the Alexander polynomial (over $\mathbb{Q}[t, t^{-1}]$) is a lower bound for the second Betti number of any Seifert solid [17]. Since the degree of the Alexander polynomial of $T(2,p)$ is $p-1$, these solids are minimal in the sense that the corresponding 2-knots cannot bound any 3-manifold with a smaller second Betti number, eg fewer $S^1 \times S^2$ summands.

For the second example, we find a Seifert solid for the 1-twist spun trefoil (which is unknotted by [23]). In Figure 5, we include a simplified triplane diagram for the 1-twist spun trefoil along with the surfaces $\widehat{F}_2$ and $\widehat{F}_3$ this diagram generates.

Next, we find the compressing curves α for $\widehat{F}_3$ and β for $\widehat{F}_2$. As in the spun trefoil example above, $P = \mathcal{D}_1 \cup \mathcal{D}_2$ is an annulus, so we view Q as being obtained by identifying the two boundary components of $\widehat{F}_3$, with this identified boundary the third curve in α^*. Figure 6 shows the curves α, β, and the union of the sets in Q, yielding the standard diagram for S^3, in which the third curve of β^* appears as a teal arc with boundary points identified (as above). Note that the existence of the curves α and β is guaranteed by Proposition 3.2; in practice, however, these curves are found using ad hoc methods.

4 Spinal Seifert solids

A natural aspect of the study of Seifert surfaces for links in the 3-sphere is the consideration of their exteriors. We call a Seifert surface F for L *canonical* if it is isotopic to a surface obtained by applying Seifert's procedure to a diagram for L. We call a Seifert surface F *free* if its exterior $S^3 \setminus \nu(F)$ is a 3-dimensional handlebody — equivalently, has free fundamental group. It is an easy exercise to see that a canonical Seifert surface is free, provided that it is connected; so every link admits a free Seifert surface, by the application of Seifert's algorithm to a nonsplit diagram. However, such a surface can be far from minimal genus. M Kobayashi and T Kobayashi showed that the difference between the genus of a knot and the minimal genus of a free Seifert surface for the knot can be arbitrarily large, and that moreover the difference between the minimal genus of a free Seifert surface for a knot and the minimal genus of a canonical Seifert surface can also be arbitrarily large [11]. (In fact, they show that both of these differences can be made arbitrarily large at the same time.)

In this section, we introduce 4-dimensional analogs of the notions of canonical and free Seifert surfaces. Going forward, let $\mathcal{S} \subset S^4$ be a surface-link admitting a Seifert solid. (This is equivalent to the condition that $\mathcal{S}$ be orientable or have normal Euler number zero.) We call a Seifert solid Y *canonical* if it is isotopic to a Seifert solid obtained by the procedure given in Section 3.1 (see Propositions 3.5 and 3.6). We call a Seifert solid Y *spinal* if $S^4 \setminus \nu(Y)$ deformation retracts onto a finite 2-complex. Equivalently, $S^4 \setminus \nu(Y)$ can be built with handles of index at most two.

Theorem 4.1 *If a surface-knot $\mathcal{S}$ admits a Seifert solid, then it admits a canonical Seifert solid that is spinal.*

Proof First, note that in the proof of Propositions 3.5 and 3.6, it is possible to arrange that each Seifert surface F_i is connected: for example, this is assured if each $\mathbb{D}_i \cup \varepsilon$ is nonsplit. Let Y be a canonical Seifert solid for $\mathcal{S}$ given by Proposition 3.5 or Proposition 3.6 such that the canonical surface $F_i = Y \cap H_i$ is connected for each $i \in \mathbb{Z}_3$. We make use of the notation of the proof of Proposition 3.5 in what follows.

Recall that $V_i = X_i \cap Y$ is a handlebody with $\partial V_i = \widehat{F}_i \cup \mathcal{D}_i$. Moreover, V_i is built relative to $\widehat{F}_i$ by attaching 3-dimensional 2- and 3-handles. It follows that $X_i \setminus \nu(V_i)$ can be built with 4-dimensional 0-, 1-, and 2-handles.

Next, recall that F_i is a canonical Seifert surface for the link $\mathbb{D}_i \cup \varepsilon$, considered in $S^3 = H_i \cup_\Sigma B^3$. Since we have assumed F_i is connected, F_i is free in $H_i \cup_\Sigma B^3$. Since $\varepsilon \subset \partial H_i$, it follows that $H_i \setminus F_i$ is also a 3-dimensional handlebody.

Finally, we can build $S^4 \setminus \nu(Y)$ by taking the $X_i \setminus \nu(V_i)$ and gluing them along the $H_i \setminus \nu(F_i)$. Since the three gluings occur along 3-dimensional handlebodies, $S^4 \setminus \nu(Y)$ is obtained from the disjoint union of the $X_i \setminus \nu(V_i)$ by attaching 4-dimensional 1- and 2-handles. Because each of the $X_i \setminus \nu(V_i)$ were built with 4-dimensional handles of index at most two, the same is true for $S^4 \setminus \nu(Y)$. This shows that Y is spinal, as desired. □

When studying Seifert surfaces, the genus of the surface is the obvious measure of complexity that one might try to minimize. In contrast, there are many ways one might try to quantify the complexity of a Seifert solid Y for a surface-knot; indeed, any complexity one might associate to a 3-manifold could be interesting to consider. Here we content ourselves to give some examples showing that there is at least one sense in which a simple Seifert solid for a surface-knot can be arbitrarily far from being spinal.

Theorem 4.2 *Given any $n \in \mathbb{N}$, there exists a 2-knot $\mathcal{K}$ that bounds a Seifert solid Y homeomorphic to $(S^1 \times S^2)^\circ$ such that $S^4 \setminus \nu(Y)$ requires at least n 4-dimensional 3-handles.*

Proof Let J be an arbitrary knot, and let $K = \mathrm{Wh}_0(J \# \overline{J})$ be the untwisted Whitehead double of the connected sum of J with its mirror. Let F be the standard genus one Seifert surface for K, and let γ be the curve on F that is isotopic to $J \# \overline{J}$. (Alternatively, F is obtained by taking a 0-framed annular thickening of a curve γ isotopic to $J \# \overline{J}$ and plumbing on a Hopf band.)

Let E be the standard ribbon disk for γ, so that $(B^4, E) = (S^3, J)^\circ \times I$. The surface F can be surgered along E in the 4-ball to get a slice disk D for K, and the trace of this surgery yields a solid torus V with $\partial V = F \cup D$.

Let $\mathcal{K} = D \cup_K \overline{D}$ be the 2-knot obtained by doubling D, and let $Y = V \cup_F \overline{V}$ be the double of V along F. Then Y is a Seifert solid for $\mathcal{K}$ and $Y \cong (S^1 \times S^2)^\circ$.

We claim that $\pi_1(S^4 \setminus \nu(Y)) \cong \pi_1(S^3 \setminus \nu(J))$. First, $\pi_1(S^4 \setminus \nu(Y)) \cong \pi_1(B^4 \setminus V)$, since the former exterior is the double of the latter exterior along the exterior of F in S^3 and $\pi_1(S^3 \setminus \nu(F))$ surjects onto $\pi_1(B^4 \setminus V)$ under inclusion. Next, by construction, V is obtained by thickening the slice disk E and attaching a trivial 3-dimensional 1-handle. It follows that

$$\pi_1(B^4 \setminus \nu(V)) \cong \pi_1(B^4 \setminus \nu(E)) \cong \pi_1(S^3 \setminus \nu(J)),$$

as desired.

To complete the proof, let $n \in \mathbb{N}$ be given, and choose J to be any knot with $\operatorname{rank}(\pi_1(S^3 \setminus \nu(J))) \geq n+2$ (eg take J to be a connected sum of $n+1$ trefoils [21]). The exterior $S^4 \setminus \nu(Y)$ can be built relative to $\partial(S^4 \setminus \nu(Y)) \cong (S^1 \times S^2) \# (S^1 \times S^2)$ with some number of 4-dimensional 1-, 2-, 3-, and 4-handles. Since the 1-handles correspond to generators of the fundamental group, at least n are required; the boundary $\partial(S^4 \setminus \nu(Y))$ contributes only two to the rank of the fundamental group. Similarly, since we can obtain another presentation of $\pi_1(S^4 \setminus \nu(Y))$ with generators corresponding to 3-handles, the number of 3-handles in this decomposition is at least $n+2$. □

We note that the construction of $\mathcal{K}$ given in the above proof is closely related to an interesting construction of 2-knots given by Cochran [3].

Next, we observe that many important examples of Seifert solids are, in fact, spinal:

(1) Every ribbon 2-knot bounds a Seifert solid Y that is homeomorphic to $\big(\#^m(S^1 \times S^2)\big)^\circ$ for some m [22]. The manifold Y is obtained by taking a Seifert surface F for some ribbon knot in an equatorial S^3, thickening it, and attaching trivial 2-handles above and below the equator. By attaching tubes to F (at the cost of increasing m), we can arrange for F to be free. Then Y is spinal.

(2) If $\mathcal{K}$ is fibered with fiber Y, then $S^4 \setminus \nu(Y) \cong Y \times I$ is spinal, since Y is a punctured 3-manifold.

(3) Connected Seifert solids arising from broken surface diagrams via the construction given by Carter and Saito [2] are spinal. Recall that a connected canonical Seifert surface is free because it deformation retracts to a graph so that on each edge, there is one local maximum and no local minima with respect to the radial height function on S^3. (Here the vertices of the graph correspond to the disks produced in Seifert's procedure while the edges correspond to the half-twisted bands.) This ensures that the exterior of a canonical surface can be built with 0- and 1-handles. Similarly, a Seifert solid constructed à la [2] deformation retracts to a 2-complex with one local maximum and no other critical points in the interior of each 1- and 2-cell. Thus, the exterior of such a Seifert solid can be built with 0-, 1-, and 2-handles.

Finally, we can formulate a question analogous to the 3-dimensional results in [11] in the setting of surface-knots.

Question 4.3 *Define the* ***genus*** *of an orientable surface-knot $\mathcal{S}$ in S^4 to be the minimal first Betti number of any Seifert solid bounded by $\mathcal{S}$, and define the* ***spinal genus*** *and* ***canonical genus*** *similarly, using spinal Seifert solids and canonical Seifert solids, respectively. Do there exist surface-knots for which these three measures of complexity differ?*

We remark that using techniques as in the proof of Theorem 4.2, one can show that for some of the known classical knots K whose genus and free genus are sufficiently different (see [18], for example), the spun knots $\mathcal{S}(K)$ admit low-complexity nonspinal Seifert solids, whereas the obvious spinal and canonical Seifert solids have greater complexity. However, it is likely to be considerably more difficult to obstruct the existence of low-complexity spinal or canonical Seifert solids, even for these examples.

5 On standardness of bridge trisections

The goal of this section is to prove Theorem 5.2, which states that a $(b; c_1, c_2, c_3)$-bridge trisection that satisfies $c_i \geq b-1$ for some $i \in \mathbb{Z}_3$ can be completely decomposed into standard pieces. This proves Conjecture 4.3 of [15], and the theorem can be viewed as the bridge trisection analog of the main result in [13], which states that every $(g; k_1, k_2, k_3)$-trisection with $k_i \geq g-1$ for some i is standard in that it decomposes into genus one summands.

We encourage the reader to recall the notions of *perturbation* and *connected summation* for bridge trisections. The former was first introduced in [15, Section 6], where it was referred to as stabilization, and the latter can be reviewed in [15, Subsection 2.2]. See also [14, Section 3] for a succinct description of these concepts.

We call a surface-link an *unlink* if it is the split union of unknotted surface-knots, though we allow the topology of each component to vary. For example, one might have a 2-component unlink that is the split union of an unknotted 2-sphere and an unknotted projective plane. (See [14, Subsection 2.2] and Section 2.3 above for a brief discussion of unknotted surface-knots.)

Before proving Theorem 5.2 in generality, we recall the case in which $c_i = b$ for some $i \in \mathbb{Z}_3$. This was addressed as [15, Proposition 4.1]. A bridge trisection is called *completely decomposable* if it is a disjoint union of perturbations of one-bridge and two-bridge trisections.

Proposition 5.1 [15, Proposition 4.1] *Let $\mathfrak{T}$ be a $(b; c_1, c_2, c_3)$-bridge trisection with $c_i = b$ for some $i \in \mathbb{Z}_3$. Then $\mathfrak{T}$ is completely decomposable, and the underlying surface-link is the unlink of $\min_i\{c_i\}$ 2-spheres.*

Note that if $c_i = b$ for some $i \in \mathbb{Z}_3$, then $c_{i-1} = c_{i+1}$. Similarly, in what follows we will see that if $c_i = b-1$ for some $i \in \mathbb{Z}_3$, then $|c_{i-1} - c_{i+1}| \leq 1$. We now present and prove the main result of this section:

Theorem 5.2 *Let $\mathfrak{T}$ be a $(b; c_1, c_2, c_3)$-bridge trisection with $c_i = b-1$ for some $i \in \mathbb{Z}_3$. Then $\mathfrak{T}$ is completely decomposable, and the underlying surface-link is either the unlink of $\min\{c_i\}$ 2-spheres or the unlink of $\min\{c_i\}$ 2-spheres and one projective plane, depending on whether $|c_{i-1} - c_{i+1}| = 1$ or $c_{i-1} = c_{i+1}$.*

The key ingredient in the proof of the theorem is a pair of results of Bleiler and Scharlemann about planar surfaces in 3-manifolds [1; 19]. We refer the reader to Section 1 of each of these papers, as we will adopt the notation of [1, Theorem 1.3; 19, Theorem 1.1] in the proof below.

Proof of Theorem 5.2 We induct on the bridge number b of the bridge trisection. When $b = 1$ or $b = 2$, there is an easy classification of b-bridge trisections [15, Subsection 4.3], which we take as the base case. Assume the theorem holds when the bridge number is less than b, and let $\mathfrak{T}$ be a $(b; c_1, c_2, c_3)$-bridge trisection. Assume without loss of generality that $c_3 = b - 1$.

Suppose that $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ are the three tangles comprising the spine of the bridge trisection. Every b-bridge splitting of a c-component unlink with $b > c$ is a perturbation of the standard c-bridge splitting of the c-component unlink, which is itself unique up to isotopy [15, Proposition 2.3]. It follows that there exist collections Δ_1 and Δ_3 of bridge disks for $\mathcal{T}_1$ and $\mathcal{T}_3$, respectively, such that the shadows $\Delta_1^* = \Delta_1 \cap \Sigma$ and $\Delta_3^* = \Delta_3 \cap \Sigma$ have the property that $\Delta_1^* \cup \Delta_3^*$ is an embedded collection of $b-2$ bigons and a single quadrilateral. Let α_0^* denote one of the arcs of Δ_1^* in the quadrilateral.

Let $L = \mathcal{T}_2 \cup \overline{\mathcal{T}}_3$, and let $\mathfrak{b}$ be the band for L that is framed by Σ and whose core is α_0^*. Then the data $(\Sigma, L, \mathfrak{b})$ encodes a banded b-bridge splitting, since the resolution $L_{\mathfrak{b}}$ is the unlink $L' = \mathcal{T}_2 \cup \overline{\mathcal{T}}_1$. (Here, we think of $\mathfrak{b}$ as being slightly perturbed to lie in the 3-ball containing $\mathcal{T}_3$.) We refer the reader to Section 3 of [15], especially Lemma 3.3, for more details about banded bridge splittings and how they arise from bridge trisections.

Assume without loss of generality that $c_2 = |L|$ is greater than or equal to $c_1 = |L'|$. We break the remainder of the proof into two cases: Either $c_2 > c_1$ or $c_2 = c_1$. Note that since there is only one band present, we must have $c_2 - c_1 \leq 1$. The proofs of the two cases are very similar, except that we apply [19, Theorem 1.1] in the first case and [1, Theorem 1.3] in the second.

Case 1 If $c_2 = c_1 + 1$, then $\mathfrak{b}$ connects distinct components K_1 and K_2 of L. Let K' denote the component of L' obtained as the resolution $(K_1 \cup K_2)_{\mathfrak{b}}$. We now translate this setup into the notation of [19, Section 1]. Let $N = \overline{\nu(K_1 \cup \mathfrak{b} \cup K_2)}$, a genus two handlebody, and let $M = S^3 \setminus \nu(L \setminus (K_1 \sqcup K_2))$. Let E_1 denote the spanning disk bounded by K_1. Let $P' = \partial\nu(E_1)$, a 2-sphere disjoint from $K_1 \sqcup K_2$ in M. Let Q' denote a spanning disk bounded by K' in M. Let $P = \overline{P' \setminus N}$, and let $Q = \overline{Q' \setminus N}$.

It is clear from this setup that $P \cap \partial N$ is a collection of m parallel separating curves A_m for some odd m, since P' was disjoint from K_1 and K_2, but intersects $\mathfrak{b}$ transversely; see [19, Figure 1]. Similarly, $Q \cap \partial N$ agrees with the curves B_n, since $\partial Q' = K'$ and Q' may crash through $\mathfrak{b}$ in arcs parallel to its core. Thus, M, N, P, and Q satisfy the hypotheses of [19, Theorem 1.1]. The relevant conclusion is that A_1 and B_0 bound embedded disks E and F in $\overline{M \setminus N}$ that intersect in a single arc; compare with the proof of [19, Main Theorem].

Translating this conclusion back into the setting of interest, the disk E is properly embedded in $S^3 \setminus \nu(\mathfrak{b})$ and F is a spanning disk for K'. This implies that the pair $(B^3, T) = (S^3, L) \setminus (\nu(\mathfrak{b}), \nu(L \cap \mathfrak{b}))$ is the split union of a trivial tangle and an unlink: the strands of the trivial tangle are parallel into pushoffs of E via the components of $F \setminus \nu(E)$, at which point they are parallel into $\partial\nu(\mathfrak{b})$ via the pushoffs of E.

The bridge sphere Σ induces a bridge splitting (B^3, T). By [24, Theorem 2.2], Σ is either minimal for (B^3, T) or perturbed.[1] If the splitting were minimal, we would have $b = c_2$, so $\mathfrak{T}$ would be completely decomposable by Proposition 5.1. If the splitting is perturbed, then $\mathfrak{T}$ is perturbed, since each bridge arc of $\mathcal{T}_3$ that is disjoint from $\nu(\mathfrak{b})$ is a strand of a 1-bridge splitting of a component of $L_3 = \mathcal{T}_3 \cup \overline{\mathcal{T}}_1$. After deperturbing $\mathfrak{T}$, we find that $\mathfrak{T}$ is completely decomposable, by the inductive hypothesis.

Case 2 If $c_2 = c_1$, then $\mathfrak{b}$ connects a component K of L to itself. Let $K' = K_{\mathfrak{b}}$. We now translate this setup into the notation of [1, Section 1], abbreviating the discourse where it is overly repetitive of the previous case. Let $M = S^3 \setminus \nu(L \setminus K)$, and let $N = \nu(K \cup \mathfrak{b})$. Let P' be a spanning disk bounded by K in M, and let Q' be a spanning disk bounded by K' in M. Let $P = \overline{P' \setminus N}$, and let $Q = \overline{Q' \setminus N}$.

It is clear from the setup that the hypotheses of [1, Theorem 1.3] are satisfied, so we can conclude that some A_0 and B_0 bound embedded disks E_P and E_Q, respectively, in $\overline{M \setminus N}$. Moreover, there is a properly embedded disk D in $\overline{M \setminus N}$, disjoint from E_P and E_Q, that runs once over one of the handles of N and is disjoint from the other handle. We can extend E_P to a spanning disk F for K; compare with the proof of [1, Theorem 1.8].

The strands of $K \setminus \nu(\mathfrak{b})$ are parallel into pushoffs of D via the components of $E_P \setminus \nu(D)$, at which point they are parallel into $\partial\nu(\mathfrak{b})$ via the pushoffs of D. It follows that the tangle $(B^3, T) = (S^3, L) \setminus (\nu(\mathfrak{b}), \nu(\mathfrak{b} \cap K))$ is the split union of a trivial tangle and an unlink, and Σ gives rise to a bridge splitting of (B^3, T). As before, this splitting is either minimal or perturbed. The case that the splitting is perturbed has the same consequence as in Case 1 above.

If the splitting is minimal, then it is a split union of a 2-bridge splitting of the trivial tangle and a $(b-2)$-bridge splitting of an unlink. It follows that the bridge trisection is a split union: $\mathfrak{T} = \mathfrak{T}' \sqcup \mathfrak{T}''$, where $\mathfrak{T}'$ is a $(2, 1)$-bridge trisection (of a projective plane, necessarily), and $\mathfrak{T}''$ is a $(b-2; c_1-1, c_2-1, b-2)$-bridge trisection (of an unlink of 2-spheres, necessarily). The latter is completely decomposable by Proposition 5.1. □

We can also use Theorem 5.2 to understand surface-links with particular banded link presentations, where a *banded link presentation* (L, v) consists of an unlink $L \subset S^3$ and a collection of bands v such that the resolution L_v of L along v is also an unlink. Every banded link presentation gives rise to a surface $\mathcal{S}$ in S^4, and conversely, every surface-link $\mathcal{S}$ in S^4 can be presented by a banded link [10].

In [15, Section 3], the authors introduced the notion of *banded bridge splitting* of (L, v), a bridge splitting of L such that the bands v are isotopic into the bridge sphere with the surface framing and are dual to a collection of bridge disks on one side. They showed that $(S^4, \mathcal{S})$ admits a $(b; \boldsymbol{c})$-bridge trisection if

[1] Although [24, Theorem 2.2], as stated, applies to a closed 3-manifold M and a link K in M, a verbatim proof establishes the more general case where the 3-manifold M is replaced by a punctured 3-manifold and the link K is a tangle.

and only if a banded link presentation (L, v) of $\mathcal{S}$ admits a banded b-bridge splitting such that $|L| = c_1$, $|v| = b - c_2$, and $|L_v| = c_3$. As a corollary to Theorem 5.2, we obtain the following, which states, in essence, that a surface is unknotted if the bands are attached in a relatively simple way to the maxima or minima disks.

Corollary 5.3 *Suppose a surface-link $\mathcal{S}$ in S^4 is presented by a banded link (L, v) with a banded b-bridge splitting such that $b = |L| + 1$ or $b = |L_v| + 1$. Then $\mathcal{S}$ is an unlink of 2-spheres or an unlink of 2-spheres and an unknotted projective plane.*

The corollary exploits a feature of trisection theory called *handle triality*: If (L, v) admits a banded bridge splitting as in the corollary, then it admits a $(b, \boldsymbol{c})$-bridge trisection such that $c_1 = b - 1$ or $c_3 = b - 1$. By the three-fold symmetry of the trisection setup, we can extract a different banded link presentation with a single band, as in the proof of Theorem 5.2, and now we rely on known results about surface-links built with a single band to classify $\mathcal{S}$. The result can be interpreted as an analog for knotted surfaces of [13, Theorem 1.2].

References

[1] **S Bleiler**, **M Scharlemann**, *A projective plane in $\mathbb{R}^4$ with three critical points is standard: strongly invertible knots have property P*, Topology 27 (1988) 519–540 MR Zbl

[2] **J S Carter**, **M Saito**, *A Seifert algorithm for knotted surfaces*, Topology 36 (1997) 179–201 MR Zbl

[3] **T Cochran**, *Ribbon knots in S^4*, J. Lond. Math. Soc. 28 (1983) 563–576 MR Zbl

[4] **I Dai**, **M Miller**, *The 0-concordance monoid admits an infinite linearly independent set*, Proc. Amer. Math. Soc. 151 (2023) 3601–3609 MR Zbl

[5] **D Gay**, **R Kirby**, *Trisecting 4-manifolds*, Geom. Topol. 20 (2016) 3097–3132 MR Zbl

[6] **H Gluck**, *The embedding of two-spheres in the four-sphere*, Trans. Amer. Math. Soc. 104 (1962) 308–333 MR Zbl

[7] **C M Gordon**, *Knots in the 4-sphere*, Comment. Math. Helv. 51 (1976) 585–596 MR Zbl

[8] **C M Gordon**, **R A Litherland**, *On the signature of a link*, Invent. Math. 47 (1978) 53–69 MR Zbl

[9] **J Joseph**, **J Meier**, **M Miller**, **A Zupan**, *Bridge trisections and classical knotted surface theory*, Pacific J. Math. 319 (2022) 343–369 MR Zbl

[10] **A Kawauchi**, **T Shibuya**, **S Suzuki**, *Descriptions on surfaces in four-space, I: Normal forms*, Math. Sem. Notes Kobe Univ. 10 (1982) 75–125 MR Zbl

[11] **M Kobayashi**, **T Kobayashi**, *On canonical genus and free genus of knot*, J. Knot Theory Ramifications 5 (1996) 77–85 MR Zbl

[12] **C Livingston**, *Surfaces bounding the unlink*, Michigan Math. J. 29 (1982) 289–298 MR Zbl

[13] **J Meier**, **T Schirmer**, **A Zupan**, *Classification of trisections and the generalized property R conjecture*, Proc. Amer. Math. Soc. 144 (2016) 4983–4997 MR Zbl

[14] **J Meier**, **A Thompson**, **A Zupan**, *Cubic graphs induced by bridge trisections*, Math. Res. Lett. 30 (2023) 1207–1231 MR Zbl

[15] **J Meier**, **A Zupan**, *Bridge trisections of knotted surfaces in S^4*, Trans. Amer. Math. Soc. 369 (2017) 7343–7386 MR Zbl

[16] **J Meier**, **A Zupan**, *Bridge trisections of knotted surfaces in* 4*-manifolds*, Proc. Natl. Acad. Sci. USA 115 (2018) 10880–10886 MR Zbl

[17] **K Miyazaki**, *On the relationship among unknotting number, knotting genus and Alexander invariant for* 2*-knots*, Kobe J. Math. 3 (1986) 77–85 MR Zbl

[18] **Y Moriah**, *On the free genus of knots*, Proc. Amer. Math. Soc. 99 (1987) 373–379 MR Zbl

[19] **M Scharlemann**, *Smooth spheres in $\mathbb{R}^4$ with four critical points are standard*, Invent. Math. 79 (1985) 125–141 MR Zbl

[20] **C M Tsau**, *A note on incompressible surfaces in solid tori and in lens spaces*, from "Knots 90", de Gruyter, Berlin (1992) 213–229 MR Zbl

[21] **R Weidmann**, *On the rank of amalgamated products and product knot groups*, Math. Ann. 312 (1998) 761–771 MR Zbl

[22] **T Yanagawa**, *On ribbon* 2*-knots: the* 3*-manifold bounded by the* 2*-knots*, Osaka Math. J. 6 (1969) 447–464 MR Zbl

[23] **E C Zeeman**, *Twisting spun knots*, Trans. Amer. Math. Soc. 115 (1965) 471–495 MR Zbl

[24] **A Zupan**, *Bridge and pants complexities of knots*, J. Lond. Math. Soc. 87 (2013) 43–68 MR Zbl

Department of Mathematics, North Carolina School of Science and Mathematics
Morganton, NC, United States

Department of Mathematics, Western Washington University
Bellingham, WA, United States

Department of Mathematics, University of Texas at Austin
Austin, TX, United States

Department of Mathematics, University of Nebraska–Lincoln
Lincoln, NE, United States

jason.joseph@ncssm.edu, jeffrey.meier@wwu.edu, maggie.miller.math@gmail.com, zupan@unl.edu

Received: 4 November 2022 Revised: 28 August 2023

Algebraic & Geometric Topology **25**:3 (2025) 1523–1544
DOI: 10.2140/agt.2025.25.1523
Published: 20 June 2025

Random Artin groups

ANTOINE GOLDSBOROUGH
NICOLAS VASKOU

We introduce a new model of random Artin groups. The two variables we consider are the rank of the Artin groups and the set of permitted coefficients of their defining graphs.

The heart of our model is to control the speed at which we make that set of permitted coefficients grow relatively to the growth of the rank of the groups, as it turns out different speeds yield very different results. We describe these speeds by means of (often polynomial) functions. In this model, we show that for a large range of such functions, a random Artin group satisfies most conjectures about Artin groups asymptotically almost surely.

Our work also serves as a study of how restrictive the commonly studied families of Artin groups are, as we compute explicitly the probability that a random Artin group belongs to various families of Artin groups, such as the classes of 2-dimensional Artin groups, FC-type Artin groups, large-type Artin groups, and others.

20F36, 20F65, 20F69, 20P05; 20F67

1 Introduction

Artin groups are a family of groups that have drawn an increasing interest in the past few decades. They are defined as follows. Let Γ be a *defining graph*, that is a simplicial graph with vertex set $V(\Gamma)$ and edge set $E(\Gamma)$, such that every edge e_{ab} of Γ connecting two vertices a and b is given a coefficient $m_{ab} \in \{2, 3, \dots\}$. Then Γ defines an *Artin group*:

$$A_\Gamma := \langle V(\Gamma) \mid \underbrace{aba\cdots}_{m_{ab}\text{ terms}} = \underbrace{bab\cdots}_{m_{ab}\text{ terms}}, \forall e_{ab} \in E(\Gamma)\rangle.$$

The cardinality of $V(\Gamma)$, that is the number of *standard generators* of A_Γ, is called the *rank* of A_Γ. When a and b are not connected by an edge we set $m_{ab} := \infty$.

One of the main reasons why Artin groups have become of such great interest is because of the amount of (often easily stated) conjectures and problems about them that are still to be solved. While some of these conjectures are algebraic (torsion, centres), some others are more geometric (acylindrical hyperbolicity, CAT(0)-ness), algorithmic (word and conjugacy problems, biautomaticity), or even topological. Although close to none of these conjectures or problems has been answered in the most general case, there has been progress on each of them. A common theme towards proving these conjectures has been to prove them for smaller families of Artin groups.

The goal of this paper is to consider Artin groups with a probabilistic approach. One might wonder what a typical Artin group looks like, and hence want to define a notion of randomness for Artin groups. By computing the different "sizes" of the most commonly studied classes of Artin groups, we give a way to quantify how restrictive these different classes really are. In light of that, our model provides a novel and explicit way of quantifying the state of the common knowledge about the aforementioned conjectures and problems about Artin groups.

Although Artin groups are defined using defining graphs, it is not known in general when two defining graphs give rise to isomorphic Artin groups. This problem, known as the *isomorphism problem*, is actually quite hard to solve even for restrictive classes of Artin groups. With our current knowledge, any (reachable) theory of randomness for Artin groups must then be based on the randomness of defining graphs, and not of the Artin groups themselves.

Random right-angled Coxeter (and Artin) groups have been studied by several authors in the literature (see Behrstock, Hagen and Sisto [1] and Charney and Farber [4]), using the Erdős–Rényi model. While in [4] the authors fix the probability of apparition of an edge as some constant $0 \leq p \leq 1$, in [1] this model is refined: $p = p(N)$ depends on the rank N of the group. That said, these models restrict to right-angled groups, where the associated defining graphs are not labelled. In [7], Deibel introduces a model of randomness for Coxeter groups in general. There are similarities between this model and ours, although the former revolves more about making the probabilities of apparition of specific coefficients vary. In particular, this model is not very well suited to provide insights on the "sizes" of the most commonly studied classes of Coxeter and Artin groups. On the contrary, this is a central goal of our model.

The two variables that come to mind when thinking about Artin groups are their rank, that is the number of vertices of the defining graph, as well as the choice of the associated coefficients. A first step in the theory is to consider what happens if we restrict ourselves to the family $\mathcal{G}^{N,M}$ of all the defining graphs with N vertices and with coefficients in $\{\infty, 2, 3, \dots, M\}$, for some $N \geq 1$ and $M \geq 2$. As we want any possible rank and any possible coefficient to eventually appear in a random Artin group, a convenient way to think about randomness is to pick a defining graph at random in the family $\mathcal{G}^{N,M}$, and then to make N and M grow to infinity. Note that isomorphic labelled graphs may be counted multiples times in $\mathcal{G}^{N,M}$.

As it turns out, randomness of defining graphs highly depends on the speed at which N and M grow. A prime example of this is that the probability for a defining graph of $\mathcal{G}^{N,M}$ to give an Artin group of large-type (meaning that none of the coefficients is 2) tends to 1 when M grows much faster than N, and tends to 0 when N grows much faster than M. To solve this problem, we decide to relate N and M through a function f such that $M := f(N)$. This way, we only have to look at the family $\mathcal{G}^{N,f(N)}$ when N goes to infinity.

If $A_{\mathcal{F}}$ is a family of Artin groups coming from a family of defining graphs $\mathcal{F}$, a way of measuring the "size" of $A_{\mathcal{F}}$ is to compute the limit

$$\lim_{N\to\infty} \frac{\#(\mathcal{F} \cap \mathcal{G}^{N,f(N)})}{\#(\mathcal{G}^{N,f(N)})}.$$

Of course, this ratio depends on the choice we make for the function f. When the above limit is 1, that is when the probability that a graph picked at random in $\mathcal{G}^{N,f(N)}$ will give an Artin group that belongs to the said family $A_{\mathcal{F}}$ tends to 1, we say that a random Artin group (with respect to f) is *asymptotically almost surely* in $A_{\mathcal{F}}$.

One may wonder why our model only considers graphs of rank N, and not all graphs with rank at most N. As it turns out, the size of the set of all graphs with at most N vertices (and coefficients in $\{\infty, 2, \dots, f(N)\}$) is asymptotically the same as the size of $\mathcal{G}^{N,f(N)}$, in the sense that the quotient of the two values tends to 1 when N approaches ∞. Thus asymptotically it is not an actual restriction to only consider graphs with precisely N vertices.

Now, there are families $A_{\mathcal{F}}$ of Artin groups for which the above limit tends to 1 no matter what (sensible) choice we make for the function f. We say that such a family is *uniformly large* (resp. *uniformly small* if that limit is always 0). Our first result concern such families of Artin groups:

Theorem 1.1 *The family of irreducible Artin groups and the family of Artin groups with connected defining graphs are uniformly large. On the other hand, the family of Artin groups of type FC is uniformly small. In particular, the same applies to the families of RAAGs and triangle-free Artin groups.*

As mentioned earlier, there are numerous families of Artin groups whose "size" depends on the choice of function f. When f is large enough, which means that the choice of possible coefficients for the defining graphs grows fast enough compared to the rank of the Artin group, we obtain much stronger results. This is made explicit in the next two theorems.

For two nondecreasing divergent functions $f, g\colon \mathbb{N} \to \mathbb{N}$ we say that $f \preccurlyeq g$ if the limit

$$\lim_{N\to\infty} \frac{f(N)}{g(N)}$$

exists and is finite. If $f \preccurlyeq g$ and $f \succcurlyeq g$ then we will write $f \simeq g$. Finally if $f \preccurlyeq g$ but $f \not\simeq g$ then we will write $f \prec g$, and similarly for $f \succ g$.

Theorem 1.2 *Let $A_{\mathcal{F}}$ be any family of Artin groups defined by forbidding a finite number of coefficients from their defining graphs, and consider a function $f\colon \mathbb{N} \to \mathbb{N}$. Let Γ be a graph picked at random in $\mathcal{G}^{N,f(N)}$.*

1. *If $f(N) \succ N^2$, then A_Γ asymptotically almost surely belongs to $A_{\mathcal{F}}$.*
2. *If $f(N) \prec N^2$, then A_Γ asymptotically almost surely does not belong to $A_{\mathcal{F}}$.*
3. *If $f(N) \simeq N^2$, then the probability that A_Γ belongs to $A_{\mathcal{F}}$ is strictly between 0 and 1.*

Note that the previous theorem applies to the families of large-type, extra-large-type, or large-type and free-of-infinity Artin groups. There are strong results in the literature about these families of Artin groups, as most of the famous conjectures and problems about Artin groups have been solved for at least one of them (see Section 2).

While these different families of Artin groups have the same threshold at $f(N) \simeq N^2$ no matter how many coefficients we forbid, the class of 2-dimensional Artin groups turns out to be substantially bigger. Studying this class, we obtain the following result:

Theorem 1.3 *Consider a nondecreasing divergent function $f\colon \mathbb{N} \to \mathbb{N}$. Let Γ be a graph picked at random in $\mathcal{G}^{N,f(N)}$.*

1. *If $f(N) \succ N^{3/2}$, then A_Γ asymptotically almost surely is 2-dimensional.*
2. *If $f(N) \prec N^{3/2}$, then A_Γ asymptotically almost surely is not 2-dimensional.*

A consequence of the two previous theorems is that we are able, when f grows fast enough, to show that a random Artin group asymptotically almost surely satisfies most of the main conjectures about Artin groups:

Theorem 1.4 *Let $f\colon \mathbb{N} \to \mathbb{N}$ be such that $f(N) \succ N^{3/2}$, and let Γ be a graph picked at random in $\mathcal{G}^{N,f(N)}$. Then asymptotically almost surely, the following properties hold:*

1. *A_Γ is torsion-free;*
2. *A_Γ has trivial centre;*
3. *A_Γ has solvable word and conjugacy problem;*
4. *A_Γ satisfies the $K(\pi,1)$-conjecture;*
5. *the set of parabolic subgroups of A_Γ is closed under (arbitrary) intersections;*
6. *A_Γ is acylindrically hyperbolic;*
7. *A_Γ satisfies the Tits alternative;*
8. *A_Γ is not virtually cocompactly cubulated.*

Moreover, if $f(N) \succ N^2$ then asymptotically almost surely the following properties also hold:

1. *A_Γ is CAT(0);*
2. *A_Γ is hierarchically hyperbolic;*
3. *A_Γ is systolic and thus biautomatic;*
4. *A_Γ is rigid;*
5. $\mathrm{Aut}(A_\Gamma) \cong A_\Gamma \rtimes \mathrm{Out}(A_\Gamma)$, *where* $\mathrm{Out}(A_\Gamma) \cong \mathrm{Aut}(\Gamma) \times (\mathbb{Z}/2\mathbb{Z})$ *is finite.*

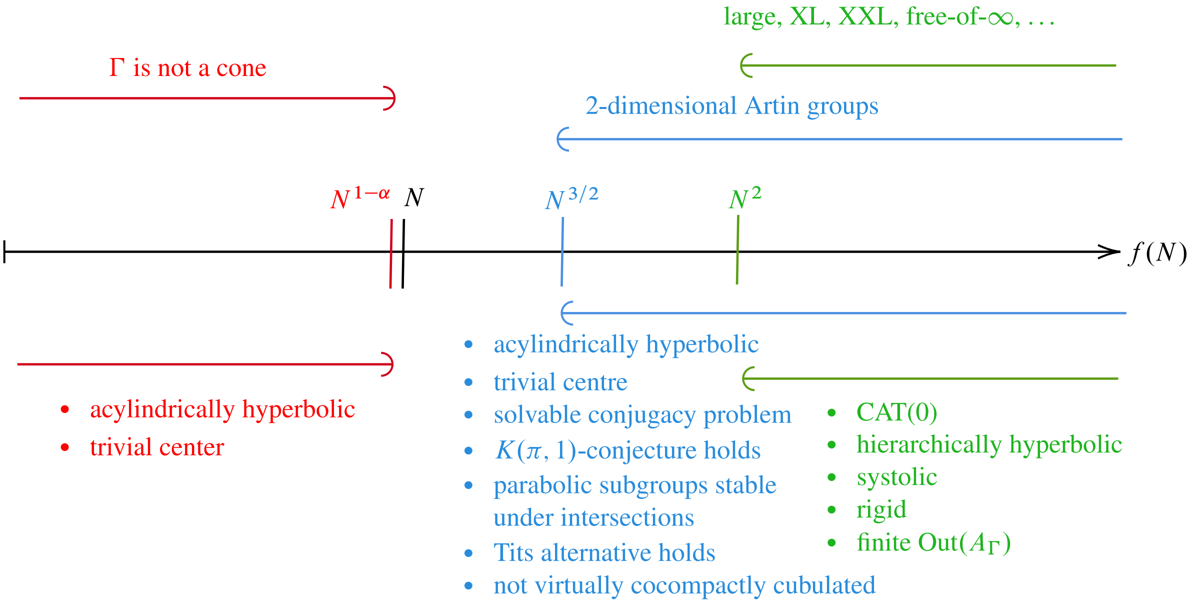

large, XL, XXL, free-of-∞, . . .
Γ is not a cone
2-dimensional Artin groups
$N^{1-\alpha}$
N
$N^{3/2}$
N^2
$f(N)$
acylindrically hyperbolic
trivial center
acylindrically hyperbolic
trivial centre
solvable conjugacy problem
$K(\pi,1)$-conjecture holds
parabolic subgroups stable under intersections
Tits alternative holds
not virtually cocompactly cubulated
CAT(0)
hierarchically hyperbolic
systolic
rigid
finite $\mathrm{Out}(A_\Gamma)$

Figure 1: The axis represents various (polynomial) functions f. Above the main axis are described the classes of Artin groups that we obtain asymptotically almost surely with respect to f, while under this axis we list the properties that we know these groups will satisfy asymptotically almost surely.

At last, we also prove interesting results for families of Artin groups in which the number M of permitted coefficients grows “slowly enough” compared to the rank N. We focus on the class of Artin groups A_Γ whose associated graphs Γ are not cones, and we prove that for most (nondecreasing divergent) functions, the probability that a random Artin group is acylindrically hyperbolic and has trivial centre tends to 1.

Theorem 1.5 *Let $\alpha \in (0,1)$ and let $f : \mathbb{N} \to \mathbb{N}$ be a nondecreasing divergent function satisfying $f(N) \prec N^{1-\alpha}$. Let now Γ be a graph picked at random in $\mathcal{G}^{N,f(N)}$. Then the associated Artin group A_Γ is acylindrically hyperbolic and has trivial centre asymptotically almost surely.*

The results of the above theorems for polynomial functions is encapsulated in Figure 1.

The previous results shows that we are very close to being able to state that “almost all Artin groups are acylindrically hyperbolic and have trivial centres”. It is conjectured that all irreducible nonspherical Artin groups are acylindrically hyperbolic; see Charney and Morris-Wright [5]. Although proving this conjecture for all Artin groups seems to be a difficult problem, some progress has been made in recent years; see Kato and Oguni [13] and Vaskou [17]. It would seem to be an interesting line of research to try to expand the spectrum of families of Artin groups for which one can prove acylindrical hyperbolicity, in order to “fill in” the gap of functions at which a random Artin group is acylindrically hyperbolic. This leads to the following question.

Question 1.6 Construct a family $A_{\mathscr{F}}$ of acylindrically hyperbolic Artin groups or of Artin groups with trivial centres for which the following holds:

There exists an $\alpha \in (0,1)$ such that for all functions $f : \mathbb{N} \to \mathbb{N}$ satisfying $N^{1-\alpha} \preccurlyeq f(N) \preccurlyeq N^{3/2}$, a graph Γ picked at random in $\mathscr{G}^{N,f(N)}$ is such that A_Γ asymptotically almost surely belongs to $A_{\mathscr{F}}$.

Acknowledgements

The authors would like to thank Alessandro Sisto for helpful conversations and for pointing out the strategy of the second moment method, and Alexandre Martin for useful comments. We would also like to thank the Institut Henri Poincaré (UAR 839 CNRS-Sorbonne Université), and LabEx CARMIN (ANR-10-LABX-59-01) for providing a warm environment for this project to begin. We thank the referee for carefully reading the article and for their detailed comments. The work of Goldsborough was supported by the EPSRC-UKRI studentship EP/V520044/1.

2 Preliminaries and first results

In this section we bring more details about some of the notions discussed in the introduction. This includes discussions about most of the commonly studied classes of Artin groups, as well as discussions regarding open conjectures related to Artin groups.

Throughout this paper, we will often call a *triangle* in a graph Γ any subgraph of Γ that is generated by 3 vertices. This notion will be convenient, although one must note that with this definition, triangles may have strictly fewer than 3 edges, as subgraphs of Γ.

Most of the main conjectures about Artin groups are still open in general. That said, many of them have been proved for smaller families of Artin groups. Two important of these families are the families of 2-dimensional Artin groups and the family of Artin groups of type FC. These two families have been extensively studied following the work of Charney and Davis [3]. The other well-studied families are usually subfamilies of these.

Before coming to these definitions, we first recall what a parabolic subgroup of an Artin group is. Let A_Γ be any Artin group, and let Γ' be a full subgraph of Γ. A standard result about Artin groups states that the subgroup of A_Γ generated by the vertices of Γ' is also an Artin group, that is isomorphic to $A_{\Gamma'}$ [14]. Such a subgroup is called a *standard parabolic subgroup* of A_Γ. The conjugates of these subgroups are called the *parabolic subgroups* of A_Γ.

Definition 2.1 (0) An Artin group A_Γ is said to be *spherical* if the associated Coxeter group W_Γ is finite.

(1) An Artin group A_Γ is said to be 2*-dimensional* if for every triplet of distinct standard generators $a, b, c \in V(\Gamma)$, the subgraph Γ' spanned by a, b and c corresponds to an Artin group $A_{\Gamma'}$ that is *not* spherical. By a result of [3], this is equivalent to requiring that

$$\frac{1}{m_{ab}} + \frac{1}{m_{ac}} + \frac{1}{m_{bc}} \leq 1.$$

We let $\mathscr{D}$ be the set of graphs Γ such that the above condition is satisfied. We let $A_{\mathscr{D}}$ be the set of 2-dimensional Artin groups. The family of 2-dimensional Artin groups contains the well-studied families of *large-type* Artin groups (every coefficient is at least 3), *extra-large-type* Artin groups (every coefficient is at least 4), or *XXL* Artin groups (every coefficient is at least 5).

(2) An Artin group A_Γ is said to be of *type FC* if every complete subgraph $\Gamma' \subseteq \Gamma$ generates an Artin group $A_{\Gamma'}$ that is spherical. Let $\mathscr{FC}$ be the set of graphs Γ that give rise to an Artin group of type FC and let $A_{\mathscr{FC}}$ be the set of Artin groups of type FC.

The family of Artin groups of type FC contains the family of right-angled Artin groups, also called *RAAGs* (the only permitted coefficients are 2 and ∞), the family of spherical Artin groups, and the family of *triangle-free* Artin groups (the Artin groups whose associated graphs don't contain any 3-cycles). Being triangle-free is actually equivalent to being both of type FC and 2-dimensional.

We now move towards the main conjectures related to Artin groups. For each conjecture, we will briefly describe the state of the common research towards proving it, by mentioning the one or two result(s) that will turn out to be the more "probabilistically relevant" in our model — in other words, the results that cover the largest classes.

Conjecture 2.2 Let A_Γ be any Artin group. Then:

(1) A_Γ is torsion-free.
↪ This was proved for 2-dimensional Artin groups [3].

(2) If A_Γ is irreducible and nonspherical, then A_Γ has trivial centre.
↪ This was proved for 2-dimensional Artin groups [17], and for Artin groups whose graph is not the cone of a single vertex [5].

(3) A_Γ has solvable word and conjugacy problems.
↪ This was proved for 2-dimensional Artin groups [11].

(4) A_Γ satisfies the $K(\pi, 1)$-conjecture.
↪ This was proved for 2-dimensional Artin groups [3].

(5) Intersections of parabolic subgroups of A_Γ give parabolic subgroups of A_Γ.
↪ This was proved for large-type Artin groups [6] and more generally for $(2, 2)$-free 2-dimensional Artin groups [2].

(6) A_Γ is CAT(0).
↪ This was proved for XXL Artin groups [9].

(7) If A_Γ is irreducible and nonspherical, then A_Γ is acylindrically hyperbolic.
$\hookrightarrow$ This was proved for 2-dimensional Artin groups [17], and for Artin groups whose graph is not the cone of a single vertex [13].

(8) A_Γ is hierarchically hyperbolic.
$\hookrightarrow$ This was proved for extra-large-type Artin groups [10].

(9) A_Γ is systolic and biautomatic.
$\hookrightarrow$ This was proved for large-type Artin groups [12].

(10) A_Γ satisfies the Tits alternative.
$\hookrightarrow$ This was proved for 2-dimensional Artin groups [15].

In addition to these conjectures, the following questions have been raised:

Question 2.3 Let A_Γ be any Artin group.

(1) When is A_Γ not virtually cocompactly cubulated?
$\hookrightarrow$ This was proved to be the case when A_Γ is 2-dimensional and satisfies the condition of [8, Conjecture B].

(2) When is $\mathrm{Out}(A_\Gamma)$ finite?
$\hookrightarrow$ This was proved to be the case for large-type free-of-infinity Artin groups [18].

(3) When is A_Γ rigid, in the sense of [16]?
$\hookrightarrow$ This was proved to be the case for large-type Artin groups that have no separating edges [16, Theorem B]. This includes the class of large-type free-of-infinity Artin groups.

Definition 2.4 Let $\mathscr{F}$ be a family of defining graphs and let $A_{\mathscr{F}}$ be the corresponding class of Artin groups. Let $f:\mathbb{N}\to\mathbb{N}$ be a nondecreasing divergent function. We say that a random Artin group (with respect to f) A_Γ belongs to $A_{\mathscr{F}}$ with probability

$$\mathbb{P}_f[A_\Gamma \in A_{\mathscr{F}}] := \lim_{N\to\infty} \mathbb{P}[\Gamma \in \mathscr{F} \mid \Gamma \in \mathscr{G}^{N,f(N)}] = \lim_{N\to\infty} \frac{\#(\mathscr{F}\cap\mathscr{G}^{N,f(N)})}{\#(\mathscr{G}^{N,f(N)})},$$

when the limit exists. Furthermore, we say that a random Artin group A_Γ (with respect to f) is *asymptotically almost surely* in $A_{\mathscr{F}}$ if $\mathbb{P}_f[A_\Gamma \in A_{\mathscr{F}}]=1$. Similarly, we say that A_Γ is *asymptotically almost surely not* in $A_{\mathscr{F}}$ if $\mathbb{P}_f[A_\Gamma \in A_{\mathscr{F}}]=0$.

Definition 2.5 Let $A_{\mathscr{F}}$ be a family of Artin groups. Then we say that $A_{\mathscr{F}}$ is *uniformly large* if for every nondecreasing divergent function $f:\mathbb{N}\to\mathbb{N}$, a random Artin group A_Γ (with respect to f) is asymptotically almost surely in $A_{\mathscr{F}}$. We say that $\mathscr{F}$ is *uniformly small* if A_Γ is asymptotically almost surely not in $A_{\mathscr{F}}$.

We now move towards our first results. The first thing we will prove is that the family of irreducible Artin groups and the family of Artin groups with connected defining graphs are uniformly large. This is important as many results regarding Artin groups assume that the corresponding groups are irreducible and/or have a connected defining graph. Our work shows that these two hypotheses are very much not restrictive.

Definition 2.6 Let Γ_1 and Γ_2 be two defining graphs. The graph $\Gamma_1 *_k \Gamma_2$ is the graph obtained by attaching every vertex of Γ_1 to every vertex of Γ_2 by an edge with label k (with $k \in \{\infty, 2, 3, \dots\}$).

Let now Γ be any defining graph. Then Γ is called a *k-join* relative to Γ_1 and Γ_2 if there are two subgraphs $\Gamma_1, \Gamma_2 \subseteq \Gamma$ such that $V(\Gamma_1) \sqcup V(\Gamma_2) = V(\Gamma)$ and such that $\Gamma = \Gamma_1 *_k \Gamma_2$.

We will denote by $\mathcal{A}_{\mathcal{J}_k}$ the class of Artin groups whose defining graphs decompose as k-joins.

Remark 2.7 (1) If $\Gamma \in \mathcal{J}_2$ then A_Γ decomposes as a direct product $A_{\Gamma_1} \times A_{\Gamma_2}$ in an obvious way. In that case, Γ is called *reducible*. The class $\mathcal{J}_2^C$ of *irreducible* defining graphs will be denoted by Irr.

(2) If $\Gamma \in \mathcal{J}_\infty$ then it is disconnected. The class $\mathcal{J}_\infty^C$ of connected defining graphs will be denoted by Con.

Lemma 2.8 *For all $k \in \{\infty, 2, 3, \dots\}$, the family $\mathcal{A}_{\mathcal{J}_k}$ is uniformly small. In particular, the classes A_{Irr} and A_{Con} of Artin groups are both uniformly large.*

Proof We will count the number of decompositions of the graph Γ as $\Gamma = \Gamma_1 *_k \Gamma_2$. Without loss of generality, we will let Γ_1 denote the subgraph with the lower rank, so that $|V(\Gamma_1)| \leq \lfloor N/2 \rfloor$. Let $f \colon \mathbb{N} \to \mathbb{N}$ be a nondecreasing divergent function and consider the family $\mathcal{J}_k$. For a given $N \geq 1$,

$$\begin{aligned}
\mathbb{P}[\Gamma \in \mathcal{J}_k \mid \Gamma \in \mathcal{G}^{N,f(N)}] &= \mathbb{P}\big[\exists\, \Gamma_1, \Gamma_2 \text{ with } |V(\Gamma_1)| \leq N/2 \text{ such that } \Gamma = \Gamma_1 *_k \Gamma_2 \mid \Gamma \in \mathcal{G}^{N,f(N)}\big] \\
&\leq \sum_{j=1}^{\lfloor N/2 \rfloor} \mathbb{P}\big[\exists\, \Gamma_1, \Gamma_2 \text{ with } |V(\Gamma_1)| = j \text{ such that } \Gamma = \Gamma_1 *_k \Gamma_2 \mid \Gamma \in \mathcal{G}^{N,f(N)}\big] \\
&= \sum_{j=1}^{\lfloor N/2 \rfloor} \binom{N}{j} \left(\frac{1}{f(N)}\right)^{j(N-j)} \\
&\leq \sum_{j=1}^{\lfloor N/2 \rfloor} \left(\frac{Ne}{jf(N)^{N/2}}\right)^{j} \leq \frac{Ne}{f(N)^{N/2}} \cdot \left(\frac{1-(Ne/f(N)^{N/2})^{N/2+1}}{1-Ne/f(N)^{N/2}}\right)
\end{aligned}$$

where we used the bound

$$\binom{N}{j} \leq \left(\frac{Ne}{j}\right)^{j}.$$

Now $\lim_{N\to\infty} Ne/f(N)^{N/2} = 0$ for any nondecreasing divergent function f, so we obtain

$$\mathbb{P}_f[A_\Gamma \in A_{\mathcal{J}_k}] = \lim_{N\to\infty} \mathbb{P}[\Gamma \in \mathcal{J}_k \mid \Gamma \in \mathcal{G}^{N,f(N)}] = 0 \cdot \left(\frac{1-0}{1-0}\right) = 0.$$

This proves the main statement of the lemma. The second then directly follows from Remark 2.7. □

Our next result concerns the class of Artin groups of type FC.

Lemma 2.9 *The family $A_{\mathcal{FC}}$ of Artin groups of type FC is uniformly small. In particular, the family of triangle-free Artin groups, the family of spherical Artin groups and the family of RAAGs are also uniformly small.*

Proof Let f be any nondecreasing divergent function, and let $\Gamma \in \mathcal{G}^{N,f(N)}$. We want to bound the probability that Γ belongs to $\mathcal{FC} \cap \mathcal{G}^{N,f(N)}$. Let a, b and c be three vertices of Γ. The probability that any of the three corresponding coefficients m_{ab}, m_{ac} and m_{bc} is not 2 or ∞ is precisely $(f(N)-2)/f(N)$, and hence the probability that the three coefficients are not 2 nor ∞ is $((f(N)-2)/f(N))^3$. Note that when this happens, the subgraph $\Gamma' \subseteq \Gamma$ spanned by a, b and c is complete but generates an Artin group $A_{\Gamma'}$ which is nonspherical (the sum of the inverses of the three corresponding coefficients is ≤ 1). In particular, Γ is not of type FC. We obtain

$$\mathbb{P}_f[A_\Gamma \notin A_{\mathcal{FC}}] = \lim_{N\to\infty} \frac{\#(\mathcal{G}^{N,f(N)} \backslash \mathcal{FC})}{\#(\mathcal{G}^{N,f(N)})} \geq \lim_{N\to\infty} \left(\frac{f(N)-2}{f(N)}\right)^3 = \lim_{N\to\infty} \left(1-\frac{2}{f(N)}\right)^3 = 1. \qquad \square$$

As mentioned in the introduction, there are interesting classes of Artin groups for which the probability that a graph taken at random will belong to the class highly depends on the choice of function f. Some examples are given through the following theorem.

Theorem 2.10 *Let $\mathcal{F}$ be any family of graphs defined by forbidding a finite number k of coefficients and let $A_{\mathcal{F}}$ be the family of corresponding Artin groups. Consider a function $f : \mathbb{N} \to \mathbb{N}$. Let A_Γ be a random Artin group (with respect to f).*

1. *If $f(N) \succ N^2$, then A_Γ asymptotically almost surely belongs to $A_{\mathcal{F}}$.*
2. *If $f(N) \prec N^2$, then A_Γ asymptotically almost surely does not belong to $A_{\mathcal{F}}$.*
3. *If $f(N) \simeq N^2$ then asymptotically we have $\mathbb{P}_f[A_\Gamma \in A_{\mathcal{F}}] \in (0,1)$. Moreover, if $f(N) = \lambda N^2$ for some $\lambda > 0$, then $\mathbb{P}_f[A_\Gamma \in A_{\mathcal{F}}] = e^{-k/2\lambda}$.*

Proof A graph with N vertices has $\frac{1}{2}N(N-1)$ pairs of vertices, each of which is given one of $f(N)$ possible coefficients. Hence, direct computations on the possible number of graphs give

$$\#\mathcal{G}^{N,f(N)} = (f(N))^{\frac{N(N-1)}{2}}.$$

Similarly, we have

$$\#(\mathcal{F} \cap \mathcal{G}^{N,f(N)}) = (f(N)-k)^{\frac{N(N-1)}{2}}.$$

And thus we obtain

$$\begin{aligned}\mathbb{P}_f[A_\Gamma \in A_{\mathcal{F}}] = \lim_{N\to\infty} \frac{\#(\mathcal{F} \cap \mathcal{G}^{N,f(N)})}{\#(\mathcal{G}^{N,f(N)})} &= \lim_{N\to\infty} \left(\frac{f(N)-k}{f(N)}\right)^{\frac{N(N-1)}{2}} \\ &= \lim_{N\to\infty} \left(\frac{f(N)-k}{f(N)}\right)^{f(N)\left(\frac{N(N-1)}{2f(N)}\right)}.\end{aligned}$$

Observe that

$$\lim_{N\to\infty}\left(\frac{f(N)-k}{f(N)}\right)^{f(N)}=e^{-k}.$$

In particular, for any $\epsilon>0$ there is a big enough N_ϵ such that for all $N\geq N_\epsilon$ we have

$$e^{-k}-\epsilon\leq\left(\frac{f(N)-k}{f(N)}\right)^{f(N)}\leq e^{-k}+\epsilon.$$

Hence for $N\geq N_\epsilon$,

$$(e^{-k}-\epsilon)^{r(N)}\leq\left(\frac{f(N)-k}{f(N)}\right)^{f(N)\left(\frac{N(N-1)}{2f(N)}\right)}\leq(e^{-k}+\epsilon)^{r(N)},$$

where $r(N)=N(N-1)/(2f(N))$.

Therefore, if $f(N)\succ N^2$, there is a function h with $\lim_{N\to\infty}h(N)=\infty$ such that $f(N)=h(N)N^2$, and hence $r(N)=(N-1)/(2Nh(N))$ which tends to 0 as $N\to+\infty$. Thus, in this case

$$\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}]=\lim_{N\to\infty}\left(\frac{f(N)-k}{f(N)}\right)^{f(N)r(N)}=1.$$

If $f(N)\prec N^2$, there exists a function h with $\lim_{N\to\infty}h(N)=\infty$ such that $f(N)h(N)=N^2$, and here $r(N)=(N-1)h(N)/(2N)$ which tends to ∞ as $N\to\infty$, so in this case

$$\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}]=\lim_{N\to\infty}\left(\frac{f(N)-k}{f(N)}\right)^{f(N)r(N)}=0.$$

If $f(N)\simeq N^2$, then $\lim_{N\to\infty}f(N)/N^2$ is a nonzero constant and hence $\lim_{N\to\infty}r(N)=M$ for some constant $M>0$. Thus in this case,

$$\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}]=e^{-kM}.$$

Finally, if $f(N)=\lambda N^2$, we obtain $r(N)\to 1/(2\lambda)=:M$ and the result follows. □

The previous theorem has many consequences, as it can be applied to the families of large-type, extra-large-type, XXL or free-of-infinity Artin groups, for which much is known. Before stating an explicit result in Corollary 2.12, we prove the following small lemma:

Lemma 2.11 *Let $A\_\mathscr{F}$ and $A\_\mathscr{H}$ be two families of Artin groups, let $f:\mathbb{N}\to\mathbb{N}$ be a nondecreasing divergent function, and suppose that $\mathbb{P}_f[A_\Gamma\in A\_\mathscr{H}]=1$. Then*

$$\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}]=\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}\cap A\_\mathscr{H}].$$

Proof This is straightforward:

$$\begin{aligned}\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}]&=\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}\cap A\_\mathscr{H}]+\underbrace{\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}\cup A\_\mathscr{H}]}_{=1}-\underbrace{\mathbb{P}_f[A_\Gamma\in A\_\mathscr{H}]}_{=1}\\&=\mathbb{P}_f[A_\Gamma\in A\_\mathscr{F}\cap A\_\mathscr{H}].\end{aligned}$$ □

Corollary 2.12 *Let $f : \mathbb{N} \to \mathbb{N}$ be a function satisfying $f(N) \succ N^2$. Then a random Artin group A_Γ (with respect to f) satisfies any of the following properties asymptotically almost surely*:

(1) *A_Γ is CAT*(0);

(2) *A_Γ is hierarchically hyperbolic*;

(3) *A_Γ is systolic and biautomatic*;

(4) *A_Γ is rigid*;

(5) $\mathrm{Aut}(A_\Gamma) \cong A_\Gamma \rtimes \mathrm{Out}(A_\Gamma)$, *where* $\mathrm{Out}(A_\Gamma) \cong \mathrm{Aut}(\Gamma) \times (\mathbb{Z}/2\mathbb{Z})$ *is finite.*

Proof Let $A_{\mathcal{X}}$ be the class of XXL free-of-infinity Artin groups, and let $A_{\mathcal{L}} := A_{\mathrm{Irr}} \cap A_{\mathrm{Con}} \cap A_{\mathcal{X}}$. Using Lemmas 2.8 and 2.11 we can see that $\mathbb{P}_f[A_\Gamma \in A_{\mathcal{L}}] = \mathbb{P}_f[A_\Gamma \in A_{\mathcal{X}}]$. The class $A_{\mathcal{X}}$ has been defined as forbidding 4 coefficients from the defining graph; hence by Theorem 2.10 a random Artin group A_Γ (with respect to f) is asymptotically almost surely in $A_{\mathcal{X}}$ and therefore asymptotically almost surely in $A_{\mathcal{L}}$. The various results given in Conjecture 2.2 concern families of Artin groups that all contain the family $A_{\mathcal{L}}$. In particular, every Artin group in $A_{\mathcal{L}}$ satisfies the ten points of Conjecture 2.2. The results given in items (6), (8) and (9) of Conjecture 2.2 are precisely those needed for items (1), (2) and (3) of Corollary 2.12. Similarly, every Artin group in $A_{\mathcal{X}}$ is rigid, as per item (3) of Question 2.3. This proves point (4) of Corollary 2.12. For item (5), this follows from [18, Theorem A] which shows that this result holds for large-type free-of-infinity Artin groups, and in particular for Artin groups in the family $A_{\mathcal{X}}$. □

Remark 2.13 The previous corollary proves the five points in the second half of Theorem 1.4. Note that at this point, we could already prove the eight points in the first half of Theorem 1.4 for $f(N) \succ N^2$. We did not include this proof as it will be extended to all functions $f(N) \succ N^{3/2}$ in the following section.

3 Two-dimensional Artin groups

This section aims at studying from our probabilistic point of view the family of 2-dimensional Artin groups. This family is particularly important in the study of Artin groups, and many authors in the literature have obtained strong results for this class (see Conjecture 2.2).

Our goal will be to show that if $f(N) \succ N^{3/2}$ then asymptotically almost surely a random Artin group (with respect to f) will be 2-dimensional and if $f(N) \prec N^{3/2}$ then asymptotically almost surely a random Artin group (with respect to f) will not be 2-dimensional. In particular, we will be able to improve the result of Corollary 2.12, thus proving Theorem 1.4.

The condition of being 2-dimensional (see Definition 2.1(1)) is quite specific, which makes it hard to compute the "size" of the family. As it turns out, the size of this family is comparable to the size of another family of Artin groups, which is slightly easier to compute (see Lemma 3.2 and Theorem 3.3). This other family resembles the family introduced in [2]. We introduce it here:

Definition 3.1 We say an Artin group A_Γ is $(2,2)$-free if Γ does not have any two adjacent edges labelled by 2. We denote by $\mathcal{B}$ the set of graphs that do not have two adjacent edges labelled by 2. We define $A_{\mathcal{B}}$ to be the family of $(2,2)$-free Artin groups.

Recall that in Definition 2.1(1), we have defined the set of graphs $\mathcal{D}$ and the set of Artin groups $A_{\mathcal{D}}$. The following lemma is a key result. It will allow us to restrict to the study of $(2,2)$-free Artin groups, as asymptotically this family has the same size as the family $A_{\mathcal{D}}$.

Lemma 3.2 *For all nondecreasing divergent functions* $f\colon \mathbb{N}\to\mathbb{N}$,

- $\mathbb{P}_f[A_\Gamma \in A_{\mathcal{D}}] \le \mathbb{P}_f[A_\Gamma \in A_{\mathcal{B}}]$;
- *further, if* $f(N) \succ N$, *then* $\mathbb{P}_f[A_\Gamma \in \mathcal{A}_{\mathcal{D}}] = \mathbb{P}_f[A_\Gamma \in A_{\mathcal{B}}]$.

Proof The probability that a random Artin group A_Γ gives rise to a 2-dimensional Artin group can be found by conditioning on the event "$\Gamma \in \mathcal{B}$":

$$(*)\quad \mathbb{P}[\Gamma\in\mathcal{D}\mid\Gamma\in\mathcal{G}^{N,f(N)}] = \mathbb{P}\big[\Gamma\in\mathcal{D}\mid(\Gamma\in\mathcal{B})\cap(\Gamma\in\mathcal{G}^{N,f(N)})\big]\mathbb{P}\big[\Gamma\in\mathcal{B}\mid\Gamma\in\mathcal{G}^{N,f(N)}\big] + \mathbb{P}\big[\Gamma\in\mathcal{D}\mid(\Gamma\notin\mathcal{B})\cap(\Gamma\in\mathcal{G}^{N,f(N)})\big]\mathbb{P}\big[\Gamma\notin\mathcal{B}\mid\Gamma\in\mathcal{G}^{N,f(N)}\big].$$

Note that once we have two adjacent edges e_1 and e_2 labelled by 2, then the probability that the triangle spanned by $\{e_1,e_2\}$ generates an Artin groups of spherical type is exactly the probability that the last edge is not labelled by ∞. This probability is $(f(N)-1)/f(N)$; hence we have

$$\mathbb{P}\big[\Gamma\in\mathcal{D}\mid(\Gamma\notin\mathcal{B})\cap(\Gamma\in\mathcal{G}^{N,f(N)})\big] \le 1-\frac{f(N)-1}{f(N)} = \frac{1}{f(N)}.$$

Whence we get the following upper bound for $(*)$, for any nondecreasing function f:

$$\mathbb{P}[\Gamma\in\mathcal{D}\mid\Gamma\in\mathcal{G}^{N,f(N)}] \le \mathbb{P}[\Gamma\in\mathcal{B}\mid\Gamma\in\mathcal{G}^{N,f(N)}] + \mathbb{P}[\Gamma\notin\mathcal{B}\mid\Gamma\in\mathcal{G}^{N,f(N)}]\cdot\frac{1}{f(N)}.$$

By noting that for any nondecreasing divergent function f we have that $1/f(N)\to 0$, we get

$$\mathbb{P}_f[A_\Gamma\in A_{\mathcal{D}}] = \lim_{N\to\infty}\mathbb{P}[\Gamma\in\mathcal{D}\mid\Gamma\in\mathcal{G}^{N,f(N)}] \le \lim_{N\to\infty}\mathbb{P}[\Gamma\in\mathcal{B}\mid\Gamma\in\mathcal{G}^{N,f(N)}] = \mathbb{P}_f[A_\Gamma\in A_{\mathcal{B}}].$$

We now deal with the lower bound. The probability that a given triangle Δ is of spherical type is the quotient

$$(**)\qquad \frac{\#\text{ ways that }\Delta\text{ can be spherical}}{\#\text{ possible coefficients on }\Delta}.$$

In general, the only triangles that give spherical Artin groups are of the form $(2,3,3)$, $(2,3,4)$, $(2,3,5)$, and $(2,2,k)$ for $k\ge 2$. In our case, it is given that A_Γ is $(2,2)$-free, so the only triangles which are of spherical type are of the form $(2,3,3)$, $(2,3,4)$ or $(2,3,5)$. When considering the possible permutations of the order of the coefficients, this gives 15 possibilities. This yields the numerator of $(**)$.

In order to find an upper bound for $(**)$, it remains to find a lower bound for the denominator. In a graph Γ that we know is $(2,2)$-free, a triangle whose edges are all labelled by coefficients other than 2 will always be a possible combination of coefficients for a triangle Δ of Γ. Hence the number of possible coefficients for a triangle Δ of a $(2,2)$-free graph is at least $(f(N)-1)^3$. This yields

$$\frac{\#\text{ ways that }\Delta\text{ can be spherical}}{\#\text{ possible coefficients on }\Delta}\le\frac{15}{(f(N)-1)^3}. \tag{$***$}$$

Hence, by an union bound we get

$$\begin{aligned}\mathbb{P}\big[\Gamma\notin\mathfrak{D}\mid(\Gamma\in\mathfrak{B})\cap(\Gamma\in\mathcal{G}^{N,f(N)})\big]&\le\sum_{\Delta\text{ triangle in }\Gamma}\mathbb{P}\big[\Delta\text{ is of spherical type}\mid(\Gamma\in\mathfrak{B})\cap(\Gamma\in\mathcal{G}^{N,f(N)})\big]\\&\le\binom{N}{3}\frac{15}{(f(N)-1)^3}.\end{aligned}$$

Therefore, by $(*)$ we get

$$\mathbb{P}[\Gamma\in\mathfrak{D}\mid\Gamma\in\mathcal{G}^{N,f(N)}]\ge\left(1-\binom{N}{3}\frac{15}{(f(N)-1)^3}\right)\mathbb{P}[\Gamma\in\mathfrak{B}\mid\Gamma\in\mathcal{G}^{N,f(N)}]. \tag{$****$}$$

Hence, if $f(N)\succ N$, we have

$$\lim_{N\to\infty}\left(\binom{N}{3}\frac{15}{(f(N)-1)^3}\right)=0.$$

This means that

$$\begin{aligned}\mathbb{P}_f[A_\Gamma\in A_{\mathfrak{D}}]&=\lim_{N\to\infty}\mathbb{P}[\Gamma\in\mathfrak{D}\mid\Gamma\in\mathcal{G}^{N,f(N)}]\\&\ge\lim_{N\to\infty}\mathbb{P}[\Gamma\in\mathfrak{B}\mid\Gamma\in\mathcal{G}^{N,f(N)}]\quad(\text{by }(****))\\&=\mathbb{P}_f[A_\Gamma\in A_{\mathfrak{B}}].\end{aligned}$$

□

We now move towards determining for which (nondecreasing divergent) functions a random Artin group is asymptotically almost surely 2-dimensional, or not 2-dimensional. In view of Lemma 3.2, looking at $(2,2)$-free Artin groups will be enough to give a conclusion for 2-dimensional Artin groups. The result we want to prove is the following:

Theorem 3.3 *Let $f:\mathbb{N}\to\mathbb{N}$; then, for a random Artin group A_Γ (with respect to f):*

(1) *If $f(N)\succ N^{3/2}$, then asymptotically almost surely A_Γ is 2-dimensional.*

(2) *If $f(N)\prec N^{3/2}$, then asymptotically almost surely A_Γ is not 2-dimensional.*

(3) *If $f(N)\simeq N^{3/2}$ then $\mathbb{P}_f[A_\Gamma\in A_{\mathfrak{D}}]<1$. Moreover, if $f(N)=N^{3/2}$ then $\mathbb{P}_f[A_\Gamma\in A_{\mathfrak{D}}]\le 2/3$.*

Proof Let f be any nondecreasing, divergent function. We need to compute $\mathbb{P}_f[A_\Gamma\in A_{\mathfrak{D}}]$. In view of Lemma 3.2, it is enough to compute $\mathbb{P}_f[A_\Gamma\in A_{\mathfrak{B}}]$, ie the probability that an Artin group A_Γ picked at random is $(2,2)$-free. To do this, we will use the second moment method.

Let us consider a graph $\Gamma\in\mathcal{G}^{N,f(N)}$. For any ordered triplet (v_1,v_2,v_3) of distinct vertices of Γ, we let $I_{(v_1,v_2,v_3)}:\mathcal{G}^{N,f(N)}\to\{0,1\}$ be the random variable which takes 1 on $\Gamma\in\mathcal{G}^{N,f(N)}$ precisely when (v_1,v_2,v_3) spans a triangle with $m_{v_1,v_2}=m_{v_1,v_3}=2$. We let

$$X=\left(\sum_{(v_1,v_2,v_3)\in V(\Gamma)^3}I_{(v_1,v_2,v_3)}\right):\mathcal{G}^{N,f(N)}\to\mathbb{N},$$

where the sum is taken over all triplets of distinct vertices. The variable X counts the number of pairs of adjacent edges labelled by a 2, twice (because of the permutation of these edges).

We can compute the expectation $\mathbb{E}[I_{(v_1,v_2,v_3)}] = f(N)^{-2}$ and hence

$$\mathbb{E}[X] = \sum_{(v_1,v_2,v_3)} \mathbb{E}[I_{(v_1,v_2,v_3)}] = N(N-1)(N-2)f(N)^{-2} \sim N^3 f(N)^{-2}.$$

Now, we use the second moment method, as in [4, Theorem 6]:

$$\mathbb{P}[X \neq 0] \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

We have already computed $\mathbb{E}[X]$, so we now compute $\mathbb{E}[X^2]$ by dividing into several cases the sum

$$X^2 = \sum I_{(v_1,v_2,v_3)} I_{(w_1,w_2,w_3)}.$$

Note that the sum is taken over all ordered triplets (v_1, v_2, v_3) and (w_1, w_2, w_3) of vertices, where the v_i are distinct, and the w_i are distinct. Also note that if one of the two triangles does not have two edges labelled by 2, then the corresponding term in the sum is trivial. In other words, it is enough to only sum over pairs of triangles that both have at least two edges labelled by 2. In a triangle (v_1, v_2, v_3) such that $m_{v_1 v_2} = m_{v_1 v_3} = 2$, we shall call v_1 the *central* vertex of the triangle. The different cases are treated below. They can be seen in Figure 2.

Case 1 Let X_1 denote the sum of products $I_{(v_1,v_2,v_3)} I_{(w_1,w_2,w_3)}$ such that no vertex appears in both triples. Then

$$\mathbb{E}[X_1] = \frac{N!}{(N-6)!} f(N)^{-4} \sim N^6 f(N)^{-4}.$$

Case 2 Let X_2 denote the sum of products $I_{(v_1,v_2,v_3)} I_{(w_1,w_2,w_3)}$ such that these two triangles share exactly one vertex and the vertex they share is central in both triangles (ie $v_1 = w_1$). Then we have

$$\mathbb{E}[X_2] = \frac{N!}{(N-5)!} f(N)^{-4} \sim N^5 f(N)^{-4}.$$

Case 3 Let X_3 denote the sum of products $I_{(v_1,v_2,v_3)} I_{(w_1,w_2,w_3)}$ such that these two triangles share exactly one vertex, where this vertex is the central vertex for one triangle and not a central vertex for the other triangle (for example $v_2 = w_1$). In this case

$$\mathbb{E}[X_3] = 4\frac{N!}{(N-5)!} f(N)^{-4} \sim 4N^5 f(N)^{-4}.$$

Case 4 Let X_4 denote the sum of products $I_{(v_1,v_2,v_3)} I_{(w_1,w_2,w_3)}$ such that these two triangles share exactly one vertex, where this vertex is not central for either triangle (for example $v_2 = w_2$). Then

$$\mathbb{E}[X_4] = 4\frac{N!}{(N-5)!} f(N)^{-4} \sim 4N^5 f(N)^{-4}.$$

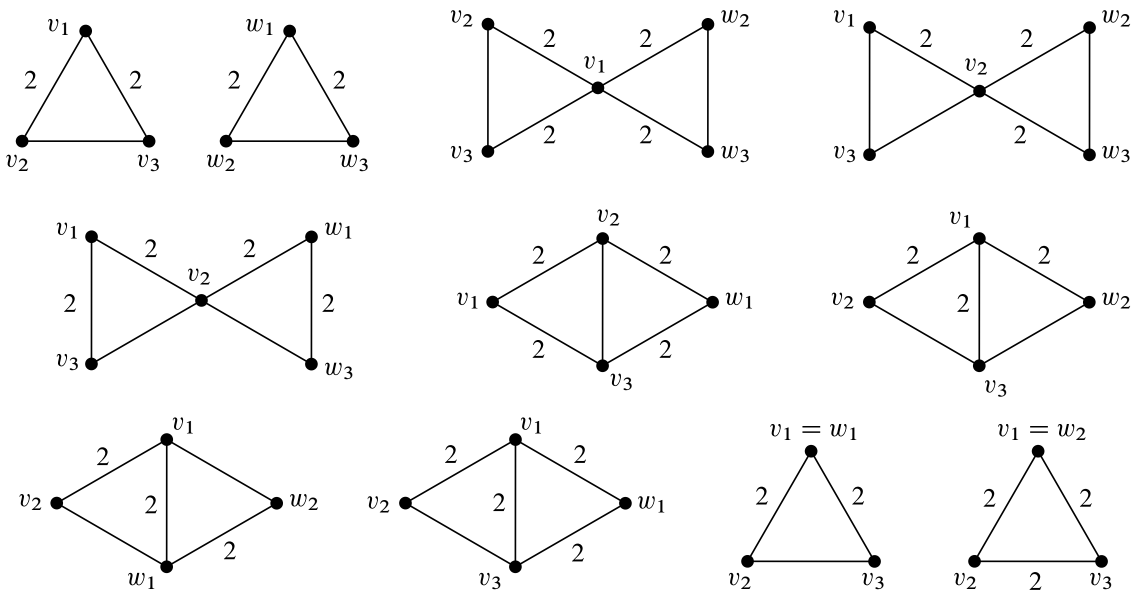


Figure 2: From top-left to bottom-right: the ten cases described in the proof of Theorem 3.3. The edges that are not explicitly labelled by 2 can be labelled by any coefficient, including ∞.

Case 5 Let X_5 denote the sum of products $I_{(v_1,v_2,v_3)}I_{(w_1,w_2,w_3)}$ such that these two triangles share exactly two vertices and these two vertices are not central for either triangle (for example $v_2 = w_2$ and $v_3 = w_3$). In this case

$$\mathbb{E}[X_5] = 2\frac{N!}{(N-4)!}f(N)^{-4} \sim 2N^4 f(N)^{-4}.$$

Case 6 Let X_6 denote the sum of products $I_{(v_1,v_2,v_3)}I_{(w_1,w_2,w_3)}$ such that these two triangles share exactly two vertices and one of these is central in both triangles and the other is not (for example $v_1 = w_1$ and $v_3 = w_2$). In this case

$$\mathbb{E}[X_6] = 4\frac{N!}{(N-4)!}f(N)^{-3} \sim 4N^4 f(N)^{-3}.$$

Case 7 Let X_7 denote the sum of products $I_{(v_1,v_2,v_3)}I_{(w_1,w_2,w_3)}$ such that these two triangles share exactly two vertices where one of these is central for the triangle (v_1, v_2, v_3) but not for (w_1, w_2, w_3), and the other vertex is central for the triangle (w_1, w_2, w_3) but not for (v_1, v_2, v_3) (for example $v_1 = w_3$ and $w_1 = v_3$). In this case

$$\mathbb{E}[X_7] = 4\frac{N!}{(N-4)!}f(N)^{-3} \sim 4N^4 f(N)^{-3}.$$

Case 8 Let X_8 denote the sum of products $I_{(v_1,v_2,v_3)}I_{(w_1,w_2,w_3)}$ such that these two triangles share exactly two vertices where one of these is central for the triangle (v_1, v_2, v_3) but none of the two vertices is central for (w_1, w_2, w_3) (for example $v_1 = w_2$ and $v_3 = w_3$). In this case

$$\mathbb{E}[X_8] = 4\frac{N!}{(N-4)!}f(N)^{-4} \sim 4N^4 f(N)^{-4}.$$

Case 9 Let X_9 denote the sum of products $I_{(v_1,v_2,v_3)}I_{(w_1,w_2,w_3)}$ such that these two triangles share all three vertices, and such the central vertices of both triangles are the same (ie $v_1 = w_1$). In this case

$$\mathbb{E}[X_9] = 2\frac{N!}{(N-3)!}f(N)^{-2} \sim 2N^3 f(N)^{-2}.$$

Case 10 Let X_{10} denote the sum of products $I_{(v_1,v_2,v_3)}I_{(w_1,w_2,w_3)}$ such that these two triangles share all three vertices, and such that the central vertex of the first triangle is not the central vertex of the second triangle (for example $v_1 = w_2$). We get

$$\mathbb{E}[X_{10}] = 4\frac{N!}{(N-3)!}f(N)^{-3} \sim 2N^3 f(N)^{-3}.$$

Therefore, we have

$$\begin{aligned}\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2} &= \sum_{i=1}^{8}\frac{\mathbb{E}[X_i]}{\mathbb{E}[X]^2}\\ &\sim \frac{N^6 f(N)^{-4} + 9N^5 f(N)^{-4} + 6N^4 f(N)^{-4} + 8N^4 f(N)^{-3} + 2N^3 f(N)^{-3} + 2N^3 f(N)^{-2}}{N^6 f(N)^{-4}}\\ &\sim 1 + \frac{9}{N} + \frac{6}{N^2} + \frac{8f(N)}{N^2} + \frac{2f(N)}{N^3} + \frac{2f(N)^2}{N^3}.\end{aligned}$$

Hence, if $f(N) \prec N^{3/2}$ then by definition there exists a nondecreasing divergent function h such that $f(N)h(N) = N^{3/2}$. In this case we get

$$\mathbb{P}[X \neq 0] \geq \left(\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2}\right)^{-1} \sim \left(1 + \frac{9}{N} + \frac{6}{N^2} + \frac{8}{h(N)N^{1/2}} + \frac{4}{h(N)N^{3/2}} + \frac{2}{h(N)^2}\right)^{-1}.$$

When $f(N) \prec N^{3/2}$, we obtain

$$\mathbb{P}_f[A_\Gamma \in A_{\mathcal{B}}] = \lim_{N\to\infty}\mathbb{P}[\Gamma \in \mathcal{B} \mid \Gamma \in \mathcal{G}^{N,f(N)}] = \lim_{N\to\infty}\mathbb{P}[X = 0] = 1 - \lim_{N\to\infty}\mathbb{P}[X \neq 0] = 0.$$

Thus asymptotically almost surely A_Γ is not $(2,2)$-free. In view of Lemma 3.2, this also means that asymptotically almost surely A_Γ is not of dimension 2, this proves item (2) in Theorem 3.3.

If $f(N) \simeq N^{3/2}$ then the quotient $f(N)/N^{3/2}$ tends to M for some constant $M > 0$. Hence in this case,

$$\mathbb{P}[X \neq 0] \gtrsim \left(1 + \frac{9}{N} + \frac{6}{N^2} + \frac{8f(N)}{N^2} + \frac{2f(N)}{N^3} + \frac{2f(N)^2}{N^3}\right)^{-1} \sim (1 + 2M^2)^{-1} > 0.$$

Therefore $\mathbb{P}_f[A_\Gamma \in A_{\mathcal{B}}] < 1$ at $f(N) \simeq N^{3/2}$ and hence by Lemma 3.2 we have that $\mathbb{P}_f[A_\Gamma \in A_{\mathcal{D}}] < 1$.

We note that the above calculation allows us to find a better upper bound for $\mathbb{P}_f[A_\Gamma \in A_{\mathcal{B}}]$ at $f(N) = N^{3/2}$. Indeed, this implies that $M = 1$ and hence we get $\mathbb{P}[X \neq 0] \gtrsim \frac{1}{3}$, and so at $f(N) = N^{3/2}$ we have $\mathbb{P}_f[A_\Gamma \in A_{\mathcal{B}}] \leq \frac{2}{3}$. Hence by Lemma 3.2, this proves item (3) in the theorem.

We note that $\mathbb{P}[\Gamma \in \mathcal{B} \mid \Gamma \in \mathcal{G}^{N,f(N)}] = 1 - \mathbb{P}[X \geq 1]$ and by the Markov inequality,

$$\mathbb{P}[X \geq 1] \leq \mathbb{E}[X] \leq N^3 f(N)^{-2}.$$

Hence if $f(N) \succ N^{3/2}$ then we can write $f(N) = N^{3/2}g(N)$ for some nondecreasing divergent function $g\colon \mathbb{N} \to \mathbb{N}$ and in this case

$$\mathbb{P}[X \geq 1] \leq \frac{1}{g(N)^2}.$$

Therefore, for $f(N) \succ N^{3/2}$ we have

$$\mathbb{P}_f[A_\Gamma \in A_{\mathcal{B}}] = \lim_{N\to\infty} \mathbb{P}[\Gamma \in \mathcal{B} \mid \mathcal{G}^{N,f(N)}] = 1 - \lim_{N\to\infty} \mathbb{P}[X \geq 1] \geq 1 - \lim_{N\to\infty} \frac{1}{g(N)^2} = 1.$$

In particular, asymptotically almost surely A_Γ is $(2,2)$-free. By applying Lemma 3.2 (as $f(N) \succ N$), we get that asymptotically almost surely A_Γ is 2-dimensional. This proves item (1) and hence Theorem 3.3. □

Before stating a corollary which will be a refinement of Corollary 2.12, we prove a small lemma which will allow us to study the problem of (virtual) cocompact cubulation of random Artin groups. We note that the class $\mathcal{P}$ defined below is point 3 in [8, Conjecture B].

Lemma 3.4 *Let $\mathcal{P}$ be the class of defining graphs Γ for which there exist 4 distinct $a, b, c, d \in V(\Gamma)$ such that $m_{ab} \notin \{2,\infty\}, m_{ac}, m_{bd} \neq \infty$ and $m_{ad}, m_{bc} \neq 2$. Then $A_{\mathcal{P}}$ is uniformly large.*

Proof Let $f\colon \mathbb{N} \to \mathbb{N}$ be any nondecreasing, divergent function. Fix a, b, c and d to be any distinct vertices. The probability that these vertices and their corresponding coefficients satisfy the defining condition of $\mathcal{P}$ is at exactly

$$\left(\frac{f(N)-1}{f(N)}\right)^4\left(\frac{f(N)-2}{f(N)}\right).$$

This tends to 1 for all nondecreasing divergent functions f. □

Corollary 3.5 *Let $f\colon \mathbb{N} \to \mathbb{N}$ be a function satisfying $f(N) \succ N^{3/2}$. Then a random Artin group A_Γ (with respect to f) satisfies any of the following properties asymptotically almost surely:*

(1) *A_Γ is torsion-free*;

(2) *A_Γ has trivial centre*;

(3) *A_Γ has solvable word and conjugacy problems*;

(4) *A_Γ satisfies the $K(\pi,1)$-conjecture*;

(5) *the set of parabolic subgroups of A_Γ is closed under arbitrary intersections*;

(6) *A_Γ is acylindrically hyperbolic*;

(7) *A_Γ satisfies the Tits alternative*;

(8) *A_Γ is not virtually cocompactly cubulated.*

Proof By Theorem 3.3, A_Γ is asymptotically almost surely 2-dimensional. Using Lemma 3.2, A_Γ is also asymptotically almost surely $(2,2)$-free. Using Lemma 2.8, we also know that A_Γ is asymptotically almost surely irreducible. By Lemma 3.4 we know that A_Γ is asymptotically almost surely in $A_{\mathcal{P}}$. Using Lemma 2.11 three times, this ensures that A_Γ is asymptotically almost surely in the class

$$A_{\mathcal{H}} := A_{\mathrm{Irr}} \cap A_{\mathcal{D}} \cap A_{\mathcal{B}} \cap A_{\mathcal{P}}.$$

Note that the results given for points (1), (2), (3), (4), (5), (7) and (10) of Conjecture 2.2 concern families of Artin groups that all contain $A_{\mathscr{H}}$. In particular, every Artin group of $A_{\mathscr{H}}$ satisfies the first seven points of the Corollary 3.5. For point (8) of Corollary 3.5, we note that by [8, Theorem E], if $A_\Gamma \in A_{\mathscr{D}} \cap A_{\mathscr{P}}$ then A_Γ is not virtually cocompactly cubulated. □

Finding out the exact probability for an Artin group to be 2-dimensional (or equivalently, $(2,2)$-free) at $f(N) = N^{3/2}$ requires more work. In Theorem 3.3, we gave an upper bound for this probability. The goal of the following lemma is to give an explicit formula for the value of $\mathbb{P}_f[A_\Gamma \in A_{\mathscr{B}}]$ at $f(N) = N^{3/2}$. Later, we give a conjecture on the exact value.

Lemma 3.6 *For all nondecreasing, divergent functions $f : \mathbb{N} \to \mathbb{N}$ we have that*

$$\mathbb{P}_f[A_\Gamma \in A_{\mathscr{B}}] = \lim_{N\to\infty} \left(\frac{f(N)-1}{f(N)}\right)^{\binom{N}{2}} \left(\sum_{k=1}^{\lfloor N/2\rfloor} \frac{N!(f(N)-1)^{-k}}{(N-2k)!\,k!\,2^k} + 1\right).$$

Proof Let E_k be the family of defining graphs that have exactly k edges labelled by a 2, and consider the associated family A_{E_k} of Artin groups. Note that each edge is attached to two vertices, so by the pigeonhole principle, if $k > N/2$ then $\mathbb{P}_f[\Gamma \in \mathscr{B} \cap E_k] = 0$. Hence

$$\mathbb{P}[\Gamma \in \mathscr{B} \mid \Gamma \in \mathscr{G}^{N,f(N)}] = \sum_{k=0}^{\lfloor N/2\rfloor} \mathbb{P}[\Gamma \in \mathscr{B} \cap E_k \mid \Gamma \in \mathscr{G}^{N,f(N)}].$$

As usual, the total number of graphs in $\mathscr{G}^{N,f(N)}$ is $f(N)^{\binom{N}{2}}$. On the other hand, we must compute how many of these graphs have exactly k edges labelled by a 2, while these edges are never adjacent.

First of all, when $k = 0$, we have $\mathbb{P}[\Gamma \in \mathscr{B} \cap E_k \mid \Gamma \in \mathscr{G}^{N,f(N)}] = ((f(N)-1)/f(N))^{\binom{N}{2}}$.

For the case when $0 < k \le \lfloor N/2\rfloor$, we look at how many ways we have of placing the k edges labelled by a 2. For the first such edge, we have $\binom{N}{2}$ choices. The two vertices of the first edge must not appear in any other edge labelled by a 2, so for the second edge we only have $\binom{N-2}{2}$ choices left. This goes on until the k^{th} edge labelled by a 2, for which we have $\binom{N-2(k-1)}{2}$ choices. As the order in which we have chosen these edges do not matter, we must divide this product by $k!$. Now for the remaining $\binom{N}{2} - k$ edges, we can use any label other than a 2. Hence we multiply the previous product by $(f(N)-1)^{\binom{N}{2}-k}$. Hence, for $0 < k \le \lfloor N/2\rfloor$, we have

$$\mathbb{P}[\Gamma \in \mathscr{B} \cap E_k \mid \Gamma \in \mathscr{G}^{N,f(N)}] = \frac{(f(N)-1)^{\binom{N}{2}-k} \cdot \prod_{i=0}^{k-1} \binom{N-2i}{2}}{f(N)^{\binom{N}{2}} \cdot k!}.$$

Therefore,

$$\begin{aligned}\mathbb{P}_f[A_\Gamma \in A_{\mathscr{B}}] &= \lim_{N\to\infty} \sum_{k=1}^{\lfloor N/2\rfloor} \mathbb{P}[\Gamma \in \mathscr{B} \cap E_k \mid \Gamma \in \mathscr{G}^{N,f(N)}] + \mathbb{P}[\Gamma \in \mathscr{B} \cap E_0 \mid \Gamma \in \mathscr{G}^{N,f(N)}] \\ &= \lim_{N\to\infty} \sum_{k=1}^{\lfloor N/2\rfloor} \frac{(f(N)-1)^{\binom{N}{2}-k} \cdot \prod_{i=0}^{k-1} \binom{N-2i}{2}}{f(N)^{\binom{N}{2}} \cdot k!} + \left(\frac{f(N)-1}{f(N)}\right)^{\binom{N}{2}}\end{aligned}$$

$$= \lim_{N\to\infty}\left(\frac{f(N)-1}{f(N)}\right)^{\binom{N}{2}}\left(\sum_{k=1}^{\lfloor N/2\rfloor}\frac{N!(f(N)-1)^{-k}}{(N-2k)!\,k!\,2^k}+1\right),$$

where we go from the second to the third line by noting that

$$\prod_{i=0}^{k-1}\binom{N-2i}{2}=\frac{1}{2^k}N(N-1)(N-2)\cdots(N-2(k-1))(N-2(k-1)-1)=\frac{N!}{(N-2k)!2^k}. \qquad \square$$

Now, by Lemma 3.2 at $f(N)=N^{3/2}$ we have $\mathbb{P}_f[A_\Gamma\in A_{\mathscr{D}}]=\mathbb{P}_f[A_\Gamma\in A_{\mathscr{B}}]$; hence Lemma 3.6 also holds for $\mathbb{P}_f[A_\Gamma\in A_{\mathscr{D}}]$. We have computed this expression in *Python* for N up to 190, which leads us to the following conjecture.

Conjecture 3.7 For $f(N)=N^{3/2}$ we have

$$\mathbb{P}_f[A_\Gamma\in A_{\boldsymbol{B}}]=1-e^{-1}.$$

In particular, we also have

$$\mathbb{P}_f[A_\Gamma\in A_{\mathscr{D}}]=1-e^{-1}.$$

4 Acylindrical hyperbolicity and centres

Two open questions in the study of Artin groups is whether all irreducible nonspherical Artin groups are acylindrically hyperbolic and have trivial centres (see Conjecture 2.2). In this section, we study these two aspects of Artin groups for another family of Artin groups, that we will denote $A_{\mathscr{C}^C}$. The families of Artin groups studied in Sections 2 and 3 are very large when $f(N)$ grows fast enough compared to N. While the spirit of this section resembles that of Sections 2 and 3, $A_{\mathscr{C}^C}$ will turn out to be very large when $f(N)$ grows slowly enough compared to N.

Definition 4.1 A graph Γ is said to be a cone if it has a join decomposition as a subgraph consisting of a single vertex v_0 and a subgraph Γ' such that $\Gamma=v_0*\Gamma'$. Let $\mathscr{C}$ be the class of defining graphs that are cones and $\mathscr{C}^C$ the class of defining graphs which are not cones.

Recall that Irr is the class of irreducible graphs. By [13, Theorem 1.4], we have that if Γ has at least 3 vertices, is irreducible and is not a cone then A_Γ is acylindrically hyperbolic. Hence it suffices to find the probability that a random Artin group is irreducible and is not a cone.

Proposition 4.2 *For all $\alpha\in(0,1)$ and all nondecreasing functions $f(N)\prec N^{1-\alpha}$ we have that*

$$\mathbb{P}_f[A_\Gamma\in A_{\mathscr{C}^C}]=1.$$

Proof Fix $\alpha\in(0,1)$ and $f(N)\prec N^{1-\alpha}$ a nondecreasing divergent function. Then, by definition, there exists a nondecreasing divergent function h such that $f(N)h(N)=N^{1-\alpha}$.

By the definition of a cone and by a union bound, we get

$$\begin{aligned}\mathbb{P}[\Gamma \in \mathscr{C} \mid \Gamma \in \mathscr{G}^{N,f(N)}] &\leq \sum_{v_0 \in V(\Gamma)} \mathbb{P}\big[\forall u \in V(\Gamma) - v_0 : m_{u,v_0} \neq \infty \mid \Gamma \in \mathscr{G}^{N,f(N)}\big] \\ &= \sum_{v_0 \in V(\Gamma)} \left(\frac{f(N)-1}{f(N)}\right)^{N-1} \\ &= N\left(\frac{f(N)-1}{f(N)}\right)^{N-1} \\ &= N\left(\left(\frac{f(N)-1}{f(N)}\right)^{f(N)}\right)^{h(N)N^\alpha}\left(\frac{f(N)}{f(N)-1}\right).\end{aligned}$$

Thus,

$$\mathbb{P}_f[A_\Gamma \in A_{\mathscr{C}}] = \lim_{N\to\infty} \mathbb{P}[\Gamma \in \mathscr{C} \mid \Gamma \in \mathscr{G}^{N,f(N)}] = \lim_{N\to\infty} Ne^{-N^\alpha h(N)} = 0.$$

Hence for $f(N) \prec N^{1-\alpha}$ we have $\mathbb{P}_f[A_\Gamma \in A_{\mathscr{C}^C}] = 1$, proving the proposition. □

Corollary 4.3 *Let $\alpha \in (0,1)$ and let $f(N) \prec N^{1-\alpha}$ be a nondecreasing divergent function. Then a random Artin group (with respect to f) asymptotically almost surely is acylindrically hyperbolic and has a trivial centre.*

Proof We note that by Lemmas 2.8 and 2.11 we have $\mathbb{P}_f[A_\Gamma \in A_{\mathrm{Irr}} \cap A_{\mathscr{C}^C}] = \mathbb{P}_f[A_\Gamma \in A_{\mathscr{C}^C}]$. As we noted above, by [13, Theorem 1.4], if Γ is irreducible and not a cone then A_Γ is acylindrically hyperbolic. Hence, by Proposition 4.2, for a function f as in the statement of the corollary, we get that a random Artin group (relatively to f) is asymptotically almost surely irreducible and a cone, hence asymptotically almost surely acylindrically hyperbolic.

Further, by [5, Theorem 3.3], we have that if Γ is irreducible and not a cone then A_Γ has trivial centre. Hence a random Artin group (relatively to f) asymptotically almost surely has a trivial centre. □

Let $\alpha \in (0,1)$, by Corollary 4.3 and Corollary 3.5-(6), we have shown that for all nondecreasing divergent functions f such that either

- $f(N) \prec N^{1-\alpha}$, or
- $f(N) \succ N^{3/2}$,

a random Artin group A_Γ (relatively to f) is asymptotically almost surely acylindrically hyperbolic and has trivial centre. This motivates the following:

Question 4.4 For which nondecreasing divergent functions f do we have that a random Artin group (relatively to f) is asymptotically almost surely acylindrically hyperbolic and has trivial centre?

References

[1] **J Behrstock**, **M F Hagen**, **A Sisto**, *Thickness, relative hyperbolicity, and randomness in Coxeter groups*, Algebr. Geom. Topol. 17 (2017) 705–740 MR Zbl

[2] **M A Blufstein**, *Parabolic subgroups of two-dimensional Artin groups and systolic-by-function complexes*, Bull. Lond. Math. Soc. 54 (2022) 2338–2350 MR Zbl

[3] **R Charney**, **M W Davis**, *The $K(\pi,1)$–problem for hyperplane complements associated to infinite reflection groups*, J. Amer. Math. Soc. 8 (1995) 597–627 MR Zbl

[4] **R Charney**, **M Farber**, *Random groups arising as graph products*, Algebr. Geom. Topol. 12 (2012) 979–995 MR Zbl

[5] **R Charney**, **R Morris-Wright**, *Artin groups of infinite type: trivial centers and acylindrical hyperbolicity*, Proc. Amer. Math. Soc. 147 (2019) 3675–3689 MR Zbl

[6] **M Cumplido**, **A Martin**, **N Vaskou**, *Parabolic subgroups of large-type Artin groups*, Math. Proc. Cambridge Philos. Soc. 174 (2023) 393–414 MR Zbl

[7] **A Deibel**, *Random Coxeter groups*, Int. J. Algebra Comput. 30 (2020) 1305–1321 MR Zbl

[8] **T Haettel**, *Virtually cocompactly cubulated Artin–Tits groups*, Int. Math. Res. Not. 2021 (2021) 2919–2961 MR Zbl

[9] **T Haettel**, XXL *type Artin groups are* CAT(0) *and acylindrically hyperbolic*, Ann. Inst. Fourier (Grenoble) 72 (2022) 2541–2555 MR Zbl

[10] **M Hagen**, **A Martin**, **A Sisto**, *Extra-large type Artin groups are hierarchically hyperbolic*, Math. Ann. 388 (2024) 867–938 MR Zbl

[11] **J Huang**, **D Osajda**, *Metric systolicity and two-dimensional Artin groups*, Math. Ann. 374 (2019) 1311–1352 MR Zbl

[12] **J Huang**, **D Osajda**, *Large-type Artin groups are systolic*, Proc. Lond. Math. Soc. 120 (2020) 95–123 MR Zbl

[13] **M Kato**, **S-i Oguni**, *Acylindrical hyperbolicity of Artin groups associated with graphs that are not cones*, Groups Geom. Dyn. 18 (2024) 1291–1316 MR Zbl

[14] **H van der Lek**, *The homotopy type of complex hyperplane complements*, PhD thesis, Katholieke Universiteit te Nijmegen (1983) Available at `https://repository.ubn.ru.nl/handle/2066/148301`

[15] **A Martin**, *The Tits alternative for two-dimensional Artin groups and Wise's power alternative*, J. Algebra 656 (2024) 294–323 MR Zbl

[16] **A Martin**, **N Vaskou**, *Characterising large-type Artin groups*, Bull. Lond. Math. Soc. (online publication August 2024)

[17] **N Vaskou**, *Acylindrical hyperbolicity for Artin groups of dimension* 2, Geom. Dedicata 216 (2022) art. id. 7 MR Zbl

[18] **N Vaskou**, *Automorphisms of large-type free-of-infinity Artin groups*, Geom. Dedicata 219 (2025) art. no. 16 MR Zbl

School of Mathematics & Computer Sciences, Heriot-Watt University
Edinburgh, United Kingdom

Department of Mathematics, University of Bristol
Bristol, United Kingdom

ag2017@hw.ac.uk, nicolas.vaskou@bristol.ac.uk

Received: 25 January 2023 Revised: 30 October 2023

Algebraic & Geometric Topology **25**:3 (2025) 1545–1560
DOI: 10.2140/agt.2025.25.1545
Published: 20 June 2025

A deformation of Asaeda–Przytycki–Sikora homology

ZHENKUN LI
YI XIE
BOYU ZHANG

We define a 1-parameter family of homology invariants for links in thickened oriented surfaces. It recovers the homology invariant of Asaeda, Przytycki and Sikora (Algebr. Geom. Topol. 4 (2004) 1177–1210) and the invariant defined by Winkeler (Michigan Math. J. 74 (2024) 1–31). The new invariant can be regarded as a deformation of Asaeda–Przytycki–Sikora homology; it is not a Lee-type deformation as the deformation is only nontrivial when the surface is not simply connected. Our construction is motivated by computations in singular instanton Floer homology. We also prove a detection property for the new invariant, which is a stronger result than our previous work (Selecta Math. 29 (2023) art. id. 84).

57K18

1 Introduction

Khovanov homology [9] is a link invariant that assigns a bigraded homology group to every oriented link in $\mathbb{R}^3$. Asaeda, Przytycki and Sikora [1] introduced a generalization of Khovanov homology for links in $(-1,1)$-bundles over surfaces, where the bundles are required to be oriented as 3-manifolds. Such $(-1,1)$-bundles are called *thickened surfaces*. When the surface is an annulus, Asaeda–Przytycki–Sikora homology is also called *annular Khovanov homology*. Khovanov homology and Asaeda–Przytycki–Sikora homology have been essential tools for the study of knots and links for decades. More recently, Winkeler [16] introduced another variation of Khovanov homology for links in thickened multipunctured disks, which is different from the invariant of Asaeda, Przytycki and Sikora.

Suppose Σ is an oriented surface. We define a 1-parameter family of homology invariants for oriented links in $(-1,1)\times\Sigma$. As bigraded modules, the new invariant recovers both Asaeda–Przytycki–Sikora homology and the invariant of Winkeler, and it can be interpreted as a 1-parameter deformation of Asaeda–Przytycki–Sikora homology. The deformation is not a Lee-type deformation as it is only nontrivial when the surface has a nontrivial fundamental group. The construction is motivated by computations from singular instanton Floer homology. We also use instanton Floer theory to prove a detection result for the deformed Asaeda–Przytycki–Sikora homology, which gives a stronger rank estimate than the main theorem of Li, Xie and Zhang [12].

The paper is organized as follows. Section 2 introduces some notation and conventions. Sections 3 and 4 define the differential map and prove that $d^2 = 0$. Section 5 defines the homology invariant and proves the invariance under Reidemeister moves. Section 6 explains the motivation from instanton Floer homology and proves the aforementioned detection result in Theorem 6.1.

Acknowledgments Xie is supported by the National Key R&D Program of China 2020YFA0712800 and NSFC 12071005.

2 Notation

Throughout this paper we use R to denote a fixed commutative ring with unit. We use Σ to denote an oriented surface, possibly with boundary and possibly noncompact.

For every embedded closed 1-manifold $c \subset \Sigma$, we assign an R-module $V(c)$ to c as follows:

(1) If γ is a contractible simple closed curve on Σ, define $V(\gamma)$ to be the free R-module generated by $\boldsymbol{v}(\gamma)_+$ and $\boldsymbol{v}(\gamma)_-$, where $\boldsymbol{v}(\gamma)_+$ and $\boldsymbol{v}(\gamma)_-$ are formal generators associated with γ.

(2) If γ is a noncontractible simple closed curve, let $\mathfrak{o}$ and $\mathfrak{o}'$ be the two orientations of γ. Define $V(\gamma)$ to be the free module generated by $\boldsymbol{v}(\gamma)_{\mathfrak{o}}$ and $\boldsymbol{v}(\gamma)_{\mathfrak{o}'}$, where $\boldsymbol{v}(\gamma)_{\mathfrak{o}}$ and $\boldsymbol{v}(\gamma)_{\mathfrak{o}'}$ are formal generators.

(3) In general, suppose the connected components of c are $\gamma_1, \dots, \gamma_k$. Define $V(c)$ to be $\bigotimes_{i=1}^k V(\gamma_i)$.

When the choice of Σ needs to be emphasized, we will write $V(c)$ as $V^\Sigma(c)$, and write $\boldsymbol{v}(\gamma)_{\mathfrak{o}}$ and $\boldsymbol{v}(\gamma)_\pm$ as $\boldsymbol{v}^\Sigma(\gamma)_{\mathfrak{o}}$ and $\boldsymbol{v}^\Sigma(\gamma)_\pm$, respectively.

If $\mathfrak{o}$ is an orientation of a curve γ, we use $\gamma_{\mathfrak{o}}$ to denote the corresponding oriented curve.

3 Band surgery homomorphisms

Suppose c is an embedded closed 1-manifold on Σ, suppose b is an embedded disk on Σ such that the interior of b is disjoint from c and the boundary of b intersects c at two arcs (see Figure 1). The surgery of c along b yields another embedded closed 1-manifold on Σ, which we denote by c_b. We will call the disk b a *band* that is *attached* to c.

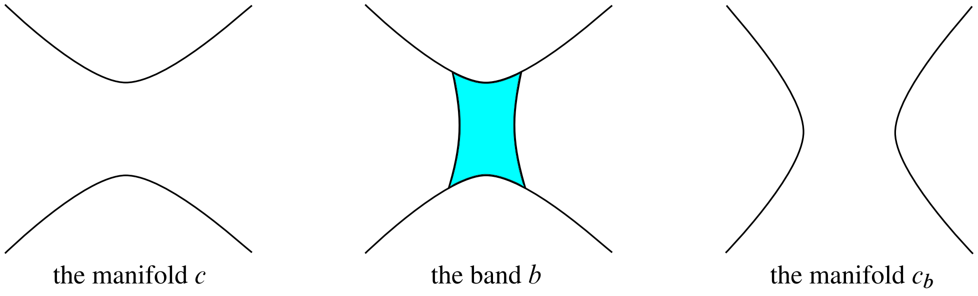


Figure 1: Band surgery.

For later reference, we record the following two elementary lemmas:

Lemma 3.1 *The change from c to c_b has three possibilities*:

(1) *two circle components of c are merged to one circle*,

(2) *one circle component of c is split to two circles*,

(3) *one circle component of c is modified by the surgery to another circle.*

Proof Since $\partial b \cap c$ contains two arcs, at most two components of c are affected by the surgery. If the arcs of $\partial b \cap c$ are on two different components of c, then the surgery merges these two components into one circle. If the arcs of $\partial b \cap c$ are on one component of c, then the boundary orientation of b defines an orientation on both components of $\partial b \cap c$, so we have two oriented arcs embedded in one component γ of c. If these two arcs induce the same orientation on γ, then the surgery splits one component of c to two circles. If these two arcs induce opposite orientations on γ, then the surgery changes this component to another circle. □

Recall that if $\mathfrak{o}$ is an orientation of a curve γ, we use $\gamma_{\mathfrak{o}}$ to denote the corresponding oriented curve.

Lemma 3.2 *Suppose γ is a simple closed curve on a connected surface Σ, and assume Σ is not diffeomorphic to S^2. Suppose $\mathfrak{o}$ and $\mathfrak{o}'$ are the two orientations of γ. Then $\gamma_{\mathfrak{o}}$ and $\gamma_{\mathfrak{o}'}$ are not isotopic on Σ.*

Proof If γ is nonseparating, there exists an oriented simple closed curve β such that the algebraic intersection number of β and γ is nonzero. Since isotopies preserve the sign of algebraic intersection numbers, the desired result follows.

If γ is separating and $\partial\Sigma \neq \varnothing$, then every orientation of γ defines an ordering of the two components of $\Sigma\backslash\gamma$, which defines an ordered partition of the components of $\partial\Sigma$. Since every isotopy of γ on Σ can be extended to an isotopy of Σ fixing the boundary, the desired result is proved.

If γ is separating and Σ is closed, then every orientation of γ defines an ordering of the two components of $\Sigma\backslash\gamma$. Suppose Σ_1 and Σ_2 are the two components of $\Sigma\backslash\gamma$ ordered by an orientation $\mathfrak{o}$ of γ. Since Σ is not a sphere, the images of $H_1(\Sigma_1;\mathbb{Z})$ and $H_1(\Sigma_2;\mathbb{Z})$ are distinct in $H_1(\Sigma;\mathbb{Z})$. The images of $H_1(\Sigma_1;\mathbb{Z})$ and $H_1(\Sigma_2;\mathbb{Z})$ are invariant under isotopies of $\gamma_{\mathfrak{o}}$, so the desired result is proved. □

Taking an arbitrary element $\lambda \in R$, we define a homomorphism

$$T_\lambda(b)\colon V(c) \to V(c_b)$$

associated with the band surgery along b. When the choice of Σ needs to be emphasized, we will write $T_\lambda(b)$ as $T_\lambda^\Sigma(b)$.

We first assume that the intersection of ∂b with every component of c is nonempty. The general case will be discussed later. By Lemma 3.1, if the intersection of ∂b with every component of c is nonempty, then there are three cases:

Case 1 (c has two components γ_1 and γ_2 and they are merged into one circle $\gamma = c_b$ after the surgery) In this case, we define $T_\lambda(b)\colon V(\gamma_1) \otimes V(\gamma_2) \to V(\gamma)$ as follows:

(1) If both γ_1 and γ_2 are contractible circles, then γ is also contractible, and we define $T_\lambda(b)$ by

$$\begin{aligned} \boldsymbol{v}(\gamma_1)_+ \otimes \boldsymbol{v}(\gamma_2)_+ &\mapsto \boldsymbol{v}(\gamma)_+, \quad & \boldsymbol{v}(\gamma_1)_+ \otimes \boldsymbol{v}(\gamma_2)_- &\mapsto \boldsymbol{v}(\gamma)_-, \\ \boldsymbol{v}(\gamma_1)_- \otimes \boldsymbol{v}(\gamma_2)_+ &\mapsto \boldsymbol{v}(\gamma)_-, \quad & \boldsymbol{v}(\gamma_1)_- \otimes \boldsymbol{v}(\gamma_2)_- &\mapsto 0. \end{aligned}$$

(2) If γ_1 is contractible and γ_2 is noncontractible, then γ_2 is isotopic to γ. The existence of noncontractible curves on Σ implies that Σ is not diffeomorphic to S^2. By Lemma 3.2, the orientations of γ_2 are canonically identified with the orientations of γ via an isotopy. This identification defines a canonical isomorphism from $V(\gamma_2)$ to $V(\gamma)$, which we denote by ι. In this case, the homomorphism $T_\lambda(b)$ is defined by

$$\boldsymbol{v}(\gamma_1)_+ \otimes x \mapsto \iota(x), \quad \boldsymbol{v}(\gamma_1)_- \otimes x \mapsto 0$$

for all $x \in V(\gamma_2)$.

(3) If γ_1 is noncontractible and γ_2 is contractible, define $T_\lambda(b)$ by requiring the map to be symmetric with respect to γ_1 and γ_2 and reducing to (2) above.

(4) If γ_1 and γ_2 are both noncontractible and γ_3 is contractible, then γ_1 and γ_2 must be isotopic. By Lemma 3.2, the orientations of γ_1 and γ_2 are canonically identified by the isotopy. Let $\mathfrak{o}$ and $\mathfrak{o}'$ be the two orientations of γ_1, and use the same notation to denote the corresponding orientations of γ_2. The map $T_\lambda(b)$ is then defined by

$$\begin{aligned} \boldsymbol{v}(\gamma_1)_{\mathfrak{o}} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}} &\mapsto 0, \quad & \boldsymbol{v}(\gamma_1)_{\mathfrak{o}'} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}'} &\mapsto 0, \\ \boldsymbol{v}(\gamma_1)_{\mathfrak{o}} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}'} &\mapsto \boldsymbol{v}(\gamma)_-, \quad & \boldsymbol{v}(\gamma_1)_{\mathfrak{o}'} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}} &\mapsto \boldsymbol{v}(\gamma)_-. \end{aligned}$$

(5) If all of γ_1, γ_2, and γ are noncontractible, let N be the regular neighborhood of $b \cup \gamma_1 \cup \gamma_2$. Then N is a sphere with three disks removed, and the three boundary components of N are parallel to γ_1, γ_2 and γ. Since $N \subset \Sigma$ is oriented, the boundary orientation of N defines an orientation on each of γ_1, γ_2 and γ, and we denote these orientations by $\mathfrak{o}_1$, $\mathfrak{o}_2$ and $\mathfrak{o}$, respectively. Denote their opposite orientations by $\mathfrak{o}_1'$, $\mathfrak{o}_2'$ and $\mathfrak{o}'$. Then $T_\lambda(b)$ is defined by

$$\begin{aligned} \boldsymbol{v}(\gamma_1)_{\mathfrak{o}_1'} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}_2'} &\mapsto \lambda \cdot \boldsymbol{v}(\gamma)_{\mathfrak{o}}, \quad & \boldsymbol{v}(\gamma_1)_{\mathfrak{o}_1'} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}_2} &\mapsto 0, \\ \boldsymbol{v}(\gamma_1)_{\mathfrak{o}_1} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}_2'} &\mapsto 0, \quad & \boldsymbol{v}(\gamma_1)_{\mathfrak{o}_1} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}_2} &\mapsto 0. \end{aligned}$$

Case 2 (c contains one component γ and c_b has two components γ_1 and γ_2) In this case, we define $T_\lambda(b)\colon V(\gamma) \to V(\gamma_1) \otimes V(\gamma_2)$ as follows:

(1) If γ_1 and γ_2 are both contractible circles, then γ is also contractible, and we define $T_\lambda(b)$ by

$$\boldsymbol{v}(\gamma)_+ \mapsto \boldsymbol{v}(\gamma_1)_+ \otimes \boldsymbol{v}(\gamma_2)_- + \boldsymbol{v}(\gamma_1)_- \otimes \boldsymbol{v}(\gamma_2)_+, \quad \boldsymbol{v}(\gamma)_- \mapsto \boldsymbol{v}(\gamma_1)_- \otimes \boldsymbol{v}(\gamma_2)_-.$$

(2) If one of $\{\gamma_1, \gamma_2\}$ is contractible and the other is noncontractible, assume without loss of generality that γ_1 is contractible and γ_2 is noncontractible. Then γ is isotopic to γ_2, and the orientations of γ and

γ_2 are canonically identified. Let $\mathfrak{o}$ and $\mathfrak{o}'$ be the two orientations of γ, and use the same notation to denote the corresponding orientations of γ_2. Define the map $T_\lambda(b)$ by

$$\boldsymbol{v}(\gamma)_{\mathfrak{o}} \mapsto \boldsymbol{v}(\gamma_1)_- \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}}, \quad \boldsymbol{v}(\gamma)_{\mathfrak{o}'} \mapsto \boldsymbol{v}(\gamma_1)_- \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}'}.$$

(3) If both γ_1 and γ_2 are noncontractible and γ is contractible, then γ_1 and γ_2 are isotopic to each other, and the orientations of γ_1 are γ_2 are canonically identified. Let $\mathfrak{o}$ and $\mathfrak{o}'$ be the orientations of γ_1 and use the same notation for the orientations of γ_2. Define the map $T_\lambda(b)$ by

$$\boldsymbol{v}(\gamma)_+ \mapsto \boldsymbol{v}(\gamma_1)_{\mathfrak{o}} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}'} + \boldsymbol{v}(\gamma_1)_{\mathfrak{o}'} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}}, \quad \boldsymbol{v}(\gamma)_- \mapsto 0.$$

(4) If all of γ, γ_1 and γ_2 are noncontractible, let N be the regular neighborhood of $b \cup \gamma$. Then N is a sphere with three disks removed, and the three boundary components of N are parallel to γ_1, γ_2 and γ. The boundary orientation of N defines an orientation on each of γ_1, γ_2 and γ, and we denote them by $\mathfrak{o}_1$, $\mathfrak{o}_2$ and $\mathfrak{o}$, respectively. Denote their opposite orientations by $\mathfrak{o}_1'$, $\mathfrak{o}_2'$ and $\mathfrak{o}'$. Define the map $T_\lambda(b)$ by

$$\boldsymbol{v}(\gamma)_{\mathfrak{o}'} \mapsto \lambda \cdot \boldsymbol{v}(\gamma_1)_{\mathfrak{o}_1} \otimes \boldsymbol{v}(\gamma_2)_{\mathfrak{o}_2}, \quad \boldsymbol{v}(\gamma)_{\mathfrak{o}} \mapsto 0.$$

Case 3 (both c and c_b have exactly one component) In this case, define $T_\lambda(b)$ to be zero.

In general, suppose $c = c^{(1)} \sqcup c^{(2)}$ such that ∂b is disjoint from $c^{(2)}$ and intersects every component of $c^{(1)}$. We define the band surgery homomorphism $T_\lambda(b)\colon V_\lambda(c) \to V_\lambda(c_b)$ to be

$$T_\lambda(b) = T_\lambda(b)|_{V(c^{(1)})} \otimes \mathrm{id}|_{V(c^{(2)})}. \tag{3-1}$$

Remark 3.3 In the above definition, the coefficient λ only appeared in Cases 1(5) and 2(4).

4 Commutativity of band surgery homomorphisms

The main result of this section is the following proposition:

Proposition 4.1 *Suppose c is an embedded closed 1-manifold on Σ, and suppose b_1 and b_2 are two disjoint bands attached to c. Then for all $\lambda \in R$,*

$$T_\lambda(b_1) \circ T_\lambda(b_2) = T_\lambda(b_2) \circ T_\lambda(b_1). \tag{4-1}$$

The key idea is to use the following two lemmas to reduce Proposition 4.1 to the case when Σ has genus 0 or 1.

Lemma 4.2 *Suppose Σ is an oriented surface, and $\Sigma' \subset \Sigma$ is an embedded surface whose orientation is induced by Σ. Suppose the embedding of Σ' in Σ is π_1-injective. Suppose c is an embedded closed 1-manifold in Σ', and b_1 and b_2 are two disjoint bands in Σ' attached to c. Then*

$$T_\lambda^{\Sigma'}(b_1) \circ T_\lambda^{\Sigma'}(b_2) = T_\lambda^{\Sigma'}(b_2) \circ T_\lambda^{\Sigma'}(b_1)$$

on $V_{\Sigma'}(c)$ if and only if

$$T_\Sigma(b_1) \circ T_\Sigma(b_2) = T_\Sigma(b_2) \circ T_\Sigma(b_1)$$

on $V_\Sigma(c)$.

Proof Since the embedding of Σ' in Σ is π_1-injective, there is a canonical isomorphism from $V_{\Sigma'}(c)$ to $V_\Sigma(c)$ for every embedded 1-manifold $c \subset \Sigma'$ which takes the generators of $V_{\Sigma'}(c)$ to the corresponding generators of $V_\Sigma(c)$, and this isomorphism intertwines with $T_\lambda^{\Sigma'}$ and T_λ^{Σ}, so the lemma is proved. □

Lemma 4.3 *Assume Proposition 4.1 holds whenever Σ is a sphere, finitely punctured sphere, torus or finitely punctured torus. Then Proposition 4.1 holds for all cases.*

Proof Without loss of generality, we may assume that every component of c intersects b_1 and b_2 nontrivially, and that $c \cup b_1 \cup b_2$ is connected.

In this case, $c \cup b_1 \cup b_2$ is homotopy equivalent to the wedge sum of three circles. Therefore its Euler characteristic is -2.

Let N be a closed regular neighborhood of $c \cup b_1 \cup b_2$ in Σ. Let Σ' be obtained from N as follows: For each component γ of ∂N, if γ is contractible in Σ but not contractible in N, then γ bounds a disk D_γ in Σ such that $D_\gamma \cap N = \gamma$. Define Σ' to be the union of N and all disks D_γ as above. Then the embedding of Σ' in Σ is π_1-injective. Note that $\chi(\Sigma') \geq \chi(N) = -2$. If $\chi(\Sigma') = \chi(N) = -2$, then no disk D_γ appears in the construction of Σ', so $\partial\Sigma' \neq \varnothing$. Therefore the genus of Σ' is 0 or 1. By assumption, (4-1) holds on Σ'. Hence by Lemma 4.2, the desired equation also holds on Σ. □

The rest of this section proves Proposition 4.1 when Σ is a sphere, finitely punctured sphere, torus or finitely punctured torus.

4.1 The genus-zero case

We first establish (4-1) when Σ is a sphere or a finitely punctured sphere. Our argument here is inspired by the work of Winkeler [16].

Lemma 4.4 *Equation* (4-1) *holds if Σ is a sphere or a finitely punctured sphere.*

Proof If Σ is a sphere or a disk, then every curve is contractible, and Lemma 3.1(3) is not possible. In this case, our definition of $T_\lambda(b)$ does not depend on λ and it coincides with the definition of the merge and split maps in standard Khovanov theory. Therefore (4-1) holds.

When Σ has $n \geq 2$ boundary components, we view Σ as a disk B with $n-1$ interior disks $B_1, \dots, B_{n-1}$ removed. Assume the orientation of Σ is defined so that the boundary orientation on ∂B is given by the counterclockwise orientation, and the boundary orientation on ∂B_i is the clockwise orientation.

Recall that when the surface Σ needs to be emphasized, we write $V(c)$, $\boldsymbol{v}(\gamma)_\circ$, $\boldsymbol{v}(\gamma)_\pm$ and $T_\lambda(b)$ as $V^\Sigma(c)$, $\boldsymbol{v}^\Sigma(\gamma)_\circ$, $\boldsymbol{v}^\Sigma(\gamma)_\pm$ and $T_\lambda^\Sigma(b)$, respectively.

For each embedded closed 1-manifold $c \subset \Sigma$, define an isomorphism $\Phi\colon V^B(c) \to V^\Sigma(c)$ as follows. For each component γ of c, if γ is contractible in Σ, define

$$\Phi(\boldsymbol{v}^B(\gamma)_\pm) = \boldsymbol{v}^\Sigma(\gamma)_\pm.$$

If γ is noncontractible in Σ, let $\mathfrak{o}$ denote the counterclockwise orientation of γ, let $\mathfrak{o}'$ denote the clockwise orientation of γ, and define

$$\Phi(\boldsymbol{v}^B(\gamma)_+) = \boldsymbol{v}^\Sigma(\gamma)_{\mathfrak{o}}, \quad \Phi(\boldsymbol{v}^B(\gamma)_-) = \boldsymbol{v}^\Sigma(\gamma)_{\mathfrak{o}'}.$$

Since $T^B_\lambda(b)$ does not depend on λ, we denote it by $T^B(b)$. Then

$$\Phi \circ T^B(b) \circ \Phi^{-1}$$

is a homomorphism from $V^\Sigma(c)$ to $V^\Sigma(c_b)$.

For each $i \in \{1, \dots, n-1\}$, define a grading on $V^\Sigma(c)$ as follows. If a circle γ is a contractible curve on Σ, define the degree of $\boldsymbol{v}^\Sigma(\gamma)_\pm$ to be zero. If γ is noncontractible, for each orientation $\mathfrak{o}$ of γ, define the degree of $\boldsymbol{v}^\Sigma(\gamma)_{\mathfrak{o}}$ to be the rotation number of $\gamma_{\mathfrak{o}}$ around B_i. Here our convention on the rotation number is defined so that counterclockwise orientations always have nonnegative rotation numbers. Define the grading of the tensor product of a set of generators to be the sum of the grading of each generator.

By checking all the cases in the definition of $T_\lambda(b)$, it is straightforward to verify that the map $T^\Sigma(b)$ preserves all the $n-1$ gradings defined above. Moreover, for each $i \in \{1, \dots, n-1\}$, the map $\Phi \circ T^B(b) \circ \Phi^{-1}$ does not increase the i^{th} grading. The components of $\Phi \circ T^B(b) \circ \Phi^{-1}$ that preserve all the $n-1$ gradings is equal to the map $T^\Sigma_1(b)$, which is the map T^Σ_λ when $\lambda = 1$. Since $T^B(b_1) \circ T^B(b_2) = T^B(b_2) \circ T^B(b_1)$ on B, we conclude that (4-1) holds for T^Σ_1.

To show that (4-1) holds for general λ, define $T^\Sigma_\delta = T^\Sigma_1 - T^\Sigma_0$. Then

$$T^\Sigma_\lambda = T^\Sigma_0 + \lambda \cdot T^\Sigma_\delta.$$

We define another grading on $V^\Sigma(-)$ as follows. If a circle γ is a contractible curve on Σ, define the degree of $\boldsymbol{v}_\Sigma(\gamma)_\pm$ to be zero. If γ is noncontractible, for each orientation $\mathfrak{o}$ of γ, define the degree of $\boldsymbol{v}^\Sigma(\gamma)_{\mathfrak{o}}$ to be 1 if $\mathfrak{o}$ is the counterclockwise orientation, and define the degree of $\boldsymbol{v}^\Sigma(\gamma)_{\mathfrak{o}}$ to be -1 if $\mathfrak{o}$ is the clockwise orientation. Define the grading of the tensor product of a set of generators to be the sum of the grading of each generator.

By checking all the cases in the definition of T^Σ_λ, it is straightforward to verify that under the above grading, the map T^Σ_0 is homogeneous with degree 0, and T^Σ_δ is homogeneous with degree -1. Since (4-1) holds for $\lambda = 1$, we have

$$\begin{gathered}
T^\Sigma_0(b_1) \circ T^\Sigma_0(b_2) = T^\Sigma_0(b_2) \circ T^\Sigma_0(b_1), \\
T^\Sigma_\delta(b_1) \circ T^\Sigma_0(b_2) + T^\Sigma_0(b_1) \circ T^\Sigma_\delta(b_2) = T^\Sigma_\delta(b_2) \circ T^\Sigma_0(b_1) + T^\Sigma_0(b_2) \circ T^\Sigma_\delta(b_1), \\
T^\Sigma_\delta(b_1) \circ T^\Sigma_\delta(b_2) = T^\Sigma_\delta(b_2) \circ T^\Sigma_\delta(b_1).
\end{gathered}$$

Therefore (4-1) holds for all $\lambda \in R$. □

4.2 The genus-one case

Now we prove Proposition 4.1 when Σ is a torus or a finitely punctured torus. Let Σ_0 be a torus and suppose $\Sigma = \Sigma_0 \backslash \{p_1, \dots, p_n\}$ with $n \geq 0$. Let c, b_1 and b_2 be as in Proposition 4.1. By the definition of T_λ, we may assume without loss of generality that every component of c intersects $\partial(b_1 \cup b_2)$ nontrivially.

Lemma 4.5 *Assume every simple closed curve $\gamma_0 \subset \Sigma_0$ that is disjoint from $c \cup b_1 \cup b_2$ is contractible in Σ_0. Then up to orientation-preserving diffeomorphisms of Σ_0, there are only eight possible configurations of c, b_1 and b_2 as subsets of Σ_0, which are shown in Figure 2.*

In each case of Figure 2, the torus Σ_0 is the quotient space obtained by gluing the two boundary components of the annulus. The blue curves denote the 1-manifold c, and the disks b_1 and b_2 are defined to be the thickening of the red arcs.

Proof We discuss the following cases:

If c contains two circles γ_1 and γ_2, and both of them are contractible, let $D_1, D_2 \subset \Sigma$ denote the disks bounded by γ_1 and γ_2. Then $D_1 \cup D_2 \cup b_1 \cup b_2$ is a disk or an annulus, and hence there exists a circle γ_0 in the complement of $c \cup b_1 \cup b_2$ that is contractible, contradicting the assumptions.

If c contains two circles γ_1 and γ_2 such that both γ_1 and γ_2 are noncontractible, then γ_1 and γ_2 must be parallel to each other. The complement $\Sigma_0 \backslash (\gamma_1 \cup \gamma_2)$ contains two components. If every simple closed curve in $\Sigma_0 \backslash (c \cup b_1 \cup b_2)$ is contractible in Σ_0, then the interior of b_1 and b_2 must be contained in different components of $\Sigma_0 \backslash (\gamma_1 \cup \gamma_2)$, and ∂b_i must intersect both components of c for each i. Therefore, up to orientation-preserving diffeomorphisms of Σ_0, the configuration is given by Figure 2(1).

If c contains two circles γ_1 and γ_2, where γ_1 is contractible and γ_2 is not contractible, let D_1 be the disk bounded by γ_1. If either b_1 or b_2 is contained in D_1, then $D_1 \cup c \cup b_1 \cup b_2$ deformation retracts onto γ_2, so there exists a noncontractible simple closed curve in Σ_0 that is disjoint from $D_1 \cup c \cup b_1 \cup b_2$, which

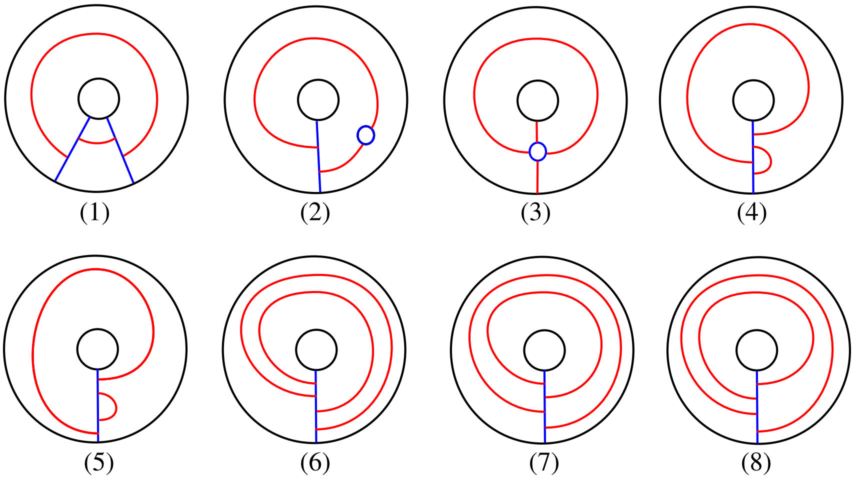


Figure 2: All possible configurations.

contradicts the assumptions. Therefore both b_1 and b_2 must be on the outside of D_1, so $b_1 \cup D_1 \cup b_2$ deformation retracts onto an arc with both endpoints on γ_2. The assumptions then imply that $c \cup b_1 \cup b_2$ is given by Figure 2(2) up to orientation-preserving diffeomorphisms of Σ_0.

If c consists of one simple closed curve γ that is contractible in Σ_0, let D be the disk bounded by γ. Then b_1 and b_2 must be the thickening of two disjoint arcs r_1 and r_2 in $\Sigma_0 \backslash D$. For $i = 1, 2$, let $\bar{r}_i$ be the circle obtained by the union of r_i with an arc in D. Since r_1 and r_2 are disjoint arcs, we may choose the arcs in D so that $\bar{r}_1$ and $\bar{r}_2$ are either disjoint or intersect transversely at one point. The assumptions then imply that $\bar{r}_1$ and $\bar{r}_2$ must intersect transversely at one point. Hence the configuration is given by Figure 2(3) up to orientation-preserving diffeomorphisms of Σ_0.

If c consists of one noncontractible simple closed curve, then the possible configurations are given by Figure 2(4)–(8). □

Lemma 4.6 *Equation* (4-1) *holds if* Σ *is a torus or a finitely punctured torus.*

Proof If there exists a noncontractible simple closed curve $\gamma_0 \subset \Sigma_0$ that is disjoint from $c \cup b_1 \cup b_2$, we may cut open Σ_0 along γ_0, and the desired result follows from Lemmas 4.4 and 4.2. Therefore, by Lemma 4.5, we only need to consider the eight cases given by Figure 2.

In (2) and (4)–(8), both sides of (4-1) are zero because Lemma 3.1(3) appears on both sides of the equations.

For (1) and (3), the complement $\Sigma \backslash (c \cup b_1 \cup b_2)$ has two connected components. Therefore, by Lemma 4.2 again, we only need to consider the cases when there is at most one puncture on each component.

Recall that n denotes the number of punctures on Σ_0. For (1) with $n = 0$ or 2, and for (3), there is an orientation-preserving diffeomorphism of Σ_0 that preserves c and Σ, is orientation-preserving on c, and switches b_1 and b_2. Therefore (4-1) holds.

For (1) with $n = 1$, it is straightforward to verify that both sides of (4-1) are zero. □

5 Khovanov homology

Suppose $L \subset (-1, 1) \times \Sigma$ is a link. For each λ, we define a homology invariant for L using the maps T_λ.

Suppose a link L is given by a diagram D on Σ with k crossings, and fix an ordering of the crossings. For $v = (v_1, v_2, \dots, v_k) \in \{0, 1\}^k$, resolving the crossings of D by a sequence of 0-smoothings and 1-smoothings (see Figure 3) by v turns D into an embedded closed 1-manifold in Σ. Denote the resolved diagram by D_v.

Whenever u is obtained from v by changing one coordinate from 0 to 1, there is a band b near the crossing such that v is obtained from u by a band surgery along b. Define $d_{vu}^\lambda \colon V(D_v) \to V(D_u)$ to be $T_\lambda(b)$. Let e_i be the i^{th} standard basis vector of $\mathbb{Z}^k$. Define

$$\mathrm{CKh}_{\Sigma,\lambda}(L) = \bigoplus_{v \in \{0,1\}^k} V(D_v),$$

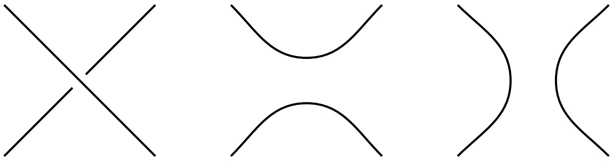


Figure 3: Two types of smoothings.

and define an endomorphisms on $\mathrm{CKh}_\Sigma(L)$ by

$$\mathcal{D}_{\Sigma,\lambda} = \sum_i \sum_{u-v=e_i} (-1)^{\sum_{i<j\le c} v_j} d_{vu}.$$

By (4-1), we have $\mathcal{D}^2_{\Sigma,\lambda} = 0$.

We define a quantum grading and a homological grading on $\mathrm{CKh}_{\lambda,\Sigma}(L)$ as follows. For each circle γ, if γ is noncontractible, define the quantum grading on $V(\gamma)$ to be zero. If γ is contractible, define the quantum grading of $\boldsymbol{v}(\gamma)_+$ to be 1 and the quantum grading of $\boldsymbol{v}(\gamma)_-$ to be -1. This grading then extends to a grading on $\mathrm{CKh}_{\lambda,\Sigma}(L)$. Define the homology grading of $V(D_v) \subset \mathrm{CKh}_{\lambda,\Sigma}(L)$ to be the sum of coordinates in v.

There is also a grading on $\mathrm{CKh}_{\lambda,\Sigma}(L)$ over $H_1(\Sigma;\mathbb{Z})$ defined as follows. For each circle γ, if γ is contractible, define the grading on $V(\gamma)$ to be zero. If γ is noncontractible, for each orientation $\mathfrak{o}$ of γ, define the grading of $\boldsymbol{v}(\gamma)_{\mathfrak{o}}$ to be the fundamental class of $\gamma_{\mathfrak{o}}$.

Following the standard convention, we use curly brackets $\{l\}$ to denote the shifting in quantum gradings by l (namely, adding the quantum grading to each homogeneous element by l); we use the square brackets $[l]$ to denote the shifting in homology gradings by l.

Theorem 5.1 *The homology of*

$$(\mathrm{CKh}_{\lambda,\Sigma}(L)[-n_-]\{n_+ - 2n_-\}, \mathcal{D}_{\Sigma,\lambda})$$

as a $\mathbb{Z}\oplus\mathbb{Z}\oplus H_1(\Sigma;\mathbb{Z})$-graded module is independent of the diagram or the ordering of the crossings, where n_+ and n_- denote the number of positive and negative crossings of the diagram.

Proof The proof is identical to the proof of the invariance of the standard Khovanov homology under Reidemeister moves in [6]. Besides (3-1) and (4-1), the only properties about the band homomorphisms $T_\lambda(b)$ needed in the proof are the following:

(1) If γ is a contractible circle, then $V(\gamma)$ is rank 2 with two generators $\boldsymbol{v}(\gamma)_\pm$.

(2) Suppose the band surgery along b merges two circles γ_1 and γ_2 to γ, where γ_1 is contractible. Then γ_2 and γ are isotopic, and this isotopy defines a canonical isomorphism $\iota\colon V(\gamma_2) \to V(\gamma)$. Then $T_\lambda(b)(\boldsymbol{v}(\gamma_1)_+ \otimes x) = \iota(x)$ for all $x \in V(\gamma_2)$.

(3) Suppose the band surgery along b splits one circle γ to circles γ_1 and γ_2, where γ_1 is contractible. Then γ_2 and γ are isotopic, and this isotopy defines a canonical isomorphism $\iota\colon V(\gamma)\to V(\gamma_2)$. Then the composition map

$$V(\gamma)\xrightarrow{T_\lambda(b)} V(\gamma_1)\otimes V(\gamma_2)\xrightarrow{/\boldsymbol{v}(\gamma_1)_+=0}\operatorname{span}\{\boldsymbol{v}(\gamma_1)_-\}\otimes V(\gamma_2)$$

is given by the tensor product with $\boldsymbol{v}(\gamma_1)_-$, where the second map above is a quotient map.

The only remark worth making is that there is a typo in the definition of the "transpose" map in [6, Section 3.5.5]. The map Υ on the top layer should map the *quotient image* of the pair (β_1,γ_1) to the *quotient image* of the pair (β_2,γ_2) *such that* $\gamma_1+\tau_1\beta_1=\gamma_2+\tau_2\beta_2$. The italicized phrases and the last equation in the previous sentence were missing in [6]. □

Definition 5.2 We define the homology of

$$(\mathrm{CKh}_{\lambda,\Sigma}(L)[-n_-]\{n_+-2n_-\},\mathcal{D}_{\Sigma,\lambda})$$

as a $\mathbb{Z}\oplus\mathbb{Z}\oplus H_1(\Sigma;\mathbb{Z})$-module to be the Khovanov invariant of $L\subset(-1,1)\times\Sigma$, and denote it by $\Sigma\mathrm{Kh}_{\lambda,\Sigma}(L;R)$.

When there is no risk of confusion on the surface Σ and the coefficient ring R, we will also denote $\Sigma\mathrm{Kh}_{\lambda,\Sigma}(L;R)$ by $\Sigma\mathrm{Kh}_\lambda(L)$.

Remark 5.3 When $\lambda=0$, the differential map $\mathcal{D}_{\Sigma,\lambda}$ is identical to the differential map of Asaeda–Przytycki–Sikora homology defined in [1]. When $R=\mathbb{Z}$, $\lambda=1$ and Σ is a punctured disk, the homology $\Sigma\mathrm{Kh}_\lambda$ recovers the invariant defined by Winkeler [16].

6 Relations with instanton Floer homology

This section explains the motivation of the definition of $T_\lambda(b)$ from instanton homology. We will also prove the following detection result:

Theorem 6.1 *Suppose that* Σ *is a surface with genus zero, and* $L\subset(-1,1)\times\Sigma$ *is a link. Then* $\operatorname{rank}_{\mathbb{Z}/2}\Sigma\mathrm{Kh}_1(L;\mathbb{Z}/2)\geq 2$, *and equality holds if and only if* L *is isotopic to an embedded knot in* Σ.

The detection problems of Khovanov homology and other quantum invariants of knots and links have attracted considerable attention since the introduction of the invariants. Kronheimer and Mrowka [10] proved that the standard Khovanov homology detects the unknot; see also [7; 8]. Since then, a large number of detection results on Khovanov homology were obtained using different versions of Floer theory, for example, by [2; 3; 4; 5; 13; 14; 19; 20]. The main theorem in [12] gave the first detection result on Khovanov homology that is valid on an infinite family of manifolds. Theorem 6.1 above is an improvement of the main theorem of [12].

In fact, by a spectral sequence of Winkeler [16, Theorem 1.3], we have

$$\operatorname{rank}_{\mathbb{Z}/2} \Sigma\mathrm{Kh}_0(L;\mathbb{Z}/2) \geq \operatorname{rank}_{\mathbb{Z}/2} \Sigma\mathrm{Kh}_1(L;\mathbb{Z}/2). \tag{6-1}$$

The main result in [12] states a classification of all links L such that $\Sigma\mathrm{Kh}_0(L;\mathbb{Z}/2)$ has the minimum possible rank. Theorem 6.1 immediately implies the result in [12] because of (6-1).

6.1 Motivation from instanton Floer homology

We start by discussing the motivation of the definition of $T_\lambda(b)$ from computations in instanton Floer homology. All instanton homology groups here will be defined with $\mathbb{C}$ coefficients. We refer the reader to [12, Section 2] for the general notation and properties of singular instanton Floer homology. In particular, we will use $\mathrm{I}(Y, L, \omega)$ to denote the instanton homology of a nonintegral triple (Y, L, ω), where Y is a closed 3-manifold, $L \subset Y$ is a link and $(\omega, \partial\omega) \subset (Y, L)$ is an embedded 1-manifold. The nonintegral condition is a technical condition to ensure that Floer homology is well-defined, and the statement of the condition can be found in [12, Section 2.3]. If $\Sigma \subset Y$ is an oriented embedded surface, then $\mathrm{I}(Y, L, \omega \mid \Sigma)$ denotes a subspace of $\mathrm{I}(Y, L, \omega)$ introduced by [18]; the complete definition can be found in [12, Definition 2.10]. If Σ is connected, one may regard $\mathrm{I}(Y, L, \omega \mid \Sigma)$ as the component of $\mathrm{I}(Y, L, \omega)$ at the maximum possible grading with respect to a grading induced by Σ.

Suppose Q is a *closed* oriented surface, and let L be a link in $(-1, 1) \times Q$. Let p be a point on Q that is disjoint from the projection of L to Q. In [12], the authors studied the instanton homology group

$$\Sigma\mathrm{HI}_{Q,p}(L) := \mathrm{I}(S^1 \times Q, L, S^1 \times \{p\} \mid \{t_*\} \times Q), \tag{6-2}$$

where S^1 is viewed as the quotient space of $[-1, 1]$ with -1 identified with 1, and $t_* \in S^1$ is a fixed basepoint.

Remark 6.2 The closed surface in (6-2) was denoted by R instead of Q in [12]. We use the notation Q here to avoid collision of notation with the coefficient ring.

Suppose c is an embedded 1-manifold in Q, and b is a band attached to c that is disjoint from p. Then the band surgery along b defines a link cobordism from c to c_b as links in $(-1, 1) \times Q$. Therefore it induces a cobordism map for Floer homology groups (up to sign)

$$\Sigma\mathrm{HI}_{Q,p}(b) \colon \Sigma\mathrm{HI}_{Q,p}(c) \to \Sigma\mathrm{HI}_{Q,p}(c_b).$$

It was proved in [12, Proposition 6.12] that the maps $\Sigma\mathrm{HI}_{Q,p}(b)$ are components of the second page of a variant of Kronheimer and Mrowka's spectral sequence which abuts to $\Sigma\mathrm{HI}_{Q,p}(L)$. In [12, Proposition 6.11], the cobordism maps $\Sigma\mathrm{HI}_{Q,p}(b)$ were computed for multiple special cases; in all the computed cases, $\Sigma\mathrm{HI}_{Q,p}(b)$ is equal to $T_\lambda(b)$ for some $\lambda \in \mathbb{C}$ in a suitable sense. This motivated our definition of the map $T_\lambda(b)$. It is natural to conjecture that the second page of Kronheimer and Mrowka's spectral sequence is isomorphic to $\Sigma\mathrm{Kh}_{\lambda,Q}(L;\mathbb{C})$ for some $\lambda \in \mathbb{C}$.

Conjecture 6.3 *Suppose Q is a closed oriented surface, and let L be a link in $(-1,1)\times Q$ given by a diagram D on Q. Let p be a fixed point on Q that is disjoint from D. Then there exist $\lambda\in\mathbb{C}$ and a spectral sequence that abuts to $\Sigma\mathrm{HI}_{Q,p}(L)$ whose second page is isomorphic to $(\mathrm{CKh}_{\lambda,Q}(L),\mathcal{D}_{\Sigma,\lambda})$ (with $\mathbb{C}$-coefficients) as a chain complex.*

6.2 Proof of Theorem 6.1

We may assume without loss of generality that Σ is connected and compact. If $\Sigma=S^2$, the desired result follows from the unknot detection theorem for the standard Khovanov homology [10]. We assume in the following that $\partial\Sigma\neq\varnothing$.

Assume F is a connected oriented surface such that ∂F equals $\partial\Sigma$ with the reversed orientation. Let $Q=\Sigma\cup_\partial F$. By [12, Proposition 6.12], there is a spectral sequence that abuts to $\Sigma\mathrm{HI}_{Q,p}(L)$ whose second page is given by maps of the form $\Sigma\mathrm{HI}_{Q,p}(b)$, where b is a band corresponding to a crossing change between different smoothings of the diagram D. By [12, Lemma 5.2], the second page of Kronheimer and Mrowka's spectral sequence, as a linear space, is isomorphic to $\mathrm{CKh}_{\lambda,Q}(L)$ (with $\mathbb{C}$-coefficients).

By [12, Proposition 6.11], after conjugating by an isomorphism from $V(-)$ to $\Sigma\mathrm{HI}_{Q,p}(-)$ defined in [12, Section 5.2], each component of the differential map on the second page has the form $i^kT_\lambda(b)$ for some $\lambda\in\mathbb{C}$ and $k\in\mathbb{Z}$. We show that it is possible to choose an isomorphism from $V(-)$ to $\Sigma\mathrm{HI}_{Q,p}(-)$ such that after conjugation, the coefficients λ are the same (up to sign) on all the components of the differential map. Moreover, we show that the coefficient λ must be nonzero.

In the following, we will denote $\Sigma\mathrm{HI}_{Q,p}(-)$ by $\Sigma\mathrm{HI}(-)$ to simplify notation.

Let $\lambda_1,\dots,\lambda_4$ be the constants from [12, Section 6]. By [12, Lemma 6.9], one can rescale the isomorphisms in [12, Section 5.2] so that $\lambda_1=\pm1$ and $\lambda_3=\pm1$.

Lemma 6.4 *Assume the generator w_0 defined in* [12, Section 5.2.1] *is chosen so that $\lambda_1=\pm1$ and $\lambda_3=\pm1$. Then $\lambda_2=\pm\lambda_4$.*

Proof Consider the two bands in Figure 4 and apply the TQFT property of $\Sigma\mathrm{HI}(b)$. □

Lemma 6.5 *The coefficients λ_2 and λ_4 are both nonzero.*

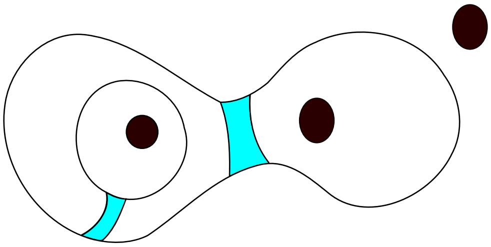


Figure 4: Two bands.

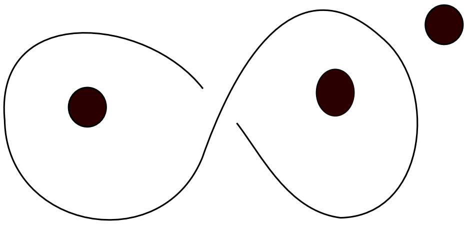


Figure 5: The diagram D on Σ.

Proof Suppose Σ is a sphere with three open disks removed, and let D be a diagram on Σ as shown in Figure 5. Let $D_0=\gamma$ be the resolution of D into one circle, let $D_1=\gamma_1\cup\gamma_2$ be the resolution of D into two circles and let b be the band relating D_0 and D_1. Let K be the knot represented by D. Then by [10, Theorem 6.8], there is an exact triangle

$$\cdots\to\Sigma\mathrm{HI}(\gamma)\xrightarrow{\Sigma\mathrm{HI}(b)}\Sigma\mathrm{HI}(\gamma_1\cup\gamma_2)\to\Sigma\mathrm{HI}(K)\to\Sigma\mathrm{HI}(\gamma)\xrightarrow{\Sigma\mathrm{HI}(b)}\Sigma\mathrm{HI}(\gamma_1\cup\gamma_2)\to\cdots.$$

By [12, Lemma 6.4], $\lambda_2=\lambda_4=0$ if and only if $\Sigma\mathrm{HI}(b)=0$ in the above exact sequence. By [12, Lemma 5.2], we have $\dim\Sigma\mathrm{HI}(\gamma_1\cup\gamma_2)=4$ and $\dim\Sigma\mathrm{HI}(\gamma)=2$. Therefore we only need to show

$$\dim\Sigma\mathrm{HI}(K)<6. \tag{6-3}$$

Let L be the link in the thickened annulus as shown in Figure 6. Pick a meridional disk in the thickened annulus which intersects L at two points. We decompose the thickened annulus along this disk and obtain a product sutured thickened disk with a tangle T in it. The sutured instanton Floer homology of this sutured manifold with tangle T is isomorphic to $\mathrm{AHI}(L,2)$ according to [11, Theorem 2.14], where $\mathrm{AHI}(L,2)$ denotes the component of the annular instanton Floer homology with Alexander grading 2. (The Alexander grading is also called the annular grading, the f-grading or the k-grading in the literature. We follow the terminology of [15, Definition 2.2] here, which agrees with the notation in [12].)

The tangle T has two product vertical components. We remove the tubular neighborhoods of the two vertical components and add a meridian suture to the boundary of each neighborhood to obtain a sutured

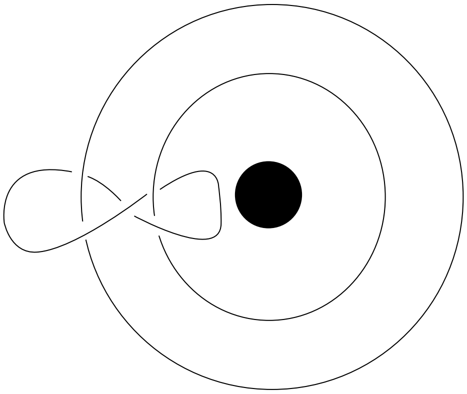


Figure 6: The annular link L.

manifold M' with a knot K' in it. Moreover, this process does not change the sutured instanton Floer homology according to [18, Lemma 7.10] and its proof. Therefore

$$\mathrm{SHI}(M', \gamma_{M'}, K') \cong \mathrm{AHI}(K, 2).$$

Notice that in the definition of sutured instanton Floer homology, the pairs (M', K') and (M, K) can be given the same closure; therefore their sutured instanton homologies are isomorphic. As a result, we have

$$\mathrm{SHI}([-1,1]\times\Sigma, \{0\}\times\Sigma, K) \cong \mathrm{AHI}(L, 2). \tag{6-4}$$

A straightforward calculation shows that

$$\mathrm{AKh}(L, 2; \mathbb{C}) \cong \mathbb{C}^4,$$

where $\mathrm{AKh}(L, 2; \mathbb{C})$ denotes the component of the annular Khovanov homology of L with Alexander grading 2 and with coefficient ring $\mathbb{C}$. According to [17, Theorem 5.16], we have

$$\dim \mathrm{AHI}(L, 2; \mathbb{C}) \le \dim \mathrm{AKh}(L, 2; \mathbb{C}) = 4.$$

Therefore, (6-4) implies

$$\dim \Sigma\mathrm{HI}(K) = \dim \mathrm{SHI}([-1,1]\times\Sigma, \{0\}\times\Sigma, K) \le 4.$$

This verifies (6-3), and hence the desired result is proved. □

Theorem 6.1 can now be proved using an argument from [12].

Proof of Theorem 6.1 When Σ is a compact surface with genus zero, there is a grading on $V(-)$ such that $T_0(b)$ is homogeneous with degree zero and $T_1(b)$ is homogeneous with degree -1. Since $\lambda_2 \neq 0$, we can rescale the map $\Theta_{w_0,\sigma}$ in [12] by a factor of λ_2^k at degree k. By the discussion in [12, Section 6], there is a spectral sequence of chain complexes in $\mathbb{C}$-coefficients that converges to $\mathrm{I}([-1,1]\times\Sigma, \{0\}\times\partial\Sigma, L)$, whose second page (E_2, d_2) is isomorphic to the chain complex $(\mathrm{CKh}_{\Sigma,1}(L), \mathcal{D}_{\Sigma,1})$ up to multiplications by integer powers of i on the components of the differential map. In other words, there exists a chain complex (C, d) defined with $\mathbb{Z}[i]$ coefficients, such that when reducing to $\mathbb{C}$ coefficients, it is isomorphic to (E_2, d_2); when reducing to $\mathbb{Z}[i]/(i-1) \cong \mathbb{Z}/2$ coefficients, it is isomorphic to the chain complex $(\mathrm{CKh}_{\Sigma,1}(L), \mathcal{D}_{\Sigma,1})$. By the universal coefficient theorem,

$$\begin{aligned}\operatorname{rank}_{\mathbb{Z}/2} \Sigma\mathrm{Kh}_{\Sigma,1}(L; \mathbb{Z}/2) &\ge \operatorname{rank}_{\mathbb{Z}[i]} H(C, d) = \dim_{\mathbb{C}} H(E_2, d_2) \\ &\ge \dim_{\mathbb{C}} \mathrm{I}([-1,1]\times\Sigma, \{0\}\times\partial\Sigma, L),\end{aligned}$$

and the desired result follows from [12, Theorem 1.3]. □

References

[1] **M M Asaeda**, **J H Przytycki**, **A S Sikora**, *Categorification of the Kauffman bracket skein module of I-bundles over surfaces*, Algebr. Geom. Topol. 4 (2004) 1177–1210 MR Zbl

[2] **J A Baldwin**, **N Dowlin**, **A S Levine**, **T Lidman**, **R Sazdanovic**, *Khovanov homology detects the figure-eight knot*, Bull. Lond. Math. Soc. 53 (2021) 871–876 MR Zbl

[3] **J A Baldwin**, **Y Hu**, **S Sivek**, *Khovanov homology and the cinquefoil*, J. Eur. Math. Soc. (online publication January 2024)

[4] **J A Baldwin**, **S Sivek**, *Khovanov homology detects the trefoils*, Duke Math. J. 171 (2022) 885–956 MR Zbl

[5] **J A Baldwin**, **S Sivek**, **Y Xie**, *Khovanov homology detects the Hopf links*, Math. Res. Lett. 26 (2019) 1281–1290 MR Zbl

[6] **D Bar-Natan**, *On Khovanov's categorification of the Jones polynomial*, Algebr. Geom. Topol. 2 (2002) 337–370 MR Zbl

[7] **J E Grigsby**, **S M Wehrli**, *On the colored Jones polynomial, sutured Floer homology, and knot Floer homology*, Adv. Math. 223 (2010) 2114–2165 MR Zbl

[8] **M Hedden**, *Khovanov homology of the 2-cable detects the unknot*, Math. Res. Lett. 16 (2009) 991–994 MR Zbl

[9] **M Khovanov**, *A categorification of the Jones polynomial*, Duke Math. J. 101 (2000) 359–426 MR Zbl

[10] **P B Kronheimer**, **T S Mrowka**, *Khovanov homology is an unknot-detector*, Publ. Math. Inst. Hautes Études Sci. 113 (2011) 97–208 MR Zbl

[11] **Z Li**, **Y Xie**, **B Zhang**, *On Floer minimal knots in sutured manifolds*, Trans. Amer. Math. Soc. Ser. B 9 (2022) 499–516 MR Zbl

[12] **Z Li**, **Y Xie**, **B Zhang**, *Instanton homology and knot detection on thickened surfaces*, Selecta Math. 29 (2023) art. id. 84 MR Zbl

[13] **R Lipshitz**, **S Sarkar**, *Khovanov homology detects split links*, Amer. J. Math. 144 (2022) 1745–1781 MR Zbl

[14] **G Martin**, *Khovanov homology detects $T(2,6)$*, Math. Res. Lett. 29 (2022) 835–849 MR Zbl

[15] **L P Roberts**, *On knot Floer homology in double branched covers*, Geom. Topol. 17 (2013) 413–467 MR Zbl

[16] **Z Winkeler**, *Khovanov homology for links in thickened multipunctured disks*, Michigan Math. J. 74 (2024) 1–31 MR Zbl

[17] **Y Xie**, *Instantons and annular Khovanov homology*, Adv. Math. 388 (2021) art. id. 107864 MR Zbl

[18] **Y Xie**, **B Zhang**, *Instanton Floer homology for sutured manifolds with tangles* (2019) arXiv 1907.00547 To appear in J. Differential Geom.

[19] **Y Xie**, **B Zhang**, *Classification of links with Khovanov homology of minimal rank*, J. Eur. Math. Soc. 27 (2025) 333–394 MR Zbl

[20] **Y Xie**, **B Zhang**, *Instantons and Khovanov skein homology on $I \times T^2$*, Quantum Topol. 16 (2025) 75–115 MR

School of Mathematics and Statistics, University of South Florida
Tampa, FL, United States

Beijing International Center for Mathematical Research, Peking University
Beijing, China

Department of Mathematics, University of Maryland at College Park
College Park, MD, United States

zhenkun@usf.edu, yixie@pku.edu.cn, bzh@umd.edu

Received: 9 March 2023 Revised: 8 November 2023

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

Algebraic & Geometric Topology **25**:3 (2025) 1561–1597
DOI: 10.2140/agt.2025.25.1561
Published: 20 June 2025

Cubulating a free-product-by-cyclic group

FRANÇOIS DAHMANI
SURAJ KRISHNA MEDA SATISH

Let $G = H_1 * \cdots * H_k * F_r$ be a finitely generated torsion-free group and ϕ an automorphism of G that preserves this free factor system. We show that when ϕ is fully irreducible and atoroidal relative to this free factor system, the mapping torus $\Gamma = G \rtimes_\phi \mathbb{Z}$ acts relatively geometrically on a hyperbolic CAT(0) cube complex. This is a generalisation of a result of Hagen and Wise for hyperbolic free-by-cyclic groups.

20E08, 20E36, 20F65, 20F67

1 Introduction

Consider a finitely generated free group F and an automorphism $\phi\colon F \to F$. Hagen and Wise showed in [HW16] that if ϕ is atoroidal and fully irreducible, then the mapping torus $F \rtimes_\phi \mathbb{Z}$ acts properly and cocompactly on a hyperbolic CAT(0) cube complex. It often happens that automorphisms of free groups are neither atoroidal nor fully irreducible, suggesting various directions of generalisation. In [HW15], Hagen and Wise relaxed the requirement of full irreducibility and, using the sophisticated machinery of relative train track maps, showed that the mapping torus still acts geometrically on a CAT(0) cube complex. They asked (see the discussion around [HW15, Problem B]) if a systematic answer to which free-by-cyclic groups admit cubulations is possible, especially in the presence of polynomially growing subgroups.

Here, by investigating *relative cubulations* instead of usual cubulations, we provide an answer in great generality as to when such groups are relatively cubulated. Let ϕ be an automorphism of F and let $F = H_1 * H_2 * \cdots * H_k * F_r$ be a free decomposition that is preserved by ϕ (up to taking conjugates of the factors). Such a free decomposition always exists for any ϕ. In particular, when ϕ is not fully irreducible, there exists a free decomposition preserved by ϕ, relative to which ϕ is fully irreducible. Let us also assume that ϕ is atoroidal relative to the free decomposition. We will clarify the meanings of both these terms more precisely, using the notion of free factor systems, in Section 2. In our setting, we allow, for instance, elements to have polynomial growth under ϕ as long as they are elliptic in the free product. Under a mild complexity condition, the first author and Li showed in [DL22] that the mapping torus $F \rtimes_\phi \mathbb{Z}$ is hyperbolic relative to the suspensions of the free factors H_i. A particular case of our main result shows that such a mapping torus acts relatively geometrically on a hyperbolic CAT(0) cube complex.

1.1 Main result

Let G be a finitely generated group, ϕ an automorphism of G and H a subgroup of G whose conjugacy class is preserved by a power of ϕ. Let n be the minimal positive power of ϕ such that $\phi^n(H) = g^{-1}Hg$. Then we say that the suspension of H by ϕ in the semidirect product $G \rtimes_\phi \langle t \rangle$ is the group $H \rtimes_{\mathrm{ad}_g \circ \phi^n} \langle t^n g^{-1} \rangle$, where $\mathrm{ad}_g : G \to G$ denotes the inner automorphism given by $h \mapsto ghg^{-1}$.

To state our main result for automorphisms of free products of groups, we need the notions of full irreducibility and atoroidality *relative to a free factor system*. These are analogous to the corresponding notions for automorphisms of free groups, with the condition that the free decomposition is preserved, but within a free factor there are no restrictions. We also need a technical notion of *no twinned subgroups*.[1] We refer the reader to Section 2 for the definitions.

We also recall the notion of relative cubulation introduced by Einstein and Groves in [EG20]: a relatively hyperbolic group $(\Gamma, \mathscr{P})$ is *relatively cubulated* if it acts cocompactly on a CAT(0) cube complex with cell stabilisers either trivial or conjugate to a finite index subgroup in $\mathscr{P}$.

Theorem 1.1 *Let G be a finitely generated torsion-free group and let $G \cong H_1 * \cdots * H_k * F_r$ be a free decomposition such that each H_i is nontrivial. Let ϕ be an automorphism that preserves the associated free factor system. Assume that $k + r \geq 3$, that ϕ is fully irreducible relative to the free factor system and atoroidal relative to the free factor system, and that there exist no twinned subgroups. Then the mapping torus $\Gamma = G \rtimes_\phi \mathbb{Z}$ admits a relative cubulation for the peripheral structure of the suspensions of the free factors H_i.*

We recover the result of [HW16] when $G = F_r$ above,[2] as well as some cases of [HW15], through a telescopic argument of Groves and Manning [GM23, Theorem D]. In general, the relatively geometric action on CAT(0) cube complexes that we obtain otherwise still has interesting consequences.

Note that any group acting on a CAT(0) cube complex admits an action on an ℓ^2-space built using characteristic functions of hyperplane-halfspaces; see Niblo and Reeves [NR97]. We thus have:

Corollary 1.2 *Let G and ϕ be as in Theorem 1.1. Then the mapping torus Γ acts on a Hilbert space with unbounded orbits, with no global fixed point for G.*

[1] This requirement can in fact be removed; see work of the authors and Mutanguha [DMM25, Lemma 3.3] or Remark 2.5.

[2] It is stated there in terms of irreducible atoroidal automorphisms, which are fully irreducible by Dowdall, Kapovich, and Leininger [DKL15, Corollary B.4].

In addition, when G is residually finite, using a generalisation by Einstein and Groves [EG22, Theorem 1.6] of a well-known result of Haglund and Wise [HW08], we have:

Corollary 1.3 *Let G and ϕ be as in Theorem 1.1. Further, let G be residually finite. Then every full relatively quasiconvex subgroup of the mapping torus Γ is separable.*

Another consequence of Theorem 1.1 can be seen in our recent work with Mutanguha [DMM25], where we showed that all hyperbolic hyperbolic-by-cyclic groups are virtually special.

1.2 Method

Our procedure to cubulate follows the scheme laid out by Hagen and Wise [HW16]. The goal is to obtain a collection of codimension-1 subgroups of Γ and then apply Sageev's dual cube complex construction [Sag95]. The codimension-1 subgroups we build will be stabilised by full relatively quasiconvex subgroups in the relatively hyperbolic group Γ. We then apply the boundary criterion of Einstein and Groves [EG20].

In order to do this, an important tool is Francaviglia and Martino's absolute train tracks for free products [FM15]. Given G and an automorphism ϕ satisfying the hypotheses of the main result, there exists a G-tree T and a train track map $f: T \to T$ representing ϕ. Taking mapping cylinders for f, one then obtains a *flow space* on which Γ acts. The flow space has the structure of a tree of spaces, where the underlying graph is a line and vertex and edge spaces are copies of the tree T. The map f "flows" a point on any tree to its image in the next tree. We describe the flow space and various properties we need in order to define walls in the flow space in Section 2.

Before explaining how we build walls in our setup, let us motivate our construction in the surface case. Let M_f be the mapping torus of a closed hyperbolic surface S_g under a pseudo-Anosov map f. Cooper, Long and Reid showed in [CLR94] that in this case there exists an immersed quasi-Fuchsian surface in M_f (and hence a quasiconvex wall in the universal cover). First, take a simple closed curve C in S_g which is disjoint from its f-image. Such a simple closed curve exists up to taking a finite cover of M_f. The required immersed surface in M_f is then obtained by cutting S_g along C and $f(C)$ and gluing $C_\pm$ to $f(C_\mp)$ (the cut-and-cross-join technique).

Hagen and Wise mimicked this construction for the setup of hyperbolic free-by-cyclic groups in the fully irreducible case. A surface with a pseudo-Anosov map is now replaced by a graph with a train track map. The analogue of cutting along a simple closed curve is cutting along a point in the graph. However, the situation here is more complicated as a train track map is only a homotopy equivalence and not a homeomorphism. A point often has multiple preimages and the cut-and-cross-join operation is performed along all points with the same image.

We use the same operation, but now our train track representative is defined not on a finite graph but on the G-tree T. The lack of local finiteness of the tree T gives rise to additional difficulties, but we were able to manage them because of the behaviour of train track maps in this setting, and considerations of angles at vertices under relative hyperbolicity.

While cocompact cubulations require walls to be (relatively) quasiconvex, relative cubulations require walls to also be full. This forced us to introduce *saturations* of our walls in order to ensure fullness. The construction of walls and their saturations can be found in Section 4.

Finally, in order to use the boundary criterion, we need sufficiently many wall saturations to not only cut biinfinite geodesics in the flow space, but also to cut pairs of *principal flow lines* that are stabilised by maximal parabolic subgroups and pairs consisting of a geodesic ray and a principal flow line. We show this in Section 5, ensuring the separation of every pair of points in the Bowditch boundary. We verify the latter in Section 6, where we also give a proof of Theorem 1.1.

1.3 Questions

We end this introduction with three questions arising from this work.

Question 1.4 Let $G = A * B$ be a torsion-free group and ϕ be an automorphism that is fully irreducible and atoroidal relative to the above free decomposition. Does the mapping torus of G admit a relative cubulation?

The above makes a case for a combination theorem of relatively cubulated groups, which is as yet a largely unexplored area of research.

Question 1.5 Let G be a free product and ϕ be an automorphism that is fully irreducible but not necessarily atoroidal, relative to the given free decomposition. Let $\mathscr{P}$ be the peripheral structure of suspensions of maximal subgroups of G on which iterations of ϕ make lengths of conjugacy classes grow at most polynomially. Does the mapping torus of G admit a relative cubulation?

Motivated by work of the first author and Li [DL22], such a construction would apply to free group automorphisms, refining the cartography of possible cubulations of free-by-cyclic groups.

In [DM23], we showed that the mapping torus of a torsion-free hyperbolic group is hyperbolic relative to the suspensions of the maximal polynomially growing subgroups. This leads to a natural question:

Question 1.6 Let G be a torsion-free hyperbolic group and ϕ be an automorphism of G. Does the mapping torus of G admit a relative cubulation?

The answer is yes when ϕ is atoroidal; see our work with Mutanguha [DMM25].

Acknowledgements Meda Satish was partially supported during this work by CEFIPRA grant 5801-1, *Interactions between dynamical systems*, *geometry and number theory* at TIFR and by ISF grant 1226/19 at the Technion. Dahmani is supported by ANR-22-CE40-0004 GoFR. Both acknowledge support by LabEx CARMIN, ANR-10-LABX-59-01. Different institutions must be thanked for hosting different stages of this work: the IHP (UAR 839 CNRS-Sorbonne Université) hosting the trimester program *Groups acting on fractals* in 2022, the Institut Fourier, the Math department of the Technion, for mutual visits, and the IHES.

We would like to thank Mark Hagen, Thomas Ng and Dani Wise for helpful conversations, and the referee for numerous comments which made the paper more pleasant.

2 The flow space

2.1 Free G-trees relative to $\mathcal{H}$, and train track maps

Let us fix G to be a finitely generated group for the rest of the paper.

A free factor system for G is a tuple $(H_1, \dots, H_k)$ of subgroups such that there exists a free subgroup $F < G$ for which $G = H_1 * H_2 * \cdots * H_k * F$. Another free factor system $(J_1, \dots, J_\ell)$ of G is strictly larger if each H_i is conjugate into some J_r, and one inclusion is strict.

A G-tree is a metric tree endowed with an isometric action of G.

A *free G-tree relative to* $\mathcal{H} = \{H_1, \dots, H_k\}$ is a G-tree which is minimal for its G-action, its edge stabilisers are trivial, and its nontrivial elliptic subgroups are exactly the conjugates of the $\{H_1, \dots, H_k\}$ in G. We may as well require that there is no vertex of valence 2 that has trivial stabiliser.

A vertex is *singular* if its stabiliser is nontrivial.

Observe that, because G is finitely generated, any such G-tree has finite quotient by G. There is a whole space of such free G-trees relative to $\{H_1, \dots, H_k\}$, as studied in [GL07].

It is convenient to have a notion of *angle* in these nonlocally finite trees. Let T be a free G-tree relative to $\mathcal{H}$. Let us choose a word metric for each H_j, which is finitely generated as G is. Let $v_j \in T$ be the unique vertex fixed by H_j, and finally choose a finite set of orbit representatives for the H_j-action of edges issuing from v_j. Define the angle between two edges e and e' issuing from the vertex fixed by H_j to be the word length of the element hh'^{-1} given by h and h' sending e and e', respectively, into our set of representatives. If e and e' are distinct edges in the finite set of orbit representatives, the angle between them is one. It is clear that the $\mathrm{Stab}(v)$-action on edges adjacent to v preserves angles. We hence complete the definition by G-equivariance: it defines the angle between two edges adjacent to singular vertices stabilised by conjugates of the H_j. In other (locally finite) vertices, we may say that angles are 0 (if edges are equal) or 1 (if they are different). Observe that only finitely many edges make a given angle with a given edge, and that angles (around a given vertex) satisfy the triangle inequality.

Let ϕ be an automorphism of G that preserves the conjugacy class of each H_i. Consider a free G-tree T relative to $\{H_1, \dots, H_k\}$. One says that a continuous map $f : T \to T$ *realises the automorphism* ϕ if it is equivariant in the sense that for all $p \in T$ and all $g \in G$, $f(gp) = \phi(g) f(p)$. Such maps realising ϕ exist in our context [FM15, Lemma 4.2]. Such a map is also, by equivariance, a quasiisometry of T.

A *turn* is a pair of edges e_1 and e_2 in T starting at a common vertex v. The pair is a *proper turn* if e_1 and e_2 are distinct. We say that e_1 and e_2 make a *legal turn* if the two paths $f(e_1)$ and $f(e_2)$ starting at $f(v)$ share no proper common subpath. In other words, by definition, f sends legal turns to proper turns. We say that f is a *train track map* if it sends edges to reduced paths without nonlegal turns, and if moreover, for all such e_1 and e_2 in T making a legal turn, their images $f(e_1)$ and $f(e_2)$ are two paths starting at $f(v)$ by two edges that themselves make a legal turn. In other words, by definition, f is a train track map if it sends legal turns to legal turns.

It is far from obvious that train track maps representing automorphisms exist. A theorem of Francaviglia and Martino [FM15] ensures that if ϕ is fully irreducible relative to $\{H_1, \dots, H_k\}$, then there exists a free G-tree relative to $\{H_1, \dots, H_k\}$, which we denote by T, and there exists $f : T \to T$ continuous, with constant speed on edges, that realises ϕ and is a train track map.

Recall that, if $(H_1, \dots, H_k)$ is a free factor system of G, and if ϕ is an automorphism permuting it (ie preserving the set of conjugacy classes $\{[H_1], \dots, [H_k]\}$), we say that ϕ is *fully irreducible relative to* $\{H_1, \dots, H_k\}$ [FM15, Definition 8.2] if no positive power of ϕ preserves any larger free factor system.

We recall for later use an equivalent formulation of irreducibility (see [FM15, Definition 8.1 and Lemma 8.3]). Let $f : T \to T$ be a map realising the automorphism ϕ. We say f is *irreducible* if for every proper subgraph W of T that is f-invariant and G-invariant, the quotient of W by G is a forest such that each component subtree contains at most one nonfree vertex. The map f is *fully irreducible* if f^i is irreducible for all $i > 0$. We say ϕ is (fully) irreducible relative to $\{H_1, \dots, H_k\}$ if every f realising ϕ is (fully) irreducible.

2.2 The flow space of an automorphism

From now on T and f are thus chosen so that f is a train track map on T representing the relatively fully irreducible ϕ.

Define for all $i \in \mathbb{Z}$ a tree T_i equivariantly isomorphic to T. Let us denote by $(p \mapsto p_i)$ the identification $T \to T_i$. Define the action of G on T_i as $g.p_i = (\phi^i(g)p)_i$. Observe that this makes T_i a free G-tree relative to $\{H_1, \dots, H_k\}$. Define $f_i : T_i \to T_{i+1}$ by $f_i(p_i) = (f(p))_{i+1}$. Observe, by the property of train tracks, how f_i sends a turn: if it is legal when seen in T, its image is a legal turn when seen in T. We will make the abuse of language that any composition of the form $f_{i+k} \circ f_{i+k-1} \circ \dots \circ f_{i+1} \circ f_i$ (from T_i to T_{i+k+1}) is called a (positive) iteration of f.

Start with the disjoint union of all the T_i for $i \in \mathbb{Z}$. For each i, and each (unoriented) edge e_i in T_i, we choose an orientation and glue a rectangle $R_{e_i} = [0, 1] \times [0, 1]$ so that $\{0\} \times [0, 1]$ is glued to e_i and $\{1\} \times [0, 1]$ is glued on the path $f_i(e_i)$ in T_{i+1} (respecting orientation). See Figure 1 for an illustration. Finally, for each vertex v_i, we glue together the sides $[0, 1] \times \{0\}$ of the rectangles R_{e_i} for all e_i starting at this vertex, and the sides $[0, 1] \times \{1\}$ of the rectangles R_{e_i} for all e_i terminating at this vertex. The resulting space, with a natural structure of a cell complex, is denoted by $\widetilde{X}$. We call this space the *flow space of* ϕ (with respect to $\mathcal{H}$, T and f). Thus the flow space is a *tree of spaces*, where the underlying tree is a biinfinite combinatorial line and vertex spaces are the trees T_i.

In $\widetilde{X}$, we will call any edge in some tree T_i a vertical edge, and the image of a side $[0, 1] \times \{1\}$ or $[0, 1] \times \{0\}$ of the rectangle $R(e_i)$ a horizontal edge. We call a path horizontal if it intersects rectangles in horizontal segments $[0, 1] \times \{h\}$.

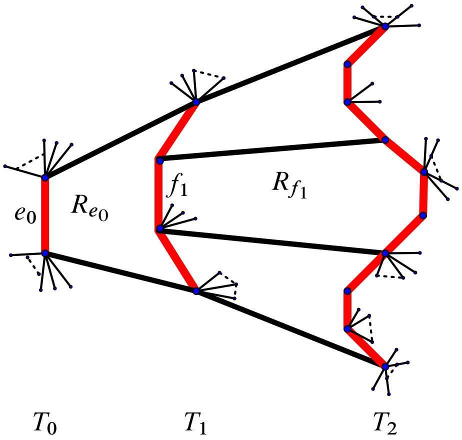


Figure 1: 2-cells in the flow space. The 2-cell R_{e_0} is bounded by the vertical edge e_0 on the left, two horizontal edges, and three vertical edges (including f_1) in T_1.

If $f : T \to T$ is a train track map realising ϕ, then for all $L \geq 1$, the map f^L is a train track map realising the automorphism ϕ^L. Keeping the same T and f we may thus produce the space $\widetilde{X}_L$ by using the map f^L realising ϕ^L. We will prefer to index the trees by $L\mathbb{Z}$ though.

Observe that $\widetilde{X}_L$ need not be isomorphic to $\widetilde{X}$. However we have a natural map $\varrho_L : \widetilde{X}_L \to \widetilde{X}$: it is obviously defined on the trees T_{Li}, and on a rectangle $R(e_{Li})$, one sends a horizontal edge onto the unique reduced concatenation of horizontal edges of $\widetilde{X}$ with the same endpoints.

2.3 Action on the flow space

Consider the semidirect product $G \rtimes_\phi \mathbb{Z}$, and write t for the generator of $\mathbb{Z}$ that induces the automorphism ϕ by conjugation on G: $G \rtimes_\phi \mathbb{Z} = \langle G, t \mid t^{-1}gt = \phi(g), \forall g \in G \rangle$. Then $G \rtimes_\phi \mathbb{Z}$ acts cofinitely on $\widetilde{X}$ by defining the action of G on each T_i as above, and defining the action of t to be the shift of indices: t maps T_i on T_{i+1} isometrically, through the identification with T.

The group G preserves each of the trees T_i, and in each of them it induces the action of G on T precomposed by ϕ^i. So the orbits of G on each of the T_i are the same, after identification to T. In particular:

Lemma 2.1 *If $x, y \in T$ are such that there is $g \in G$ for which $gx = y$, and if $i_1, i_2 \in \mathbb{Z}$, then there exists $\gamma \in G \rtimes_\phi \mathbb{Z}$ for which $\gamma x_{i_1} = y_{i_2}$.* □

2.4 Forward flow, backward flow and principal flow lines

For each i, from each point x_i of T_i, there is a unique horizontal ray starting at x_i, containing all its images by positive iterates of f. We denote it by $\sigma(x_i)$, and $\sigma_k(x_i)$ is its initial subsegment of length k. Its endpoint is denoted (slightly abusively) by $f^k(x_i)$ as already mentioned. The segment $\sigma_1(x_i)$ of this ray whose endpoints are x_i and $f_i(x_i) \in T_{i+1}$ is called the midsegment at x_i. We call the ray $\sigma(x_i)$ the *forward flow* (*ray*) from x_i (forward path in [HW16]). It is sometimes useful to use the forward flow

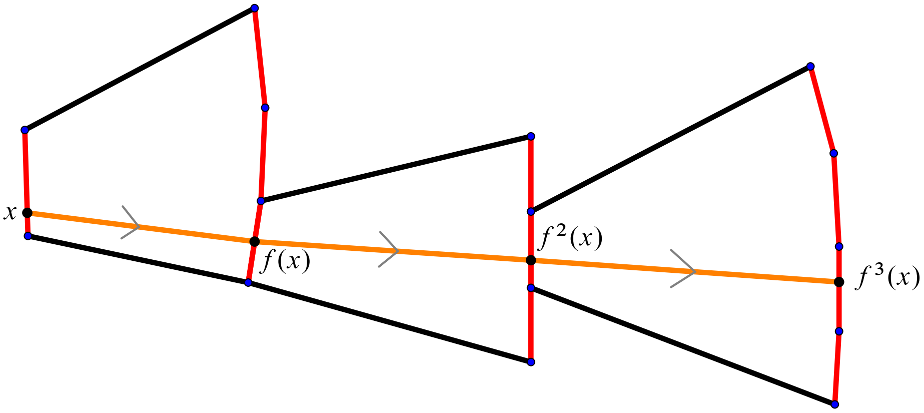


Figure 2: The orange segment between x and $f^3(x)$ is the forward flow of length 3 from x. Its forward ladder is the union of the 2-cells in the picture.

of length L, which is the initial subsegment of length L of the forward flow ray (see Figure 2 for an illustration). The *forward ladder* of a forward flow (ray or segment) σ, denoted by $N(\sigma)$, is the smallest subcomplex of the cell complex $\widetilde{X}$ containing σ.

We emphasise that the forward flow is different from the action of t.

In the backward direction, we note the following:

Lemma 2.2 *For any $x_i \in T_i$ with x_i not a singular vertex, its preimage $f_{i-1}^{-1}(x_i)$ in T_{i-1} is finite. In particular, every vertical edge of $\widetilde{X}$ is contained in finitely many 2-cells.*

Proof It suffices to prove that each edge of T is contained in finitely many images of edges of T under f. Assume on the contrary that there are infinitely many such edges e_k, whose images meet x_i. Since there are finitely many G-orbits of edges in T, we may assume that all edges e_k are images of e_0 by some elements $g_k \in G$, all different. The map f being a quasiisometry of T, all edges e_k are at bounded distances, so the g_k have bounded displacement. This forces that for large k, the g_k and their images by ϕ have a contribution in some of the free factors H_i that is larger than the maximal angle of the path $f(e_0)$. However $f(e_k) = f(g_k(e_0)) = \phi(g_k) f(e_0)$. This forces $f(e_k)$ and $f(e_0)$ to be disjoint, a contradiction. □

In the case of the preimage of an infinite-valence vertex of T_i, we have an even clearer picture. For $j \in \mathbb{N}$, let us denote by $f^{-j}(v)$ the set $(f^j)^{-1}(v)$.

Lemma 2.3 *Let $v \in T_i$ be a singular vertex of T_i. Then for each $j \in \mathbb{N}$, the set $f^{-j}(v)$ is finite and contains a unique singular vertex.*

Proof We may assume that $v \in T$. The set $f^{-j}(v)$ lies in T. We will first show that there is a unique singular vertex in $f^{-j}(v)$ and then show that the set is finite.

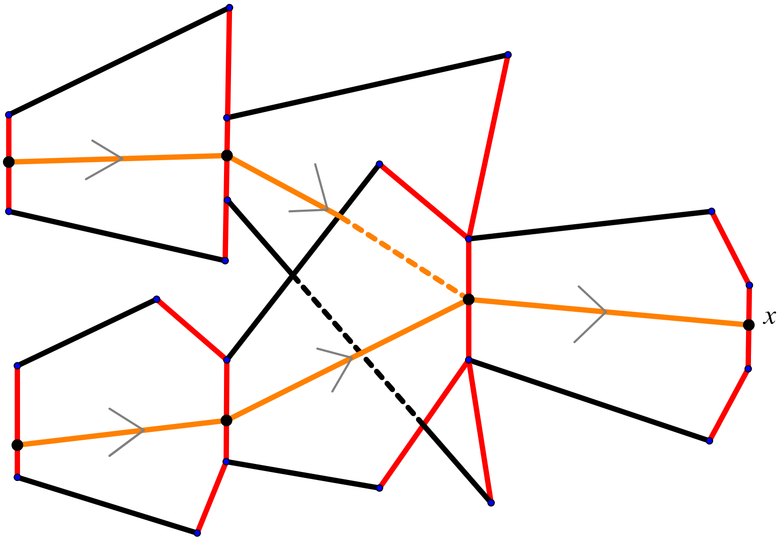


Figure 3: The backward flow of length 3 from x (in orange).

Assume that it contains two different singular vertices w_1 and w_2 such that $f^j(w_1) = f^j(w_2) = v$. By equivariance of f, the stabilisers G_{w_1} and G_{w_2} of w_1 and w_2 are sent by ϕ^j inside the stabiliser of G_v. But they are also sent onto stabilisers of vertices, since ϕ preserves the conjugacy classes of the H_k. Thus $\phi^j(G_{w_1}) = \phi^j(G_{w_2})$, and therefore $G_{w_1} = G_{w_2}$. In particular, the tree-geodesic between the vertices w_1 and w_2 is pointwise fixed by G_{w_1}, and since $w_1 \neq w_2$, this contradicts the triviality of stabilisers of edges of the tree.

Assume now that the preimage $f^{-j}(v)$ has infinitely many nonsingular points. Denote by $(e_i)_{i\in\mathbb{N}}$ the collection of edges containing a preimage. Since f is a quasiisometry, they are all at a bounded distance from each other. Up to extracting a subsequence, we may assume they are in the same G-orbit: write $e_i = g_i e_0$. Being at bounded distance from each other, the displacement of g_i remains bounded. Consider the segment s_i between e_0 and $e_i = g_i e_0$ containing both. Taking a subsequence if necessary, we may assume that there is a vertex w for which the s_i have the same prefix until w and then start having angles going to infinity at this vertex. Consider $f(e_i) = \phi(g_i) f(e_0)$. The displacement of $\phi(g_i)$ also remains bounded, and the angle of $f(s_i)$ at the image $f(w)$ tends to infinity. After w, all these images of the $f(s_i)$ are thus disjoint. It follows that all the images e_i are either disjoint, or share possibly only $f(w)$. In other words, the $f(e_i \setminus \{w\})$ are disjoint. Since at w, angles are arbitrarily large, it is a singular point, contradicting the initial assumption that the e_i each contain a nonsingular preimage of a point. □

We may now define the backward flow. For a point $x_i \in T_i$ the *backward flow* (level in [HW16]) $\tau_L(x_i)$ from x_i of length L is the union of length L forward flows from each point in $f^{-L}(x_i)$; see Figure 3. Note that $\tau_L(x_i)$ is a rooted tree, rooted at x_i. When $x \notin \widetilde{X}^0$, by Lemma 2.2, $\tau_L(x)$ is a finite rooted tree. In particular, [HW16, Proposition 2.5] gives the following observation:

Proposition 2.4 *Let $x \notin \widetilde{X}^0$. Then for any $L \geq 0$, there exists a topological embedding $\tau_L(x) \times [-1,1] \to \widetilde{X}$ such that $\tau_L(x) \times \{0\}$ maps isomorphically onto $\tau_L(x)$.* □

Finally, by Lemma 2.3, there is a unique singular vertex v_j in each T_j for $j < i$ that is in the backward flow of a singular vertex v_i of T_i. Hence we can construct the *principal flow line* of v_i to be the direct limit of the forward flow rays of v_j, for $j \to -\infty$. It is well defined, biinfinite, as all v_j as above, and all images of v_i by positive iterates of f have the same principal flow line.

2.5 Geometry of the flow space

In order to have relative hyperbolicity for the group $\Gamma := G \rtimes_\phi \mathbb{Z}$, following [DL22], we will need the following three additional properties, which we will assume to hold from now on:

- Any fundamental domain of T contains at least two edges, ie $k + r \geq 3$ or $r \geq 2$.
- The automorphism ϕ is *atoroidal relative to* $\{H_1, \dots, H_k\}$: given any element $g \in G$ such that g is not contained in any conjugate of any H_i, then for all $n \in \mathbb{N}$, $[\phi^n(g)] \neq [g]$.
- The automorphism ϕ has *no twinned subgroups*: given two subgroups $H \neq K$ such that $[H], [K] \in \mathcal{H}$, then given any $g \in G$ and any $m \in \mathbb{N}$, $\phi^m(H) \neq gHg^{-1}$ whenever $\phi^m(K) = gKg^{-1}$.

Remark 2.5 We in fact do not need to assume the no twinning property, as by [DMM25, Lemma 3.3], it automatically holds whenever ϕ is relatively fully irreducible and $k + r \geq 3$. We clarify that the case $r \geq 2$ is redundant if $k = 0$ as there are no atoroidal maps of F_2.

Let us recall that the *cone-off* of a graph Z over a family of subgraphs $\mathcal{L}$, which we denote here by $\widehat{Z}$, is the graph obtained by adding to Z a vertex v_L for each $L \in \mathcal{L}$ and an edge between each vertex of L and v_L. Usually, each such edge is assigned to have length $\frac{1}{2}$.

It can help to picture T as equivariantly quasiisometric (by a collapse map) to a cone-off of the Cayley graph of G over the left cosets of the free factors, and to picture $\widetilde{X}$ as equivariantly quasiisometric to a cone-off of the Cayley graph of Γ over the left cosets of the same free factors of $G < \Gamma$. Finally, one can see the cone-off of $\widetilde{X}$ over principal flow lines as equivariantly quasiisometric to the cone-off of the Cayley graph of Γ over the cosets of the suspensions of the free factors of G.

Theorem 2.6 *The 1-skeleton $\widetilde{X}^1$ is δ-hyperbolic for some $\delta > 0$. Moreover, the cone-off of $\widetilde{X}^1$ over the principal flow lines is also hyperbolic.*

This is essentially proved in [DL22]. Let us cover how to obtain it. We want to use [MR08, Theorem 4.5]. However, a little care is in order. The assumption of this theorem is that one has a tree (denoted by T in [MR08], and temporarily denoted by T_{MR} here) of relatively hyperbolic spaces S_v, for $v \in T_{MR}^{(0)}$, with some properties, and the conclusion is that the whole space, denoted by X in [MR08] and temporarily denoted by X_{MR}, is itself hyperbolic relative to maximal cone subtrees. Here the tree T_{MR} is just a biinfinite line $\mathbb{Z}$ indexing our trees T_i, and the spaces S_i for $i \in \mathbb{Z}$ are indeed the trees T_i, all isometric to T, on which G acts through ϕ^i, and the total space X_{MR} is $\widetilde{X}$. The different assumptions of the theorem were checked in [DL22, Section 2.3], and the conclusion is that $\widetilde{X}$ is hyperbolic relative to a collection of quasiconvex lines (the principal flow lines of singular vertices). It follows that $\widetilde{X}$ is hyperbolic itself. It also implies that its cone-off over those lines is hyperbolic as well.

Recall that a graph is said to be *fine* if for each $n \in \mathbb{N}$, every edge of the graph belongs to only finitely many simple cycles of length n. Recall also that a finitely generated group Γ is hyperbolic relative to a finite collection of finitely generated subgroups $\mathscr{P}$ if the cone-off of a Cayley graph of Γ over the left cosets of elements of $\mathscr{P}$ is hyperbolic and fine [Bow12].

The main result of [DL22] (Theorem 0.2) was actually a related statement, that the group $G \rtimes_\phi \langle t \rangle$ is itself relatively hyperbolic with respect to the collection $\mathscr{P}$ consisting of mapping tori of the subgroups $H_1, \dots, H_k$. Since the stabiliser of a singular vertex in T_0 is a conjugate of some H_i, each principal flow line is in fact cocompactly stabilised by the suspension of a conjugate of some H_i (this can, for instance, be deduced from Lemma 2.15(2)).

2.6 Quasiconvexity and divergence of forward flow rays

We first observe, following [HW16], that forward flow rays are uniformly quasiconvex (Proposition 2.7).

From $x_i \in T_i$ to $f^n(x_i)$ any path intersects each tree T_{i+k} and therefore has at least n horizontal edges. However, the forward-path has exactly n horizontal edges and no vertical contribution. If D is an upper bound to the diameter of 2-cells, we see that the intersection of the forward flow ray with $\bigcup_i T_i$ is a D-bilipschitz embedding of $\mathbb{N}$ in the 1-skeleton of $\widetilde{X}$. Since the latter is hyperbolic, we have quasiconvexity. We thus have:

Proposition 2.7 [HW16, Proposition 2.3] *There exists $\lambda \geq 0$ such that the 1-skeleton $N(\sigma)^1$ of any forward ladder is λ-quasiconvex in $\widetilde{X}^1$.* □

Second, two flow rays starting at different points in the same edge of T_i diverge from each other:

Lemma 2.8 *Given $\epsilon > 0$, there exists $N > 0$ such that for any two points x and y contained in any vertical edge $e_i \subset T_i$ with $d(x, y) \geq \epsilon$, the distance in T_{i+N} between $f^N(x)$ and $f^N(y)$ is at least $e^{100(\delta+\lambda)}$ (and thus the forward rays from x and y diverge).*

Proof The forward images of a single edge are all legal paths, and by [DL22, Lemma 1.11] f applies a uniform stretching factor > 1 on all legal paths. □

Finally, two flow rays starting at different points in the star of a vertex either uniformly diverge from each other or fellow-travel forever. More precisely we have the following:

Proposition 2.9 *For all $\epsilon > 0$, there exists $N > 0$ such that, for all singular vertices $v \in T_i$, for all edges $e_1 \neq e_2$ in T_i starting at v, and for all x_1 and x_2 in e_1 and e_2, respectively, at distance at least ϵ from v, either the distance in T_{i+N} between $f^N(x_1)$ and $f^N(x_2)$ is at least $e^{100(\delta+\lambda)}$, or the forward rays fellow-travel until infinity.*

Proof We fix ϵ and v. For each edge e issued at v, and $x \in e$ at distance at least ϵ from v, we know by Lemma 2.8 that there is N_ϵ such that, for all $n \geq N_\epsilon$, $f^n(x)$ is at distance (in T_{i+n}) at least $e^{1000(\delta+\lambda)}$ from the principal flow line of v.

By [DL22, Lemma 2.6] f induces a quasiisometry on angles: there exists $\theta_1 > 1$ such that, if e_1 and e_2 issued at v make an angle θ, then the paths $f(e_1)$ and $f(e_2)$ make an angle at least $\theta/\theta_1 - \theta_1$ at their common initial point $f(v)$ — while it is slightly stronger than the stated lemma, the claim is still true with same proof, however, the stated lemma is also sufficient to be used in a similar way. Call $\rho(\theta) = \theta/\theta_1 - \theta_1$. Calculus ensures that it is possible to find θ_0 such that $\rho^{N_0}(\theta_0) > 0$. If e_1 and e_2 issued at v make an angle greater than θ_0, the paths $f^{N_0}(e_1)$ and $f^{N_0}(e_2)$ make a positive angle at their initial common point $f^{N_0}(v)$. In particular, they do not overlap. It follows that, measured in T_{i+N_0}, the distance $d(f^{N_0}(x_1), f^{N_0}(x_2))$ is equal to $d(f^{N_0}(x_1), f^{N_0}(v)) + d(f^{N_0}(v), f^{N_0}(x_2))$. It is then greater than $2e^{1000(\delta+\lambda)}$. Since the two rays are quasiconvex in the hyperbolic space $\widetilde{X}$ and have started to diverge after N_0 edges, they will diverge onward after that point.

It remains to treat the case of two edges making an angle less than θ_0. There are finitely many $\mathrm{Stab}(v)$-orbits of such pairs of edges. Partition them into classes of those whose forward rays fellow-travel, and those whose forward rays exponentially diverge. Since there are finitely many, one can choose N that is suitable for all of those in the second class, and larger than N_0. □

Observe that this does not prevent some points in different edges from having fellow-travelling forward rays, with the same endpoint at infinity. However, if the origins of the flow rays are translates of each other by an elliptic element g in T_i, they diverge, provided that g is not a torsion element:

Lemma 2.10 *For any point x_i in T_i, and for each $g \in G$ elliptic in T_i such that $g^n x_i \neq x_i$ for any $n \in \mathbb{N}$, the flow rays from gx_i and x_i do not fellow-travel until infinity.*

Proof Note that the midpoint of the segment of T_i between x_i and gx_i is fixed by g, and is hence a singular vertex v_i. If the flow rays from x_i and gx_i fellow-travel until infinity, then by equivariance, so do the flow rays from $g^n x_i$ and $g^{n'} x_i$, for all $n, n' \in \mathbb{Z}$. But for large enough powers, the segments from v_i to $g^n x_i$ and x_i make arbitrarily large angles. By arguing as in Proposition 2.9, we can conclude that the flow rays from x_i and $g^n x_i$ cannot fellow-travel for arbitrarily large powers, a contradiction. □

Henceforth, we will assume that the group G is torsion-free, so the above result always holds for elliptic elements.

Let us also record a basic observation about principal flow lines:

Lemma 2.11 *Let $v_1 \neq v_2$ be two singular vertices in $\widetilde{X}$. Then either the principal flow lines of v_1 and v_2 coincide or uniformly diverge from each other in both the forward and backward directions, ie for each $R > 0$ there exists $B_R > 0$ such that for any pair of singular vertices $v_1 \neq v_2$, either the principal flow lines Λ_{v_i} through v_i coincide, or $N_R(\Lambda_{v_1}) \cap \Lambda_{v_2}$ has diameter at most B_R.*

Proof Observe that by Lemma 2.3, for each singular vertex there exists a unique principal flow line going through this vertex. Since the automorphism ϕ has no twinned subgroups for the (finite) collection $\mathcal{H}$, if two principal flow lines are different, they must uniformly diverge as required. □

2.7 Periodic points and forward flow rays

In the case of certain special points called periodic points, we have no fellow-travelling of flow rays between any translates, which we prove below. We will also show that periodic points with flow rays (lines, in fact) diverging from every principal flow line are dense in edges of T (Lemma 2.18).

A point $x \in T$ is a *periodic point* if there exist $g = g_x \in G$ and $n = n_x > 0$ such that $f^n(x) = gx$. If x is periodic, then we will also call each $x_i \in T_i$ a periodic point. We say that the periodic point $x \in T$ has *period n* if $f^n(x) = gx$ as above and for all $0 < k < n$, $f^k(x) \neq hx$ for any $h \in G$.

Proposition 2.12 *For any periodic point x_i in T_i, and for each $g \in G$ with $gx_i \neq x_i$, the flow rays from gx_i and x_i diverge.*

Proof Observe that by Lemma 2.10, the statement has to be proved only for loxodromic elements g. Since x_i is periodic, there is a minimal positive k_i and $g_i \in G$ such that $g_i x_{i+k_i} = (f^{k_i}(x_i))$ (which is in T_{i+k_i}). Thus the element $g_i t^{k_i}$ sends x_i to $f^{k_i}(x_i)$ in the flow space. This flow space is hyperbolic, and this element is loxodromic, since for any $x \in T_0$, $d_{\widetilde{X}^1}((g_i t^{k_i})^r x, x) \geq r k_i$ (the distance separating T_0 from T_{rk_i}). It follows that the forward flow from x_i remains at bounded distance from $((g_i t^{k_i})^n x_i)_{n \in \mathbb{N}}$, and hence from the axis of $(g_i t^{k_i})$. Now, from gx_i, the forward flow remains close to the axis of $g(g_i t^{k_i})g^{-1}$.

Since the element g is loxodromic on T_i, g and $g_i t^{k_i}$ generate a nonelementary subgroup of isometries of $\widetilde{X}$, because their axes eventually diverge. It follows that the limit points of $g_i t^{k_i}$ and of g in the boundary of $\widetilde{X}^1$ cannot be the same. One concludes that g does not fix the limit points of $g_i t^{k_i}$, and $g(g_i t^{k_i})g^{-1}$ and $(g_i t^{k_i})$ thus have divergent axes. □

Lemma 2.13 *For all edges e and e' in T, there exist $n \geq 0$ and $g \in G$ such that $f^n(e)$ contains ge'.*

Proof Assume that for some e and e', for every n, the path $f^n(e)$ does not contains any G-translate of e'. Note that this means that no G-translate of $f^n(e)$ contains any G-translate of e'. Let W be the union of all G-translates of all paths $f^n(e)$. Then W is a proper subgraph which is both f-invariant and G-invariant. By the irreducibility of f, the quotient of W by G is a (bounded) forest with at most one nonfree vertex in each component. Thus the quotient of each $f^n(e)$ is a segment with at most one nonfree vertex. This is not possible as the lengths of $f^n(e)$ are arbitrarily long. □

Lemma 2.14 *For each $\epsilon > 0$ and each closed subinterval d of any edge e in T of length ϵ, there exists a periodic point x in d.*

Proof Lemma 1.11 of [DL22] ensures that there exists a growth factor $\lambda > 1$ such that every subsegment of every edge of T expands under f by λ. Thus, given $d \subset e$, there exists n such that $f^n(d)$ contains an edge. The result then follows from Lemma 2.13, by applying Brouwer's fixed-point theorem. □

Let us now observe the following useful facts about periodic points. Given a periodic point x of period n, the preimage $f^{-n}(x)$ contains a G-translate of x (Lemma 2.15(1)). One can then define a *periodic flow line* through x as the direct limit of forward flow rays from the G-translates of x at the various $f^{-nk}(x)$. We will show in Lemma 2.15(2) that a periodic flow line is really *periodic*, ie there is an infinite subgroup of Γ that stabilises the line and acts cocompactly on the line.

Lemma 2.15 *Let $x \in T$ be a periodic point of period n. Then*:

(1) *There exists $g' \in G$ such that $g'x \in f^{-n}(x)$.*

(2) *The periodic flow line through x is periodic. In particular, principal flow lines are periodic and stabilised by suspensions of the relevant free factors of G.*

Proof (1) Let $g \in G$ be such that $f^n(x) = gx$. Observe that, by equivariance of f, for $g' = \phi^{-n}(g^{-1})$, $g'x \in f^{-n}(x)$. Indeed, $f^n(\phi^{-n}(g^{-1})x) = \phi^n(\phi^{-n}(g^{-1}))f^n(x) = g^{-1}f^n(x) = x$.

(2) Consider the flow segment $\sigma_n(x)$ from $x \in T_0$ to $f_{n-1} \circ \cdots \circ f_0(x) \in T_n$. Recall that the element t acts by shifting indices, and therefore, $f_{n-1} \circ f_{n-2} \circ \cdots \circ f_0(x) = t^n f^n(x)$. This implies that $t^n g(\sigma_n(x))$ is the flow segment from $t^n gx \in T_n$ to $t^n g t^n f^n(x) \in T_{2n}$. Using the facts that $gt^n = t^n \phi^n(g)$ and $gx = f^n(x)$, we conclude that $t^n g(\sigma_n(x))$ is the flow segment from $t^n f^n(x)$ to $t^{2n}\phi^n(g) f^n(x)$. Since $\phi^n(g) f^n(x) = f^n(gx) = f^{2n}(x)$, we have that $\sigma_n(x) \cdot t^n g(\sigma_n(x)) = \sigma_{2n}(x)$. It is also easy to check that the flow segment of length n from $t^{-n}g'(x)$ is equal to $(t^n g)^{-1}(\sigma_n(x))$. Thus the periodic flow line through x is the union over all $k \in \mathbb{Z}$ of $(t^n g)^k(\sigma_n(x))$. □

Lemma 2.16 *If a periodic flow line is asymptotic to another periodic flow line in one direction, they are asymptotic in both directions.*

Proof We will prove the contrapositive. Assume Λ' and Λ'' are periodic flow lines that are nonasymptotic in the forward direction (the other case is similar). Since they are periodic, there are elements γ' and γ'' in the group Γ (Lemma 2.15(2)) that fix the endpoints of Λ' and Λ'', respectively, in the boundary $\partial \widetilde{X}^1$ of the hyperbolic space $\widetilde{X}^1$. In particular, γ' and γ'' cannot be elements of the same parabolic subgroup of Γ. Assume that the backward directions of Λ' and Λ'' converge to the same point of $\partial \widetilde{X}^1$. It would follow that the commutator $[\gamma', \gamma'']$ is in G and has small displacement on infinitely many T_{-i} for $i \in \mathbb{N}$, realised near the ray Λ'. Since γ' and γ'' have different axes, they do not commute. However, by hyperbolicity of the automorphism, $[\gamma', \gamma'']$ must be elliptic, and hence it is in a unique conjugate of a free factor of the free product G. It hence fixes nonfree vertices in each T_{-i}, at bounded distance from Λ', and all in the same principal flow line. It then follows that both Λ' and Λ'' are asymptotic to the same principal flow line, and therefore that the elements γ' and γ'' preserving them fix a point fixed by a parabolic group, thus ensuring that they are in the same parabolic subgroup, contradicting what we previously had. □

Given two flow lines that intersect the tree T_0 in x and y, we say that the two flow lines are separated by the segment $[x, y]$ in T_0. The next statement says that a periodic flow line either diverges from every principal flow line, or it is asymptotic to a principal flow line which is Hausdorff-close to it in terms of both distance and angle.

Lemma 2.17 *There exist constants δ_0 and θ_0 such that the following holds: Let Λ be a periodic flow line in $\widetilde{X}$ and x be its intersection with T_0. If a principal flow line Λ', with intersection y at T_0, is asymptotic to Λ, then $d_{T_0}(x, y) \le \delta_0$. Further, there exists a principal flow line Λ'', asymptotic to Λ and Λ', such that if $z \in \Lambda'' \cap T_0$, then for every vertex v in the interior of the segment $[x, z]_{T_0}$ the angle subtended by $[x, z]_{T_0}$ at v is less than θ_0.*

Proof Consider a geodesic $[x, y]_{\widetilde{X}^1}$ in $\widetilde{X}^1$. By asymptoticity of the biinfinite lines Λ and Λ' in the hyperbolic space $\widetilde{X}^1$, its length is at most δ. Denote by $e_1, \dots, e_r$ its consecutive vertical edges, let T_{k_i} be the tree containing e_i and write $e_i = (v_i^{k_i}, w_i^{k_i})$. Observe that $r \le \delta$ and $|k_i| \le \delta$ for all i.

Since e_1 is the first vertical edge, $x \in f^{-k_1}(v_1^{k_1})$, and similarly, $y \in f^{-k_r}(w_r^{k_r})$. Take, for all $i < r$, $x_i \in f^{-k_i}(w_i^{k_i})$. Such a point x_i is in T_0 and is in $f^{-k_{i+1}}(v_{i+1}^{k_{i+1}})$. Therefore, for all $i < r$, $d_{T_0}(x_i, x_{i+1}) \le K^{|k_{i+1}|}$, where K is the Lipschitz constant of f. Since for all i, $|k_{i+1}| \le \delta$, and $r \le \delta$, it follows that $d_{T_0}(x, y) \le \delta K^{\delta}$.

We turn to the second part of the statement. By Lemma 2.10, for any edge e incident to a vertex w, and any $g \in \mathrm{Stab}(w)$, e and ge form a legal turn and so their iterates under f subtend a positive angle at all $f^n(w)$. Since there are finitely many orbits of edges incident to w, there are only finitely many edges incident to w that make an illegal turn with e. Since there are finitely many orbits of vertices, there exists a maximal angle, which we call $\theta_0 - 1$, such that any illegal turn between any pair of edges at any vertex subtends an angle of at most $\theta_0 - 1$.

Assume now that there exists $v \in [x, y]_{T_0}$ such that $\mathrm{Ang}_v[x, y]_{T_0} \ge \theta_0$. We choose v to be closest to x with this property. Note that, after changing θ_0 if necessary, v is a singular vertex. Let Λ_v be the associated principal flow line.

For all $n \in \mathbb{N}$, let us denote by x_{-n} the point $f^{-n}(x) \cap \Lambda$ and by y_{-n} the point $f^{-n}(y) \cap \Lambda'$. Let us also write $x_n = f^n(x)$, $y_n = f^n(y)$, $v_n = f^n(v)$ and v_{-n} the point $f^{-n}(v) \cap \Lambda_v$.

Since $\theta_0 - 1$ is the maximal angle at an illegal turn, for all $n \in \mathbb{N}$, $\mathrm{Ang}_{v_n}[x_n, y_n]_{T_n} \geq 1$. In particular, v_n is between x_n and y_n in T_n for all $n \in \mathbb{N}$, and so the principal flow line Λ_v is asymptotic to Λ in one and hence both directions. □

Lemma 2.18 *For each $\epsilon > 0$ and each closed subinterval d of any edge e in T of length ϵ, there exists a periodic point x in d such that the periodic flow line through x diverges from every principal flow line in both the forward and backward directions.*

Proof There are finitely many orbits of vertices, and hence finitely many orbits of periodic lines containing a vertex. Let n_0 be the maximum of their period. If a subinterval d of e is given, one may take a sufficiently small subinterval d' such that its images by f^k for $k = 1, \dots, n_0$ do not meet any vertex. This, together with the periodicity of periodic points, guarantees that any periodic point in this subinterval will have a periodic flow line that misses all vertices.

There are only finitely many vertices in the tree T_0 that are at distance $\leq \delta K^\delta + 1$ from d with a path that makes no angle greater than θ_1. Let N be this number. This produces N principal flow lines.

Consider $2N + 1$ periodic points in the subinterval d'. Since the interval between any two of them is a legal path, their forward flow rays diverge from each other. Therefore at least one of these periodic points, say x, belongs to a periodic flow line that diverges from the N principal flow lines in both directions. By Lemma 2.17 it diverges from all principal flow lines. □

3 Relative cubulation

In this section, we will recall the notion of relative cubulation and the boundary criterion for relative cubulation. We will construct walls in Section 4 for the flow space $\widetilde{X}$ and show in Sections 5 and 6 that the stabilisers in $\Gamma = G \rtimes_\phi \mathbb{Z}$ of "wall saturations" satisfy the hypotheses required for the boundary criterion to hold.

Definition 3.1 [EG20] Let Γ be hyperbolic relative to a collection of subgroups $\mathscr{P}$. Then $(\Gamma, \mathscr{P})$ is *relatively cubulated* if there exists a CAT(0) cube complex C such that

(1) there is a cubical cocompact action of Γ on C,

(2) each element of $\mathscr{P}$ is elliptic, and

(3) the stabiliser of any cube of C is either finite or conjugate to a finite-index subgroup of an element of $\mathscr{P}$.

We call such an action of Γ on C a *relatively geometric* action.

Let us recall a few definitions before stating the boundary criterion for relative cubulation. A subgroup $H \leq \Gamma$ is *relatively quasiconvex* (with respect to $\mathscr{P}$) if there exists a constant K such that given $h_1, h_2 \in H$, any geodesic path γ between h_1 and h_2 in the cone-off $\widehat{\Gamma}$ (with respect to $\mathscr{P}$) is such that every noncone vertex of γ is at Γ-distance at most K from a vertex of H; see [Hru10; Osi06].

A subgroup H of $(\Gamma, \mathscr{P})$ is *full* if the intersection of any conjugate of H with any element $P \in \mathscr{P}$ is either finite or is of finite index in P.

A subgroup H of Γ is a *codimension*-1 *subgroup* if for some $r \geq 0$, $\Gamma \setminus \mathcal{N}_r(H)$ contains at least two orbits of components that are not contained in any finite neighbourhood of H.

Let us also recall a definition (of the many equivalent ones) of the *Bowditch boundary* of a relatively hyperbolic group $(\Gamma, \mathscr{P})$: it is a compact metrisable space $\partial_B(\Gamma, \mathscr{P})$ such that

(1) Γ acts properly discontinuously on the space of distinct triples of $\partial_B(\Gamma, \mathscr{P})$,

(2) each point of $\partial_B(\Gamma, \mathscr{P})$ is either a conical point or a bounded parabolic point, and

(3) the stabiliser of a bounded parabolic point is a conjugate of an element of $\mathscr{P}$.

As a set, $\partial_B(\Gamma, \mathscr{P})$ is the union of the Gromov boundary of the cone-off $\widehat{\Gamma}$ relative to $\mathscr{P}$ and the set of cone vertices of $\widehat{\Gamma}$. We refer to [Bow12, Section 9] for more details, in particular for the definitions of conical and bounded parabolic points, which will not be needed here.

The following theorem is implicit in the work of Bergeron and Wise [BW12] and explicitly stated and proved by Einstein and Groves [EG20].

Theorem 3.2 (boundary criterion for relative cubulation) *Let $(\Gamma, \mathscr{P})$ be relatively hyperbolic with one-ended parabolics. Suppose that for each pair of distinct points $u, v \in \partial_B(\Gamma, \mathscr{P})$ there exists a full relatively quasiconvex codimension-1 subgroup H of Γ such that u and v lie in H-distinct components of $\partial_B(\Gamma, \mathscr{P}) \setminus \Lambda H$. Then there exists a finite collection of full relatively quasiconvex codimension-1 subgroups such that the action of Γ on the dual cube complex is relatively geometric.*

We note that since our parabolics are suspensions of the infinite free factors of G, they are always one-ended.

4 Walls in the flow space

4.1 Constructing immersed walls

Our construction of walls starts with that of Hagen and Wise in [HW16], although they work in a locally finite setup. However, the stabilisers of walls thus obtained would not be full subgroups in general, so we later "saturate" these walls to obtain full codimension-1 subgroups.

Let us explain the construction of Hagen and Wise. Fix a fundamental domain $D \subset T$ for the G-action on T, ie a smallest subtree that contains exactly one edge from each orbit of edges. In order to construct immersed walls in $\widetilde{X}$, we need the following data: a choice of a *tunnel length* $L \in \mathbb{N}$ and a choice of a subinterval of each open edge of D, called a *primary bust*.

Choose the tunnel length $L \in \mathbb{N}$. The immersed wall will be first constructed in $\widetilde{X}_L$ and then pushed to $\widetilde{X}$ via $\varrho_L \colon \widetilde{X}_L \to \widetilde{X}$.

Choose a closed subinterval in the interior of each edge of D. This choice extends equivariantly to a choice of a subinterval d_k in every edge e_k of T. For each $i \in \mathbb{Z}$, the copy $d_{k,Li}$ in T_{Li} of each $d_k \subset e_k$ is a *primary bust*. Note that $f^{-L}(d_k)$ is a union of finitely many subintervals $\{d_{kj}\} \subset T$, by Lemma 2.2. For each $i \in \mathbb{Z}$, the copy $d_{kj,Li}$ in T_{Li} of each $\{d_{kj}\}$ is a *secondary bust*. We will always choose primary busts satisfying the following:

Lemma 4.1 *Let $L \in \mathbb{N}$ and $\epsilon > 0$. Let $\{x_k\}$ be a collection of points in D, with exactly one point in any edge of D. The collection of primary busts in $\widetilde{X}_L$ can be chosen so that*

1. *every primary bust is disjoint from every secondary bust in the collection,*
2. *the primary busts in D lie in the ϵ-neighbourhood of $\{x_k\}$,*
3. *the flow rays from the endpoints of the primary busts do not meet any vertex of $\widetilde{X}_L$,*
4. *if the flow ray from x_k does not meet a vertex, then one of the endpoints of the primary bust d_k in D can be chosen to be x_k,*
5. *the f^L-image of any primary bust does not contain two points in the same G-orbit,*
6. *if $f^L(x_k) \neq f^L(x_j)$ whenever $k \neq j$, then $f^L(d_k) \cap f^L(d_j) = \varnothing$.*

The lemma is a reproduction of [HW16, Lemma 3.5], whose proof can be replicated in our case. Since there are several properties to check, the proof is long, but not difficult, based on Brouwer's fixed-point theorem. We thus refer the reader to [HW16] for the proof. In fact, (3) can be obtained either by the proof of Hagen and Wise [HW16], or by applying Lemma 2.18.

Denote by $T_{Li+1/2} = T'_{Li} \subset \widetilde{X}_L$ the parallel copy of T at distance $\frac{1}{2}$ from T_{Li} and distance $L - \frac{1}{2}$ from $T_{(i+1)L}$. We will denote by $d'_{k,Li}$ and $d'_{kj,Li}$, respectively, the copies of $d_{k,Li}$ and $d_{kj,Li}$ in T'_{Li}. We refer the reader to Figure 4 for an illustration of what follows.

(1) Consider the subspace $\boldsymbol{C}_{Li}$ of T'_{Li} obtained by removing each open primary bust and each open secondary bust in T'_{Li}. A component $\boldsymbol{N}_{Li}$ of $\boldsymbol{C}_{Li}$ is a *nucleus*. Note that each nucleus is either a subinterval of some edge or contained in the star of a vertex.

(2) The endpoints of the nuclei are endpoints of either primary busts or of secondary busts. To each endpoint of each secondary bust $d'_{kj,L(i-1)}$ in a nucleus, glue a forward flow segment of length $L - \frac{1}{2}$. Note that this flow segment ends at the primary bust $d_{k,Li}$. The union of all flow segments which end at a common endpoint (an endpoint of some $d_{k,Li}$) is a *level* $\boldsymbol{L}$. Such an endpoint is a *forward endpoint*.

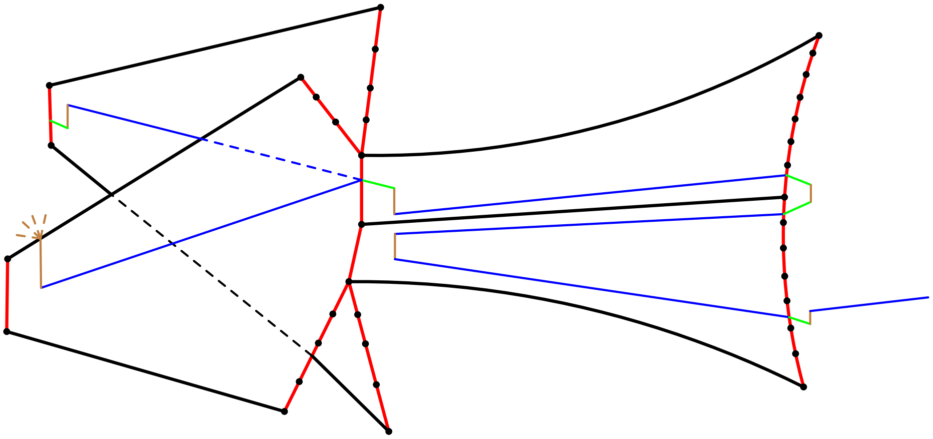


Figure 4: A part of a wall in $\widetilde{X}_L$. Brown paths are nuclei, blue paths are levels and green paths are slopes.

(3) Join the endpoints $d^{\pm}_{k,Li}$ of each primary bust $d_{k,Li}$ to the endpoints $d'^{\mp}_{k,Li}$ (note the change in order) of its parallel copy $d'_{k,Li}$ by segments called *slopes*. A slope is denoted by $\boldsymbol{S}$.

The graph $\overline{W}_L$ is constructed as a quotient of the disjoint union of the nuclei, levels and slopes, with the gluing performed in a natural way: an endpoint of a primary bust of a nucleus is glued to an endpoint of the corresponding slope, an endpoint of a secondary bust in a nucleus is glued to the initial point of the relevant flow segment of a level, while a forward endpoint is glued to the relevant endpoint of a slope. There exists a noncombinatorial immersion of the graph $\overline{W}_L$ in $\widetilde{X}_L$, with loss of injectivity at midpoints of slopes. Let $\overline{W}$ be the graph obtained from $\overline{W}_L$ by "folding" its levels according to ϱ_L so that the following diagram commutes:

$$\begin{tikzcd}
\overline{W}_L \arrow[r] \arrow[d] & \widetilde{X}_L \arrow[d, "\varrho_L"] \\
\overline{W} \arrow[r] & \widetilde{X}
\end{tikzcd}$$

A component W_u of $\overline{W}$ is an *immersed wall.*[3]

Remark 4.2 The immersion $W_u \to \widetilde{X}$ extends to a local homeomorphism $W_u \times [-1,1] \to \widetilde{X}$, with W_u identified with $W_u \times \{0\}$. Indeed, this is because the embedding of each nucleus, level and slope of W_u in $\widetilde{X}$ extends to a local homeomorphism of the above type.

We will denote by $N(W_u)^1 \subset \widetilde{X}^1$ the 1-skeleton of the smallest subcomplex of $\widetilde{X}$ that contains W_u. When the tunnel length L is sufficiently large, it turns out that $N(W_u)^1 \hookrightarrow \widetilde{X}^1$ is a quasiconvex embedding (Proposition 4.6) and that W_u separates $\widetilde{X}$ into exactly two components (Proposition 4.8). Thus the local homeomorphism of Remark 4.2 is in fact a homeomorphism in that case.

[3] The subscript u in W_u denotes "unsaturated", as we will soon saturate W_u in order to obtain a full codimension-1 subgroup.

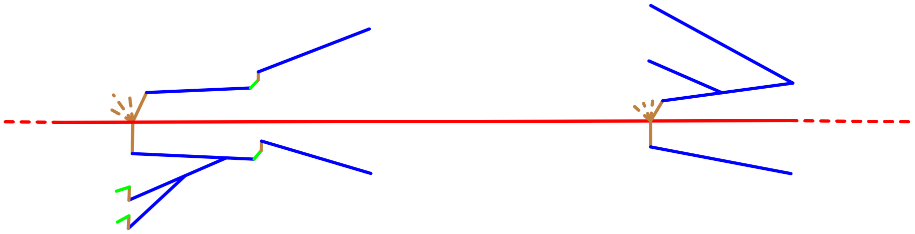


Figure 5: A part of a wall saturation. The red line is a principal flow line, with two unsaturated components (on the left and right) attached.

4.2 Wall saturations

Let W_u be an immersed wall as above, which is a component of the graph $\overline{W}$. Fix a *saturation length* $M \in L\mathbb{N}$. We will define below the M-saturation W of W_u as a union of components of $\overline{W}$ and principal flow lines. We start with a definition:

Let $v \in T_{Li}$ be a singular vertex. The *M-saturation of* v is the set $\{f^{Mk}(v) \in T_{Mk+Li}\}_{k\in\mathbb{Z}}$ of singular vertices. (Recall that for each $k \in \mathbb{Z}$, there is a unique singular vertex $f^{Mk}(v)$, by Lemma 2.3.) By abuse of notation, for a vertex v in a "fractional" copy $T_{1/2+Li}$, we will denote by the M-saturation of v the intersection points of the principal flow line of v in the trees $T_{1/2+Mk+Li}$.

Let $v \in T_{1/2+Li}$ be a singular vertex in a nucleus of $\overline{W}$. Denote by $W(v) \subset \overline{W}$ the smallest union of components of $\overline{W}$ such that each vertex in the M-saturation of v is contained in the nucleus of some component of $W(v)$.

Let W_u be an immersed wall. Let $W'_u \subset \overline{W}$ be the smallest subgraph of $\overline{W}$ satisfying

- $W_u \subset W'_u$, and
- for each singular vertex $v \in W'_u$, either the principal flow line of v intersects a component infinitely many times, or $W(v) \subset W'_u$ (but not both).

The *M-saturation W of W_u* is defined as the union of W'_u with the principal flow lines of all singular vertices of W'_u. We refer the reader to Figure 5 for an illustration.

Lemma 4.3 *Let W_u be an immersed wall. For any $M \in L\mathbb{N}$, the M-saturation W of W_u is a connected graph.*

Proof We will prove the lemma by showing that a certain connected subgraph containing W_u in W is in fact equal to W. Consider the subgraph W' of W built inductively as the ascending union of subgraphs W^n as follows: $W^1 = W_u$. W^2 is the union of each principal flow line intersecting W_u with the union of all $W(v)$, where v is a singular vertex in W^1. W^n is the union of each principal flow line intersecting W^{n-1} along with the union of all $W(v)$, where v is a singular vertex in W^{n-1} with the property that

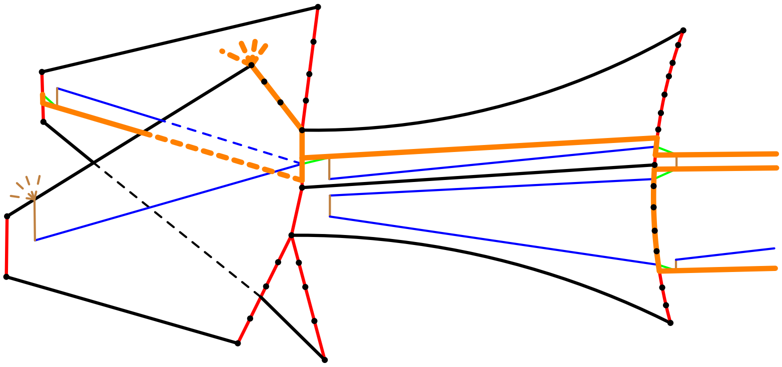


Figure 6: The approximation of the part of the wall from Figure 4 is indicated in thick orange.

$W(v) \subset W$. Note that W' is a connected subgraph of W, by construction. Further, the subgraph W'_u (introduced in the definition of the M-saturation W of W_u) is contained in W'. This gives the reverse containment $W' \supset W$. □

4.3 Approximations of walls

In order to show that families of immersed walls are uniformly quasiconvex, Hagen and Wise use a technical construction called approximations. Approximations are not walls in $\widetilde{X}$ (they are walls in $\widetilde{X}_L$, as it turns out), but have the advantage that, unlike immersed walls, they do not have backward flows that can have long fellow-travelling subpaths.

We will first state the definition of approximations given in [HW16] and then extend their definition to saturations. Let $W_u \to \widetilde{X}$ be an immersed wall of tunnel length L. The *approximation* $A\colon W_u \to \widetilde{X}$ is defined as below. We refer the reader to Figure 6 for an illustration.

(1) For each $x' \in W_u$ such that x' lies in a nucleus $\boldsymbol{N}_{Li} \subset T_{Li+1/2}$, let $x \in T_{Li}$ denote the parallel copy of x' behind it at distance $\frac{1}{2}$. Then $A(x') := f^L(x) \in T_{Li+L} \subset \widetilde{X}$.

(2) For each x in a level of W_u, $A(x)$ is the unique forward endpoint of the level.

(3) Let $\boldsymbol{S}$ be a slope in W_u associated to the primary bust d. Recall that $\boldsymbol{S}$ is a segment from an endpoint d^+ (say) of d to the other endpoint d'^- of d', where d' is the parallel copy of d at distance $\frac{1}{2}$. A maps $\boldsymbol{S}$ homeomorphically to the concatenation of d and the forward flow segment of length L from d^-.

Let $M \in L\mathbb{N}$ and let $W \to \widetilde{X}$ be an M-saturation of W_u. The *approximation of the M-saturation* $A\colon W \to \widetilde{X}$ is defined the same way on nuclei, slopes and levels. On each point x of a principal flow line, by abuse of notation, $A(x) := f^{L-1/2}(x)$.

We will denote by $N(A(W_u))^1$ and $N(A(W))^1$ the 1-skeletons of the smallest subcomplexes of $\widetilde{X}$ containing $A(W_u)$ and $A(W)$, respectively.

4.4 Large tunnel length and quasiconvexity

In this subsection, we recall results from [HW16, Section 4] towards quasiconvexity of immersed walls. We note that their methods do not require local finiteness of $\widetilde{X}^1$ and thus work in our case as well, both for immersed walls and their saturations.

The main lemma that we will repeatedly use is the following. Recall that $\mathcal{N}_r(Y)$ denotes the r-neighbourhood of Y.

Lemma 4.4 [HW16, Lemma 4.3] *Let Z be a δ-hyperbolic space and let $P = \alpha_0\beta_1\alpha_1 \dots \beta_k\alpha_k$ be a path such that each α_i is a (λ_1, λ_2)-quasigeodesic and each β_i is a (μ_1, μ_2)-quasigeodesic. Suppose that for each $R \geq 0$ there exists a $B_R \geq 0$ such that for all i each intersection below has diameter $\leq B_R$:*

$$\mathcal{N}_{3\delta+R}(\beta_i) \cap \beta_{i+1}, \quad \mathcal{N}_{3\delta+R}(\beta_i) \cap \alpha_i, \quad \mathcal{N}_{3\delta+R}(\beta_i) \cap \alpha_{i-1}.$$

Then there exists L_0 such that if each β_i is of length at least L_0, then the path P is a $\left(4\lambda_1\mu_1, \frac{1}{2}\mu_2\right)$-quasigeodesic.

Let $\mathcal{W} := \{W_u \to \widetilde{X}\}$ be a family of immersed walls in $\widetilde{X}$. In order to use Lemma 4.4, we need the following property to be satisfied:

Definition 4.5 [HW16, Definition 3.15] The family of immersed walls $\mathcal{W}$ has the *ladder overlap property* if there exists $B \geq 0$ such that for all $W_u \in \mathcal{W}$ and all distinct slopes $\boldsymbol{S}_1, \boldsymbol{S}_2 \subset W_u$,

$$\operatorname{diam}(\mathcal{N}_{3\delta+2\lambda}(N(A(\boldsymbol{S}_1))^1) \cap \mathcal{N}_{3\delta+2\lambda}(N(A(\boldsymbol{S}_2)))^1) \leq B,$$

where λ is the quasiconvexity constant for forward ladders (Proposition 2.7).

In practice, any family of immersed walls is constructed in the following way: We will first choose a finite set of periodic points in the fundamental domain D of T such that the flow rays from these points diverge and do not meet any vertex. The various immersed walls of the family will then be constructed by choosing tunnel lengths and by choosing primary busts in small neighbourhoods of these periodic points, with the size of the neighbourhoods depending on the chosen tunnel lengths. The ladder overlap property will then hold because the finitely many flow rays from the periodic points have bounded overlaps.

Proposition 4.6 *Let $\mathcal{W}$ be a family of immersed walls satisfying the ladder overlap property. Then there exist L_0, κ_1 and κ_2 such that for all $W_u \in \mathcal{W}$ with tunnel length at least L_0, the inclusion $N(A(W_u))^1 \hookrightarrow \widetilde{X}^1$ is a (κ_1, κ_2)-quasiisometric embedding.*

We refer to [HW16, Proposition 4.1] for a proof. They show that when tunnels are long enough, the hypotheses of Lemma 4.4 are satisfied. The same proof works in our case, as Lemma 4.4 works for any hyperbolic space, not necessarily proper. The main point is that geodesic paths in $N(A(W_u))^1$ can be written as alternating paths $P = \alpha_0\beta_1\alpha_1 \dots \beta_k\alpha_k$ satisfying the conditions of Lemma 4.4, where each β_i is a geodesic fellow-travelling with a forward flow segment of length $L \geq L_0$.

We note that the constants depend only on the following data: L_0 depends on the quasiconvexity constant μ of nuclei of walls (there is a uniform quasiconvexity constant μ as nuclei are either subintervals of edges or stars of vertices) and the constant B of the ladder overlap property. The constants κ_1 and κ_2 depend on L_0 and μ, but not on the tunnel length $L \geq L_0$ of individual immersed walls in $\mathcal{W}$.

Proposition 4.7 *Let $\mathcal{W}$ be a family of immersed walls with the ladder overlap property. Then there exists $L_1 \geq L_0$ such that for each $W_u \in \mathcal{W}$ with tunnel length at least L_1, $A(W_u)$ is a tree.*

This is a consequence of Proposition 4.6: Indeed, any closed path in $A(W_u)$ has length at most $\kappa_1\kappa_2$. Thus if $L_1 \geq L_0, \kappa_1\kappa_2 + 1$, then no closed path can contain a forward flow segment from a slope approximation. This implies that the closed path has to be contained in the intersection of a tree T_i in the flow space with the approximation of W_u, which is impossible. See [HW16, Proposition 4.4] for a precise proof.

Recall that a *wall* is a subspace $Y \subset \widetilde{X}$ such that $\widetilde{X} \setminus Y$ has exactly two components, neither of which is contained in any finite neighbourhood of Y (ie components are deep). We now state the main result of this subsection:

Proposition 4.8 *Let $\mathcal{W}$ be a family of immersed walls with the ladder overlap property and let L_1 be the constant from Proposition 4.7. Then for each $W_u \in \mathcal{W}$ with tunnel length at least L_1, W_u is a wall.*

Again, a detailed proof is given in [HW16, Proposition 4.6]. Since the flow space $\widetilde{X}$ is simply connected, it is enough to show that W_u locally separates (a small neighbourhood) into two components. Using Remark 4.2, the only place where this can fail is when two distinct slopes intersect in their interior. But when the tunnel length L is at least L_1, such a scenario would contradict the fact that $A(W_u)$ is a tree.

Proposition 4.9 (approximations of saturations are quasiconvex) *Let $\mathcal{W}$ be a family of immersed walls satisfying the ladder overlap property. Then there exists $L_2 \geq L_1$ such that for all $W_u \in \mathcal{W}$ with tunnel length $L \geq L_2$, there exist M_0, Θ_1 and Θ_2 such that for any M-saturation W of W_u with $M \geq M_0 L$, the inclusion $N(A(W))^1 \hookrightarrow \widetilde{X}^1$ is a (Θ_1, Θ_2)-quasiisometric embedding.*

Proof We will prove the statement by again using Lemma 4.4. Let P be a path in $N(A(W))^1$ such that P is a concatenation $\alpha_0\beta_1\alpha_1 \ldots \beta_k\alpha_k$, with each α_i a geodesic in the approximation of a component subgraph of $\overline{W}$ and each β_i a geodesic segment in a principal flow line. By Proposition 4.6, each α_i is a (κ_1, κ_2)-quasigeodesic, while Proposition 2.7 ensures that each β_i is a (λ_1, λ_2)-quasigeodesic. We choose tunnel length L large enough so that, in length smaller than L, α_i and β_j diverge from each other: recall that primary bust endpoints are chosen so that their flow rays are not asymptotic to any principal flow line (Lemma 2.18). By choosing L larger than the minimal divergence distance between the flow ray of any primary bust endpoint and a principal flow line, we obtain the required bounds. □

We also observe that since any immersed wall W_u (respectively its M-saturation W) of tunnel length L is at Hausdorff distance L from its approximation $A(W_u)$ (respectively $A(W)$), we have:

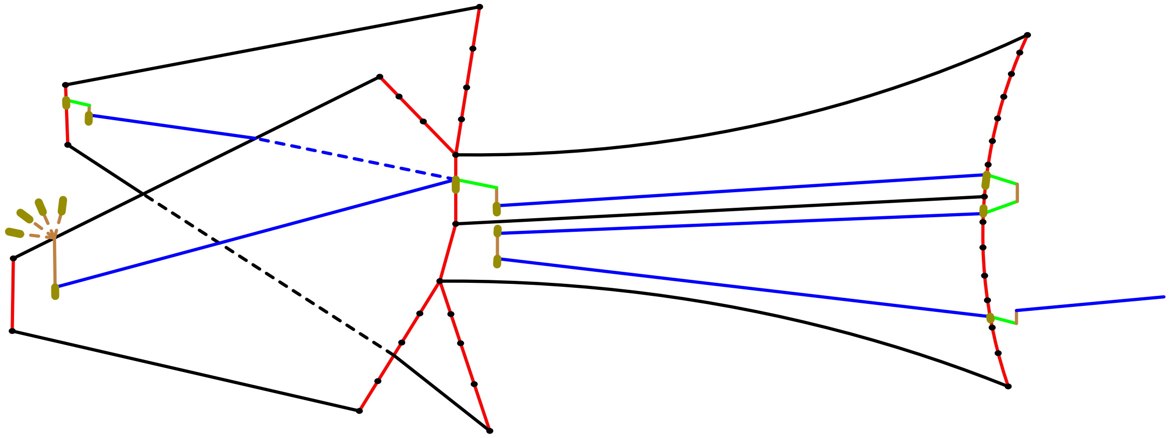


Figure 7: The busts in olive are all in a single complementary component.

Lemma 4.10 *For any wall W_u (with M-saturation W), the limit sets in $\partial\widetilde{X}^1$ of $N(W_u)^1$ and $N(A(W_u))^1$ coincide, as do those of $N(W)^1$ and $N(A(W))^1$.* □

Lemma 4.11 *Let d and d' be primary busts such that for some $m, m' \in \mathbb{Z}$, we have $d \subset T_{mL}$ and $d' \subset T_{m'L}$, and one of the endpoints of each of d and d' is contained in W_u. Then the interiors of d and d' lie in the same complementary component of W_u.*

Proof For the purposes of this argument, we endow W_u with a graph structure where the vertices are vertices of nuclei and the endpoints of various primary and secondary busts. Edges consist of forward flow segments in levels, slopes and segments in nuclei between vertices. Let us assume that $d \subset T_0$ and let P be a (geodesic) path in W_u between the endpoints of d and d'. The edge of P with endpoint d^+ (say), which is either a slope starting at d^+ or a forward flow segment ending at d^+, leads to a nucleus $\boldsymbol{N}$ in a "fractional" tree $T_{1/2}$ or $T_{-L+1/2}$. Except for the central vertex, every vertex of this nucleus is the endpoint of either a primary bust or of a secondary bust. In the latter case, note that each such secondary bust (in this fractional copy of T) meets the same complementary component C as d (see Figure 7), and thus flowing such a secondary bust to a primary bust in T_L or T_0 (along an edge of W_u without crossing it) will not lead to a change of components. In the former case, the primary bust in the fractional copy, with an endpoint in $\boldsymbol{N}$, is not in C, but travelling back along the slope will lead to its parallel copy (in either T_0 or T_{-L}), and this primary bust is contained in C. One can now proceed by induction on the distance between d and d' to obtain the desired result. □

Proposition 4.12 *Let $W_u \in \mathcal{W}$ be an immersed wall with tunnel length at least L_2 (from Proposition 4.8). Then $\mathrm{Stab}(W_u) < \Gamma$ is a codimension-1 subgroup that stabilises each complementary component of W_u and acts cocompactly on W_u.*

Proof Complementary components are not flipped: as a consequence of Lemma 4.11, if $\gamma \in \mathrm{Stab}(W_u)$ and C is a complementary component of W_u, we have that $\gamma C = C$.

Let us show cocompactness. Recall (from Section 4.1) that W_u is a component of the graph $\overline{W}$. For a point $x \in \widetilde{X}$ such that $x \in \overline{W}$ (the image of $\overline{W}$, to be precise), denote by $W(x)$ the component of $\overline{W}$ that contains x. Observe that for two points $x, y \in \overline{W}$, $W(x) = W(y)$ if and only if $y \in W(x)$ (and $x \in W(y)$).

Let L be the tunnel length of W_u. Denote by Γ_L the subgroup of Γ generated by G and t^L. We first note that, by the equivariant choice of primary busts, and the fact that no primary bust intersects any secondary bust (Lemma 4.1(1)), the action of Γ_L on the flow space restricts to an action on $\overline{W}$.

Let $x \in W_u$. Let us denote by $H < \Gamma$ the stabiliser of W_u. For an element $\gamma \in \Gamma_L$, we observe that $\gamma \in H$ if and only if $\gamma x \in W_u$. Indeed, $\gamma x \in W_u$ if and only if $W_u = W(x) = W(\gamma x)$, with the latter being equal to $\gamma W(x)$. Let us denote by X the compact quotient of $\widetilde{X}$ by the action of Γ. Observe that X is the mapping torus of the compact graph $G\backslash T$ by the map induced by f. There are finitely many images in X of busts and nuclei of W_u, and hence the image of W_u in X is compact. This implies that H acts cocompactly on W_u.

Finally, we show that $\mathrm{Stab}(W_u)$ has codimension 1. There is a continuous equivariant collapse map from a Cayley graph of Γ to $\widetilde{X}^1$ which crushes free factors of G to points. Since W_u separates $\widetilde{X}$ into deep components, so does its preimage in Γ, giving us the desired result. □

Note that if W_u is a wall, then any M-saturation of W_u separates $\widetilde{X}^1$ into several deep components and therefore separates the boundary $\partial\widetilde{X}^1$. In fact:

Proposition 4.13 *Let $W_u \in \mathcal{W}$ be a wall with tunnel length at least L_2 (from Proposition 4.9). Then there exists $L_3 \geq L_2$ such that for all $L > L_3$ and for all M large enough, the stabiliser of the M-saturation W of W_u is a relatively quasiconvex full subgroup. Further, W separates the flow space into at least two deep components.*

Proof To prove separation with deep components, since W_u itself is of codimension 1 by Proposition 4.12, it suffices to find, for each deep component of $\widetilde{X} \setminus W_u$, a ray starting at a point of W_u, entering the given component, and never encountering W again. To do that, it suffices to find a ray avoiding all the principal flow lines issued from W_u, and diverging from each of them within $\frac{1}{10}M$ of distance from W_u. Since $\widetilde{X}$ is nonelementary hyperbolic and principal flow lines are disjoint and diverge from each other in bounded time, it is possible to find such a ray, provided M is sufficiently large compared to δ.

Since W is quasiconvex, an application of [EGN21, Theorem 1.2] implies that the stabiliser of W is relatively quasiconvex in Γ. Let us show that $\mathrm{Stab}(W)$ is also full. To this end, consider a principal flow line Λ whose associated parabolic group (a suspension of a vertex stabiliser in T_0 by an element translating along Λ) intersects $\mathrm{Stab}(W)$. We want to show that $\mathrm{Stab}(W)$ contains a finite-index subgroup of the parabolic group. For this it is sufficient to show that Λ intersects W, since in that case, the saturation property of W ensures that W is stabilised by such a subgroup.

Thus, let $\mathrm{Stab}(W)$ intersect $\mathrm{Stab}(\Lambda)$. We distinguish whether the intersection contains an element translating along Λ or an element fixing a vertex in Λ.

First assume that $\mathrm{Stab}(W)$ contains an element of Γ which acts loxodromically on Λ. This implies that a finite neighbourhood of W contains Λ. Since every pair of principal flow lines diverge, either a vertex of Λ is contained in W (and we are done, by the property of saturations), or there exists a periodic union of tunnels, slopes and nuclei of W that fellow-travels with Λ. Assume the latter. If there exists a singular vertex (of some nucleus) in this periodic union, then the no twinning property ensures a contradiction. Let γ be translating along Λ, and such that γ^k stabilises W (for some large k). Let $x \in W$ be near Λ, and consider a shortest path p in W from x to $\gamma^k x \in W$. Since W is assumed not to cross Λ, the path p does not contain an arc of Λ. Thus p consists of nuclei, slopes and levels, and is a quasigeodesic that must remain close to Λ. Note that p does not contain any piece of a principal flow line by the no twinning assumption. Since p moves from a tree T_r to a tree T_{r+k} for some large k, there must be a level in p. As a subpath of p, it travels close to Λ. But a level is a piece of a periodic flow line, and by our choice of primary bust endpoints (Lemma 2.18) and large tunnel length, levels diverge from any principal flow line Λ. Hence p does not stay close to Λ, as promised.

Assume now that the stabiliser of W contains an elliptic element g, fixing some vertex v of T_0. Let $x \in T_0$ be in W so that for each $n \in \mathbb{Z}$, $g^n x \in W$. Since the T_0-distance between $g^m x$ and $g^n x$ is bounded for all m and n (it equals twice the T_0-distance between x and v), there exists $K \geq 0$ such that any geodesic path $\alpha_{m,n}$ in the quasiconvex space W between $g^m x$ and $g^n x$ does not meet T_i for $i > K$. We flow each such geodesic to T_K to obtain a path in T_K between $f^K(g^m x)$ and $f^K(g^n x)$. Observe that there is a uniform bound on the length of each such path. Thus, up to taking a subsequence, there is an infinite valence vertex of T_K contained in each of the flowed paths between $f^K(g^m x)$ and $f^K(g^n x)$. By equivariance of f, this vertex w is fixed by $\phi^K(g^n)$. By uniqueness of vertex stabilisers under the G-action on T_K, we have $w = f^K(v)$. We claim that this implies that infinitely many of the paths $\alpha_{m,n}$ intersect the principal flow line Λ_v of v. Indeed, by Lemma 2.3 the intersection of the backward flow of w with the trees T_i intersects the infinite collection $\alpha_{m,n}$ in a finite set. Let $w' \in T_i$ be one such point that lies in infinitely many $\alpha_{m,n}$. If w' is in Λ_v, we are done. If not, then by equivariance, we again have that w' is stabilised by $\phi^i(g^k)$ for infinitely many k, a contradiction. □

4.5 Many effective walls

Recall that $D \subset T$ denotes the fixed fundamental domain.

Definition 4.14 Using the terminology of [HW16], we say $\widetilde{X}$ has *many effective walls* if the following two conditions are satisfied:

(1) For each $y \in D \subset T_0$ such that the flow ray from y does not meet a vertex, there exists a set $\mathcal{W}$ of immersed walls with the ladder overlap property such that for every $\epsilon > 0$, there exists $W_u \in \mathcal{W}$ of arbitrarily large tunnel length L and a primary bust in the ϵ-neighbourhood of y. We will call such a set of walls a *regular effective set of walls*.

(2) For each periodic point $a \in D$, there exists $k = k(a)$ and a set of immersed walls $\mathcal{W}_a$ with the ladder overlap property such that for each primary bust $d' \subset T' = T_{1/2}$ that is joined to $a' = a_{1/2} \in T'$ by a path in W_u disjoint from the slopes of W_u, we have that $d_{\widetilde{X}^1}(f^n(a), f^n(d)) \geq 3\delta + 2\lambda$, for all $n \geq k$. We will call such a set of walls a *periodic effective set of walls.*

Let us comment on why one needs the property of many effective walls. The full details are available in [HW16, Section 5]. For a hyperbolic group to admit a proper cocompact cubulation, the boundary criterion of [BW12] stipulates that every pair of distinct points in the boundary of the group (equivalently, the limit set of every biinfinite geodesic of $\widetilde{X}^1$ in the setup of [HW16]) should be separated by the limit set of a quasiconvex codimension-1 subgroup (equivalently, a quasiconvex wall in $\widetilde{X}^1$). Having many effective walls assures that this can be done in $\widetilde{X}^1$. Indeed, there are two types of biinfinite geodesics in $\widetilde{X}^1$, "horizontal" geodesics and "nonhorizontal" geodesics, which we define below. If $\widetilde{X}^1$ has many effective walls, then each horizontal geodesic is cut by a wall from a periodic effective set, while each nonhorizontal geodesic is cut by a wall from a regular effective set.

Definition 4.15 (geodesic classification) Let $N > 0$. A biinfinite geodesic γ in $\widetilde{X}^1$ is *N-horizontal* (N-ladderlike in [HW16]) if there exists a forward flow segment σ of length N such that a geodesic of the forward ladder $N(\sigma)$ joining the endpoints of σ fellow-travels with a subpath of γ at distance at most $2\delta + \lambda$. Here δ is the hyperbolicity constant of $\widetilde{X}^1$ and λ is the quasiconvexity constant of forward ladders (Proposition 2.7). Otherwise, γ is *N-nonhorizontal* (N-deviating in [HW16]).

Theorem 4.16 *The flow space $\widetilde{X}$ has many effective walls.*

This is proved in [HW16, Theorem 6.16], to which we refer the reader for a detailed proof. The main ingredients of their proof, which are available in our case as well, are the facts that periodic points are dense in T (Lemma 2.14) and that flow rays starting from translates of a periodic point diverge (Proposition 2.12). With these ingredients at hand, here is a brief sketch of the proof. The first condition of Definition 4.14 can be verified by first choosing a periodic point in an ϵ-neighbourhood of the given point y and then choosing one periodic point in each edge of D so that pairwise, their flow lines diverge. This will ensure that the ladder overlap property holds when primary busts are chosen in small enough neighbourhoods of these points, depending on the tunnel length. The family of walls is now constructed by choosing tunnel lengths to be large common multiples of the periods of the chosen periodic points, and choosing primary busts as above. To verify the second condition, we choose one periodic point per edge of D as above, including one for the edge containing a, while ensuring that pairwise, the flow lines of the periodic points and of a diverge. The family of walls is then constructed as done for the first condition.

Let us conclude this section with another property that will be useful in the next section. Following [HW16], we say $\widetilde{X}$ is *level-separated* if for each $N > 0$ and $K \geq 0$, and each N-nonhorizontal geodesic γ, there exists a point $y \in \widetilde{X}$ such that the backward flow of length n meets γ in an odd-cardinality set for all large n, and the intersection is at distance at least $N + K$ from both y and the leaves of the backward flow.

Lemma 4.17 *The flow space $\widetilde{X}$ is level-separated.*

In order to prove this statement, we will make use of an $\mathbb{R}$-tree T_∞ obtained as a limit of the trees T_m in the flow space with metrics $d_m = d|_{T_m}/\lambda^m$, where λ is the (maximal) stretching factor of the train track map f. We will not go into details, but refer the reader to [HW16, Section 6; Hor17] . Formally, T_∞ is an ultralimit; its points are equivalence classes of sequences of points in the T_m. In particular, to each point in T_0 (and in $\widetilde{X}$), its flow ray defines a point in T_∞. This defines a tautological map $\rho\colon \widetilde{X} \to T_\infty$.

The main points we will use are easy facts: that the map $\rho\colon \widetilde{X} \to T_\infty$ is continuous, restricts to an isometric embedding on any vertical edge (as f is a train track map) and that the preimage of any point in T_∞ consists of points whose forward flow rays either intersect or remain at bounded Hausdorff distances from each other.

Proof Let $\gamma\colon \mathbb{R} \to \widetilde{X}^1$ be an N-nonhorizontal geodesic. Observe that the flow-like parts of γ are horizontal segments starting and ending at vertices and that the vertical parts are paths in trees T_k. Since γ is N-nonhorizontal, it maps infinitely many unit subintervals of $\mathbb{R}$ to vertical edges, which we call "vertical" subintervals.

First, we make a slight change. Let γ_{vert} be the union of paths defined on the vertical subintervals of γ, but now reparametrised on $\mathbb{R}$.

Let $\gamma^\infty = \rho \circ \gamma_{\text{vert}}\colon \mathbb{R} \to T_\infty$ be the image of γ_{vert} (and also of γ, in fact) in T_∞. Call $\gamma_{\text{vert}}(s)$ a "flat" point if s is a point of discontinuity for γ_{vert}. Observe that this happens only when $\gamma(s)$ is the starting point in γ of a horizontal segment, and that there are only countably many flat points. Since all points in any horizontal segment have the same flow, all such points have the same image in T_∞. Therefore the map $\gamma^\infty\colon \mathbb{R} \to T_\infty$ is continuous, and in fact, 1-Lipschitz.

Further, there exists D such that the preimage of a point in γ^∞ is of diameter $\leq D$. Indeed, if s_1 and s_2 have the same γ^∞-image, observe that $\gamma_{\text{vert}}(s_1)$ and $\gamma_{\text{vert}}(s_2)$ correspond to points in the image of γ which fellow-travel a forward flow segment. Since γ is N-nonhorizontal, there exists a D as required.

In fact, the same argument shows that $\gamma^\infty(\mathbb{R}_+)$ and $\gamma^\infty(\mathbb{R}_-)$ have infinite diameter and go to ends γ^∞_+ and γ^∞_- of T_∞, respectively. Note that, by the bound on the diameter of preimages, $\gamma^\infty_+ \neq \gamma^\infty_-$. In particular, there are uncountably many points that cut $\gamma^\infty(\mathbb{R})$ into two unbounded components, with each such point contained in $\gamma^\infty(\mathbb{R})$ (as T_∞ is an $\mathbb{R}$-tree).

Let us now show that there exists D' such that the preimage of a point under γ^∞ has cardinality at most D'. Indeed, if there are many points in the preimage, there must be many in the same tree T_k, and their diameter being bounded, either they accumulate or have to be placed so that there is a vertex and a huge angle to reach them. Accumulation is not permitted since all edges are stretched by the scaling factor (more precisely, we are guaranteed that two points in the same edge are flown to different points of T_∞). Also for two points separated by a huge angle, we already noticed that their flow lines cannot fellow-travel (see Proposition 2.9), and hence they must diverge with speed defined by the stretch factor λ, thus defining different points in the limit tree T_∞.

Call $\gamma^\infty(s)$ a "backtrack" point if there are two intervals $[s_-, s)$ and $(s, s+]$ that are sent by γ^∞ to the same component of $T_\infty \setminus \{\gamma^\infty(s)\}$.

Observe that since the map ρ is injective on edges, if $\gamma^\infty(s)$ is a backtrack point, then $\gamma_{\text{vert}}(s)$ is a vertex. Therefore there are countably many backtrack points in γ^∞.

There are two kinds of backtrack points: those such that the component of the intervals $[s_-, s)$ and $(s, s+]$ contains both the ends γ_+^∞ and γ_-^∞, and those that don't. For the former, take the centre of the tripod $(\gamma_+^\infty, \gamma_-^\infty, \gamma^\infty(s))$, and call it a "crossroad" point.

There are countably many backtrack points, so the union of flat points, backtrack points and crossroad points is a countable set.

Pick a point ξ in $\gamma^\infty(\mathbb{R})$ that is not a backtrack point nor a crossroad point nor a flat point, and such that ξ separates $\gamma^\infty(\mathbb{R})$ in two infinite components. Since it is not a backtrack point, there is a well-defined direction of crossing ξ at each preimage s (a small interval around s such that (s_-, s) and (s, s_+) are sent on different sides of ξ, each containing only one of the ends γ_+^∞ and γ_-^∞).

Since preimages in γ_{vert} are finite, for ξ there can only be an odd number of preimages in γ_{vert}. Since ξ is not a flat point, it has the same (odd) number of preimages in both γ and γ_{vert}. Then at least one of the points x in the preimage is such that for a large n, if $y = f^n(x)$, then the backward flow of large enough length from x meets γ in an odd-cardinality set, with the intersection at a large distance from y and the leaves of the backward flow, as required. □

5 Cutting principal flow lines and geodesics by wall saturations

In this section, we will show that three types of subsets of $\widetilde{X}^1$ are separated by the saturations of walls. In [HW16], they show that all biinfinite geodesics are separated by (unsaturated) walls. For relative cubulation (see Section 3), we will need more: that saturations of walls also separate principal flow lines from geodesic rays and pairs of principal flow lines.

Let us begin with a tautological observation in preparation for the future coning-off of principal flow lines:

Lemma 5.1 *Let W_u be a quasiconvex wall in $\widetilde{X}$ and W an M-saturation of W_u. Then for each principal flow line Λ of a singular vertex, either $\Lambda \subset W$ or the two points of $\partial\Lambda$ lie in a single component of $\partial\widetilde{X}^1 \setminus \partial W$.* □

We will also need the notion of "cut" in order to disconnect points in the Bowditch boundary:

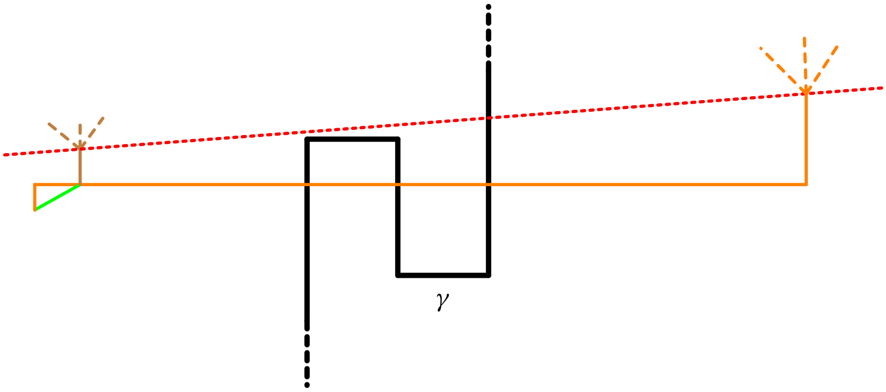


Figure 8: Proof of Proposition 5.3. The geodesic γ hits the approximation of a slope of W_u. The red dotted line indicates a flow line in the saturation of W_u.

Definition 5.2 Let W_u be a quasiconvex wall in $\widetilde{X}$ and let W be an M-saturation of W_u. Let A be a quasiconvex subspace of $\widetilde{X}^1$ such that ∂A has at least two points. We say that W *cuts* A if

(1) $\partial A \cap \partial W = \varnothing$ (as subsets of $\partial \widetilde{X}^1$), and

(2) ∂W separates ∂A, ie ∂A nontrivially intersects at least two components of $\partial \widetilde{X}^1 \setminus \partial W$.

5.1 Cutting geodesics

Recall that there are two types of geodesics (see Definition 4.15) in $\widetilde{X}^1$. In this subsection, we will show that for each geodesic line that is not a principal flow line, there exists a wall whose saturation cuts the geodesic.

Proposition 5.3 *Let $\gamma : \mathbb{R} \to \widetilde{X}^1$ be an N-nonhorizontal biinfinite geodesic in the flow space. Then there exists a quasiconvex wall $W_u \to \widetilde{X}$ and an M-saturation W of W_u such that W cuts γ.*

Proof Proposition 5.18 of [HW16] shows that in this case, there exists a quasiconvex wall W_u which separates $\partial\gamma$. Let us quickly explain the idea behind their proof. We refer the reader to Figure 8 for an illustration of what follows. We choose a wall W_u from a regular effective set of walls (Definition 4.14(1)) with tunnel length L sufficiently larger than N, while ensuring that the approximation $A(S)$ of some slope S of W_u intersects a vertical segment of γ in an odd number of points (in the interior of some vertical edges), and far away from the endpoints of $A(S)$ (arguing exactly as in [HW16] while using Lemma 4.17 for level separatedness). Such a W_u exists because of Theorem 4.16. The fact that γ is N-nonhorizontal will ensure that γ does not have long subpaths which $(3\delta+2\lambda)$-fellow-travel with W_u (as the tunnel length L of W_u is much larger than N). This implies, by Lemma 4.4, that the union of γ and $N(A(W_u))^1$ quasiisometrically embeds in $\widetilde{X}^1$ and thus ∂W_u is disjoint from, and separates, $\partial\gamma$ in $\partial\widetilde{X}^1$.

We now choose $M \in L\mathbb{N}$ sufficiently large, and let W be the M-saturation of W_u. Since ∂W_u separates $\partial\gamma$, and $W_u \subset W$, in order to show that ∂W also separates $\partial\gamma$, it suffices to show that $\partial\gamma$ is disjoint from ∂W,

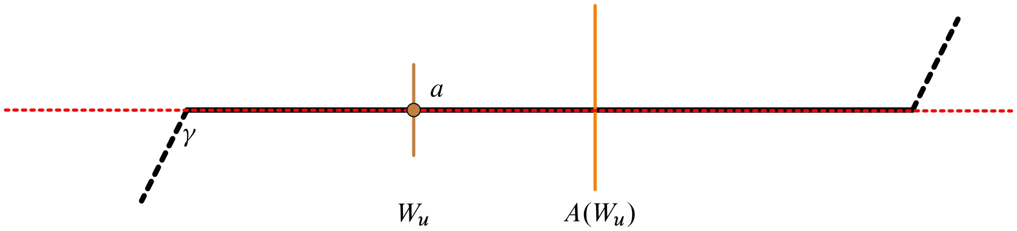


Figure 9: Proof of Proposition 5.4. The geodesic γ is cut along a nucleus of W_u at a. The red dotted line indicates the principal flow line Λ_a.

and for that it suffices to show that the union of γ and $N(A(W))^1$ quasiisometrically embeds in $\widetilde{X}^1$. Since we already know that the union of γ and $N(A(W_u))^1$ quasiisometrically embeds in $\widetilde{X}^1$ and that the saturation consists of translates of W_u that are at distance larger than M from each other, and of principal flow lines between them, a failure of quasiconvexity for the union of γ and $N(A(W))^1$ would mean that γ fellow-travels with a principal flow line for a long distance (at least M by definition of saturation). This contradicts the N-nonhorizontality of γ. □

Proposition 5.4 *Let $\gamma\colon \mathbb{R} \to \widetilde{X}^1$ be a biinfinite geodesic in the flow space such that*

(1) *γ is N-horizontal for all large N, and*

(2) *no ray of γ is asymptotic to any ray of a principal flow line.*

Then there exists a quasiconvex wall $W_u \to \widetilde{X}$ and an M-saturation W of W_u such that W cuts γ.

Proof Proposition 5.19 of [HW16] ensures that there exists a quasiconvex wall W_u which separates $\partial\gamma$; the wall W_u is chosen from a periodic effective set of walls $\mathcal{W}_a$ (Definition 4.14(2)) so that, up to translation, the point a in the nucleus of W_u intersects γ roughly in the middle of a long horizontal subpath of γ, whose length is much bigger than the tunnel length of W_u (Figure 9). Such a W_u exists, again, by Theorem 4.16. By the way W_u was chosen, slope approximations of W_u do not have long $(3\delta+2\lambda)$-fellow-travelling subpaths with the flow line of a, and therefore with γ. Lemma 4.4 now ensures that the union of γ and $N(A(W_u)^1)$ quasiisometrically embeds in $\widetilde{X}^1$. Therefore ∂W_u is disjoint from, and separates, $\partial\gamma$ in $\partial\widetilde{X}^1$.

In order to prove the proposition, we will now choose an M-saturation W of W_u with M much larger than the length of the maximal subpath of γ that $(3\delta+2\lambda)$-fellow-travels with the principal flow line Λ_a through a. Observe that this maximal subpath is bounded as $\partial\gamma$ is disjoint from $\partial\Lambda_a$. An application of Lemma 4.4 now shows that the union of $N(A(W))^1$ with γ is a quasiisometric embedding and thus W cuts γ. □

5.2 Separating geodesics from principal flow lines

In this subsection, we will show that the union of a principal flow line and a geodesic ray is cut by a wall. More precisely:

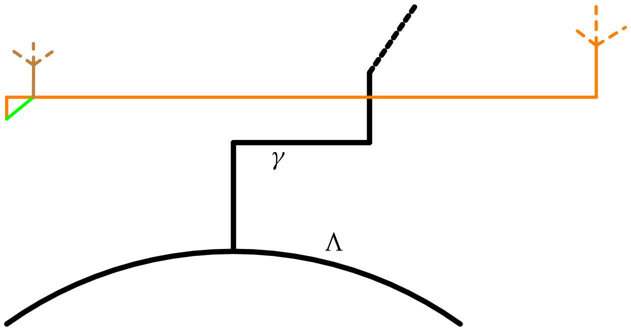


Figure 10: The wall W_u is chosen so as to hit the geodesic ray γ far from the principal flow line Λ.

Proposition 5.5 *Let Λ be a principal flow line and γ be a geodesic ray starting at some point on Λ. Suppose that γ and Λ are not asymptotic. Then there exists a quasiconvex wall $W_u \to \widetilde{X}$ and an M-saturation W of W_u, such that*

(1) *W does not separate $\partial\Lambda$, and*

(2) *W cuts $\Lambda \cup \gamma$.*

Proof Let $p \in \Lambda$ be the starting point of γ. Denote by Λ_+ and Λ_- the two geodesic subrays of Λ starting at p. Denote the endpoints at infinity of these rays by $\Lambda_+(\infty)$, $\Lambda_-(\infty)$ and $\gamma(\infty)$. Let ϵ be the maximal Gromov product at $p = \gamma(0)$ of $\gamma(\infty)$ and $\Lambda_\pm(\infty)$. Then both the concatenations $\gamma \cup \Lambda_\pm$ are $(1, 2\epsilon+10\delta)$-quasigeodesic lines. We will now choose a quasiconvex wall W_u (see Figure 10) such that

(1) $W_u \cap (\gamma \cup \Lambda_-) = W_u \cap (\gamma \cup \Lambda_+) \subset \gamma$,

(2) the union of the quasigeodesic lines $\gamma \cup \Lambda_\pm$ with $N(A(W_u)^1)$ embeds quasiisometrically, and

(3) ∂W_u separates $\gamma(\infty)$ from $\Lambda_+(\infty)$ and $\Lambda_-(\infty)$.

The choice of W_u that will work for us is a wall given by either Proposition 5.3 or 5.4 depending on whether or not γ is N-horizontal for all N.

Let λ be the quasiconvexity constant from Proposition 2.7, and ϵ as before. Let κ and L be the quasiconvexity constants for $N(A(W_u))^1$ and the tunnel length of W_u, respectively, for such walls, as guaranteed by Proposition 4.6. We choose W_u so that W_u intersects γ at points at distance much larger than $2\epsilon + 20\delta + 2\kappa + \lambda + L$ from p.

Thanks to the large distance between the point p and $W_u \cap \gamma$, Lemma 4.4 says that the union of $N(A(W_u)^1)$ with $\gamma \cup \Lambda$ embeds quasiisometrically and both W_u and $A(W_u)$ are disjoint from Λ. This implies that ∂W_u does not separate $\partial\Lambda$. We now choose an M-saturation W of W_u with M much larger than the maximal $(3\delta+2\lambda)$-fellow-travelling length between pairs of principal flow lines. This choice ensures, yet again by Lemma 4.4, that W is disjoint from Λ and therefore cuts $\gamma \cup \Lambda$ but does not separate $\partial\Lambda$. □

5.3 Cutting pairs of principal flow lines

Proposition 5.6 *Let $\Lambda_1 \neq \Lambda_2$ be two principal flow lines. Then there exists a quasiconvex wall $W_u \to \widetilde{X}$ and an M-saturation W of W_u such that*

(1) *W cuts $\Lambda_1 \cup \Lambda_2$, and*

(2) *∂W does not separate $\partial\Lambda_i$.*

Proof Let γ be a geodesic segment between Λ_1 and Λ_2. Note that γ contains at least one vertical edge e, since $\Lambda_1 \neq \Lambda_2$. By Lemma 2.18, there exists a periodic point x in the edge e such that the periodic flow line Λ_x is disjoint from the lines Λ_i and diverges from Λ_i in both the forward and backward directions. We now choose an immersed wall W_u from a regular effective set (Definition 4.14(1)) and satisfying the following:

(1) W_u has tunnel length L much larger than the $(3\delta+2\lambda)$-fellow-travelling length of Λ_x with each Λ_i,

(2) the intersection of Λ_x with the approximation $A(W_u)$ is a segment of length L with centre x,

(3) the approximation $A(W_u)$ intersects the edge e at the singleton $\{x\}$ and the geodesic γ in an odd cardinality set, and

(4) the intersection of the union $\Lambda_1 \cup \Lambda_2 \cup \gamma$ with $A(W_u)$ coincides with the intersection of γ with $A(W_u)$.

Such a W_u exists as $\widetilde{X}$ is level-separated (Lemma 4.17) and has many effective walls (Theorem 4.16). Property (1) above allows us to apply Lemma 4.4 to the union of $\Lambda_1 \cup \Lambda_2 \cup \gamma$ with $N(A(W))^1$ and conclude that it is a quasiisometric embedding in $\widetilde{X}^1$. Therefore ∂W_u is disjoint from, and separates, $\partial(\Lambda_1 \cup \Lambda_2)$. Further, by (4), it does not separate $\partial\Lambda_i$. We now choose an M-saturation W of W_u for M much larger than L. By applying Lemma 4.4 again and arguing as above for W_u, we obtain the desired result. □

6 Separating pairs of points in the Bowditch boundary

The goal of this section is to show that the mapping torus $\Gamma = G \rtimes_\phi \mathbb{Z}$ satisfies the hypothesis of Theorem 3.2. Recall that Γ is hyperbolic relative to the collection $\mathscr{P}$ consisting of the suspensions of the free factors H_i of G.

We first make two observations connecting the Gromov boundary of the flow space $\widetilde{X}^1$ with the Bowditch boundary $\partial_B(\Gamma, \mathscr{P})$. We denote by $\widehat{\widetilde{X}^1}$ the cone-off of the (1-skeleton of the) flow space $\widetilde{X}^1$ over the set of principal flow lines. Recall also that $\widehat{\Gamma}$ denotes the cone-off of Γ relative to $\mathscr{P}$.

As a consequence of the relative Švarc–Milnor lemma [CC07, Theorem 5.1], we obtain:

Proposition 6.1 *The cone-off $\widehat{\widetilde{X}^1}$ is Γ-equivariantly quasiisometric to the cone-off $\widehat{\Gamma}$.*

Let $\beta\colon \widetilde{X}^1 \to \widehat{\widetilde{X}^1}$ denote the inclusion map from the flow space to its cone-off. We will denote by $\partial\beta$ the induced map from $\partial\widetilde{X}^1$ to the Bowditch boundary $\partial_B(\Gamma, \mathscr{P})$. Observe that $\partial\beta$ maps the limit points of any principal flow line Λ to the corresponding cone vertex v_Λ. In fact:

Corollary 6.2 *The map $\partial\beta : \partial\widetilde{X}^1 \to \partial_B(\Gamma, \mathscr{P})$ is continuous and surjective. Further, for all $\xi \in \partial_B(\Gamma, \mathscr{P})$, $\partial\beta^{-1}(\xi)$ is either a singleton or contains two points. The latter arises if and only if ξ is a cone vertex.* □

Proposition 6.3 *If ξ and ζ are points in $\partial\widetilde{X}^1$ that are preimages of conical limit points of the Bowditch boundary, then there exists a geodesic line in $\widetilde{X}^1$ joining them.*

If $\xi \in \partial\widetilde{X}^1$ is the preimage of a conical limit point of the Bowditch boundary, and z_0 a vertex in $\widetilde{X}$, then there exists a geodesic ray from z_0 that converges to ξ.

Proof Let (x_n) and (y_n) be sequences of vertices in $\widetilde{X}$ going to ξ and ζ, respectively, and let $[x_n, y_n]$ be a geodesic in $\widetilde{X}^1$. Let $z_0 \in T_0$ minimise the Gromov product $(\xi, \zeta)_{z_0} \in \mathbb{N}$, and let R_0 be the minimum.

Let $B(z_0, R)$ be the ball of $\widetilde{X}^1$ of radius R centred at z_0. We aim to prove that for all $R > R_0$, there exists $m \in \mathbb{N}$ such that $\{B(z_0, R) \cap [x_n, y_n] \mid n > m\}$ is a finite collection of segments of length $\geq 2R - 2R_0 - 10\delta$.

First, for n large enough, $B(z_0, R) \cap [x_n, y_n]$ is indeed such a segment, otherwise one easily gets that $(\xi, \zeta)_{z_0} > R_0$.

If for some R the collection is infinite, let k be such that in T_k the intersection $T_k \cap B(z_0, R) \cap [x_n, y_n]$ is an infinite collection of segments as n varies (since horizontal segments are determined by their endpoints, such a k exists). Let $z_k \in T_k$, and let u_n be in $T_k \cap B(z_0, R) \cap [x_n, y_n]$. Assume that the maximal angle of $[z_k, u_n]_{T_k}$ goes to infinity with n. Let v_k be the closest vertex to z_k at which the angle goes to infinity. Then the same is true for the initial vertical segment of either $[z_k, x_n]$ or $[z_k, y_n]$. It follows that either (x_n) or (y_n) converges to a point that is an image of the parabolic point of the principal flow line of v_k, contrary to our assumption.

The number of subsegments of $[x_n, y_n]$ around z_0 of a given length is therefore bounded when n varies. One can thus diagonally extract subsequences such that, inductively, for each given length, the subsegments of this length of $[x_n, y_n]$ around z_0 make a constant sequence. We thus obtain a biinfinite geodesic between ξ and ζ in $\widetilde{X}$.

The second assertion is obtained with a similar argument. □

Proposition 6.4 *Let ξ and ζ be two distinct points in the Bowditch boundary, and let W be the M-saturation of a quasiconvex wall W_u such that W cuts a geodesic between $\beta^{-1}(\xi)$ and $\beta^{-1}(\zeta)$. Then $\partial\beta(\partial W) = \partial\,\mathrm{Stab}(W)$ separates ξ and ζ and there exists a subgroup K of index at most 2 of $\mathrm{Stab}(W)$ such that ξ and ζ are in K-distinct components of $\partial_B(\Gamma, \mathscr{P}) \setminus \partial\beta(\partial W)$.*

Let us mention a word of caution here. Though $\mathrm{Stab}(W)$ is a codimension-1 subgroup (Proposition 4.13) of Γ, the statement above is necessary because, in general, a codimension-1 subgroup need not separate the Bowditch boundary. As an example, take the cyclic subgroup $P = \langle aba^{-1}b^{-1} \rangle$ of the free group $F(a, b)$ with peripheral structure $\mathscr{P} = P$. Here P is a maximal parabolic subgroup that is full, relatively quasiconvex and codimension-1, but does not separate the Bowditch boundary, a circle.

Proof Let ξ and ζ, and W_u be as in the statement, and $\tilde{\xi}$ and $\tilde{\zeta}$ be two points in $\widetilde{X}^1$ in the preimages $\partial\beta^{-1}(\xi)$ and $\partial\beta^{-1}(\zeta)$, respectively. Let $\mathcal{U}_\xi$ and $\mathcal{U}_\zeta$ be a clopen partition of $\partial\widetilde{X}^1 \setminus \partial W$ containing $\tilde{\xi}$ and $\tilde{\zeta}$, respectively. Consider points in $\partial\widetilde{X}^1$ identified in $\partial_B(\Gamma, \mathcal{P})$. By Corollary 6.2 they are at the end of the same principal flow line, and by Lemma 5.1, if one is in $\mathcal{U}_\xi$, so is the other one. It follows that the map $\partial\beta$ sends $\mathcal{U}_\xi$ and $\mathcal{U}_\zeta$ on disjoints subsets $\partial\beta(\mathcal{U}_\xi)$ and $\partial\beta(\mathcal{U}_\zeta)$ of $\partial_B(\Gamma, \mathcal{P}) \setminus \partial\beta(\partial W)$. By surjectivity, they provide a partition of $\partial_B(\Gamma, \mathcal{P}) \setminus \partial\beta(\partial W)$, each containing ξ and ζ, respectively.

We need to check that both $\partial\beta(\mathcal{U}_\xi)$ and $\partial\beta(\mathcal{U}_\zeta)$ are closed. By symmetry, it is enough to check that one is closed. If a sequence (x_n) in $\partial\beta(\mathcal{U}_\xi)$ converges in $\partial_B(\Gamma, \mathcal{P}) \setminus \partial\beta(\partial W)$, lift it to a sequence (x'_n) in $\partial\widetilde{X}^1 \setminus \partial W$. Since $\partial\widetilde{X}^1$ is not compact, the sequence (x'_n) may or may not have an accumulation point. If it does, then the fact that $\mathcal{U}_\xi$ is closed, along with the continuity of $\partial\beta$, gives us the necessary conclusion. Now consider the case where it has no accumulation point in $\partial\widetilde{X}^1$. If x'_n is a point represented by a ray ρ'_n in $\widetilde{X}^1 \setminus W$ (from a given basepoint) that lives in the component of which ξ is adherent, it means that, up to taking a subsequence, all the rays ρ'_n pass through a singular vertex, and make an angle θ_n at that vertex such that $\theta_n \to \infty$ with n. Take v to be the closest singular vertex to the basepoint with this property. The limit of (x_n) in $\partial_B(\Gamma, \mathcal{P}) \setminus \partial\beta(\partial W)$ is then the parabolic point associated to the principal flow line of this vertex v. Since the rays live in the component of $\widetilde{X}^1 \setminus W$ adherent to ξ, this principal flow line has at least one endpoint not in $\mathcal{U}_\zeta$. By the property of saturations, either both endpoints are in $\mathcal{U}_\xi$, or both are in ∂W, and in that case the limit of (x_n) is in $\partial\beta(\partial W)$, which we assumed otherwise. So both points of the line are in $\mathcal{U}_\xi$, and their images, which are the limit of (x_n), are in $\partial\beta\mathcal{U}_\xi$. Therefore this set is closed.

Therefore ξ and ζ are separated by $\partial\beta(\partial W)$, which is $\partial\,\mathrm{Stab}(W)$, since the action of the latter is cocompact on W.

Fix a basepoint in the complement of W. Let A be the union of components C of $\widetilde{X} \setminus W$ such that there is a path from the basepoint to C that crosses the set W'_u an even number of times. Recall that W'_u is the complement in W of the set of its principal flow lines. We note that $\widetilde{X} \setminus W$ is thus partitioned into two subsets: A and its complement, which we denote here by B. Since the stabiliser of W sends complementary components of W to complementary components, it preserves this partition of W into A and B. Therefore there exists a subgroup of index at most 2 which preserves A (and hence also B).

By the way the saturation W of W_u cuts the preimages of ξ and ζ (as in Proposition 5.3, 5.4, 5.5 or 5.6), it is easy to see that ξ and ζ do not both lie in the limit set of A, or in the limit set of B, which gives us the desired result. □

We now prove the main result Theorem 1.1, namely, in our current notation, that the relatively hyperbolic group $(\Gamma, \mathscr{P})$ is relatively cubulated:

Proof of Theorem 1.1 We will show that the boundary criterion Theorem 3.2 holds. Let $\xi \neq \zeta$ be two points in the Bowditch boundary $\partial_B(\Gamma, \mathscr{P})$. We have three cases:

Case 1 (both ξ and ζ are conical points) By abuse of notation, we denote by ξ and ζ the unique preimages in $\partial\widetilde{X}^1$ of ξ and ζ (Corollary 6.2). Let γ be a geodesic in $\widetilde{X}^1$ joining ξ and ζ as given by Proposition 6.3. Then Propositions 5.3 and 5.4 ensure that there exists a wall saturation W that cuts γ. An application of Proposition 6.4 gives the desired result.

Case 2 (without loss of generality, ξ is conical while ζ is parabolic) We consider a geodesic ray γ in $\widetilde{X}^1$, between a vertex of the principal flow line Λ associated to ζ and the preimage of ξ, as given by Proposition 6.3. Proposition 5.5 ensures that the union of γ and Λ is cut by a wall saturation W. Proposition 6.4 then gives the result.

Case 3 (both ξ and ζ are parabolic points) Proposition 5.6 ensures that the principal flow lines associated to ξ and η are cut by a wall saturation W. The result then follows from Proposition 6.4. □

References

[Bow12] **B H Bowditch**, *Relatively hyperbolic groups*, Int. J. Algebra Comput. 22 (2012) art. id. 1250016 MR Zbl

[BW12] **N Bergeron**, **D T Wise**, *A boundary criterion for cubulation*, Amer. J. Math. 134 (2012) 843–859 MR Zbl

[CC07] **R Charney**, **J Crisp**, *Relative hyperbolicity and Artin groups*, Geom. Dedicata 129 (2007) 1–13 MR Zbl

[CLR94] **D Cooper**, **D D Long**, **A W Reid**, *Bundles and finite foliations*, Invent. Math. 118 (1994) 255–283 MR Zbl

[DKL15] **S Dowdall**, **I Kapovich**, **C J Leininger**, *Dynamics on free-by-cyclic groups*, Geom. Topol. 19 (2015) 2801–2899 MR Zbl

[DL22] **F Dahmani**, **R Li**, *Relative hyperbolicity for automorphisms of free products and free groups*, J. Topol. Anal. 14 (2022) 55–92 MR Zbl

[DM23] **F Dahmani**, **S Krishna M S**, *Relative hyperbolicity of hyperbolic-by-cyclic groups*, Groups Geom. Dyn. 17 (2023) 403–426 MR Zbl

[DMM25] **F Dahmani**, **S K Meda Satish**, **J P Mutanguha**, *Hyperbolic hyperbolic-by-cyclic groups are cubulable*, Geom. Topol. 29 (2025) 259–268 MR Zbl

[EG20] **E Einstein**, **D Groves**, *Relative cubulations and groups with a* 2*-sphere boundary*, Compos. Math. 156 (2020) 862–867 MR Zbl

[EG22] **E Einstein**, **D Groves**, *Relatively geometric actions on* CAT(0) *cube complexes*, J. Lond. Math. Soc. 105 (2022) 691–708 MR Zbl

[EGN21] **E Einstein**, **D Groves**, **T Ng**, *Separation and relative quasiconvexity criteria for relatively geometric actions*, Groups Geom. Dyn. 18 (2024) 649–676 MR Zbl

[FM15] **S Francaviglia**, **A Martino**, *Stretching factors, metrics and train tracks for free products*, Illinois J. Math. 59 (2015) 859–899 MR

[GL07] **V Guirardel**, **G Levitt**, *The outer space of a free product*, Proc. Lond. Math. Soc. 94 (2007) 695–714 MR Zbl

[GM23] **D Groves**, **J F Manning**, *Hyperbolic groups acting improperly*, Geom. Topol. 27 (2023) 3387–3460 MR Zbl

[Hor17] **C Horbez**, *The boundary of the outer space of a free product*, Israel J. Math. 221 (2017) 179–234 MR Zbl

[Hru10] **G C Hruska**, *Relative hyperbolicity and relative quasiconvexity for countable groups*, Algebr. Geom. Topol. 10 (2010) 1807–1856 MR Zbl

[HW08] **F Haglund**, **D T Wise**, *Special cube complexes*, Geom. Funct. Anal. 17 (2008) 1551–1620 MR Zbl

[HW15] **M F Hagen**, **D T Wise**, *Cubulating hyperbolic free-by-cyclic groups: the general case*, Geom. Funct. Anal. 25 (2015) 134–179 MR Zbl

[HW16] **M F Hagen**, **D T Wise**, *Cubulating hyperbolic free-by-cyclic groups: the irreducible case*, Duke Math. J. 165 (2016) 1753–1813 MR Zbl

[MR08] **M Mj**, **L Reeves**, *A combination theorem for strong relative hyperbolicity*, Geom. Topol. 12 (2008) 1777–1798 MR Zbl

[NR97] **G Niblo**, **L Reeves**, *Groups acting on* CAT(0) *cube complexes*, Geom. Topol. 1 (1997) 1–7 MR Zbl

[Osi06] **D V Osin**, *Relatively hyperbolic groups: intrinsic geometry, algebraic properties, and algorithmic problems*, Mem. Amer. Math. Soc. 843, Amer. Math. Soc., Providence, RI (2006) MR Zbl

[Sag95] **M Sageev**, *Ends of group pairs and non-positively curved cube complexes*, Proc. Lond. Math. Soc. 71 (1995) 585–617 MR Zbl

Institut Fourier, University Grenoble Alpes
Grenoble, France

Department of Mathematics, Ashoka University
Haryana, India

francois.dahmani@univ-grenoble-alpes.fr, suraj.meda@ashoka.edu.in

Received: 9 April 2023 Revised: 1 December 2023

Algebraic & Geometric Topology **25**:3 (2025) 1599–1666
DOI: 10.2140/agt.2025.25.1599
Published: 20 June 2025

Virtual domination of 3-manifolds, III

HONGBIN SUN

We prove that for any oriented cusped hyperbolic 3-manifold M and any compact oriented 3-manifold N with tori boundary, there exists a finite cover M' of M that admits a degree-8 map $f\colon M' \to N$, ie M virtually 8-dominates N.

57K32, 57M10; 30F40

1 Introduction

We assume all manifolds are compact, connected and oriented, unless otherwise indicated. By a cusped hyperbolic 3-manifold, we mean a compact 3-manifold with nonempty tori boundary, such that its interior admits a complete hyperbolic structure with finite volume, unless otherwise indicated.

For two closed oriented n-manifolds M and N, and a map $f\colon M \to N$, a natural quantity associated to f is its mapping degree. The mapping degree of f is $d \in \mathbb{Z}$ if $f_*([M]) = d[N]$ for oriented fundamental classes $[M] \in H_n(M;\mathbb{Z})$ and $[N] \in H_n(N;\mathbb{Z})$. The notion of mapping degree can be generalized to proper maps between manifolds with boundary. For two compact oriented n-manifolds M and N with boundary, a map $f\colon M \to N$ is *proper* if $f^{-1}(\partial N) = \partial M$. The mapping degree of a proper map $f\colon M \to N$ is $d \in \mathbb{Z}$ if $f_*([M,\partial M]) = d[N,\partial N]$ for oriented relative fundamental classes $[M,\partial M] \in H_n(M,\partial M;\mathbb{Z})$ and $[N,\partial N] \in H_n(N,\partial N;\mathbb{Z})$. In either of the above cases, f is a *nonzero degree map* if the degree of f is not zero. If the mapping degree $d \neq 0$, we say that M *d-dominates* N, and we say that M *dominates* N if M d-dominates N for some nonzero integer d. In this paper, we will work on 3-manifolds with nonempty boundary, and all maps $f\colon M \to N$ between such 3-manifolds are proper, unless otherwise indicated.

Roughly speaking, if M dominates (or 1-dominates) N, then M is topologically more complicated than N. For certain invariants of manifolds, eg ranks of fundamental groups, Betti numbers, simplicial volumes, representation volumes, etc, this impression on behavior of topological invariants under nonzero degree maps (or degree-1 maps) forms classical results. However, for some other invariants, eg Heegaard genera and Heegaard Floer homology of 3-manifolds, it is unknown whether the above impression is correct.

Carlson and Toledo [7] asked whether there is an easily described class $\mathscr{C}$ of closed oriented n-manifolds, such that any closed oriented n-manifold is dominated by some $M \in \mathscr{C}$. Gaifullin [9] proved that, for any positive integer n, there exists a closed oriented n-manifold M_0, such that any closed oriented n-manifold is dominated by a finite cover of M_0 (virtually dominated by M_0), ie we can take $\mathscr{C}$ to be the set of all finite covers of M_0. In [14; 19; 23], the author and Liu proved the following result.

Theorem 1.1 [14; 19; 23] *For any closed oriented* 3*-manifold* M *with positive simplicial volume and any closed oriented* 3*-manifold* N, *there exists a finite cover* M' *of* M *that admits a degree-*1 *map* $f\colon M' \to N$.

So for any closed oriented 3-manifold M with positive simplicial volume, we can take $\mathscr{C}$ to be the set of all finite covers of M.

Note the condition that M has positive simplicial volume is necessary for Theorem 1.1, since a manifold with zero simplicial volume does not dominate any manifold with positive simplicial volume, and the simplicial volume has the covering property.

In this paper, we generalize the above virtual domination result from closed 3-manifolds to 3-manifolds with tori boundary. The following theorem is the main result of this paper.

Theorem 1.2 *For any oriented cusped hyperbolic* 3*-manifold* M *and any compact oriented* 3*-manifold* N *with nonempty tori boundary*, *there exists a finite cover* M' *of* M *that admits a proper map* $f\colon M' \to N$ *with* $\deg(f) = 8$.

The proof of Theorem 1.2 can also be applied to prove a similar result on certain mixed 3-manifolds. Here a mixed 3-manifold is a compact oriented irreducible 3-manifold with empty or tori boundary, such that it has nontrivial JSJ decomposition and at least one hyperbolic JSJ piece.

Theorem 1.3 *For any compact oriented mixed* 3*-manifold* M *with nonempty tori boundary such that a hyperbolic piece of* M *intersects with* ∂M, *and any compact oriented* 3*-manifold* N *with nonempty tori boundary*, *there exists a finite cover* M' *of* M *that admits a proper map* $f\colon M' \to N$ *with* $\deg(f) = 8$.

We cannot prove virtual 1-domination for Theorems 1.2 and 1.3. Although we can prove virtual 1-, 2-, or 4-domination in certain special cases, we do need to state our result as virtual 8-domination.

Moreover, if M d-dominates N, then M (kd)-dominates N for any positive integer k, by taking a degree-k cyclic cover of M (since M has nonempty tori boundary). In some sense, the above degree-(kd) map is not significantly different from the degree-d map, and one may only be interested in (virtually) π_1-surjective domination maps. In fact, the virtual 8-domination maps in Theorems 1.2 and 1.3 can be π_1-surjective. The virtual 2-domination map in Proposition 4.5 is always π_1-surjective. For Theorem 5.1, the statement only gives a virtual domination of degree 1, 2 or 4, because we did not work hard enough to make the 2-complex $\mathscr{X}$ connected. Actually, we can work harder to make $\mathscr{X}$ connected and obtain a π_1-surjective virtual 4-domination.

For technical reasons, we cannot prove Theorem 1.3 for other mixed 3-manifolds with tori boundary, although we do expect the virtual domination result still holds in that case. To fully resolve this problem, it remains to study mixed 3-manifolds such that all of their boundary components are contained in Seifert pieces.

Question 1.4 Let M be a compact oriented 3-manifold with nonempty tori boundary and positive simplicial volume. Does M virtually (1-)dominate all compact oriented 3-manifolds with tori boundary?

For a statement as Theorem 1.2, we do not have to restrict to compact oriented 3-manifolds with tori boundary, and we ask what happens for all compact oriented 3-manifolds with nonempty (possibly higher genus) boundary.

Question 1.5 Which compact oriented 3-manifold M with boundary virtually dominates all compact oriented 3-manifolds with boundary?

Two necessary conditions for Question 1.5 are: M has a boundary component of genus at least 2, and the double of M has positive simplicial volume. If the boundary of M only consists of 2-spheres and tori, so does any finite cover M' of M. Then M' does not dominate any 3-manifold with higher genus boundary, by considering the restriction map on the boundary. Moreover, if M virtually dominates N and both manifolds have boundary, then $D(M)$ virtually dominates $D(N)$. Since we can choose N so that $D(N)$ has positive simplicial volume, then so does $D(M)$.

Before we sketch the proof of Theorem 1.2, let's first recall the proof of virtual domination results (Theorem 1.1) of closed 3-manifolds in [14; 19; 23]. These three proofs roughly follow the same circle of ideas, and we sketch the proof of the most general result in [23] here. First, by Boileau and Wang [4], we can assume the target manifold N is a closed hyperbolic 3-manifold, and we take a geometric triangulation of N. Since M has positive simplicial volume, let M_0 be a hyperbolic JSJ piece of a prime summand of M. Then we construct a map $j^1\colon N^{(1)} \to M_0$ from the 1-skeleton $N^{(1)}$ of N to M_0, such that j^1 maps the boundary of each triangle Δ in N to a null-homologous closed curve in M_0. For each triangle Δ in N, we construct a compact orientable surface S_Δ with connected boundary and a map $S_\Delta \looparrowright M_0$ that maps ∂S_Δ to $j^1(\partial\Delta)$, so that S_Δ is mapped to a nearly geodesic subsurface in M_0. Then the maps $j^1\colon N^{(1)} \to M_0$ and $\{S_\Delta \looparrowright M_0\}$ together give a map $j\colon Z \looparrowright M_0$ from a 2-complex Z to M_0. If we construct the maps $\{S_\Delta \looparrowright M_0\}$ carefully enough, $j\colon Z \looparrowright M_0$ induces an injective homomorphism on π_1. Since $j_*(\pi_1(Z)) < \pi_1(M_0) < \pi_1(M)$ is a separable subgroup in $\pi_1(M)$ (by [21], which generalizes Agol's celebrated result on LERF-ness of hyperbolic 3-manifold groups in [2]), the map $j\colon Z \looparrowright M$ lifts to an embedding $j'\colon Z \hookrightarrow M'$ into a finite cover M' of M. A neighborhood of Z in M' is a compact oriented 3-manifold $\mathscr{Z}$ with boundary, and is homeomorphic to the manifold obtained from a neighborhood $\mathcal{N}(N^{(2)})$ of $N^{(2)}$ in N, by replacing each $\Delta \times I$ by $S_\Delta \times I$. Then there is a proper degree-1 map $g\colon \mathscr{Z} \to \mathcal{N}(N^{(2)})$ that maps each $S_\Delta \times I \subset \mathscr{Z}$ to $\Delta \times I \subset \mathcal{N}(N^{(2)})$. This proper degree-1 map $g\colon \mathscr{Z} \to \mathcal{N}(N^{(2)})$ extends to a degree-1 map $f\colon M' \to N$, by mapping each component of $M' \setminus \mathscr{Z}$ to the union of some components of $N \setminus \mathcal{N}(N^{(2)})$ (each component is a 3-ball) and a finite graph in N.

In the context of manifolds with boundary, the above proof fails in the last step, but we need to fix it from the very first step. For example, if we apply the above approach to manifolds with boundary, it is possible that some component C of $M' \setminus \mathscr{Z}$ does not intersect with $\partial M'$, but a component of $N \setminus \mathcal{N}(N^{(2)})$ intersecting $g(\partial C)$ may contain some component of ∂N. In this case $g\colon \mathscr{Z} \to \mathcal{N}(N^{(2)})$ does not extend to a proper map $f\colon M' \to N$. Moreover, even if each component C of $M' \setminus \mathscr{Z}$ intersects with $\partial M'$, it is also difficult to construct the desired extension $f\colon M' \to N$. So we need to take a more careful construction for proving Theorem 1.2, which is sketched in the following.

In Section 4, we reduce the proof of Theorem 1.2 to 3-manifolds M and N satisfying the following extra assumptions.

- M has two components T_1 and T_2, such that the kernel of $H_1(T_1 \cup T_2; \mathbb{Z}) \to H_1(M; \mathbb{Z})$ contains an element with nontrivial components in both $H_1(T_1; \mathbb{Z})$ and $H_1(T_2; \mathbb{Z})$.
- N is a finite volume hyperbolic 3-manifold with a single cusp.

In Section 5.1, we take a geometric cellulation of a compact core N_0 of N which has extra edges than a geometric triangulation, such that each triangle contained in ∂N_0 is almost an equilateral triangle. In Section 5.2, we construct two maps $j_s^{(1)}\colon N^{(1)} \to M$ for $s = 1, 2$ such that the following hold:

(1) For each triangle Δ of N_0 contained in ∂N_0, $j_s^{(1)}(\partial\Delta)$ bounds a geodesic triangle in M.

(2) For each $s = 1, 2$, the union of geodesic triangles in M bounded by $j_s^{(1)}(\partial\Delta)$ in item (1) gives a mapped-in torus $T \to M$ homotopic into T_s.

(3) For each triangle or bigon Δ of N_0 not contained in ∂N_0, $j_1^{(1)}(\partial\Delta) \cup j_2^{(1)}(\partial\Delta)$ is null-homologous in M.

For each triangle Δ of N_0 as in item (3), we construct a compact orientable surface S_Δ and a nearly geodesic immersion $S_\Delta \looparrowright M$ bounded by two copies of $j_1^{(1)}(\partial\Delta) \cup j_2^{(1)}(\partial\Delta)$. Then two copies of $j_s^{(1)}\colon N^{(1)} \to M$ with $s = 1, 2$, two copies of the tori in item (2) and the maps $\{S_\Delta \looparrowright M\}$ together give a 2-complex Z and a map $j\colon Z \looparrowright M$. In Section 6, we prove that if the construction is done carefully, $j\colon Z \looparrowright M$ is π_1-injective. After this step, the construction of the virtual domination (proper) map is similar to the closed manifold case. We first use Agol's result [2] that $j_*(\pi_1(Z)) < \pi_1(M)$ is a separable subgroup to lift Z to an embedded 2-complex in a finite cover M' of M, and take a neighborhood of Z in M' denoted by $\mathscr{Z}$. Then we have a proper degree-4 map $g\colon \mathscr{Z} \to \mathcal{N}(N^{(2)})$, such that the following holds.

- For the component $T' = \partial N_0$ of $\partial\mathcal{N}(N^{(2)})$, each component of $g^{-1}(T')$ is a torus in M' parallel to a component of $\partial M'$.

This key property implies that g can be extended to a proper degree-4 map $f\colon M' \to N$, as desired (see Section 5.3).

Note that the π_1-injectivity of $j\colon Z \looparrowright M$ cannot be proved by exactly the same way as in [14; 19; 23]. In [14; 19; 23], we equipped Z with a natural metric and proved that the map $\tilde{j}\colon \widetilde{Z} \to \widetilde{M} = \mathbb{H}^3$ on universal covers is a quasi-isometric embedding. However, in the current case, $\tilde{j}\colon \widetilde{Z} \to \widetilde{M}$ is not a quasi-isometric embedding anymore, since $j(Z)$ contains some tori homotopic into ∂M. To prove the π_1-injectivity of j, we modify Z as follows. For each torus T in Z as in item (2) above (that is homotopic into a horotorus in M), we add the cone of T to Z with the cone point deleted, and get an ideal 3-complex Z^3 (a 3-complex with certain vertices deleted). The map $j\colon Z \looparrowright M$ extends to a map $j_1\colon Z^3 \looparrowright M$ that maps ideal vertices of Z^3 to corresponding ends of M. In Section 6, we prove the π_1-injectivity of $j\colon Z \looparrowright M$ by proving that $\tilde{j}_1\colon \widetilde{Z}^3 \to \widetilde{M} = \mathbb{H}^3$ is a quasi-isometric embedding.

Although the above description of $j\colon Z \looparrowright M$ is mostly topological, we actually need geometric methods to construct it. Our main geometric tool for constructing various geometric objects is the good pants construction. Roughly speaking, the good pants construction is a tool box that uses so called good curves, good pants and other good objects to construct geometrically nice objects in hyperbolic 3-manifolds. The good pants construction was initiated by Kahn and Markovic [10], for constructing nearly geodesic π_1-injective immersed closed subsurfaces in closed hyperbolic 3-manifolds, with good pants as building blocks. Then in [11], Kahn and Wright generalized Kahn and Markovic's work to construct nearly geodesic π_1-injective immersed closed subsurfaces in cusped hyperbolic 3-manifolds. These geometrically nice subsurfaces of Kahn–Wright are basic pieces for constructing our 2-complex $j\colon Z \looparrowright M$ in cusped hyperbolic 3-manifolds. More details on the good pants construction can be found in Section 2.

Now we summarize the organization of this paper. In Section 2, we review the good pants construction in closed and cusped hyperbolic 3-manifolds, including works in [10; 11; 13; 22]. In Section 3, we review and prove some elementary geometric estimates in hyperbolic geometry. In Section 4, we prove preparational results that reduce the domain and target manifolds in Theorems 1.2 and 1.3. The technical heart of this paper is in Sections 5 and 6. In Section 5, we construct the mapped-in 2-complex $j\colon Z \looparrowright M$ and the virtual domination map from M to N, modulo the π_1-injectivity of $j\colon Z \looparrowright M$ (Theorem 5.17). The π_1-injectivity of j will be proved in Section 6.

Acknowledgements The author thanks the referee for very helpful suggestions for improving this paper. The author is partially supported by Simons Collaboration Grant 615229.

2 Preliminaries on the good pants construction

In this section, we review the good pants construction on finite-volume hyperbolic 3-manifolds, including constructions of nearly geodesic subsurfaces [10; 11], works on panted cobordism groups [13; 22] and the connection principle of cusped hyperbolic 3-manifolds [22].

2.1 Constructing nearly geodesic subsurfaces in finite-volume hyperbolic 3-manifolds

In [10], Kahn and Markovic proved the following surface subgroup theorem. This work initiates the development of the good pants construction, and it was the first step of Agol's proof of Thurston's virtual Haken and virtual fibering conjectures [2].

Theorem 2.1 (surface subgroup theorem [10]) *For any closed hyperbolic 3-manifold M, there exists an immersed closed hyperbolic subsurface $f\colon S \looparrowright M$, such that $f_*\colon \pi_1(S) \to \pi_1(M)$ is injective.*

The immersed subsurface of Kahn and Markovic is geometrically nice, and it is built by pasting a large collection of (R,ϵ)-*good pants* along (R,ϵ)-*good curves* in a nearly geodesic way. These terminologies are summarized in the following.

We fix a closed oriented hyperbolic 3-manifold M, a small number $\epsilon > 0$ and a large number $R > 0$.

Definition 2.2 An (R,ϵ)-*good curve* is an oriented closed geodesic in M with complex length satisfying $|\boldsymbol{l}(\gamma) - 2R| < 2\epsilon$. The (finite) set consisting of all such (R,ϵ)-good curves is denoted by $\boldsymbol{\Gamma}_{R,\epsilon}$.

Here the complex length of γ is defined by $\boldsymbol{l}(\gamma) = l + i\theta \in \mathbb{C}/2\pi i\mathbb{Z}$, where $l \in \mathbb{R}_{>0}$ is the length of γ, and $\theta \in \mathbb{R}/2\pi\mathbb{Z}$ is the rotation angle of the loxodromic isometry of $\mathbb{H}^3$ corresponding to γ. In this paper, we adopt the convention in [11] that good curves have length close to $2R$, instead of the convention in [10] that good curves have length close to R.

Definition 2.3 We use $\Sigma_{0,3}$ to denote the oriented topological pair of pants. A pair of (R,ϵ)-*good pants* is a homotopy class of immersion $\Sigma_{0,3} \looparrowright M$, denoted by Π, such that all three cuffs of $\Sigma_{0,3}$ are mapped to (R,ϵ)-good curves $\gamma_1, \gamma_2, \gamma_3 \in \boldsymbol{\Gamma}_{R,\epsilon}$, and the complex half length $\mathbf{hl}_\Pi(\gamma_i)$ of each γ_i with respect to Π satisfies

$$|\mathbf{hl}_\Pi(\gamma_i) - R| < \epsilon.$$

We use $\boldsymbol{\Pi}_{R,\epsilon}$ to denote the finite set of all (R,ϵ)-good pants.

Here the complex half length $\mathbf{hl}_\Pi(\gamma_i)$ measures the complex distance between two vectors $\vec{v}_{i-1}$ and $\vec{v}_{i+1}$ along γ_i, where $\vec{v}_{i-1}$ and $\vec{v}_{i+1}$ are tangent vectors of oriented common perpendicular segments (seams) from γ_i to γ_{i-1} and γ_{i+1} respectively. See [10, Section 2.1] for the precise definition of complex half length. If $\gamma \in \boldsymbol{\Gamma}_{R,\epsilon}$ is a cuff of $\Pi \in \boldsymbol{\Pi}_{R,\epsilon}$, then $\mathbf{hl}_\Pi(\gamma)$ is uniquely determined by $\boldsymbol{l}(\gamma)$, and we denote this value by $\mathbf{hl}(\gamma)$ if no confusion is caused.

For $\gamma \in \boldsymbol{\Gamma}_{R,\epsilon}$, we can identify its unit normal bundle as $N^1(\gamma) = \mathbb{C}/(\boldsymbol{l}(\gamma)\mathbb{Z} + 2\pi i\mathbb{Z})$, then its half-unit normal bundle is defined to be

$$N^1(\sqrt{\gamma}) = \mathbb{C}/(\mathbf{hl}(\gamma)\mathbb{Z} + 2\pi i\mathbb{Z}).$$

Given $\Pi \in \boldsymbol{\Pi}_{R,\epsilon}$ with one cuff $\gamma = \gamma_i$, the pair of normal vectors $\vec{v}_{i-1}$ and $\vec{v}_{i+1}$ used to define $\mathbf{hl}_\Pi(\gamma)$ gives a unique vector $\mathbf{foot}_\gamma(\Pi) \in N^1(\sqrt{\gamma})$, called the *formal foot* of Π on γ.

In [10], to obtain the nearly geodesic subsurface, (R,ϵ)-good pants are pasted along (R,ϵ)-good curves with nearly 1-shifts, rather than exactly matching seams along common cuffs. More precisely, in the nearly geodesic subsurface $S \looparrowright M$, for any two (R,ϵ)-good pants $\Pi_1 \in \mathbf{\Pi}_{R,\epsilon}$ and $\Pi_2 \in \mathbf{\Pi}_{R,\epsilon}$ in S pasted along $\gamma \in \mathbf{\Gamma}_{R,\epsilon}$, such that γ is an oriented boundary of Π_1, after identifying $N^1(\sqrt{\gamma})$ with $N^1(\sqrt{\bar{\gamma}})$ naturally, it is required that

$$|\mathbf{foot}_\gamma(\Pi_1) - \mathbf{foot}_{\bar{\gamma}}(\Pi_2) - (1+\pi i)| < \frac{\epsilon}{R} \quad \text{in } N^1(\sqrt{\gamma}).$$

This nearly 1-shift is a crucial condition to guarantee the injectivity of $f_*\colon \pi_1(S) \to \pi_1(M)$. Kahn and Markovic showed that, for any (R,ϵ)-good curve γ, the formal feet of (R,ϵ)-good pants with cuff γ are nearly evenly distributed along γ. So M contains a large collection of (R,ϵ)-good pants, and they can be pasted together by nearly 1-shifts. Therefore, the asserted π_1-injective immersed closed subsurface can be constructed.

In [11], Kahn and Wright generalized Kahn and Markovic's surface subgroup theorem in closed hyperbolic 3-manifolds (Theorem 2.1) to cusped hyperbolic 3-manifolds.

Theorem 2.4 [11, Theorem 1.1] *Let $\Gamma < \mathrm{PSL}_2(\mathbb{C})$ be a Kleinian group and assume that $\mathbb{H}^3/\Gamma$ has finite volume and is not compact. Then for all $K > 1$, there exist K-quasi-Fuchsian* (*closed*) *surface subgroups in Γ.*

The main difficulty for proving Theorem 2.4 is that, for cusped hyperbolic 3-manifolds, good pants are not evenly distributed along good curves, especially for those good curves that run into cusps very deeply (with high heights).

We first define the height function on a cusped hyperbolic 3-manifold M. By the Margulis lemma, there exists $\epsilon_0 > 0$ such that the subset of M consisting of points of injectivity radii at most ϵ_0 is a disjoint union of solid tori and cusp neighborhoods of ends (simply called cusps). For any point in M not belonging to any cusps, we define its height to be 0. For any point p in a cusp $C \subset M$, we define the height of p to be the distance between p and the boundary of C. For a compact geodesic segment or a closed geodesic in M, we define its *height* to be the maximal height of points on it. For a pair of (R,ϵ)-good pants in M, we define its height to be the maximal height of its three cuffs.

For any $h > 0$, we use $\mathbf{\Gamma}_{R,\epsilon}^{<h}$ (resp. $\mathbf{\Pi}_{R,\epsilon}^{<h}$) to denote the set of all (R,ϵ)-good curves (resp. the set of all (R,ϵ)-good pants) in M with height less than h. We can define $\mathbf{\Gamma}_{R,\epsilon}^{\geq h}$ and $\mathbf{\Pi}_{R,\epsilon}^{\geq h}$ similarly.

To construct nearly geodesic subsurfaces in cusped hyperbolic 3-manifolds, Kahn and Wright introduced a new geometric object called (R,ϵ)*-good hamster wheel*. For a positive integer R, let Q_R be the oriented hyperbolic pants with cuff lengths 2, 2 and $2R$. The *R-perfect hamster wheel H_R* is the cyclic R-sheet regular cover of Q_R with $R+2$ boundary components, such that all cuffs of H_R have length $2R$. An (R,ϵ)*-good hamster wheel* (or simply an (R,ϵ)-hamster wheel) $\boldsymbol{H}$ is a map $f\colon H_R \to M$ up to homotopy, such that the image of each cuff of H_R lies in $\mathbf{\Gamma}_{R,\epsilon}$, and f is approximately a totally geodesic immersion. For each (R,ϵ)-good hamster wheel $\boldsymbol{H}$ and each cuff $\gamma \in \mathbf{\Gamma}_{R,\epsilon}$ of $\boldsymbol{H}$, a foot $\mathbf{foot}_\gamma(\Pi) \in N^1(\sqrt{\gamma})$ can be defined that approximates the tangent direction of $\boldsymbol{H}$. See [11, Section 2.9] for the precise definition of (R,ϵ)-hamster wheels and their feet.

In [11], Kahn and Wright defined the (R,ϵ)*-well-matched condition* for pasting finitely many (R,ϵ)-good pants and (R,ϵ)-good hamster wheels together in a nearly geodesic manner. A *good assembly* in a cusped hyperbolic 3-manifold is a compact oriented subsurface (possibly with boundary) obtained by pasting finitely many (R,ϵ)-good pants and (R,ϵ)-good hamster wheels according to the (R,ϵ)-well-matched condition. Then Kahn and Wright proved that an immersed subsurface in a cusped hyperbolic 3-manifold arising from a good assembly is π_1-injective.

To construct a closed subsurface in a cusped hyperbolic 3-manifold arising from a good assembly, Kahn and Wright defined a more complicated geometric object called an *umbrella*. An umbrella $\boldsymbol{U}$ consists of a compact planar surface U decomposed as a finite union of subsurfaces homeomorphic to $H_{\boldsymbol{R}}$ and a map $f: U \to M$, such that the restriction of f on each $H_{\boldsymbol{R}}$ subsurface (under the decomposition) gives an (R,ϵ)-good hamster wheel, and these (R,ϵ)-good hamster wheels are (R,ϵ)-well-matched with each other. For each umbrella $\boldsymbol{U}$ and each cuff $\gamma \in \boldsymbol{\Gamma}_{R,\epsilon}$ of $\boldsymbol{U}$, we define $\mathbf{foot}_{\gamma}(\boldsymbol{U}) \in N^1(\sqrt{\gamma})$ to be the foot of the (R,ϵ)-hamster wheel in $\boldsymbol{U}$ containing γ. Umbrellas are used to take care of the undesired property that feet of good pants are not evenly distributed on $N^1(\sqrt{\gamma})$ for some $\gamma \in \boldsymbol{\Gamma}_{R,\epsilon}$, especially when γ has high height.

In [11], Kahn and Wright took constants $h_T \geq 6\log R$ and $h_c \geq h_T + 44\log R$. Then they considered the collection of all (R,ϵ)-good pants Π with at least one cuff of height less than h_c. For any (Π,γ) such that $\Pi \in \boldsymbol{\Pi}_{R,\epsilon}^{\geq h_c}$ and $\gamma \in \boldsymbol{\Gamma}_{R,\epsilon}^{<h_c}$ is a cuff of Π, in [11, Theorem 4.15], Kahn and Wright constructed a $\mathbb{Q}_+$-combination of umbrellas $\widehat{U}(\Pi,\gamma)$ with coefficients sum to 1 such that the following hold:

(1) As a $\mathbb{Q}_+$-linear combination of umbrellas, the boundary of $\widehat{U}(\Pi,\gamma)$ contains one copy of γ, and all of its other boundary components have height less than h_T.

(2) $\widehat{U}(\Pi,\gamma)$ is (R,ϵ)-well-matched with any (R,ϵ)-good pants that is (R,ϵ)-well-matched with Π along γ.

Then they used $\widehat{U}(\Pi,\gamma)$ to replace Π in the above collection of good pants.

After the above replacement process, we obtain two finite linear combinations of (R,ϵ)-good objects. The first one is the sum of all (R,ϵ)-good pants in $\boldsymbol{\Pi}_{R,\epsilon}^{<h_c}$, and the second one is the sum of $\mathbb{Q}_+$-linear combinations of umbrellas constructed above:

$$A_0 = \sum_{\Pi \in \boldsymbol{\Pi}_{R,\epsilon}^{<h_c}} \Pi, \quad A_1 = \sum_{\gamma \in \boldsymbol{\Gamma}_{R,\epsilon}^{<h_c}} \sum_{\Pi \in \boldsymbol{\Pi}_{R,\epsilon}^{\geq h_c}, \gamma \subset \partial\Pi} \widehat{U}(\Pi,\gamma).$$

Then Kahn and Wright proved that, for any $\gamma \in \boldsymbol{\Gamma}_{R,\epsilon}^{<h_c}$, the feet of (R,ϵ)-pants and umbrellas in $A = A_0 + A_1$ are evenly distributed on $N^1(\sqrt{\gamma})$. After eliminating denominators in A by multiplying a large integer, they could paste good pants and umbrellas in $A \cup \overline{A}$ ($\overline{A}$ denotes the orientation reversal of A) to get the desired nearly geodesic closed subsurface.

2.2 Panted cobordism groups of finite volume hyperbolic 3-manifolds

In [13], Liu and Markovic introduced panted cobordism groups of closed oriented hyperbolic 3-manifolds and computed these groups. In [22], the author generalized some results in [13] to oriented cusped hyperbolic 3-manifolds. In this section, we review these results and their consequence Proposition 2.11, which is the main input from the good pants construction to this work.

We first fix a closed oriented hyperbolic 3-manifold M, a small number $\epsilon > 0$ and a large number $R > 0$. Let $\mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}$ be the free abelian group generated by $\mathbf{\Gamma}_{R,\epsilon}$, modulo the relation $\gamma + \bar{\gamma} = 0$ for all $\gamma \in \mathbf{\Gamma}_{R,\epsilon}$. Here $\bar{\gamma}$ denotes the orientation reversal of γ. Let $\mathbb{Z}\mathbf{\Pi}_{R,\epsilon}$ be the free abelian group generated by $\mathbf{\Pi}_{R,\epsilon}$, modulo the relation $\Pi + \overline{\Pi} = 0$ for all $\Pi \in \mathbf{\Pi}_{R,\epsilon}$. By taking the oriented boundary of (R,ϵ)-good pants, we get a homomorphism $\partial\colon \mathbb{Z}\mathbf{\Pi}_{R,\epsilon} \to \mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}$. The panted cobordism group $\Omega_{R,\epsilon}(M)$ is defined as the following in [13].

Definition 2.5 The *panted cobordism group* $\Omega_{R,\epsilon}(M)$ is defined to be the cokernel of the homomorphism ∂, ie $\Omega_{R,\epsilon}(M)$ fits into the exact sequence

$$\mathbb{Z}\mathbf{\Pi}_{R,\epsilon} \xrightarrow{\partial} \mathbb{Z}\mathbf{\Gamma}_{R,\epsilon} \to \Omega_{R,\epsilon}(M) \to 0.$$

To state the result in [13], we need the following definition.

Definition 2.6 For an oriented hyperbolic 3-manifold M and a point $p \in M$, a *special orthonormal frame* (or simply a frame) of M at p is a triple of unit tangent vectors $(\vec{t}_p, \vec{n}_p, \vec{t}_p \times \vec{n}_p)$ such that $\vec{t}_p, \vec{n}_p \in T_p^1 M$ with $\vec{t}_p \perp \vec{n}_p$, and $\vec{t}_p \times \vec{n}_p \in T_p^1 M$ is the cross product with respect to the orientation of M. We use $\mathrm{SO}(M)$ to denote the *frame bundle* of M consisting of all special orthonormal frames of M.

For simplicity, we denote each element in $\mathrm{SO}(M)$ by its basepoint and the first two vectors of the frame, as $(p, \vec{t}_p, \vec{n}_p)$, since the third vector is determined by the first two. We call $\vec{t}_p$ and $\vec{n}_p$ the tangent vector and the normal vector of this frame, respectively.

In [13], Liu and Markovic proved the following result on $\Omega_{R,\epsilon}(M)$.

Theorem 2.7 [13, Theorem 5.2] *For any closed oriented hyperbolic* 3*-manifold* M*, small enough* $\epsilon > 0$ *depending on* M*, and large enough* $R > 0$ *depending on* ϵ *and* M*, there is a natural isomorphism*

$$\Phi\colon \Omega_{R,\epsilon}(M) \to H_1(\mathrm{SO}(M);\mathbb{Z}).$$

In [22], the author generalized Theorem 2.7 to oriented cusped hyperbolic 3-manifolds. The corresponding result in [22] has some height conditions on involved curves and pants, and we need the following definition.

For any $h' > h > 0$, $\mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<h}$ is naturally a subgroup of $\mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<h'}$. For the boundary homomorphism $\partial\colon \mathbb{Z}\mathbf{\Pi}_{R,\epsilon}^{<h'} \to \mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<h'}$, we use $\mathbb{Z}\mathbf{\Pi}_{R,\epsilon}^{h,h'}$ to denote the ∂-preimage of $\mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<h} < \mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<h'}$ in $\mathbb{Z}\mathbf{\Pi}_{R,\epsilon}^{<h'}$. We first recall the following definition in [22].

Definition 2.8 For an oriented cusped hyperbolic 3-manifold M and any $h' > h > 0$, we define the (R,ϵ)-*panted cobordism group of height* (h,h'), denoted by $\Omega_{R,\epsilon}^{h,h'}(M)$, to be the cokernel of the homomorphism $\partial|\colon \mathbb{Z}\mathbf{\Pi}_{R,\epsilon}^{h,h'} \to \mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<h}$. Thus $\Omega_{R,\epsilon}^{h,h'}(M)$ fits into the exact sequence

$$\mathbb{Z}\mathbf{\Pi}_{R,\epsilon}^{h,h'} \xrightarrow{\partial|} \mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<h} \to \Omega_{R,\epsilon}^{h,h'}(M) \to 0.$$

In [22], the author proved the following analogy of Theorem 2.7 for oriented cusped hyperbolic 3-manifolds.

Theorem 2.9 [22, Theorem 1.1] *For any oriented cusped hyperbolic 3-manifold M, any numbers $\beta > \alpha \geq 4$ with $\beta - \alpha \geq 3$ and any $\epsilon \in (0, 10^{-2})$, there exists $R_0 = R_0(M,\epsilon) > 0$, such that for any $R > R_0$, we have a natural isomorphism*

$$\Phi\colon \Omega_{R,\epsilon}^{\alpha \log R, \beta \log R}(M) \to H_1(\mathrm{SO}(M);\mathbb{Z}).$$

Moreover, for Theorem 2.9 (and Theorem 2.7), if we compose the isomorphism

$$\Phi\colon \Omega_{R,\epsilon}^{\alpha \log R, \beta \log R}(M) \to H_1(\mathrm{SO}(M);\mathbb{Z})$$

with the homomorphism $\pi_*\colon H_1(\mathrm{SO}(M);\mathbb{Z}) \to H_1(M;\mathbb{Z})$ induced by the bundle projection, $\pi_* \circ \Phi$ maps the equivalent class of each (R,ϵ)-multicurve to its homology class in $H_1(M;\mathbb{Z})$.

To give the geometric meaning of $\Omega_{R,\epsilon}^{h,h'}(M)$, we define the following two types of subsurfaces in an oriented cusped hyperbolic 3-manifold M.

Definition 2.10 For any small $\epsilon > 0$ and large number $R > 0$, we define the following terms.

(1) An (R,ϵ)-*panted subsurface* in a hyperbolic 3-manifold M consists of a (possibly disconnected) compact oriented surface F with a pants decomposition and an immersion $i\colon F \looparrowright M$, such that the restriction of j to each pair of pants in the pants decomposition of F gives a pair of (R,ϵ)-good pants.

(2) If R is also an integer, an (R,ϵ)-*nearly geodesic subsurface* in a cusped hyperbolic 3-manifold M consists of a compact oriented surface F decomposed as pants and R-hamster wheels (by a family of disjoint essential curves $\mathscr{C}$), and an immersion $i\colon F \looparrowright M$, such that the following hold. The restriction of i on each pants or R-hamster wheel subsurface of F is an (R,ϵ)-good pants or an (R,ϵ)-good hamster wheel respectively, and these (R,ϵ)-good components are pasted together by the (R,ϵ)-well-matched condition.

The (R,ϵ)-panted subsurface was originally defined by Liu and Markovic [13], and it does not require any feet-matching condition when two (R,ϵ)-good pants are pasted along an (R,ϵ)-good curve. Geometrically, an (R,ϵ)-multicurve $L \in \mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<h}$ represents the trivial element in $\Omega_{R,\epsilon}^{h,h'}(M)$ if and only if it bounds an (R,ϵ)-panted subsurface of height at most h'. The definition of an (R,ϵ)-nearly geodesic subsurface is same as an (R,ϵ)-good assembly in [11], but we stick to this terminology since we have been using it

throughout [14; 19; 23]. Theorem 2.2 of [11] implies that, if $\epsilon > 0$ is small enough and $R > 0$ is large enough, an (R,ϵ)-nearly geodesic subsurface is π_1-injective.

In [23, Proposition 3.11], the author proved the following result, which generalizes [19, Corollary 2.11].

Proposition 2.11 [23, Proposition 3.11] *Let M be an oriented cusped hyperbolic 3-manifold. Then for any constant $\alpha \geq 4$, any small $\epsilon > 0$ depending on M and any large real number $R > 0$ depending on M and ϵ, the following statement holds. For any null-homologous oriented (R,ϵ)-multicurve $L \in \mathbb{Z}\mathbf{\Gamma}_{R,\epsilon}^{<\alpha \log R}$, there is a nontrivial invariant $\sigma(L) \in \mathbb{Z}_2$ such that $\sigma(L_1 \cup L_2) = \sigma(L_1) + \sigma(L_2)$ and the following hold.*

If $\sigma(L) = 0$, for any integer $R' \geq R$, L is the oriented boundary of an immersed subsurface $f\colon S \looparrowright M$ satisfying the following conditions.

(1) *If we write L as a union of its components $L = L_1 \cup \cdots \cup L_k$, then S is decomposed as oriented subsurfaces $S = \left(\bigcup_{i=1}^{k} \Pi_i\right) \cup S'$ with disjoint interior, such that $\Pi_i \cap \partial S$ is a single curve c_i that is mapped to L_i.*

(2) *The restriction $f|_{\Pi_i}\colon \Pi_i \looparrowright M$ is a pair of pants such that $|\mathbf{hl}_{\Pi_i}(L_i) - R| < \epsilon$, and $|\mathbf{hl}_{\Pi_i}(s) - R'| < \epsilon$ holds for any other component $s \subset \partial \Pi_i$.*

(3) *If we fix a normal vector $\vec{v}_i \in N^1(\sqrt{L_i})$ for each component L_i of L, then we can make sure $|\mathbf{foot}_{L_i}(\Pi_i) - \vec{v}_i| < \epsilon$ holds for all i.*

(4) *The restriction $f|_{S'}\colon S' \looparrowright M$ is an oriented (R',ϵ)-nearly geodesic subsurface.*

(5) *For any component $s \subset S' \cap \Pi_i$ that is mapped to $\gamma \in \mathbf{\Gamma}_{R',\epsilon}$, we take its orientation induced from Π_i, then we have*
$$|\mathbf{foot}_{\gamma}(\Pi_i) - \mathbf{foot}_{\bar{\gamma}}(S') - (1 + \pi i)| < \frac{\epsilon}{R}.$$

We call the immersed pants in condition (2) (R, R', ϵ)*-good pants*, and we call the immersed subsurface $f\colon S \looparrowright M$ constructed in Proposition 2.11 *an (R',ϵ)-nearly geodesic subsurface with (R,ϵ)-good boundary*. We can also assume that S has no closed component.

Remark 2.12 For the immersed subsurface $S \looparrowright M$ constructed in Proposition 2.11, the collection of curves $\mathscr{C} \subset S$ (giving the decomposition of S) gives a graph-of-space structure on S with dual graph Γ: each component of $S \setminus \mathscr{C}$ gives a vertex of Γ, and each component of $\mathscr{C}$ gives an edge of Γ. Let v_i be the vertex of Γ corresponding to $\Pi_i \subset S$. We can further modify the nearly geodesic subsurface $S \looparrowright M$ as in [20, Section 3.1, Step IV], such that the combinatorial length of any topological essential path in Γ from v_i to v_j (possibly $i = j$) is at least $R'e^{R'/2}$.

Moreover, if we endow S with a hyperbolic metric such that all ∂-curves of S have length $2R$ and all curves in $\mathscr{C}$ have length $2R'$. Since all seams (shortest geodesic segments between boundary components) have lengths at least $e^{-R'/2}$ and each geodesic segment in a pair of pants or a hamster wheel from a cuff to itself has length at least R, any proper essential path in S from L_i to L_j (possibly $i = j$) has length at least R.

2.3 The connection principle of finite-volume hyperbolic 3-manifolds

The connection principle is a fundamental tool that constructs geometric segments and ∂-framed segments in finite volume hyperbolic 3-manifolds. The idea of connection principle was initiated in [10], and the first officially stated connection principle is given in [13, Lemma 4.15]. In [22], the author proved a version of connection principle for oriented cusped hyperbolic 3-manifolds, which is the connection principle will be used in this paper. In [12; 23], connection principles with homological control in frame bundles are obtained for oriented closed and cusped hyperbolic 3-manifolds respectively.

At first, we recall the definition of oriented *∂-framed segments* and associated objects [13, Definition 4.1]. They are the geometric objects constructed by our connection principle.

Definition 2.13 An *oriented ∂-framed segment* in M is a triple

$$\mathfrak{s} = (s, \vec{n}_{\mathrm{ini}}, \vec{n}_{\mathrm{ter}})$$

such that s is an immersed oriented compact geodesic segment (simply called a geodesic segment), $\vec{n}_{\mathrm{ini}}$ and $\vec{n}_{\mathrm{ter}}$ are unit normal vectors of s at its initial and terminal points respectively.

We have the following objects associated to an oriented ∂-framed segment $\mathfrak{s}$:

- The *carrier segment* of $\mathfrak{s}$ is the (oriented) geodesic segment s, and the *height* of $\mathfrak{s}$ is the height of s.
- The *initial endpoint* $p_{\mathrm{ini}}(\mathfrak{s})$ and the *terminal endpoint* $p_{\mathrm{ter}}(\mathfrak{s})$ are the initial and terminal points of s respectively.
- The *initial framing* $\vec{n}_{\mathrm{ini}}(\mathfrak{s})$ and the *terminal framing* $\vec{n}_{\mathrm{ter}}(\mathfrak{s})$ are the unit normal vectors $\vec{n}_{\mathrm{ini}}$ and $\vec{n}_{\mathrm{ter}}$ respectively.
- The *initial direction* $\vec{t}_{\mathrm{ini}}(\mathfrak{s})$ and the *terminal direction* $\vec{t}_{\mathrm{ter}}(\mathfrak{s})$ are the unit tangent vectors in the direction of s at $p_{\mathrm{ini}}(\mathfrak{s})$ and $p_{\mathrm{ter}}(\mathfrak{s})$ respectively.
- The *initial frame* and the *terminal frame* of $\mathfrak{s}$ are $(p_{\mathrm{ini}}(\mathfrak{s}), \vec{t}_{\mathrm{ini}}(\mathfrak{s}), \vec{n}_{\mathrm{ini}}(\mathfrak{s}))$ and $(p_{\mathrm{ter}}(\mathfrak{s}), \vec{t}_{\mathrm{ter}}(\mathfrak{s}), \vec{n}_{\mathrm{ter}}(\mathfrak{s}))$ respectively.
- The *length* $l(\mathfrak{s}) \in (0, \infty)$ of $\mathfrak{s}$ is the length of its carrier s, the *phase* $\varphi(\mathfrak{s}) \in \mathbb{R}/2\pi\mathbb{Z}$ of $\mathfrak{s}$ is the angle from the parallel transport of $\vec{n}_{\mathrm{ini}}$ along s to $\vec{n}_{\mathrm{ter}}$.
- The *orientation reversal* of $\mathfrak{s} = (s, \vec{n}_{\mathrm{ini}}, \vec{n}_{\mathrm{ter}})$ is defined to be
$$\bar{\mathfrak{s}} = (\bar{s}, \vec{n}_{\mathrm{ter}}, \vec{n}_{\mathrm{ini}}).$$
- For any angle $\phi \in \mathbb{R}/2\pi\mathbb{Z}$, the frame rotation of $\mathfrak{s}$ by ϕ is defined to be
$$\mathfrak{s}(\phi) = \big(s, \cos\phi \cdot \vec{n}_{\mathrm{ini}} + \sin\phi \cdot (\vec{t}_{\mathrm{ini}} \times \vec{n}_{\mathrm{ini}}), \cos\phi \cdot \vec{n}_{\mathrm{ter}} + \sin\phi \cdot (\vec{t}_{\mathrm{ter}} \times \vec{n}_{\mathrm{ter}})\big).$$

Now we state the connection principle in [23, Theorem 3.7]. Since we do not need a homological statement in frame bundles, we only state a weaker version of condition (3) here.

Theorem 2.14 [23, Theorem 3.7] *Let M be an oriented cusped hyperbolic 3-manifold, and let*

$$\boldsymbol{p} = (p, \vec{t}_p, \vec{n}_p), \boldsymbol{q} = (q, \vec{t}_p, \vec{n}_p) \in \mathrm{SO}(M)$$

be two frames based at $p, q \in M$ respectively. Let $\xi \in H_1(M, \{p, q\}; \mathbb{Z})$ be a relative homology class with boundary $\partial\xi = [q] - [p]$.

Then for any $\delta \in (0, 10^{-2})$, there exists $T = T(M, \xi, \delta)$ depending on M, ξ and δ, such that for any $t > T$, there is a ∂-framed segment $\mathfrak{s}$ from p to q such that the following hold.

(1) *The heights of p and q are at most $\log t$, and the height of $\mathfrak{s}$ is at most $2 \log t$.*

(2) *The length and phase of $\mathfrak{s}$ are δ-close to t and 0 respectively. The initial and terminal frames of $\mathfrak{s}$ are δ-close to $\boldsymbol{p}$ and $\boldsymbol{q}$ respectively.*

(3) *The relative homology class of the carrier of $\mathfrak{s}$ equals $\xi \in H_1(M, \{p, q\}; \mathbb{Z})$.*

3 Preliminaries on hyperbolic geometry

In this section, we give some geometric estimates on ∂-framed segments and geodesic segments, by using elementary hyperbolic geometry. Most of these results can be found in [22, Section 3], while some of them were originally proved in [13]. We have a new result (Proposition 3.5) that estimates the length of a consecutive chain of geodesic segments (see definition below), where some involved geodesic segments can be short.

We first need a few geometric definitions on ∂-framed segments from [13, Section 4].

Definition 3.1 Let $0 < \delta < \frac{\pi}{3}$, $L > 0$ and $0 < \theta < \pi$ be three constants.

(1) Two oriented ∂-framed segments $\mathfrak{s}$ and $\mathfrak{s}'$ are *δ-consecutive* if the terminal point of $\mathfrak{s}$ is the initial point of $\mathfrak{s}'$, and the terminal framing of $\mathfrak{s}$ is δ-close to the initial framing of $\mathfrak{s}'$. The *bending angle* between $\mathfrak{s}$ and $\mathfrak{s}'$ is the angle between the terminal direction of $\mathfrak{s}$ and the initial direction of $\mathfrak{s}'$.

(2) A *δ-consecutive chain* of oriented ∂-framed segments is a finite sequence $\mathfrak{s}_1, \dots, \mathfrak{s}_m$ such that each $\mathfrak{s}_i$ is δ-consecutive to $\mathfrak{s}_{i+1}$ for $i = 1, \dots, m-1$. It is a *δ-consecutive cycle* if furthermore $\mathfrak{s}_m$ is δ-consecutive to $\mathfrak{s}_1$. A δ-consecutive chain or cycle is *(L, θ)-tame* if each $\mathfrak{s}_i$ has length at least $2L$ and each bending angle is at most θ.

(3) For an (L, θ)-tame δ-consecutive chain $\mathfrak{s}_1, \dots, \mathfrak{s}_m$, the *reduced concatenation*, denoted by $\mathfrak{s}_1 \cdots \mathfrak{s}_m$, is the oriented ∂-framed segment defined as the following. The carrier segment of $\mathfrak{s}_1 \cdots \mathfrak{s}_m$ is homotopic to the concatenation of carrier segments of $\mathfrak{s}_1, \dots, \mathfrak{s}_m$, with respect to endpoints. The initial and terminal framings of $\mathfrak{s}_1 \cdots \mathfrak{s}_m$ are the closest unit normal vectors to the initial framing of $\mathfrak{s}_1$ and the terminal framing of $\mathfrak{s}_m$ respectively.

(4) For an (L, θ)-tame δ-consecutive cycle $\mathfrak{s}_1, \dots, \mathfrak{s}_m$, the *reduced cyclic concatenation*, denoted by $[\mathfrak{s}_1 \cdots \mathfrak{s}_m]$, is the oriented closed geodesic freely homotopic to the cyclic concatenation of carrier segments of $\mathfrak{s}_1, \dots, \mathfrak{s}_m$, assuming it is not null-homotopic.

Without considering initial and terminal framings, we can also talk about the following terms on geodesic segments: consecutive geodesic segments and their bending angles, a consecutive chain and a consecutive cycle of geodesic segments and their (L,θ)-tameness, the reduced concatenation of a consecutive chain of geodesic segments, and the reduced cyclic concatenation of a consecutive cycle of geodesic segments.

The following lemma from [13] is very useful for estimating length and phase of a concatenation of oriented ∂-framed segments. The function $I(\,\cdot\,)$ is defined by $I(\theta) = 2\log\bigl(\sec\frac{1}{2}\theta\bigr)$.

Lemma 3.2 [13, Lemma 4.8] *Given positive constants δ, θ and L with $0<\theta<\pi$ and $L \geq I(\theta)+10\log 2$, the following statements hold in any oriented hyperbolic 3-manifold.*

(1) *If $\mathfrak{s}_1,\dots,\mathfrak{s}_m$ is an (L,θ)-tame δ-consecutive chain of oriented ∂-framed segments, denoting the bending angle between $\mathfrak{s}_i$ and $\mathfrak{s}_{i+1}$ by $\theta_i \in [0,\theta)$, then*

$$\left| l(\mathfrak{s}_1\cdots\mathfrak{s}_m) - \sum_{i=1}^{m} l(\mathfrak{s}_i) + \sum_{i=1}^{m-1} I(\theta_i) \right| < \frac{(m-1)e^{(-L+10\log 2)/2}\sin(\theta/2)}{L-\log 2}$$

and

$$\left| \varphi(\mathfrak{s}_1\cdots\mathfrak{s}_m) - \sum_{i=1}^{m} \varphi(\mathfrak{s}_i) \right| < (m-1)(\delta + e^{(-L+10\log 2)/2}\sin(\theta/2)),$$

where $|\cdot|$ on $\mathbb{R}/2\pi\mathbb{Z}$ is understood as the distance from zero valued in $[0,\pi]$.

(2) *If $\mathfrak{s}_1,\dots,\mathfrak{s}_m$ is an (L,θ)-tame δ-consecutive cycle of oriented ∂-framed segments, denoting the bending angle between $\mathfrak{s}_i$ and $\mathfrak{s}_{i+1}$ by $\theta_i \in [0,\theta)$ with $\mathfrak{s}_{m+1}$ equal to $\mathfrak{s}_1$ by convention, then*

$$\left| l([\mathfrak{s}_1\cdots\mathfrak{s}_m]) - \sum_{i=1}^{m} l(\mathfrak{s}_i) + \sum_{i=1}^{m} I(\theta_i) \right| < \frac{me^{(-L+10\log 2)/2}\sin(\theta/2)}{L-\log 2}$$

and

$$\left| \varphi([\mathfrak{s}_1\cdots\mathfrak{s}_m]) - \sum_{i=1}^{m} \varphi(\mathfrak{s}_i) \right| < m(\delta + e^{(-L+10\log 2)/2}\sin(\theta/2)),$$

where $|\cdot|$ on $\mathbb{R}/2\pi\mathbb{Z}$ is understood as the distance from zero valued in $[0,\pi]$.

For an (L,θ)-tame δ-consecutive chain of ∂-framed segments $\mathfrak{s}_1,\dots,\mathfrak{s}_m$, we need the following lemma in [23] to bound the difference between initial frames of $\mathfrak{s}_1$ and $\mathfrak{s}_1\cdots\mathfrak{s}_m$.

Lemma 3.3 [23, Lemma 3.4] *Let δ, θ and L be positive constants with $0<\theta<\pi$ and $L \geq I(\theta)+10\log 2$. If $\mathfrak{s}_1,\dots,\mathfrak{s}_m$ is an (L,θ)-tame δ-consecutive chain of oriented ∂-framed segments, then the distance between the initial frames of $\mathfrak{s}_1$ and $\mathfrak{s}_1\cdots\mathfrak{s}_m$ in $\mathrm{SO}(M)_{p_{\mathrm{ini}}(\mathfrak{s}_1)}$ is at most $8e^{-L}$.*

The following lemma in [22] bounds the distance between a δ-consecutive cycle of geodesic segments and the corresponding closed geodesic, which is useful for bounding heights of closed geodesics arising from geometric constructions.

Lemma 3.4 [22, Lemma 3.7] *Given positive constants θ and L with $0<\theta<\pi$ and $L\ge 4(I(\theta)+10\log 2)$, the following statement holds in any oriented hyperbolic 3-manifold. If $s_1,\dots,s_m$ is an (L,θ)-tame cycle of geodesic segments with $m\le L$, and the bending angle between s_i and s_{i+1} lies in $[0,\theta)$ for each i, with s_{m+1} equal to s_1 by convention, then the closed geodesic $[s_1\cdots s_m]$ lies in the 1-neighborhood of the union $\bigcup_{i=1}^m s_i$.*

The following result generalizes Lemma 3.2(1), which estimates the length of a consecutive chain of geodesic segments where some involved geodesic segments are short.

Proposition 3.5 *Given any positive constants θ and L with $0<\theta<\frac{\pi}{2}$ and*

$$L\ge \max\bigl\{12I(\pi-\theta)+80\log 2, 24\log 2-16\log\bigl(\tfrac{\pi}{2}-\theta\bigr)\bigr\},$$

the following statement holds in any hyperbolic 3-manifold. Let $s_1,\dots,s_m$ be a consecutive chain of geodesic segments such that one of the following hold for each $i=1,\dots,m-1$:

(1) *either both s_i and s_{i+1} have length at least L, and the bending angle between s_i and s_{i+1} lies in $[0,\pi-\theta]$, or*

(2) *exactly one of s_i and s_{i+1} has length at least L, and the bending angle between s_i and s_{i+1} lies in $\bigl[0,\frac{\pi}{2}-\theta\bigr]$.*

Then we have

$$l(s_1\cdots s_m)\ge \frac{1}{2}\sum_{i=1}^m l(s_i).$$

Proof We lift the consecutive chain of geodesic segments $s_1,\dots,s_m$ to the universal cover, and work on a consecutive chain of geodesic segments in $\mathbb{H}^3$. If $m=1$ or 2, the result follows directly from the cosine law of hyperbolic geometry (see also estimates below), so we assume that $m\ge 3$.

Let x_1 be the initial point of s_1 and let x_{m+1} be the terminal point of s_m. For any $i=2,\dots,m$, let x_i be the terminal point of s_{i-1}, which is also the initial point of s_i. Let $y_1=x_1$, $y_m=x_{m+1}$, and let y_i be the middle point of s_i for each $i=2,\dots,m-1$. For any $i=1,\dots,m-1$, let t_i be the geodesic segment from y_i to y_{i+1}. We will use Lemma 3.2(1) to estimate $l(t_1\cdots t_{m-1})=l(s_1\cdots s_m)$.

We need to estimate lengths of t_i and bending angles between t_i and t_{i+1}. We claim that

$$l(t_i)\ge \tfrac{2}{3}(d(y_i,x_{i+1})+d(x_{i+1},y_{i+1}))\ge \tfrac{1}{3}L. \tag{3-1}$$

Case I Both $l(s_i)$ and $l(s_{i+1})$ are at least $\frac{1}{4}L$. Then $d(y_i,x_{i+1}), d(x_{i+1},y_{i+1})\ge \frac{1}{8}L$ and by assumption at least one of them is greater than $\frac{1}{2}L$. By [13, Lemma 4.10(2)], we have

$$\begin{aligned} l(t_i)=d(y_i,y_{i+1}) &\ge d(y_i,x_{i+1})+d(x_{i+1},y_{i+1})-I(\pi-\angle y_i x_{i+1} y_{i+1})\\ &\ge d(y_i,x_{i+1})+d(x_{i+1},y_{i+1})-I(\pi-\theta)\\ &\ge \tfrac{2}{3}(d(y_i,x_{i+1})+d(x_{i+1},y_{i+1}))\ge \tfrac{1}{3}L. \end{aligned}$$

Case II Otherwise, one of $l(s_i), l(s_{i+1})$ is at most $\frac{1}{4}L$, and we assume that $l(s_i) \leq \frac{1}{4}L$. By assumption of this lemma, we have $l(s_{i+1}) > L$ and $\angle y_i x_{i+1} y_{i+1} > \frac{\pi}{2} + \theta$. So we have $d(y_i, x_{i+1}) \leq l(s_i) \leq \frac{1}{4}L$ and $d(x_{i+1}, y_{i+1}) \geq \frac{1}{2}l(s_{i+1}) \geq \frac{1}{2}L$. Since $\angle y_i x_{i+1} y_{i+1} > \frac{\pi}{2}$, we have

$$l(t_i) = d(y_i, y_{i+1}) \geq d(x_{i+1}, y_{i+1}) \geq \tfrac{2}{3}(d(y_i, x_{i+1}) + d(x_{i+1}, y_{i+1})) \geq \tfrac{1}{3}L.$$

So (3-1) holds in both cases.

Then we claim that $\angle y_i y_{i+1} x_{i+1} < \frac{\pi}{2} - \theta$. In Case I above, we apply [13, Lemma 4.10(1)] to get

$$\angle y_i y_{i+1} x_{i+1} < e^{(-\frac{1}{8}L + 3\log 2)/2} < \tfrac{\pi}{2} - \theta.$$

In Case II above, $\angle y_i x_{i+1} y_{i+1} > \frac{\pi}{2} + \theta$ implies $\angle y_i y_{i+1} x_{i+1} < \frac{\pi}{2} - \theta$ directly.

The same argument implies $\angle y_{i+2} y_{i+1} x_{i+2} < \frac{\pi}{2} - \theta$. So we get

$$\angle y_i y_{i+1} y_{i+2} \geq \pi - \angle y_i y_{i+1} x_{i+1} - \angle y_{i+2} y_{i+1} x_{i+2} \geq \pi - 2\big(\tfrac{\pi}{2} - \theta\big) = 2\theta. \tag{3-2}$$

For the consecutive chain of geodesic segments $t_1, \dots, t_{m-1}$, by (3-1) and (3-2), each segment has length at least $\frac{1}{3}L$ and each bending angle is at most $\pi - 2\theta$. By Lemma 3.2(1), we have

$$\begin{aligned} l(s_1 \cdots s_m) = l(t_1 \cdots t_{m-1}) &\geq \sum_{i=1}^{m-1} l(t_i) - (m-2)I(\pi - 2\theta) - (m-2)\frac{e^{(-\frac{1}{6}L + 10\log 2)/2}}{\frac{1}{6}L - \log 2} \\ &\geq \sum_{i=1}^{m-1} l(t_i) - (m-2)(I(\pi - 2\theta) + 1) \geq \frac{3}{4}\sum_{i=1}^{m-1} l(t_i). \end{aligned}$$

Here the last inequality holds since $\frac{1}{4}l(t_i) \geq \frac{1}{12}L \geq I(\pi - 2\theta) + 1$ for each t_i, by (3-1). By (3-1) again, we have

$$\begin{aligned} l(s_1 \cdots s_m) \geq \frac{3}{4}\sum_{i=1}^{m-1} l(t_i) &\geq \frac{1}{2}\sum_{i=1}^{m-1}(d(y_i, x_{i+1}) + d(x_{i+1}, y_{i+1})) \\ &= \frac{1}{2}\bigg(d(y_1, x_2) + \sum_{i=2}^{m-1}(d(x_i, y_i) + d(y_i, x_{i+1})) + d(x_m, y_m)\bigg) \\ &= \frac{1}{2}\sum_{i=1}^{m} l(s_i). \end{aligned}$$

□

4 Reduction of the domain and target in Theorem 1.2

In this section, we prove a few preparational results that reduce the domain and target manifolds M and N in Theorem 1.2 to some convenient form. Recall that when talking about a cusped hyperbolic 3-manifold, we mean a compact 3-manifold with tori boundary whose interior admits a complete hyperbolic metric with finite volume, unless otherwise indicated.

4.1 Reducing the domain manifold M

At first, we prove that any cusped hyperbolic 3-manifold M has a finite cover that satisfies a convenient homological condition.

Proposition 4.1 *For any oriented cusped hyperbolic* 3*-manifold* M*, it has a finite cover* M' *with two distinct boundary components* $T_1, T_2 \subset \partial M'$*, such that the kernel of*

$$H_1(T_1 \cup T_2; \mathbb{Z}) \to H_1(M'; \mathbb{Z})$$

contains an element $\alpha_1 + \alpha_2 \in H_1(T_1 \cup T_2; \mathbb{Z})$ *such that* $0 \neq \alpha_1 \in H_1(T_1; \mathbb{Z})$ *and* $0 \neq \alpha_2 \in H_1(T_2; \mathbb{Z})$*.*

Note that we do need at least two boundary components of M' in Proposition 4.1. If all boundary components of M are H_1-injective (eg M is the complement of a two-component hyperbolic link with nonzero linking number), then any boundary component of any finite cover of M is H_1-injective.

Proof We take a boundary component T of M, a slope c on T, and a slope l on T that intersects c once.

By [17, Proposition 4.6] and its proof, there is a geometrically finite π_1-injective connected oriented immersed subsurface $i: S \looparrowright M$, such that the following hold:

- S has exactly two oriented boundary components, C_1 and C_2.
- i maps both C_1 and C_2 to T, and $i_*[C_1] = -i_*[C_2] = d[c]$ for some positive integer d.

For any positive integer D, let T_D be the covering space of T corresponding to $\langle dc, Dl \rangle < \langle c, l \rangle \cong \pi_1(T)$. Let S_D be the 2-complex obtained by pasting S and two copies of T_D (denoted by $T_{D,1}$ and $T_{D,2}$), such that $C_1, C_2 \subset \partial S$ are pasted with curves in $T_{D,1}$ and $T_{D,2}$ corresponding to $\pm dc$, respectively. Then the map $i: S \looparrowright M$ and covering maps $T_{D,1}, T_{D,2} \to T \subset M$ together induce a map $i_D: S_D \looparrowright M$. By [15, Theorem 1.1], when D is large enough, i_D is a π_1-injective map.

Since hyperbolic 3-manifold groups are LERF [2] and S_D embeds into the covering space of M corresponding to $(i_D)_*(\pi_1(S_D)) < \pi_1(M)$, there is a finite cover M' of M, such that $i_D: S_D \looparrowright M$ lifts to an embedding $\tilde{i}_D: S_D \hookrightarrow M'$. Since $\tilde{i}_D$ is an embedding, $T_1 = \tilde{i}_D(T_{D,1})$ and $T_2 = \tilde{i}_D(T_{D,2})$ are two distinct boundary components of M'. Since $\tilde{i}_D|_S: S \hookrightarrow M'$ gives an embedded oriented subsurface of M' whose boundary consists of a pair of essential curves on T_1 and T_2 respectively,

$$H_1(T_1 \cup T_2; \mathbb{Z}) \to H_1(M'; \mathbb{Z})$$

is not injective. Let α_1 be the intersection of $\partial S \cap T_1$, and let α_2 be the intersection of $\partial S \cap T_2$, then $\alpha_1 + \alpha_2 \in H_1(T_1 \cup T_2; \mathbb{Z})$ is an element in the kernel with the desired form. □

A similar argument as in Proposition 4.1 proves a similar result on certain mixed 3-manifolds.

Proposition 4.2 *Let M be an oriented mixed 3-manifold with tori boundary such that ∂M intersects with a hyperbolic piece M_0 of M. Then M has a finite cover M' with a hyperbolic piece M_0' of M', such that $M_0' \cap \partial M'$ contains two components T_1 and T_2, and the kernel of*

$$H_1(T_1 \cup T_2; \mathbb{Z}) \to H_1(M_0'; \mathbb{Z})$$

contains an element $\alpha_1 + \alpha_2 \in H_1(T_1 \cup T_2; \mathbb{Z})$ such that $0 \neq \alpha_1 \in H_1(T_1; \mathbb{Z})$ and $0 \neq \alpha_2 \in H_1(T_2; \mathbb{Z})$.

Proof Let T be a component of $M_0 \cap \partial M$. By the proof of Proposition 4.1, there is a π_1-injective 2-complex $i_D \colon S_D \looparrowright M_0 \hookrightarrow M$, such that S_D is a union of three surfaces, S, $T_{D,1}$ and $T_{D,2}$, and both $T_{D,1}$ and $T_{D,2}$ are mapped to T via covering maps.

Since $i_D \colon S_D \looparrowright M$ is mapped into a hyperbolic piece M_0 of M, by [21], $(i_D)_*(\pi_1(S_D))$ is a separable subgroup of $\pi_1(M)$. Since S_D embeds into the covering space of M corresponding to $(i_D)_*(\pi_1(S_D)) < \pi_1(M)$, there is a finite cover M' of M, such that $i_D \colon S_D \looparrowright M$ lifts to an embedding $\tilde{i}_D \colon S_D \hookrightarrow M'$. Since S_D is connected, the image of $\tilde{i}_D$ is contained in a hyperbolic piece $M_0' \subset M'$. Since $\tilde{i}_D$ is an embedding, $T_1 = \tilde{i}_D(T_{D,1})$ and $T_2 = \tilde{i}_D(T_{D,2})$ are two distinct boundary components of M_0' and they are both contained in $M_0' \cap \partial M'$. Similar to the proof of Proposition 4.1, the existence of the subsurface $\tilde{i}_D|_S \colon S \to M'$ implies that

$$H_1(T_1 \cup T_2; \mathbb{Z}) \to H_1(M'; \mathbb{Z})$$

is not injective, and the kernel contains an element in the desired form. □

4.2 Reducing the target manifold N

To reduce the target manifold N, we first need to prove two topological lemmas. The first one is quite elementary, while the second one uses results on branched coverings between 3-manifolds.

Lemma 4.3 *For any compact oriented 3-manifold N with nonempty tori boundary, there exists a proper map $g \colon N \to D^2 \times S^1$ such that the following hold.*

(1) *The degree of g is at least 3.*

(2) *g induces a surjective homomorphism on π_1.*

(3) *The restriction of g to each boundary component of N is a covering map to $S^1 \times S^1 \subset D^2 \times S^1$ of positive degree.*

Conditions (2) and (3) in this lemma imply that g is an "allowable primitive map", according to the terminology in [8].

Proof Let the boundary components of N be $T_1, \dots, T_k$. For each $i = 1, \dots, k$, let $j_i \colon T_i \to N$ be the inclusion map, then the free rank of the image of $(j_i)_* \colon H_1(T_i; \mathbb{Z}) \to H_1(N; \mathbb{Z})$ is at least 1.

So there exists a surjective homomorphism $\alpha\colon H_1(N;\mathbb{Z})\to\mathbb{Z}$ such that $\alpha\circ(j_i)_*\colon H_1(T_i;\mathbb{Z})\to\mathbb{Z}$ is nontrivial for all i. We consider α as an element in $\mathrm{Hom}(H_1(N;\mathbb{Z});\mathbb{Z})\cong H^1(N;\mathbb{Z})$, then the dual of α in $H_2(N,\partial N;\mathbb{Z})$ is represented by a compact oriented (possibly disconnected) proper subsurface $\Sigma\subset N$. By doing surgery, we can assume that for each i, $\Sigma\cap T_i$ consists of parallel essential circles with consistent orientation. By our choice of α, Σ intersects with each T_i nontrivially.

We take a proper map $h_0\colon\Sigma\to D^2=D^2\times\{\mathrm{pt}\}$ of degree at least 3, such that for any T_i, the restriction of $h_0|\colon\Sigma\cap T_i\to S^1=\partial D^2$ on each component of $\Sigma\cap T_i$ has the same positive degree d_i. This map h_0 can be constructed by first pinching $\Sigma\setminus\mathcal{N}(\partial\Sigma)$ to a point, with the resulting space being a one-point union of discs, then each disc is mapped to D^2 by a branched cover of positive degree. The restriction $h_0|\colon\Sigma\cap T_i\to S^1=S^1\times\{\mathrm{pt}\}$ can be extended to a covering map $h_i\colon T_i\to S^1\times S^1$ of positive degree as following. Since each component A of $T_i\setminus(\Sigma\cap T_i)$ is an annulus, and its two boundary components are mapped to $S^1\times\{\mathrm{pt}\}$ with the same degree d_i, we define $h_i|_A$ to be a (orientation preserving) covering map to $(S^1\times S^1)\setminus(S^1\times\{\mathrm{pt}\})$ of degree d_i. Then we have $h_i^{-1}(S^1\times\{\mathrm{pt}\})=\Sigma\cap T_i$ and $\deg(h_i)=\deg(h_0|\colon\Sigma\cap T_i\to S^1)$.

The maps h_0 and h_i, $i=1,\dots,k$ together give a map

$$h\colon\Sigma\cup\left(\bigcup_{i=1}^{k}T_i\right)\to(D^2\times\{\mathrm{pt}\})\cup(S^1\times S^1)\subset D^2\times S^1.$$

The map h extends to a proper map $g\colon N\to D^2\times S^1$, since it extends to a neighborhood of $\Sigma\cup(\cup T_i)\subset N$ and $(D^2\times S^1)\setminus\big((D^2\times\{\mathrm{pt}\})\cup(S^1\times S^1)\big)$ is a 3-ball.

By construction, we have $\deg(g)=\sum_{i=1}^{k}\deg(h_i)=\deg(h_0)\geq 3$; thus condition (1) holds. For the composition $N\xrightarrow{g}D^2\times S^1\to S^1$, the preimage of $\mathrm{pt}\in S^1$ is exactly Σ, so the induced homomorphism $H_1(N;\mathbb{Z})\to H_1(S^1;\mathbb{Z})$ is the same as $\alpha\colon H_1(N;\mathbb{Z})\to\mathbb{Z}$. Since α is surjective, g induces a surjective homomorphism on both H_1 and π_1; thus condition (2) holds. Since $g\colon N\to D^2\times S^1$ is an extension of $h_i\colon T_i\to S^1\times S^1$, condition (3) holds. □

Lemma 4.4 *For any compact oriented* 3*-manifold N with nonempty tori boundary, there exists a compact oriented* 3*-manifold M with connected torus boundary, such that M virtually properly* 2*-dominates N.*

Proof We first take a proper map $g\colon N\to D^2\times S^1$ satisfying the conclusion of Lemma 4.3. These conditions make [8, Theorem 4.1] applicable, so $g\colon N\to D^2\times S^1$ is homotopic to a branched covering map relative to the boundary, such that the branching locus is a link (a disjoint union of circles) in $D^2\times S^1$. By [3, Theorem 6.5], $g\colon N\to D^2\times S^1$ is further homotopic to a simple branched covering, and we still denote this map by g. Here by simple branched covering, we mean that g is a branched covering of degree $d\in\mathbb{Z}_{>0}$, such that for any $p\in D^2\times S^1$, $g^{-1}(p)$ consists of at least $d-1$ points.

Note that if $g\colon N\to D^2\times S^1$ is a nonbranched cover, then $N=D^2\times S^1$ and we can simply take $M=D^2\times S^1$. So we can assume that g is a genuine branched cover.

Let $L \subset D^2 \times S^1$ be the branched locus of the simple branched covering $g\colon N \to D^2 \times S^1$, and write $L = L_1 \cup \dots \cup L_k$ as a union of its components. We take a tubular neighborhood

$$\mathcal{N}(L) = \mathcal{N}(L_1) \cup \dots \cup \mathcal{N}(L_k)$$

of $L = L_1 \cup \dots \cup L_k$. On the torus $\partial\mathcal{N}(L_i)$, we take an oriented meridian m_i that bounds the meridian disc of $\mathcal{N}(L_i)$, and take an oriented longitude l_i that intersects with m_i exactly once. For each L_i, we take M_i to be a copy of $\Sigma_{1,1} \times S^1$, and take a linear homeomorphism

$$\phi_i\colon \partial M_i = \partial\Sigma_{1,1} \times S^1 \to \partial\mathcal{N}(L_i)$$

that maps oriented curves $\partial\Sigma_{1,1} \times \{*\}$ and $\{*\} \times S^1$ to m_i and l_i respectively. Then we take M to be

$$M = (D^2 \times S^1 \setminus \mathcal{N}(L)) \bigcup_{\{\phi_i\}_{i=1}^k} \left(\bigcup_{i=1}^k M_i\right).$$

Now we need to construct a finite cover $M' \to M$ and a degree-2 map $M' \to N$.

Let $\mathscr{L} = g^{-1}(L) \subset N$. Since $g\colon N \to D^2 \times S^1$ is a branched covering, there is a tubular neighborhood $\mathcal{N}(\mathscr{L})$ of $\mathscr{L}$ in N, such that the restriction map $g|\colon N \setminus \mathcal{N}(\mathscr{L}) \to D^2 \times S^1 \setminus \mathcal{N}(L)$ is a covering map of degree $d = \deg(g)$.

Let $\mathscr{L}_i = g^{-1}(L_i)$. Since $g\colon N \to D^2 \times S^1$ is a simple branched covering, there is a unique component $\mathscr{L}_i^0$ of $\mathscr{L}_i$ such that g is locally a 2-to-1 map near $\mathscr{L}_i^0$, and g is a local homeomorphism near any point in $\mathscr{L}_i \setminus \mathscr{L}_i^0 = \bigcup_{j=1}^{n_i} \mathscr{L}_i^j$. The restriction of g to $\partial\mathcal{N}(\mathscr{L}_i^0) \to \partial N(L_i)$ is a finite cover corresponding to a subgroup of $\pi_1(\partial N(L_i))$ in one of the following two types:

(1) $\langle 2m_i, k_i l_i\rangle < \pi_1(\partial\mathcal{N}(L_i))$ or

(2) $\langle 2m_i, k_i l_i + m_i\rangle < \pi_1(\partial\mathcal{N}(L_i))$

for some positive integer k_i. For any $j \in \{1, \dots, n_i\}$, the restriction of g to $\partial\mathcal{N}(\mathscr{L}_i^j) \to \partial\mathcal{N}(L_i)$ is a finite cover corresponding to subgroup

(3) $\langle m_i, k_i^j l_i\rangle < \pi_1(\partial\mathcal{N}(L_i))$

for some positive integer k_i^j.

In case (3), let $\tilde{m}_i^j, \tilde{l}_i^j \subset \partial\mathcal{N}(\mathscr{L}_i^j)$ be one (oriented) component of the preimage of $m_i, l_i \subset \partial\mathcal{N}(L_i)$, respectively. Then $\tilde{m}_i^j \to m_i$ and $\tilde{l}_i^j \to l_i$ are covering maps of degree 1 and k_i^j respectively, and these two curves intersect once on $\partial\mathcal{N}(\mathscr{L}_i^j)$. We take $M_i^j = \Sigma_{1,1} \times S^1$ and a degree-k_i^j covering map

$$p_i^j\colon M_i^j = \Sigma_{1,1} \times S^1 \to M_i = \Sigma_{1,1} \times S^1,$$

that is the product of $\mathrm{id}\colon \Sigma_{1,1} \to \Sigma_{1,1}$ and the degree-k_i^j covering map $S^1 \to S^1$. Let

$$\psi_i^j\colon \partial M_i^j = \partial\Sigma_{1,1} \times S^1 \to \partial N(\mathscr{L}_i^j)$$

be the linear homeomorphism that maps oriented curves $\partial\Sigma_{1,1}\times\{*\}$ and $\{*\}\times S^1$ to $\tilde{m}_i^j$ and $\tilde{l}_i^j$ respectively. Then we have the following commutative diagram:

(4-1)
$$\begin{tikzcd}
\partial M_i^j = \partial\Sigma_{1,1}\times S^1 \arrow[r, "\psi_i^j"] \arrow[d, "p_i^j|"'] & \partial\mathcal{N}(\mathscr{L}_i^j) \subset N\setminus N(\mathscr{L}) \arrow[d, "g|"'] \\
\partial M_i = \partial\Sigma_{1,1}\times S^1 \arrow[r, "\phi_i"] & \partial\mathcal{N}(L_i) \subset D^2\times S^1\setminus N(L)
\end{tikzcd}$$

In case (1) above, let $\tilde{m}_i^0, \tilde{l}_i^0 \subset \partial\mathcal{N}(\mathscr{L}_i^0)$ be one (oriented) component of the preimage of $m_i, l_i \subset \partial\mathcal{N}(L_i)$, respectively. Then $\tilde{m}_i^0 \to m_i$ and $\tilde{l}_i^0 \to l_i$ are covering maps of degree 2 and k_i respectively, and these two curves intersect once on $\partial N(\mathscr{L}_i^0)$. We take $M_i^0 = \Sigma_{2,2}\times S^1$, and take a degree-$4k_i$ covering map

$$p_i^0 \colon M_i^0 = \Sigma_{2,2}\times S^1 \to M_i = \Sigma_{1,1}\times S^1.$$

It is the product of the degree-4 covering map $\Sigma_{2,2}\to\Sigma_{1,1}$ that factors through $\Sigma_{1,2}$ (which restricts to a degree-2 cover on each boundary component), and the degree-k_i covering map $S^1\to S^1$. Let

$$\psi_i^0 \colon \partial M_i^0 = \partial\Sigma_{2,2}\times S^1 \to \partial\mathcal{N}(\mathscr{L}_i^0)$$

be a linear map that restricts to a homeomorphism on each component of ∂M_i^0, such that it maps each oriented boundary component of $\partial\Sigma_{2,2}\times\{*\}$ to $\tilde{m}_i^0$, and maps oriented curves $\{*\}\times S^1$ on both components of ∂M_i^0 to $\tilde{l}_i^0$. Then we have the following commutative diagram:

(4-2)
$$\begin{tikzcd}
\partial M_i^0 = \partial\Sigma_{2,2}\times S^1 \arrow[r, "\psi_i^0"] \arrow[d, "p_i^0|"'] & \partial N(\mathscr{L}_i^0) \subset N\setminus N(\mathscr{L}) \arrow[d, "g|"'] \\
\partial M_i = \partial\Sigma_{1,1}\times S^1 \arrow[r, "\phi_i"] & \partial N(L_i) \subset D^2\times S^1\setminus N(L)
\end{tikzcd}$$

In case (2) above, let $\tilde{m}_i^0, \tilde{l}_i^0 \subset \partial N(\mathscr{L}_i^0)$ be one (oriented) component of the preimage of $m_i, l_i \subset \partial N(L_i)$, respectively. There is actually a unique $\tilde{l}_i^0$. Then $\tilde{m}_i^0 \to m_i$ and $\tilde{l}_i^0 \to l_i$ are covering maps of degree 2 and $2k_i$ respectively, and these two curves intersect (algebraically) twice on $\partial\mathcal{N}(\mathscr{L}_i^0)$. Here $\tilde{l}_i^0$ corresponds to $2(k_il_i+m_i)-(2m_i)=2k_il_i\in\pi_1(\partial\mathcal{N}(L_i))$. We take $M_i^0=\Sigma_{2,2}\times I/(x,0)\sim(\phi(x),1)$, where $\phi\colon\Sigma_{2,2}\to\Sigma_{2,2}$ is the nontrivial deck transformation of the double cover $q\colon\Sigma_{2,2}\to\Sigma_{1,2}$. Note that q restricts to a degree-2 cover on each boundary component of $\Sigma_{2,2}$. Then we take the degree-$4k_i$ covering map

$$p_i^0 \colon M_i^0 = \Sigma_{2,2}\times I/{\sim} \to M_i = \Sigma_{1,1}\times S^1$$

that is a composition $\Sigma_{2,2}\times I/{\sim} \to \Sigma_{1,2}\times S^1 \to \Sigma_{1,1}\times S^1$. Here the first map takes the double covering map $q\colon\Sigma_{2,2}\to\Sigma_{1,2}$ on each fiber and takes the identity map on the base S^1, the second map is the product of the double cover $\Sigma_{1,2}\to\Sigma_{1,1}$ and the degree-k_i covering map $S^1\to S^1$. Then each oriented component of the p_i^0-preimage of $\partial\Sigma_{1,1}\times\{*\}$ is a component of $\partial\Sigma_{2,2}\times\{*\}$. Each oriented component of the p_i^0-preimage of $\{*\} \times S^1$ is a flow line of M_i^0 along the I-direction, and it intersects the corresponding component of $\partial\Sigma_{2,2} \times \{*\}$ algebraically twice.

There exists a linear map

$$\psi_i^0 \colon \partial M_i^0 = \partial\Sigma_{2,2} \times I/{\sim} \to \partial\mathcal{N}(\mathscr{L}_i^0)$$

that restricts to a homeomorphism on each component of ∂M_i^0, such that it maps each oriented boundary component of $\partial\Sigma_{2,2} \times \{*\}$ to $\tilde{m}_i^0$, and maps an I-flow line on each component of ∂M_i^0 to $\tilde{l}_i^0$. Then we have the following commutative diagram:

$$\begin{tikzcd}
\partial M_i^0 \arrow[r, "\psi_i^0"] \arrow[d, "p_i^0|"'] & \partial N(\mathscr{L}_i^0) \subset N \setminus \mathcal{N}(\mathscr{L}) \arrow[d, "g|"'] \\
\partial M_i \arrow[r, "\phi_i"] & \partial N(L_i) \subset D^2 \times S^1 \setminus \mathcal{N}(L)
\end{tikzcd} \tag{4-3}$$

Now we take two copies of $N \setminus N(\mathscr{L})$ and denote them by $(N \setminus N(\mathscr{L}))_1$ and $(N \setminus N(\mathscr{L}))_2$ respectively. For any $i = 1, \dots, k$ and $j = 1, \dots, n_i$, we take two copies of M_i^j and denote them by $(M_i^j)_1$ and $(M_i^j)_2$. For any $i = 1, \dots, k$, M_i^0 has two boundary components, and we denote them by $(\partial M_i^0)_1$ and $(\partial M_i^0)_2$ respectively. Then we take M' to be the union of manifolds

$$(N \setminus \mathcal{N}(\mathscr{L}))_1, \quad (N \setminus \mathcal{N}(\mathscr{L}))_2, \quad (M_i^j)_1, \quad (M_i^j)_2, \quad M_i^0 \quad \text{for } i = 1, \dots, k,\ j = 1, \dots, n_i,$$

by pasting maps

$$(\psi_i^j)_1 \colon \partial(M_i^j)_1 \to (\partial\mathcal{N}(\mathscr{L}_i^j))_1 \subset (N \setminus \mathcal{N}(\mathscr{L}))_1, \quad (\psi_i^j)_2 \colon \partial(M_i^j)_2 \to (\partial\mathcal{N}(\mathscr{L}_i^j))_2 \subset (N \setminus \mathcal{N}(\mathscr{L}))_2,$$
$$(\psi_i^0|)_1 \colon (\partial M_i^0)_1 \to (\partial\mathcal{N}(\mathscr{L}_i^0))_1 \subset (N \setminus \mathcal{N}(\mathscr{L}))_1, \quad (\psi_i^0|)_2 \colon (\partial M_i^0)_2 \to (\partial\mathcal{N}(\mathscr{L}_i^0))_2 \subset (N \setminus \mathcal{N}(\mathscr{L}))_2.$$

The homeomorphisms $(\psi_i^j)_1$ and $(\psi_i^j)_2$ denote copies of the map ψ_i^j on the corresponding copy of ∂M_i^j, while $(\psi_i^0|)_1$ and $(\psi_i^0|)_2$ denote the restriction of ψ_i^0 on the corresponding component of ∂M_i^0.

The covering map $\pi \colon M' \to M$ is defined by the following covering maps on pieces of M':

- $g|$ maps $(N \setminus \mathcal{N}(\mathscr{L}))_1, (N \setminus \mathcal{N}(\mathscr{L}))_2 \subset M'$ to $D^2 \times S^2 \setminus \mathcal{N}(L) \subset M$,
- p_i^j maps $(M_i^j)_1, (M_i^j)_2 \subset M'$ to $M_i \subset M$ for $i = 1, \dots, k$ and $j = 1, \dots, n_i$,
- p_i^0 maps $M_i^0 \subset M'$ to $M_i \subset M$ for $i = 1, \dots, k$.

Here π is a well-defined map because of three commutative diagrams (4-1), (4-2) and (4-3).

The degree-2 map $f \colon M' \to N$ is defined by the following maps on pieces of M':

- The identity map that maps $(N \setminus \mathcal{N}(\mathscr{L}))_1, (N \setminus \mathcal{N}(\mathscr{L}))_2 \subset M'$ to $N \setminus \mathcal{N}(\mathscr{L}) \subset N$.
- A pinching map that maps $(M_i^j)_1, (M_i^j)_2 = \Sigma_{1,1} \times S^1 \subset M'$ to $\mathcal{N}(\mathscr{L}_i^j) = D^2 \times S^1 \subset N$. This map is the product of a degree-1 pinching map $\Sigma_{1,1} \to D^2$ and the identity map $S^1 \to S^1$.

- A pinching map that maps $M_i^0 = \Sigma_{2,2} \times I/{\sim} \subset M'$ to $\mathcal{N}(\mathcal{L}_i^0) = D^2 \times S^1 \subset N$, where $\sim$ is induced by the identity map of $\Sigma_{2,2}$ or the nontrivial deck transformation of $\Sigma_{2,2} \to \Sigma_{1,2}$. Here we take a fixed degree-2 pinching map $\Sigma_{2,2} \to D^2$ on each fiber, such that it commutes with monodromy homeomorphisms of M_i^0 and $\mathcal{N}(\mathcal{L}_i^0)$, and restricts to a homeomorphism on each boundary component.

Then f is a degree-2 proper map from M' to N. □

Now we are ready to prove the following result.

Proposition 4.5 *For any compact oriented* 3*-manifold* N *with nonempty tori boundary*, *there exists a one-cusped oriented hyperbolic* 3*-manifold* M, *such that* M *virtually properly* 2*-dominates* N.

Proof By Lemma 4.4, N is virtually 2-dominated by a compact oriented 3-manifold N' with connected torus boundary. By [23, Lemma 4.1], N' is 1-dominated by a compact oriented irreducible 3-manifold N'' with connected torus boundary.

This result follows from the proof of Proposition 3.2 of [4], although the result in [4] is only stated for closed 3-manifolds. By [16, Theorem 7.2], there exists a hyperbolic knot $K \subset N''$ that is null-homotopic in N''. We take M to be a hyperbolic Dehn-filling of $N'' \setminus \mathcal{N}(K)$, then M is a one-cusped hyperbolic 3-manifold. The proof of Proposition 3.2 of [4] constructs a degree-1 map $f\colon M \to N''$. More precisely, the map f is identity on $N'' \setminus \mathcal{N}(K)$, it extends to the meridian disc of the filled-in solid torus since K is null-homotopic in N'', and it extends to the whole solid torus since N'' is irreducible. □

5 Topological construction of virtual domination

In this section, we give the topological part of the proof of Theorem 1.2, and we also point out how to modify the works to prove Theorem 1.3.

To prove Theorem 1.2, it suffices to prove the following result.

Theorem 5.1 *Let* M *be a compact oriented hyperbolic* 3*-manifold*, *such that* ∂M *has two components* T_1 *and* T_2 *and the kernel of* $H_1(T_1 \cup T_2; \mathbb{Z}) \to H_1(M; \mathbb{Z})$ *contains an element* $\alpha_1 + \alpha_2 \in H_1(T_1 \cup T_2; \mathbb{Z})$ *with* $0 \neq \alpha_1 \in H_1(T_1; \mathbb{Z})$ *and* $0 \neq \alpha_2 \in H_1(T_2; \mathbb{Z})$. *Let* N *be a compact oriented hyperbolic* 3*-manifold with connected torus boundary. Then* M *has a finite cover* M', *such that there is a proper map* $f\colon M' \to N$ *with* $\deg(f) \in \{1, 2, 4\}$.

We first prove that Theorem 5.1 implies Theorem 1.2.

Proof of Theorem 1.2 (by assuming Theorem 5.1) Let M be a compact oriented cusped hyperbolic 3-manifold, and let N be a compact oriented 3-manifold with tori boundary, as in Theorem 1.2. By

Lemmas 4.1 and 4.5, there are compact oriented 3-manifold M_1 and N_1 with tori boundary, such that the following hold:

(1) M_1 is a finite cover of M and ∂M_1 contains two components, T_1 and T_2, such that the kernel of $H_1(T_1 \cup T_2;\mathbb{Z}) \to H_1(M_1;\mathbb{Z})$ contains an element $\alpha_1 + \alpha_2 \in H_1(T_1 \cup T_2;\mathbb{Z})$ with $0 \neq \alpha_1 \in H_1(T_1;\mathbb{Z})$ and $0 \neq \alpha_2 \in H_1(T_2;\mathbb{Z})$.

(2) N_1 is an oriented one-cusped hyperbolic 3-manifold that admits a finite cover $p\colon N_2 \to N_1$ and a degree-2 map $g\colon N_2 \to N$.

Theorem 5.1 implies that M_1 has a finite cover M_2 and there is a proper map $h\colon M_2 \to N_1$ such that $\deg(h) \in \{1,2,4\}$.

Let $q\colon M_3 \to M_2$ be the covering space of M_2 corresponding to $(h_*)^{-1}(p_*(\pi_1(N_2)))$, then we have the following commutative diagram:

$$\begin{tikzcd}
M_3 \arrow[r, "h'"] \arrow[d, "q"'] & N_2 \arrow[d, "p"] \\
M_2 \arrow[r, "h"] & N_1
\end{tikzcd}$$

Here

$$\deg(q) = [\pi_1(M_2) : q_*(\pi_1(M_3))] = [\pi_1(M_2) : (h_*)^{-1}(p_*(\pi_1(N_2)))]$$

is a factor of $\deg(p) = [\pi_1(N_1) : p_*(\pi_1(N_2))]$. Since $\deg(h)\cdot\deg(q) = \deg(p)\cdot\deg(h')$, $\deg(h')$ is a factor of $\deg(h) \in \{1,2,4\}$. So $f' = g \circ h'\colon M_3 \xrightarrow{h'} N_2 \xrightarrow{g} N$ is a map such that $\deg(f') \in \{2,4,8\}$.

Since M_3 has tori boundary, $b_1(M_3) \geq 1$ holds. So M_3 has a cyclic cover $r\colon M' \to M_3$ of degree $8/\deg(f')$. Then $f = f' \circ r\colon M' \to N$ is a proper map of degree 8, as desired. □

The following three subsections are devoted to prove Theorem 5.1, modulo a π_1-injectivity result (Theorem 5.17). We always assume that M and N satisfy the assumption of Theorem 5.1.

5.1 Initial data of the construction

In this section, we first give some geometric data deduced from N, then we give some related geometric notions on M.

We first prove the following lemma on triangulation of flat tori. The resulting triangulation will be the restriction of our desired triangulation of a cusped hyperbolic 3-manifold to its horotorus.

Lemma 5.2 *For any flat torus T, any primitive closed geodesic l on T, and any $\epsilon \in (0, 0.1)$, T has a geometric triangulation such that the following hold.*

(1) *l is contained in the 1-skeleton of this triangulation.*

(2) *There exists $r \in (0,\epsilon)$, such that all edges have length in $[r,(1+2\epsilon)r)$.*

(3) *Any inner angle of any triangle is ϵ-close to $\frac{\pi}{3}$.*

Proof Up to multiplying the flat Riemannian metric on T with a positive real number, we can assume that T is isometric to $\mathbb{C}/\mathbb{Z}\oplus\mathbb{Z}z_0$ for some complex number $z_0\in\mathbb{C}$ with $\mathrm{Im}(z_0)>0$, and the closed geodesic l corresponds to the edge from 0 to 1.

We consider the lattice $\Lambda_0=\mathbb{Z}\oplus\mathbb{Z}\omega_0$ of $\mathbb{C}$ for $\omega_0=\frac{1}{2}(1+\sqrt{3}i)$. For large $N\in\mathbb{N}$, we take ω to be the point in Λ_0 closest to Nz_0. Note that $\mathrm{Im}(\omega)>0$ holds if $N>1/\mathrm{Im}(z_0)$.

Let $T:\mathbb{C}\to\mathbb{C}$ be the linear transformation that maps $N,\omega\in\Lambda_0$ to $N, Nz_0\in\mathbb{Z}\oplus\mathbb{Z}z_0$ respectively. Then $\frac{1}{N}T$ maps Λ_0 to a lattice of $\mathbb{C}$ that contains $\mathbb{Z}\oplus\mathbb{Z}z_0$. The equilateral triangulation of $\mathbb{C}$ corresponding to Λ_0 induces a triangulation of T, and l is contained in the 1-skeleton.

It is straight forward to check that, if N is large enough (say $N>6/(\epsilon\cdot\mathrm{Im}(z_0))$), then all inner angles of the above triangulation of T are ϵ-close to $\frac{\pi}{3}$. So we get an ϵ-almost-equilateral geodesic triangulation of T, such that all triangles are isometric to each other.

For each triangle in this triangulation of T, we take middle points of its edges and divide it into four smaller triangles, to get a finer ϵ-almost-equilateral geodesic triangulation, such that all smaller triangles are similar to the original ones. We do this process repeatedly so that all edges have length at most ϵ. Let r be the length of the shortest edge, then the Euclidean sine law implies that all edges have length in $[r,(1+2\epsilon)r)$. □

Let $\epsilon_0>0$ be a constant smaller than the 3-dimensional Margulis constant. For the one-cusped hyperbolic 3-manifold N as in Theorem 5.1 (considered as an open complete Riemannian manifold), let N_c be the complement of the cusp end with injectivity radius at most $\epsilon_0/10$, and let T_c be the boundary of N_c. By a classical application of the Lefschetz duality, there is a primitive closed geodesic l on T_c that spans $\ker(H_1(T_c;\mathbb{R})\to H_1(N_c;\mathbb{R}))$.

Construction 5.3 We construct a geometric triangulation of a compact submanifold of N containing N_c, whose geometry near ∂N_c is quite special. In the process, we also construct two submanifolds of N, which are denoted by N_{collar} and N_0. These notations will be used several times in the remaining of this paper.

Let $\epsilon\in(0,\epsilon_0/100)$ be a constant smaller than the injectivity radius of N_c.

(1) By Lemma 5.2 (applied to $\epsilon/100$), the horotorus T_c (also called the outside torus) has a geometric triangulation, such that l is contained in the 1-skeleton, all edges have length in $[r,(1+2\epsilon/100)r)$ for some $r\in(0,\epsilon/100)$, and all inner angles of triangles are $\epsilon/100$-close to $\frac{\pi}{3}$.

(2) Let T_c' be the horotorus in N_c that has distance $(\sqrt{6}/3)r$ from T_c, which is also called the inside torus. For any triangle Δ in T_c, we take its circumcenter, and let v_Δ be its closest point on T_c'. For any vertex n of the triangulation of T_c, let its closest point on T_c' be v_n.

(3) We connect v_Δ to the three vertices of Δ and obtain a (hyperbolic) tetrahedron in N. All inner angles of all triangles on the boundary of this tetrahedron are ϵ-close to $\frac{\pi}{3}$. Note that the triangle Δ

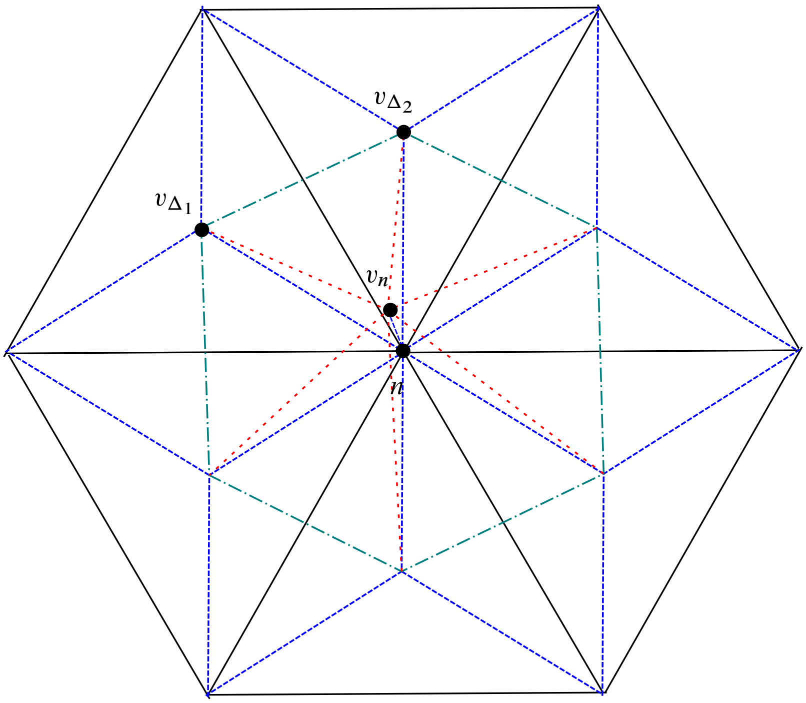


Figure 1: A picture of the triangulation near ∂N_0, viewing from the outside of N_0. Each black triangle lies in the outside torus T_c. Up to homotopy, blue edges connect T_c (the outside torus) and T_c' (the inside torus), while red and green edges lie in T_c'.

contained in T_c is not a face of this tetrahedron, since Δ is only a Euclidean triangle but not a hyperbolic one.

(4) For any two triangles Δ_1 and Δ_2 contained in T_c that share an edge, we add an edge connecting v_{Δ_1} and v_{Δ_2}. This edge and the edge $\Delta_1 \cap \Delta_2$ together give a hyperbolic tetrahedron in N.

(5) For any vertex n and two triangles Δ_1 and Δ_2 contained in T_c, such that $\Delta_1 \cap \Delta_2$ is an edge containing n, we get a hyperbolic tetrahedron with vertices n, v_n, v_{Δ_1} and v_{Δ_2}. A picture of the tetrahedra we have constructed can be found in Figure 1, which gives a triangulation of a compact submanifold containing T_c.

(6) Let N_{collar} be the union of tetrahedra (with disjoint interior) constructed in previous steps, and let N_0 be the union of N_c and N_{collar}, which is compact and is a deformation retract of N. Then we extend the above triangulation of N_{collar} to a geometric triangulation of N_0.

For the above geometric triangulation of N_0, we use $V_N = \{n_1, n_2, \dots, n_l\}$ to denote the set of vertices, and let $V_{N,\partial} = V_N \cap \partial N_0$. If there is an oriented edge from n_i to n_j, we denote it by e_{ij} and denote its orientation reversal by e_{ji}. For each triangle with vertices n_i, n_j and n_k, we denote the corresponding marked triangle by Δ_{ijk} (with an order on its vertices). We can naturally identify ∂N_0 with $\partial N_c = T_c$, and identify their triangulations.

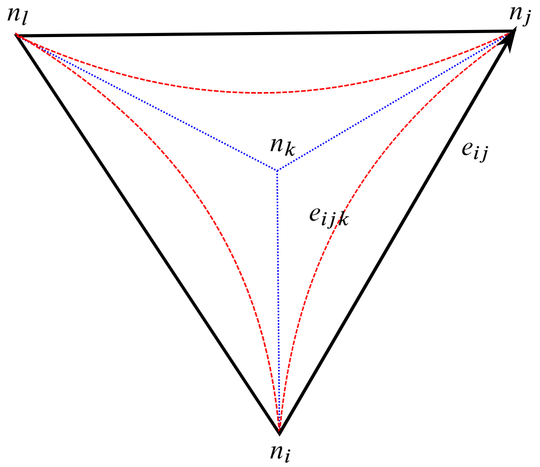


Figure 2: Construction of the new edge e_{ijk} near ∂N_0, where vertex n_k is to the left of e_{ij}. Here the black and blue edges are the same as in Figure 1, while the red edges are the new edges constructed in Construction 5.4.

Instead of directly working with the above geometric triangulation of N_0, we add more edges to get a cellulation of N_0.

Construction 5.4 For any triangle Δ_{ijk} that only intersects with ∂N_0 along an edge e_{ij}, we add an oriented path e_{ijk} in Δ_{ijk} from n_i to n_j of constant geodesic curvature, such that the tangent vector of e_{ijk} at n_i is $\epsilon/200$-close to the average of tangent vectors of e_{ij} and e_{ik}, and the same for the tangent vector of e_{ijk} at n_j. See Figure 2 for a picture of e_{ijk}. After this construction, there are two edges from n_i to n_j.

The new edge e_{ijk} divides Δ_{ijk} to a bigon and a triangle. We denote the bigon by B_{ijk}, and abuse notation to denote the new triangle by Δ_{ijk}. For a triangle Δ_{ijk} obtained by this modification process, it is called a *modified triangle*. The original triangle Δ_{ijk} (defined in Construction 5.3) will not be used anymore.

After adding these edges to the triangulation of N_0 in Construction 5.3, we get a cellulation of N_0, which is called a geometric cellulation.

We use $N^{(1)}$ and $N^{(2)}$ to denote the 1- and 2-skeletons of the above geometric cellulation of N respectively. This cellulation also gives a handle structure of a neighborhood of N_0 in N, and we use $\mathcal{N}^{(1)}$ and $\mathcal{N}^{(2)}$ to denote the union of 0-, 1-handles, and 0-, 1-, 2-handles.

Let m be a simple closed curve on T_c that intersects with l exactly once. We isotope m to a curve γ in $N_0 \setminus N_{\text{collar}}$, so that it is disjoint from $N^{(1)}$, and intersects with all triangles in $N^{(2)}$ transversely. Let $\mathcal{N}(\gamma)$ be the union of all tetrahedra that intersect with γ, then we can assume that $\mathcal{N}(\gamma)$ is a neighborhood of γ homeomorphic to the solid torus. Let N_γ be $N_0 \setminus \text{int}(\mathcal{N}(\gamma))$. The above cellulation of N_0 induces a cellulation of N_γ, and we denote the 2-skeleton of N_γ by $N_\gamma^{(2)}$.

Since the 2-skeleton carries the first homology group, we have the following commutative diagram:

$$\begin{tikzcd} H_1(N_\gamma^{(2)};\mathbb{Z}) \arrow[r] \arrow[d] & H_1(N_\gamma;\mathbb{Z}) \arrow[d] & \\ H_1(N^{(2)};\mathbb{Z}) \arrow[r] & H_1(N_0;\mathbb{Z}) \arrow[r] & H_1(N;\mathbb{Z}) \end{tikzcd} \tag{5-1}$$

Here all homomorphisms are induced by inclusions, and all horizontal homomorphisms are isomorphisms.

The following lemma provides some elementary properties of vertical homomorphisms in diagram (5-1).

Lemma 5.5 *Let $i : N_\gamma^{(2)} \to N$ be the inclusion map, and let c be the meridian of $\mathcal{N}(\gamma)$, then the following hold.*

(1) *$i_* \colon H_1(N_\gamma^{(2)};\mathbb{Z}) \to H_1(N;\mathbb{Z})$ is surjective.*

(2) *The kernel of i_* is spanned by a nontorsion element $[c] \in H_1(N_\gamma^{(2)};\mathbb{Z})$, and $[c]-[l]$ is a torsion element in $H_1(N_\gamma^{(2)};\mathbb{Z})$.*

(3) *The inclusion $\partial N_0 \to N_\gamma^{(2)}$ induces an injective homomorphism on $H_1(\,\cdot\,;\mathbb{Z})$.*

Proof Since horizontal homomorphisms in diagram (5-1) are isomorphisms, it suffices to study the inclusion $N_\gamma \to N_0$. We consider the Mayer–Vietoris sequence given by $N_0 = N_\gamma \cup \mathcal{N}(\gamma)$:

$$H_1(\partial\mathcal{N}(\gamma);\mathbb{Z}) \to H_1(N_\gamma;\mathbb{Z}) \oplus H_1(\mathcal{N}(\gamma);\mathbb{Z}) \to H_1(N_0;\mathbb{Z}) \to 0.$$

Item (1) follows from the surjectivity of $H_1(N_\gamma;\mathbb{Z}) \oplus H_1(\mathcal{N}(\gamma);\mathbb{Z}) \to H_1(N_0;\mathbb{Z})$ and the surjectivity of $H_1(\partial\mathcal{N}(\gamma);\mathbb{Z}) \to H_1(\mathcal{N}(\gamma);\mathbb{Z})$.

Now we prove item (2). By the Mayer–Vietoris sequence, the kernel of $H_1(N_\gamma;\mathbb{Z}) \to H_1(N_0;\mathbb{Z})$ is spanned by the meridian c of $\mathcal{N}(\gamma)$. Since $[l]$ spans $\ker(H_1(\partial N_0;\mathbb{R}) \to H_1(N_0;\mathbb{R}))$, we take the minimal $d \in \mathbb{Z}_{>0}$ such that $d[l] = 0 \in H_1(N_0;\mathbb{Z})$. So dl bounds a compact oriented surface S in N_0, and the algebraic intersection number between $m \subset \partial N_0$ and S is d. Since γ is isotopic to m in N_0, the algebraic intersection number between γ and S is d. So $S \cap N_\gamma$ is a compact oriented surface in N_γ and it implies that $d[l]-d[c] = 0 \in H_1(N_\gamma;\mathbb{Z})$, so $[c]-[l]$ is a torsion element in $H_1(N_\gamma^{(2)};\mathbb{Z})$. Moreover, the above argument also shows that kl does not bound a compact oriented surface in N_γ for any $k \in \mathbb{Z}_{>0}$. So $[l]$ is not a torsion element in $H_1(N_\gamma;\mathbb{Z})$, and $[c]$ is not a torsion element either.

For the composition $H_1(\partial N_0;\mathbb{Z}) \to H_1(N_\gamma;\mathbb{Z}) \to H_1(N_0;\mathbb{Z})$, its kernel is spanned by a multiple of $[l]$. Since item (2) implies that $[l]$ is not an torsion element in $H_1(N_\gamma;\mathbb{Z})$, the homomorphism $H_1(\partial N_0;\mathbb{Z}) \to H_1(N_\gamma;\mathbb{Z})$ is injective. □

Recall that M has two boundary components, T_1 and T_2, such that the kernel of

$$H_1(T_1 \cup T_2;\mathbb{Z}) \to H_1(M;\mathbb{Z})$$

contains $\alpha_1 + \alpha_2 \in H_1(T_1 \cup T_2; \mathbb{Z})$ with $0 \neq \alpha_1 \in H_1(T_1; \mathbb{Z})$ and $0 \neq \alpha_2 \in H_1(T_2; \mathbb{Z})$. Now we treat M as a noncompact open hyperbolic 3-manifold, and consider T_1 and T_2 as two horotori of M. The following lemma gives some data that will instruct us to construct the mapped-in 2-complex $j\colon Z \looparrowright M$.

Lemma 5.6 *For any $\epsilon \in (0, 10^{-2})$, there exist $R_0 > 0$, such that for any $R > R_0$, there exist the following maps and homomorphisms*:

- *up to rechoosing the horotori T_1 and T_2 in M (by changing their heights), we have maps $i_{\partial,1}\colon \partial N_0 \to T_1$ and $i_{\partial,2}\colon \partial N_0 \to T_2$,*
- *$\boldsymbol{i}_1\colon H_1(N_\gamma^{(2)}; \mathbb{Z}) \to H_1(M; \mathbb{Z})$ and $\boldsymbol{i}_2\colon H_1(N_\gamma^{(2)}; \mathbb{Z}) \to H_1(M; \mathbb{Z})$,*
- *$\boldsymbol{i}\colon H_1(N; \mathbb{Z}) \to H_1(M; \mathbb{Z})$,*
- *$i_1, i_2\colon N_\gamma^{(1)} \to M$ (note that $N^{(1)} = N_\gamma^{(1)}$ holds),*

such that the following properties hold.

(1) *For any $s = 1, 2$, $i_{\partial,s}$ maps each triangle in ∂N_0 to a Euclidean geometric triangle in T_s such that each inner angle is ϵ-close to $\frac{\pi}{3}$, and the length of each edge lies in $[R, (1+\epsilon)R]$.*

(2) *For any $s = 1, 2$, the following diagram commutes*:

$$\begin{tikzcd}
H_1(\partial N_0; \mathbb{Z}) \arrow[r] \arrow[d, "(i_{\partial,s})_*"'] & H_1(N_\gamma^{(2)}; \mathbb{Z}) \arrow[d, "\boldsymbol{i}_s"] \\
H_1(T_s; \mathbb{Z}) \arrow[r] & H_1(M; \mathbb{Z})
\end{tikzcd}$$

(3) *The following diagram commutes*:

$$\begin{tikzcd}
H_1(N_\gamma^{(2)}; \mathbb{Z}) \arrow[r, "\boldsymbol{i}_1+\boldsymbol{i}_2"] \arrow[d] & H_1(M; \mathbb{Z}) \\
H_1(N; \mathbb{Z}) \arrow[ur, "\boldsymbol{i}"'] &
\end{tikzcd}$$

(4) *For any $s = 1, 2$, $i_s|_{N^{(0)}}$ is an embedding, $i_s|_{\partial N_0^{(1)}} = i_{\partial,s}|_{\partial N_0^{(1)}}$ and the following diagram commutes*:

$$\begin{tikzcd}
H_1(N_\gamma^{(1)}; \mathbb{Z}) \arrow[r, "(i_s)_*"] \arrow[d] & H_1(M; \mathbb{Z}) \\
H_1(N_\gamma^{(2)}; \mathbb{Z}) \arrow[ur, "\boldsymbol{i}_s"'] &
\end{tikzcd}$$

Here all undefined homomorphisms are induced by inclusions.

Proof By Lemma 5.5(3), the inclusion induced homomorphism $H_1(\partial N_0; \mathbb{Z}) \to H_1(N_\gamma; \mathbb{Z})$ is injective. So $H_1(N_\gamma; \mathbb{Z})$ has a direct summand $H \cong \mathbb{Z}^2$ that contains $H_1(\partial N_0; \mathbb{Z})$ as a finite index subgroup, with $H_1(N_\gamma; \mathbb{Z}) \cong H \oplus H'$ and let $[H : H_1(\partial N_0; \mathbb{Z})] = k$.

Recall that by Lemma 5.2, ∂N_0 is equipped with a triangulation induced by a geometric triangulation of the Euclidean torus ∂N_c, such that each triangle of ∂N_c is almost an equilateral triangle with length at most 2ϵ. We identify ∂N_0 and ∂N_c with the quotient of $\mathbb{R}^2$ by a lattice

$$\Lambda = A\mathbb{Z} + (B + C\omega_0)\mathbb{Z},$$

such that each triangle in ∂N_0 and ∂N_c corresponds to an equilateral triangle in $\mathbb{R}^2$ of edge length 1 (this identification is not an isometry). Here $A, B, C \in \mathbb{Z}$ with $A, C \neq 0$ and $\omega_0 = \frac{1}{2}(1 + \sqrt{3}i)$. Moreover, we can assume that the $\mathbb{R}$-coefficient null-homologous curve $l \subset \partial N_0$ corresponds to $A \in \Lambda$.

Similarly, by Lemma 5.2, for any $s = 1, 2$, T_s has a geodesic triangulation such that the following hold.

- Any inner angle of a triangle is $\frac{\epsilon}{10}$-close to $\frac{\pi}{3}$.
- There exists $r_s \in \big(0, \frac{\epsilon}{10}\big)$, such that all edges of triangles have length in $\big[r_s, \big(1 + \frac{\epsilon}{5}\big)r_s\big)$.
- The homology class $\alpha_s \in H_1(T_s; \mathbb{Z})$ is represented by the $d_s{}^{\text{th}}$ power of a simple closed geodesic l_s for some $d_s \in \mathbb{Z}_{>0}$, and l_s is contained in the 1-skeleton of this triangulation.

By the same process as above, we identify T_s with the quotient of $\mathbb{R}^2$ by a lattice

$$\Lambda_s = A_s\mathbb{Z} + (B_s + C_s\omega_0)\mathbb{Z},$$

with $A_s, C_s \neq 0$ and l_s corresponds to $A_s \in \Lambda_s$. By subdividing triangles, we can assume that $A_1d_1 = A_2d_2$ holds. Moreover, by rechoosing horotori parallel to T_1 and T_2 respectively, we can assume that $r_1 = r_2$ holds, and let $r = r_1 = r_2$.

Let D be the least common multiple of $A_1C_1d_1$ and $A_2C_2d_2$. For any $a \in \mathbb{Z}_{>0}$, we construct a map $i_{\partial,s} \colon \partial N_0 \to T_s$ as following. We have

$$\Lambda_s = A_s\mathbb{Z} + (B_s + C_s\omega)\mathbb{Z} > akD \cdot \Lambda = (akDA)\mathbb{Z} + (akD(B + C\omega_0))\mathbb{Z},$$

since

$$akDA = \Big(akA\frac{D}{A_s}\Big)A_s \tag{5-2}$$

and

$$akD(B + C\omega_0) = \Big(akB\frac{D}{A_s} - akCB_s\frac{D}{A_sC_s}\Big)A_s + \Big(akC\frac{D}{C_s}\Big)(B_s + C_s\omega_0). \tag{5-3}$$

So the scaling by akD gives a map from $\mathbb{R}^2/\Lambda$ to $\mathbb{R}^2/\Lambda_s$, and it maps each equilateral triangle of length 1 to an equilateral triangle of length akD. Since we identified ∂N_0 and T_s with $\mathbb{R}^2/\Lambda$ and $\mathbb{R}^2/\Lambda_s$ respectively, the akD-scaling map induces $i_{\partial,s} \colon \partial N_0 \to T_s$, such that it maps each triangle in ∂N_0 to a triangle in T_s of inner angle ϵ-close to $\frac{\pi}{3}$, and with edge length contained in $\big[akDr, \big(1 + \frac{\epsilon}{5}\big)akDr\big)$. If $R > R_0 = 2kDr/\epsilon$, there exists a positive integer a, such that $\big[akDr, \big(1 + \frac{\epsilon}{5}\big)akDr\big) \subset [R, (1 + \epsilon)R]$. So we can choose a such that item (1) holds for both $i_{\partial,1}$ and $i_{\partial,2}$.

Note that the simple closed curve l in ∂N_0 corresponds to $A \in \Lambda$, which is mapped to $akDA \in \Lambda_s$ via the akD-scaling map. Since $\alpha_s = d_s l_s$ corresponds to $d_s A_s \in \Lambda_s$, $(i_{\partial,s})_*$ maps l to

$$\Bigl(akA\frac{D}{d_s A_s}\Bigr)\alpha_s \in H_1(T_s;\mathbb{Z}).$$

Since we assumed $A_1 d_1 = A_2 d_2$ and $\alpha_1 + \alpha_2 = 0 \in H_1(M;\mathbb{Z})$, we have

$$(i_{\partial,1})_*(l) + (i_{\partial,2})_*(l) = akA\frac{D}{d_1 A_1}(\alpha_1+\alpha_2) = 0 \in H_1(M;\mathbb{Z}). \tag{5-4}$$

Now we define $\boldsymbol{i}_s \colon H_1(N_\gamma^{(2)};\mathbb{Z}) \to H_1(M;\mathbb{Z})$ for $s = 1, 2$. By (5-2) and (5-3),

$$(i_{\partial,s})_* \colon H_1(\partial N_0;\mathbb{Z}) \to H_1(T_s;\mathbb{Z})$$

maps each element to a k-multiple of an element in $H_1(T_s;\mathbb{Z})$. Since

$$H_1(N_\gamma^{(2)};\mathbb{Z}) \cong H_1(N_\gamma;\mathbb{Z}) = H \oplus H'$$

for some H containing $H_1(\partial N_0;\mathbb{Z})$ with $[H : H_1(\partial N_0;\mathbb{Z})] = k$, the homomorphism

$$(i_{\partial,s})_* \colon H_1(\partial N_0;\mathbb{Z}) \to H_1(T_s;\mathbb{Z})$$

uniquely extends to a homomorphism $\boldsymbol{h}_s \colon H \to H_1(T_s;\mathbb{Z})$, and we define $\boldsymbol{i}_s \colon H_1(N_\gamma^{(2)};\mathbb{Z}) \to H_1(M;\mathbb{Z})$ to be

$$H_1(N_\gamma^{(2)};\mathbb{Z}) \to H \xrightarrow{\boldsymbol{h}_s} H_1(T_s;\mathbb{Z}) \to H_1(M;\mathbb{Z}).$$

Here the first homomorphism is the projection to the direct summand H, and the third homomorphism is induced by inclusion. It is straight forward to check that the commutative diagram in item (2) holds.

Note that N deformation retracts to $N_0 = N_\gamma \cup \mathcal{N}(\gamma)$ and $\mathcal{N}(\gamma)$ is a solid torus. Once we prove that the meridian c of $\mathcal{N}(\gamma)$ lies in the kernel of $\boldsymbol{i}_1 + \boldsymbol{i}_2 \colon H_1(N_\gamma^{(2)};\mathbb{Z}) \to H_1(M;\mathbb{Z})$, then $\boldsymbol{i}_1 + \boldsymbol{i}_2$ induces a homomorphism $\boldsymbol{i} : H_1(N;\mathbb{Z}) \to H_1(M;\mathbb{Z})$ and the commutative diagram in item (3) holds. Recall that Lemma 5.5(2) implies that $[l]-[c]$ is a torsion element in $H_1(N_\gamma^{(2)};\mathbb{Z}) = H \oplus H'$. Since $H \cong \mathbb{Z}^2$, we have $[c]-[l] \in H'$. By the definition of $\boldsymbol{i}_1$ and $\boldsymbol{i}_2$ above, $\boldsymbol{i}_1([c]-[l]) = \boldsymbol{i}_2([c]-[l]) = 0$ holds. So we have

$$(\boldsymbol{i}_1+\boldsymbol{i}_2)([c]) = (\boldsymbol{i}_1+\boldsymbol{i}_2)([l]) + (\boldsymbol{i}_1+\boldsymbol{i}_2)([c]-[l]) = (\boldsymbol{i}_1+\boldsymbol{i}_2)([l]) = (i_{\partial,1})_*([l]) + (i_{\partial,2})_*([l]) = 0.$$

Here the third equation follows from item (2) and the fourth equation follows from (5-4).

To define $i_s \colon N_\gamma^{(1)} \to M$, we first define $i_s|_{\partial N_0^{(1)}} = i_{\partial,s}|_{\partial N_0^{(1)}}$, and arbitrarily extend i_s to a maximal subcomplex K of $N_\gamma^{(1)}$ that deformation retracts to ∂N_0. Then since edges in $N_\gamma^{(1)} \setminus K$ form a basis of $H_1(N_\gamma^{(1)};\mathbb{Z})/H_1(\partial N_0^{(1)};\mathbb{Z})$, we can extend i_s to $N_\gamma^{(1)}$ so that the commutative diagram in item (4) holds. Finally, we slightly perturb i_s if necessary, so that $i_s|_{N^{(0)}}$ is an embedding. □

Now we give some notation on the geometry of N. Most of the items (except item (3)) are similar to those of [23, Notation 4.4].

Notation 5.7 (1) For any oriented edge e_{ij} (or e_{ijk}) in $N^{(1)}$, let $\vec{v}_{ij}$ ($\vec{v}_{ijk}$) be the unit tangent vector of e_{ij} (e_{ijk}) based at n_i. By Construction 5.4, $\vec{v}_{ijk}$ lies in the plane in $T_{v_i}M$ containing $\vec{v}_{ij}$ and $\vec{v}_{ik}$, and is $\epsilon/200$-close to $(\vec{v}_{ij}+\vec{v}_{ik})/|\vec{v}_{ij}+\vec{v}_{ik}|$. For any marked geodesic triangle Δ_{ijk} in $N^{(2)}$, let

$$\vec{n}_{ijk}=\frac{\vec{v}_{ij}\times\vec{v}_{ik}}{|\vec{v}_{ij}\times\vec{v}_{ik}|},$$

then it is a normal vector of Δ_{ijk} at n_i, and we have a frame

$$\boldsymbol{F}_{ijk}=(n_i,\vec{v}_{ij},\vec{n}_{ijk})\in \mathrm{SO}(N)_{n_i}.$$

For any marked modified triangle Δ_{ijk} defined in Construction 5.4, with e_{ij} contained in ∂N_0 (which is not an edge of Δ_{ijk}), the frames $\boldsymbol{F}_{kij}$ and $\boldsymbol{F}_{kji}$ are defined as in the previous case. For $\boldsymbol{F}_{ijk}$, let

$$\vec{n}_{ijk}=\frac{\vec{v}_{ijk}\times\vec{v}_{ik}}{|\vec{v}_{ijk}\times\vec{v}_{ik}|}=\frac{\vec{v}_{ij}\times\vec{v}_{ik}}{|\vec{v}_{ij}\times\vec{v}_{ik}|},$$

and we get a frame

$$\boldsymbol{F}_{ijk}=(n_i,\vec{v}_{ijk},\vec{n}_{ijk})\in \mathrm{SO}(N)_{n_i}.$$

For each frame $\boldsymbol{F}_{ijk}$, we denote $-\boldsymbol{F}_{ijk}=(n_i,-\vec{v}_{ij},-\vec{n}_{ijk})$, then we have a finite collection of frames in N:

$$\mathscr{F}_N=\{\pm\boldsymbol{F}_{ijk}\mid\Delta_{ijk}\text{ is a marked triangle in }N^{(2)}\}.$$

(2) For any $s=1,2$, let $m_{k,s}=i_s(n_k)$ and $V_{M,s}=i_s(V_N)$. We take an isometry $t_s\colon TN|_{V_N}\to TM|_{V_{M,s}}$ that descends to $i_s\colon V_N\to V_{M,s}$, such that the following holds for any $n_k\in V_{N,\partial}$. At $n_k\in V_{N,\partial}$, there is a frame $(\vec{v}_k,\vec{n}_k)$ such that $\vec{v}_k$ is tangent to the direction of $l\subset\partial N_c$, and $\vec{v}_k\times\vec{n}_k$ points up straightly into the cusp. We require that t_s maps $\vec{v}_k$ and $\vec{n}_k$ to $\vec{v}_k'$ and $\vec{n}_k'$ based at $m_{k,s}$ respectively, such that $\vec{v}_k'$ is tangent to the direction of $l_s\subset T_s$ and $\vec{v}_k'\times\vec{n}_k'$ points up straightly into the cusp. Then t_s induces an SO(3)-equivariant isomorphism $t_s\colon \mathrm{SO}(N)|_{V_N}\to\mathrm{SO}(M)|_{V_{M,s}}$, denoted by the same notation. We denote $\boldsymbol{F}^M_{ijk,s}=(m_{i,s},\vec{v}^M_{ij,s},\vec{n}^M_{ijk,s})=t_s(\boldsymbol{F}_{ijk})\in\mathrm{SO}(M)|_{m_{i,s}}$, and let

$$\mathscr{F}_{M,s}=t_s(\mathscr{F}_N)\subset\mathrm{SO}(M).$$

(3) Since $N^{(2)}$ contains finitely many 2-cells, there exists $\phi_0\in(0,\pi)$, such that all inner angles of bigons and triangles in $N^{(2)}$ and all dihedral angles between adjacent 2-cells of $N^{(2)}$ lie in $[\phi_0,\pi]$.

Remark 5.8 Let n_i and n_j be two vertices of ∂N_0 such that e_{ij} is contained in ∂N_0, and let n_k be the vertex not contained in ∂N_0 such that n_i, n_j and n_k span an triangle in the original triangulation of N_0 and it lies to the left of e_{ij}, as in Figure 2. We give coordinates of $T^1_{n_i}N$ such that $\vec{v}_{ij}\approx(1,0,0)$ with vanishing second coordinate and the vector pointing to the cusp is $(0,0,1)$.

Let

$$\vec{v}_0=\left(\frac{1}{2},\frac{1}{2\sqrt{3}},\frac{r}{6}-\frac{1-e^{-\frac{2\sqrt{6}}{3}r}}{2r}\right).$$

Assuming all triangles in $T_c \subset N$ are equilateral triangles of length r, an elementary computation gives

$$\vec{v}_{ik} = \vec{v}_1 = \frac{\vec{v}_0}{\|\vec{v}_0\|} \approx \left(\frac{1}{2}, \frac{1}{2\sqrt{3}}, -\frac{\sqrt{6}}{3}\right), \quad \vec{v}_{ijk} = \vec{v}_2 \approx \frac{\vec{v}_1 + (1,0,0)}{\|\vec{v}_1 + (1,0,0)\|} \approx \left(\frac{\sqrt{3}}{2}, \frac{1}{6}, -\frac{\sqrt{2}}{3}\right),$$
$$\vec{n}_{ijk} = \vec{v}_3 = \frac{\vec{v}_2 \times \vec{v}_1}{\|\vec{v}_2 \times \vec{v}_1\|} \approx \left(0, \frac{2\sqrt{2}}{3}, \frac{1}{3}\right).$$

In this remark, the actual vectors are all $\frac{\epsilon}{20}$-close to their numerical approximations above.

Remark 5.9 Although the frame bundle of a compact orientable 3-manifold N with connected torus boundary is trivial, there may not be a trivialization of $\mathrm{SO}(N)$ such that its restriction to ∂N has third vector pointing outward. So we do not have a homological instruction as good as [23, Proposition 4.5], which reduces the degree of virtual domination by a half.

5.2 Construction of the immersion $j : Z \looparrowright M$

In this section, we construct the π_1-injective immersion $j : Z \looparrowright M$. Since Z is a 2-complex, we will inductively construct the 0-, 1-, 2-skeletons of Z and the restrictions of j on these skeletons. Throughout this section, we fix a small number $\epsilon \in (0, 10^{-2})$ and a sufficiently large number $R \in \left(\frac{1}{\epsilon}, +\infty\right)$ such that all (finitely many) constructions below (invoking Theorems 2.11 and 2.14) are applicable.

Construction 5.10 We define Z^0 to be a finite set $\{v_{1,1}, v_{1,2}, v_{2,1}, v_{2,2}, \dots, v_{l,1}, v_{l,2}\}$, whose cardinality doubles the cardinality of $V_N = N^{(0)}$. Then we define $j^0 : Z^0 \to M$ by $j^0(v_{k,s}) = m_{k,s} = i_s(n_k) \in M$ for any $k \in \{1, 2, \dots, l\}$ and $s \in \{1, 2\}$.

Here we take two copies of $N^{(0)}$ since we work with two boundary components T_1 and T_2 of M. Now we construct the 1-complex Z^1 of Z.

Construction 5.11 For any unoriented edge e_{ij} (or e_{ijk}) in $N^{(1)}$, it gives two edges $e^Z_{ij,1}$ and $e^Z_{ij,2}$ (or $e^Z_{ijk,1}$ and $e^Z_{ijk,2}$) in Z^1, such that $e^Z_{ij,s}$ (or $e^Z_{ijk,s}$) connects $v_{i,s}$ and $v_{j,s}$ for $s = 1, 2$. So Z^1 consists of two isomorphic components Z^1_1 and Z^1_2, and each of them is isomorphic to $N^{(1)}$. For the vertices and edges of Z^1 corresponding to vertices and edges in $N^{(1)} \cap \partial N_0$, they form a subcomplex of Z^1 and we denote it by $\partial_p Z^1$.

A picture of Z^1 near a vertex of $\partial_p Z^1$ is shown in Figure 1.

The indices of vertices induce a total order on the set of vertices in $N^{(0)}$, and also induce total orders on the vertex set of Z^1_1 and the vertex set of Z^1_2. Any edge e_{ij} (or e_{ijk}) in $N^{(1)}$ between n_i and n_j with $i < j$ has an orientation that goes from n_i to n_j, and we always fix such a preferred orientation. Edges of Z^1_1 and Z^1_2 have identical orientations.

The map $j^1 : Z^1 \to M$ on 1-skeleton is given in the following construction, which consists of two maps $j^1_1 : Z^1_1 \to M$ and $j^1_2 : Z^1_2 \to M$ on the two (identical) components of Z^1.

Construction 5.12 For any $s = 1, 2$, we do the following construction.

(1) For each oriented edge $e^Z_{ij,s} \subset Z^1_s$ with $i < j$ contained in $\partial_p Z^1_s$, we map it to the oriented geodesic segment homotopic to $i_s(e_{ij})$ relative to endpoints, via a homeomorphism. Note that these geodesic segments have length contained in $[2\log R, 2\log R + 4\epsilon]$ (by Lemma 5.6(1)).

(2) For each oriented edge $e^Z_{ij,s} \subset Z^1_s$ with $i < j$ not contained in $\partial_p Z^1_s$, we apply Theorem 2.14 to construct a ∂-framed segment $\mathfrak{s}_{ij,s}$ in M from $m_{i,s}$ to $m_{j,s}$ such that the following conditions hold, and we map $e^Z_{ij,s}$ to the carrier of $\mathfrak{s}_{ij,s}$ via a homeomorphism.

   (a) The length and phase of $\mathfrak{s}_{ij,s}$ are $\frac{\epsilon}{10}$-close to $2R$ and 0 respectively, and the height of $\mathfrak{s}_{ij,s}$ is at most $2\log R + 2$.

   (b) Let k be the smallest index such that n_i, n_j and n_k form a triangle in $N^{(2)}$, the initial and terminal frames of $\mathfrak{s}_{ij,s}$ are $\frac{\epsilon}{10}$-close to $\boldsymbol{F}^M_{ijk,s}$ and $-\boldsymbol{F}^M_{jik,s}$ respectively.

   (c) The relative homology class of the carrier of $\mathfrak{s}_{ij,s}$ in $H_1(M, \{m_{i,s}, m_{j,s}\}; \mathbb{Z})$ equals the relative homology class of $i_s(e_{ij})$.

(3) For any oriented edge $e^Z_{ijk,s} \subset Z^1_s$ with $i < j$ that corresponds to $e_{ijk} \subset N^{(1)}$, we apply Theorem 2.14 to construct a ∂-framed segment $\mathfrak{s}_{ijk,s}$ in M from $m_{i,s}$ to $m_{j,s}$ such that the following conditions hold, and we map $e^Z_{ijk,s}$ to the carrier of $\mathfrak{s}_{ijk,s}$ via a homeomorphism.

   (a) The length and phase of $\mathfrak{s}_{ijk,s}$ are $\frac{\epsilon}{10}$-close to $2R$ and 0 respectively, and the height of $\mathfrak{s}_{ijk,s}$ is at most $2\log R + 2$.

   (b) The initial and terminal frames of $\mathfrak{s}_{ijk,s}$ are $\frac{\epsilon}{10}$-close to $\boldsymbol{F}^M_{ijk,s}$ and $-\boldsymbol{F}^M_{jik,s}$ respectively.

   (c) The relative homology class of the carrier of $\mathfrak{s}_{ijk,s}$ in $H_1(M, \{m_{i,s}, m_{j,s}\}; \mathbb{Z})$ equals the relative homology class of $i_s(e_{ijk})$.

Figure 1 shows the geometry of $j^1(Z^1_s)$ near a vertex $v_{i,s}$ corresponding to $n_i \in \partial N_0$.

Remark 5.13 (1) In Construction 5.12(1), the tangent vector of $j^1(e^Z_{ij,s})$ at $m_{i,s}$ is almost $(0,0,1)$ (with respect to the preferred coordinate system), while the tangent vector of $\boldsymbol{F}^M_{ijk,s}$ is almost $(1,0,0)$. This is the crucial reason why we need extra edges (e_{ijk} and $e^Z_{ijk,s}$) in $N^{(1)}$ and Z^1, which takes care of this difference. We map $e^Z_{ij,s} \subset \partial_p Z^1_s$ to the geodesic segment in the relative homotopy class of $i_s(e_{ij})$, instead of prescribing its tangent vectors at initial and terminal points, since we need to construct proper maps between 3-manifolds with tori boundary.

(2) In Construction 5.12(2), if we take another vertex $n_{k'}$ of N such that n_i, n_j and $n_{k'}$ also form a triangle in $N^{(2)}$, then we can rechoose frames of $\mathfrak{s}_{ij,s}$ so that it still satisfies item (2), with respect to $\boldsymbol{F}^{M,s}_{ijk'}$ and $-\boldsymbol{F}^{M,s}_{jik'}$ in item (2)(b). The reason is that $\boldsymbol{F}_{ijk'} = \boldsymbol{F}_{ijk} \cdot A$ and $-\boldsymbol{F}_{jik'} = (-\boldsymbol{F}_{jik}) \cdot A$ for the same $A \in \mathrm{SO}(3)$, while t_s is SO(3)-equivariant.

(3) In Construction 5.12(3), by our construction of the triangulation of N near ∂N_0 in Construction 5.3(1), the third coordinate of $\vec{v}^M_{ij,s}$ is at most $\epsilon/100$. Up to changing coordinate, we assume that $\vec{v}^M_{ij,s}$ has trivial

second coordinate, then it is $\epsilon/100$-close to $(1,0,0)$. We suppose that n_k lies to the left of e_{ij}, as shown in Figure 2, then $\vec{v}^{M}_{ik,s}$ is $\frac{\epsilon}{20}$-close to $\bigl(\frac{1}{2},\frac{\sqrt{3}}{6},-\frac{\sqrt{6}}{3}\bigr)$, and the initial frame of $\mathfrak{s}_{ijk,s}$ is $\frac{\epsilon}{5}$-close to

$$\biggl(\biggl(\frac{\sqrt{3}}{2},\frac{1}{6},-\frac{\sqrt{2}}{3}\biggr),\biggl(0,\frac{2\sqrt{2}}{3},\frac{1}{3}\biggr)\biggr).$$

See also Remark 5.8.

Moreover, the common perpendicular vector of $j_1(e^{Z}_{ij,s})$ and $j^1(e^{Z}_{ijk,s})$ at $m_{i,s}$ is $\frac{\epsilon}{5}$-close to $\bigl(\frac{-1}{2\sqrt{7}},\frac{3\sqrt{3}}{2\sqrt{7}},0\bigr)$. If we consider Figure 2 as a picture of $j^1_s(Z^1)$ in M, the dihedral angle between the hyperplane determined the above vector and the geodesic triangle homotopic to T_s (relative to vertices) is ϵ-close to

$$\arccos\biggl(\biggl(-\frac{1}{2\sqrt{7}},\frac{3\sqrt{3}}{2\sqrt{7}},0\biggr)\cdot\biggl(-\frac{\sqrt{3}}{2},-\frac{1}{2},0\biggr)\biggr)=\arccos\biggl(-\frac{\sqrt{3}}{2\sqrt{7}}\biggr)\approx 0.606\pi.$$

Note that this computation will be crucial for our proof of Theorem 5.17 in Section 6.

Before we construct the 2-complex Z, we need the following lemma that proves certain closed curves arising from Construction 5.12 are good curves.

Lemma 5.14 *Under the conditions in Construction 5.12, if R is large enough, we have the following good curves.*

(1) *For any bigon B_{ijk} in $N^{(2)}$, the concatenation of $\overline{j^1(e^{Z}_{ij,s})}, j^1(e^{Z}_{ijk,s})$ is homotopic to a null-homologous (R_{ij},ϵ)-good curve $\gamma_{ijk,s}$ of height at most $2\log R+3$ in M, with*

$$R_{ij}=R+\log l_{ij}-\log\frac{6}{3+\sqrt{2}}.$$

Here l_{ij} denotes the length of the Euclidean geodesic segment $i_{\partial,s}(e_{ij})$ and $l_{ij}\in[R,(1+\epsilon)R]$.

(2) *For any triangle Δ_{ijk} in $N^{(2)}$ with vertices n_i, n_j and n_k, the concatenation of $j^1(e^{Z}_{ij,s})$, $j^1(e^{Z}_{jk,s})$ and $j^1(e^{Z}_{ki,s})$ is an (R_{ijk},ϵ)-good curve $\gamma_{ijk,s}$ of height at most $2\log R+3$ in M, with*

$$R_{ijk}=3R-(I(\pi-\theta_{ijk})+I(\pi-\theta_{jki})+I(\pi-\theta_{kij})).$$

(Here $j^1(e^{Z}_{ij,s})$ is replaced by $j^1(e^{Z}_{ijk,s})$ if Δ_{ijk} is a modified triangle.) Here θ_{ijk} is the inner angle of the triangle Δ_{ijk} at vertex n_i. Moreover, if Δ_{ijk} is contained in $N^{(2)}_{\gamma}$, $\gamma_{ijk,s}$ is null-homologous in M; if Δ_{ijk} is not contained in $N^{(2)}_{\gamma}$, then $\gamma_{ijk,1}\cup\gamma_{ijk,2}$ is null-homologous in M.

Note that if R is large enough, all good curves in this lemma have length contained in $[2R,6R]$.

Proof (1) Note that $j^1(e^{Z}_{ijk,s})$ is the carrier of $\mathfrak{s}_{ijk,s}$, and we assumed that $\vec{v}^{M}_{ij,s}$ has trivial second coordinate as in Remark 5.13(2). So by Remark 5.13(3), the initial and terminal frames of $\mathfrak{s}_{ijk,s}$ are $\frac{\epsilon}{5}$-close to

$$\biggl(\biggl(\frac{\sqrt{3}}{2},\frac{1}{6},-\frac{\sqrt{2}}{3}\biggr),\biggl(0,\frac{2\sqrt{2}}{3},\frac{1}{3}\biggr)\biggr)\quad\text{and}\quad\biggl(\biggl(\frac{\sqrt{3}}{2},-\frac{1}{6},\frac{\sqrt{2}}{3}\biggr),\biggl(0,\frac{2\sqrt{2}}{3},\frac{1}{3}\biggr)\biggr)$$

respectively (with respect to preferred coordinates). For $\phi = \pi - \arcsin \frac{1}{\sqrt{7}}$, the initial and terminal frames of the frame rotation $\mathfrak{s}_{ijk,s}(\phi)$ are $\frac{\epsilon}{5}$-close to

$$\left(\left(\frac{\sqrt{3}}{2}, \frac{1}{6}, -\frac{\sqrt{2}}{3}\right), \left(\frac{1}{2\sqrt{7}}, -\frac{3\sqrt{3}}{2\sqrt{7}}, 0\right)\right) \quad \text{and} \quad \left(\left(\frac{\sqrt{3}}{2}, -\frac{1}{6}, \frac{\sqrt{2}}{3}\right), \left(-\frac{1}{2\sqrt{7}}, -\frac{3\sqrt{3}}{2\sqrt{7}}, 0\right)\right)$$

respectively.

For $j^1(e^Z_{ij,s})$, its initial and terminal directions are $(2/R)$-close to $(0,0,1)$ and $(0,0,-1)$ respectively. Since $j^1(e^Z_{ij,s})$ is a geodesic segment in a cusp, it parallel transports $(0,1,0)$ to $(0,1,0)$. We obtain a ∂-framed segment $\mathfrak{t}$ with phase 0, by equipping $j^1(e^Z_{ij,s})$ with initial and terminal framings $(0,1,0)$. Then for $\phi' = \pi + \arcsin \frac{1}{2\sqrt{7}}$, its ϕ'-rotation $\mathfrak{t}(\phi')$ is a ∂-framed segment with 0-phase, with initial and terminal frames $(4/R)$-close to

$$\left((0,0,1), \left(\frac{1}{2\sqrt{7}}, -\frac{3\sqrt{3}}{2\sqrt{7}}, 0\right)\right) \quad \text{and} \quad \left((0,0,-1), \left(-\frac{1}{2\sqrt{7}}, -\frac{3\sqrt{3}}{2\sqrt{7}}, 0\right)\right)$$

respectively.

If $R > \frac{20}{\epsilon}$, then $\overline{\mathfrak{t}(\phi')}, \mathfrak{s}_{ijk,s}(\phi)$ is a $\left(\frac{2}{5}\epsilon\right)$-consecutive cycle of ∂-framed segments, with both bending angles $\left(\frac{2}{5}\epsilon\right)$-close to $\arccos \frac{\sqrt{2}}{3}$. By elementary hyperbolic geometry, the length of $\overline{\mathfrak{t}(\phi')}$ equals

$$2 \log \frac{\sqrt{l_{ij}^2 + 4} + l_{ij}}{2}.$$

Lemma 3.2(2) implies that the concatenation of $\overline{j^1(e^Z_{ij,s})}, j^1(e^Z_{ijk,s})$ is homotopic to a closed geodesic $\gamma_{ijk,s}$ with complex length 2ϵ-close to

$$2R_{ij} = 2R + 2\log l_{ij} - 2\log \frac{6}{3+\sqrt{2}}.$$

So $\gamma_{ijk,s}$ is an (R_{ij}, ϵ)-good curve.

We take large enough R, so that heights of T_1 and T_2 are at most $\log R$. Since the heights of $\mathfrak{s}_{ijk,s}$ and $\mathfrak{t}$ are at most $2\log R + 2$, Lemma 3.4 implies the height of $\gamma_{ijk,s}$ is at most $2\log R + 3$. By the homological conditions in Construction 5.12(1) and (3)(c), $j^1(e^Z_{ij,s})$ and $j^1(e^Z_{ijk,s})$ represent the same relative homology class, so $\gamma_{ijk,s}$ is a null-homologous closed geodesic in M.

(2) We prove this result in the case that Δ_{ijk} is not a modified triangle, and the case of modified triangles can be proved similarly.

By our constructions of $j^1(e^Z_{ij,s})$, $j^1(e^Z_{jk,s})$ and $j^1(e^Z_{ki,s})$ in Construction 5.12(2), we equip them with initial and terminal frames as following to get three ∂-framed segments.

- Equip $j^1(e^Z_{ij,s})$ with initial and terminal frames that are $\frac{\epsilon}{10}$-close to $\vec{n}^M_{ijk,s}$ and $-\vec{n}^M_{jik,s}$ respectively.
- Equip $j^1(e^Z_{jk,s})$ with initial and terminal frames that are $\frac{\epsilon}{10}$-close to $\vec{n}^M_{jki,s}$ and $-\vec{n}^M_{kji,s}$ respectively.
- Equip $j^1(e^Z_{ki,s})$ with initial and terminal frames that are $\frac{\epsilon}{10}$-close to $\vec{n}^M_{kij,s}$ and $-\vec{n}^M_{ikj,s}$ respectively.

By Construction 5.12(2) and Remark 5.13(2), the phases of these ∂-framed segments are $\frac{\epsilon}{10}$-close to 0. Since $\vec{n}^M_{ijk,s} = -\vec{n}^M_{ikj,s}$, these three ∂-framed segments form a $\frac{\epsilon}{5}$-consecutive cycle. Then Lemma 3.2(2) implies that the concatenation is homotopic to a closed geodesic $\gamma_{ijk,s}$ with complex length 2ϵ-close to

$$2R_{ijk} = 6R - 2(I(\pi - \theta_{ijk}) + I(\pi - \theta_{jki}) + I(\pi - \theta_{kij})).$$

So $\gamma_{ijk,s}$ is an (R_{ijk}, ϵ)-good curve. The height bound of $\gamma_{ijk,s}$ follows from the argument in (1).

By Construction 5.12(2), $\gamma_{ijk,s}$ is homologous to the concatenation of $i_s(e_{ij})$, $i_s(e_{jk})$ and $i_s(e_{ki})$. If Δ_{ijk} is a triangle in $N_\gamma^{(2)}$, Lemma 5.6(4) implies that $\gamma_{ijk,s}$ is null-homologous in M. If Δ_{ijk} is not contained in $N_\gamma^{(2)}$, it is a meridian disc of $\mathcal{N}(\gamma)$, then Lemma 5.6(4) implies that $\gamma_{ijk,1} \cup \gamma_{ijk,2}$ is homologous to $\boldsymbol{i}_1(\partial\Delta_{ijk}) + \boldsymbol{i}_2(\partial\Delta_{ijk})$, which is null-homologous in M by Lemma 5.6(3). □

Now we construct the 2-complex Z, by adding surfaces to two copies of Z^1. Rigorously speaking, two copies of Z^1 are not the 1-skeleton of Z as a CW-complex, but we still call it the 1-skeleton of Z, for our convenience. The map $j^1\colon Z^1 \to M$ and the construction of Z below automatically give the desired immersion $j\colon Z \looparrowright M$.

Construction 5.15 We take R' to be an integer greater than all the R_{ij} and R_{ijk} in Lemma 5.14.

(1) Recall that Z^1 has two components: Z^1_1 and Z^1_2, and each of them is isomorphic to $N^{(1)}$. The 1-skeleton $Z^{(1)}$ of Z consists of two copies of Z^1, so we have $Z^{(1)} = Z_1^{1,1} \cup Z_2^{1,1} \cup Z_1^{1,2} \cup Z_2^{1,2}$. For any $s = 1, 2$, the restriction of j to $Z_s^{1,1}$ and $Z_s^{1,2}$ equals $j^1|_{Z^1_s}$. We denote the two copies of $\partial_p Z^1$ in $Z^{(1)}$ by $\partial_p Z^{(1)}$.

(2) For any triangle Δ_{ijk} in ∂N_0 with vertices n_i, n_j and n_k, and any component $Z_s^{1,t}$ of $Z^{(1)}$ with $s, t \in \{1, 2\}$, we paste a triangle $\Delta^{Z,t}_{ijk,s}$ to $Z_s^{1,t}$ along the concatenation of edges $e^Z_{ij,s}$, $e^Z_{jk,s}$ and $e^Z_{ki,s}$ in $Z_s^{1,t}$. Since $e^Z_{ij,s}$ is mapped to a path homotopic to $i_s(e_{ij})$ (Construction 5.12(1)) and $i_s|_{\partial N_0^{(1)}}$ extends to $i_{\partial,s}\colon \partial N_0 \to T_s$ (Lemma 5.6(4)), the j^1-image of this concatenation is null-homotopic in M, so we map the triangle $\Delta^{Z,t}_{ijk,s}$ to the corresponding totally geodesic triangle in M.

(3) For any bigon B_{ijk} in $N_\gamma^{(2)}$ containing an edge $e_{ij} \subset \partial N_0$ and any $s = 1, 2$, we do the following construction.

By Lemma 5.14(1), the concatenation of $\overline{j^1(e^Z_{ij,s})}$, $j^1(e^Z_{ijk,s})$ is homotopic to a null-homologous (R_{ij}, ϵ)-good curve $\gamma_{ijk,s}$ in M, via a nearly geodesic two-cornered annulus $A_{ijk,s}$ (see Figure 3, left). Let $\vec{w}_{ijk,s}$ be the tangent vector of the shortest geodesic in $A_{ijk,s}$ from $\gamma_{ijk,s}$ to $m_{i,s}$, and let

$$\vec{v}_{ijk,s} = \vec{w}_{ijk,s} + (1 + \pi i) \in N^1(\sqrt{\gamma_{ijk,s}}).$$

By Proposition 2.11 and Remark 2.12, two copies of $\gamma_{ijk,s}$ bound an (R_{ij}, R', ϵ)-nearly geodesic subsurface $S_{ijk,s} \looparrowright M$, such that the following hold.

(a) The two feet of $S_{ijk,s}$ on two copies of $\gamma_{ijk,s}$ are both (ϵ/R)-close to $\vec{v}_{ijk,s}$.

(b) Any essential path in $S_{ijk,s}$ with end points in $\partial S_{ijk,s}$ must have combinatorial length (with respect to the decomposition of $S_{ijk,s}$ to pants and hamster wheels) at least $R'e^{R'/2}$.

We identify the two boundary components of $S_{ijk,s}$ with the two copies of concatenations of $e^Z_{ij,s}$, $e^Z_{jk,s}$ and $e^Z_{ki,s}$ in $Z_s^{1,1}$ and $Z_s^{1,2}$ respectively. The restriction of j on $S_{ijk,s}$ is naturally defined by two copies of the nearly geodesic 2-cornered annulus $A_{ijk,s}$ and the above (R_{ij}, R', ϵ)-nearly geodesic subsurface $S_{ijk,s} \looparrowright M$.

(4) For any triangle Δ_{ijk} in $N_\gamma^{(2)}$ not contained in ∂N_0 and any $s = 1, 2$, we do the following construction.

By Lemma 5.14(2), the concatenation of $j^1(e^Z_{ij,s})$, $j^1(e^Z_{jk,s})$ and $j^1(e^Z_{ki,s})$ is homotopic to a null-homologous (R_{ijk}, ϵ)-good curve $\gamma_{ijk,s}$ in M, via a nearly geodesic three-cornered annulus $A_{ijk,s}$ (see Figure 3, right). Let $\vec{w}_{ijk,s}$ be the tangent vector of the shortest geodesic in $A_{ijk,s}$ from $\gamma_{ijk,s}$ to $m_{i,s}$, and let $\vec{v}_{ijk,s} = \vec{w}_{ijk,s} + (1 + \pi i) \in N^1(\sqrt{\gamma_{ijk,s}})$. By Proposition 2.11 and Remark 2.12, two copies of $\gamma_{ijk,s}$ bound an (R_{ijk}, R', ϵ)-nearly geodesic subsurface $S_{ijk,s} \looparrowright M$, such that conditions (a) and (b) in item (3) hold.

We identify the two boundary components of $S_{ijk,s}$ with the two copies of concatenations $e^Z_{ij,s}$, $e^Z_{jk,s}$ and $e^Z_{ki,s}$ in $Z_s^{1,1}$ and $Z_s^{1,2}$ respectively. The restriction of j on $S_{ijk,s}$ is naturally defined by two copies of the nearly geodesic 3-cornered annulus $A_{ijk,s}$ and the above (R_{ijk}, R', ϵ)-nearly geodesic subsurface $S_{ijk,s} \looparrowright M$.

(5) Up until now, the 2-complex we have constructed has (at least) two components, one containing $Z_1^{1,1} \cup Z_1^{1,2}$ and one containing $Z_2^{1,1} \cup Z_2^{1,2}$. For any triangle Δ_{ijk} in $N^{(2)}$ not contained in $N_\gamma^{(2)}$, we do the following construction.

By Lemma 5.14(3), for any $s = 1, 2$, the concatenation $j^1(e^Z_{ij,s})$, $j^1(e^Z_{jk,s})$, $j^1(e^Z_{ki,s})$ is homotopic to an (R_{ijk}, ϵ)-good curve $\gamma_{ijk,s}$ via a three-cornered annulus $A_{ijk,s}$ in M, and $\gamma_{ijk,1} \cup \gamma_{ijk,2}$ is null-homologous in M. Let $\vec{w}_{ijk,s}$ be the tangent vector of the shortest geodesic in $A_{ijk,s}$ from $\gamma_{ijk,s}$ to $m_{i,s}$, and let $\vec{v}_{ijk,s} = \vec{w}_{ijk,s} + (1 + \pi i) \in N^1(\sqrt{\gamma_{ijk,s}})$. By Proposition 2.11 and Remark 2.12, two copies of $\gamma_{ijk,1} \cup \gamma_{ijk,2}$ bound an (R_{ijk}, R', ϵ)-nearly geodesic surface $S_{ijk} \looparrowright M$, such that conditions (a) and (b) in item (3) hold, with $S_{ijk,s}$ replaced by S_{ijk}.

We identify the four boundary components of S_{ijk} with two copies of concatenations of $e^Z_{ij,1}, e^Z_{jk,1}, e^Z_{ki,1}$ and $e^Z_{ij,2}, e^Z_{jk,2}, e^Z_{ki,2}$ in $Z_1^{1,1}, Z_1^{1,2}$ and $Z_2^{1,1}, Z_2^{1,2}$ respectively. The restriction of j on S_{ijk} is naturally defined by two copies of the nearly geodesic 3-cornered annuli $A_{ijk,1} \cup A_{ijk,2}$, and the above (R_{ijk}, R', ϵ)-nearly geodesic subsurface $S_{ijk} \looparrowright M$.

Note that the surfaces $S_{ijk,s}$ and S_{ijk} in Construction 5.15(3), (4) and (5) may not be connected, so Z may not be connected and has at most four connected components. One can actually work harder as in [13] to make sure these surfaces are connected, but we choose to save some work here.

Alternatively, Z is obtained from four copies of $N^{(2)}$, denoted by $N_1^{(2),1}$, $N_2^{(2),1}$, $N_1^{(2),2}$ and $N_2^{(2),2}$ respectively, by making the following modifications.

(1) Each triangle in $N^{(2)} \cap \partial N_0$ is not modified.

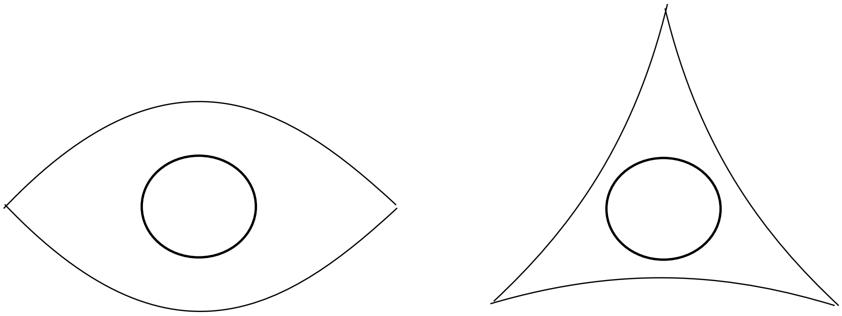


Figure 3: The two-cornered annulus and three-cornered annulus in Construction 5.15.

(2) For any bigon B_{ijk} or triangle Δ_{ijk} in $N_\gamma^{(2)} \setminus \partial N_0$ and any $s = 1, 2$, the two copies of B_{ijk} or Δ_{ijk} in $N_s^{(2),1}$, $N_s^{(2),2}$ are replaced by a compact orientable surface $S_{ijk,s}$ with two boundary components.

(3) For each triangle Δ_{ijk} in $N^{(2)}$ not contained in $N_\gamma^{(2)}$, its four copies are replaced by a compact orientable surface S_{ijk} with four boundary components.

5.3 Construction of virtual proper domination

In this section, we construct the desired finite cover M' of M and the proper nonzero map $f\colon M' \to N$, modulo a π_1-injectivity result (Theorem 5.17).

Recall that the 1-skeleton $Z^{(1)}$ of Z has four identical components. So Z has at most four components, and we take a component of Z that contains the least number (one, two or four) of components of the 1-skeleton. We abuse notation and still denote this component by Z, and we still denote the restriction of j on this component by j.

The next section is devoted to prove the π_1-injectivity of $j\colon Z \looparrowright M$ and some further refinements. To state this result, we first need to define a compact 3-manifold $\mathscr{X}$ with boundary.

Recall that the cellulation of N_0 in Construction 5.4 induces a handle structure on a compact neighborhood $\mathscr{N}(N_0)$ of N_0. Let $\mathscr{N}^{(1)}$ be the union of 0- and 1-handles of this handle structure. Then we define a compact oriented 3-manifold $\mathscr{X}$ as following.

Construction 5.16 (1) We start with four copies of $\mathscr{N}^{(1)}$, denoted by $\mathscr{N}_1^{(1),1}$, $\mathscr{N}_2^{(1),1}$, $\mathscr{N}_1^{(1),2}$ and $\mathscr{N}_2^{(1),2}$ respectively.

(2) For any triangle Δ_{ijk} contained in $N^{(2)} \cap \partial N_0$, the corresponding 2-handle is pasted to $\mathscr{N}^{(1)}$ along an annulus L_{ijk}. Then for each copy of $\mathscr{N}^{(1)}$, we attach a 2-handle along the same annulus L_{ijk}.

(3) For any bigon B_{ijk} or triangle Δ_{ijk} in $N_\gamma^{(2)} \setminus \partial N_0$, the corresponding 2-handle is pasted to $\mathscr{N}^{(2)}$ along an annulus L_{ijk}. For any $s = 1, 2$, we take the surface $S_{ijk,s}$ from Construction 5.15(3) or (4), which has two boundary components. Then we attach $S_{ijk,s} \times I$ to $\mathscr{N}_s^{(1),1} \cup \mathscr{N}_s^{(1),2}$ via an orientation reversing homeomorphism from $\partial S_{ijk,s} \times I$ to copies of L_{ijk} in $\mathscr{N}_s^{(1),1} \cup \mathscr{N}_s^{(1),2}$.

(4) For any triangle Δ_{ijk} in $N^{(2)}$ not contained in $N_{\gamma}^{(2)}$, the corresponding 2-handle is pasted to $\mathcal{N}^{(1)}$ along an annulus L_{ijk}. We take the surface S_{ijk} constructed in Construction 5.15(5), which has four boundary components. Then we paste $S_{ijk} \times I$ to the four copies of $\mathcal{N}^{(1)}$ via an orientation reversing homeomorphism from $\partial S_{ijk} \times I$ to the four copies of L_{ijk}.

Then we take $\mathcal{Z}$ to be the component of the resulting manifold containing the least number $(1, 2$ or $4)$ of components of $\mathcal{N}^{(1)}$.

In the construction of $\mathcal{Z}$ (Construction 5.16), after the second step, we obtain four copies of the same 3-manifold, which is homeomorphic to the union of $\mathcal{N}^{(1)}$ and a neighborhood of ∂N_0. Here each copy has a unique boundary component homeomorphic to the torus. Since further constructions in Construction 5.16 do not affect these four tori, we obtain at most four tori components of $\partial\mathcal{Z}$ and we denote their union by $\partial_p\mathcal{Z}$.

Now we can state the result to be proved in Section 6.

Theorem 5.17 *For any small $\epsilon \in (0, 10^{-2})$, there exists $R_0 > 0$ depending on M and ϵ, such that the following statement holds for any $R > R_0$. If the construction of $j\colon Z \looparrowright M$ satisfies all conditions in Constructions 5.11, 5.12 and 5.15 (involving ϵ and R), then $j\colon Z \looparrowright M$ is π_1-injective and the π_1-image $j_*(\pi_1(Z)) < \pi_1(M)$ is a geometrically finite subgroup.*

Moreover, the convex core of the covering space $\widetilde{M}$ of M corresponding to $j_*(\pi_1(Z)) < \pi_1(M)$ is homeomorphic to the 3-manifold $\mathcal{Z} \setminus \partial_p\mathcal{Z}$ as oriented manifolds, and the cusp ends of $\widetilde{M}$ corresponding to $\partial_p\mathcal{Z}$ are mapped to $T_1 \cup T_2 \subset \partial M$ via the covering map.

The proof of Theorem 5.17 is more complicated than proofs of corresponding π_1-injectivity results in [19; 23], since the construction of the mapped-in 2-complex $j\colon Z \looparrowright M$ is more complicated. The proof of Theorem 5.17 is more geometric, which is in different flavor from other constructions in this section, so we defer its proof to Section 6.

For $\mathcal{N}^{(2)}$, it has a unique boundary component homeomorphic to the torus, and we denote it by $\partial_p\mathcal{N}^{(2)}$. The following elementary property of $\mathcal{Z}$ is important for the construction of virtual domination.

Lemma 5.18 *There exists a proper map $g\colon (\mathcal{Z}, \partial\mathcal{Z}) \to (\mathcal{N}^{(2)}, \partial\mathcal{N}^{(2)})$ of $\deg(g) \in \{1, 2, 4\}$, such that $g^{-1}(\partial_p\mathcal{N}^{(2)}) = \partial_p\mathcal{Z}$, and the restriction of g on each component of $\partial_p\mathcal{Z}$ is an orientation preserving homeomorphism to $\partial_p\mathcal{N}^{(2)}$.*

Proof We construct g by following the steps in Construction 5.16 that constructs $\mathcal{Z}$.

First, g maps one, two or four copies of $\mathcal{N}^{(1)}$ in $\mathcal{Z}$ to $\mathcal{N}^{(1)} \subset \mathcal{N}^{(2)}$ by identity. Then g maps each 2-handle in $\mathcal{Z}$ corresponding to $\Delta_{ijk} \subset \partial N_0$ to the corresponding 2-handle in $\mathcal{N}^{(2)}$ by homeomorphism. So g maps $\partial_p\mathcal{Z}$ to $\partial_p\mathcal{N}^{(2)}$, and the restriction of g on each component of $\partial_p\mathcal{Z}$ is an orientation preserving homeomorphism to $\partial_p\mathcal{N}^{(2)}$.

In steps (3) and (4) of Construction 5.16, each 2-handle in $\mathcal{N}^{(2)}$ not contained in ∂N_0 corresponds to a bigon B_{ijk} or triangle Δ_{ijk}, and it gives $(S_{ijk,1} \times I) \cup (S_{ijk,2} \times I)$ or $S_{ijk} \times I$ in $\mathscr{X}$. Then each component of $(S_{ijk,1} \times I) \cup (S_{ijk,2} \times I)$ or $S_{ijk} \times I$ in $\mathscr{X}$ is mapped to the corresponding 2-handle $B_{ijk} \times I$ or $\Delta_{ijk} \times I$ in $\mathcal{N}^{(2)}$, via the product of a pinching map on surfaces and the identity on I. Then we get a proper map $g\colon (\mathscr{X}, \partial\mathscr{X}) \to (N^{(2)}, \partial N^{(2)})$ that maps $\partial\mathscr{X} \setminus \partial_p\mathscr{X}$ to $\partial\mathcal{N}^{(2)} \setminus \partial_p\mathcal{N}^{(2)}$.

We have $\deg(g) \in \{1, 2, 4\}$ since $g^{-1}(\mathcal{N}^{(1)})$ consists of 1, 2 or 4 copies of $\mathcal{N}^{(1)}$, and the restriction of g on each such component is identity. □

Now we are ready to prove Theorem 5.1.

Proof of Theorem 5.1 We first consider M as a noncompact 3-manifold with cusp ends. We take small $\epsilon > 0$ and large enough $R > 0$ such that Constructions 5.12, 5.15 and Theorem 5.17 hold, and we construct a mapped in 2-complex $j\colon Z \looparrowright M$. By Theorem 5.17, $j_*\colon \pi_1(Z) \to \pi_1(M)$ is injective and the convex core of $j_*(\pi_1(Z)) < \mathrm{Isom}_+(\mathbb{H}^3)$ is homeomorphic to $\mathscr{X} \setminus \partial_p\mathscr{X}$. Let $\widetilde{M}$ be the covering space of M corresponding to $j_*(\pi_1(Z)) < \pi_1(M)$, then it contains a submanifold homeomorphic to $\mathscr{X} \setminus \partial_p\mathscr{X}$, such that ends of $\mathscr{X} \setminus \partial_p\mathscr{X}$ correspond to cusp ends of $\widetilde{M}$.

Now we chop off cusp ends of M and consider it as a compact 3-manifold with boundary. As a manifold with boundary, $\widetilde{M}$ contains a compact submanifold homeomorphic to $\mathscr{X}$ such that $\mathscr{X} \cap \partial\widetilde{M} = \partial_p\mathscr{X}$.

By Agol's celebrated result that hyperbolic 3-manifold groups are LERF [2], the covering map $\widetilde{M} \to M$ factors through a finite cover M' of M, such that $\mathscr{X}$ is mapped into M' via embedding, we have $\mathscr{X} \cap \partial M' = \partial_p\mathscr{X}$.

Recall that we have a handle decomposition of a neighborhood $\mathcal{N}(N_0)$ of N_0, and we can identify $\mathcal{N}(N_0)$ with N. By Lemma 5.18, there is a proper map $g\colon (\mathscr{X}, \partial\mathscr{X}) \to (\mathcal{N}^{(2)}, \partial\mathcal{N}^{(2)})$ such that $g^{-1}(\partial_p\mathcal{N}^{(2)}) = \partial_p\mathscr{X}$ and $\deg(g) \in \{1, 2, 4\}$. Note that each component of $\partial N^{(2)} \setminus \partial_p N^{(2)}$ is the boundary of a 3-cell, which is homeomorphic to S^2. Then we extend g to a proper map $f\colon M' \to N$ as following.

Let K be a component of $M' \setminus \mathscr{X}$, then K is disjoint from $\partial_p\mathscr{X}$ and ∂K is the union of $\partial_p K = K \cap \partial M'$ and $\partial_i K = K \cap \mathscr{X}$. Then $\partial_i K$ has a neighborhood $\partial_i K \times I$ in K. Since g maps each component of $\partial_i K \subset \partial\mathscr{X}$ to an S^2-component of $\partial\mathcal{N}^{(2)}$ that bounds a 3-cell, we first map $\partial_i K \times I \subset K$ to a union of 3-cells in N_0 that maps $\partial_i K \times 0 = \partial_i K$ via g and maps $\partial_i K \times 1$ to centers of 3-cells. Then we map $K \setminus (\partial_i K \times I)$ to a graph in $\mathcal{N}(N_0)$ such that each component of $\partial_i K \times 1$ is mapped to the corresponding center of 3-cell and each component of $\partial_p K$ is mapped to a point in $\partial(\mathcal{N}(N_0))$. Moreover, we can assume that this graph misses $\mathcal{N}^{(1)}$. By the above definition of f on components of $M' \setminus \mathscr{X}$, we get a proper map $f\colon (M', \partial M') \to (\mathcal{N}(N_0), \partial\mathcal{N}(N_0)) = (N, \partial N)$.

For any point $p \in \mathcal{N}^{(1)}$, we have $f^{-1}(p) = g^{-1}(p)$, while f and g have the same local mapping degree at $f^{-1}(p)$. So $\deg(f) = \deg(g) \in \{1, 2, 4\}$ holds. □

5.4 Proof of Theorem 1.3

The proof of Theorem 1.3 is similar to the proof of Theorem 1.2, and we sketch the proof in the following.

We start with a compact oriented mixed 3-manifold M with tori boundary such that its boundary intersects with a hyperbolic JSJ piece, and a compact oriented 3-manifold N with tori boundary.

Propositions 4.2 and 4.5 imply the following hold.

- M has a finite cover M' with two boundary components T_1 and T_2 contained in the same hyperbolic JSJ piece $M_0' \subset M'$, such that the kernel of $H_1(T_1 \cup T_2; \mathbb{Z}) \to H_1(M_0'; \mathbb{Z})$ induced by inclusion contains an element that has nontrivial components in both $H_1(T_1; \mathbb{Z})$ and $H_1(T_2; \mathbb{Z})$.
- N is virtually 2-dominated by a one-cusped oriented hyperbolic 3-manifold N'.

By the argument at the beginning of Section 5 (Theorem 5.1 implies Theorem 1.2), it suffices to prove that M' virtually dominates N' with virtual mapping degree in $\{1, 2, 4\}$. By abusing notation, we still use M and N to denote M' and N' respectively, and use M_0 to denote the hyperbolic piece of M containing $T_1, T_2 \subset \partial M$.

Then we take a geometric cellulation of N as in Constructions 5.3 and 5.4, and construct instructional maps and homomorphisms as in Lemma 5.6, with M replaced by M_0. Then we choose small enough $\epsilon > 0$ and large enough $R > 0$, and follow Constructions 5.10, 5.11, 5.12 and 5.15 to construct a map $j\colon Z \looparrowright M_0$. By Theorem 5.17 (to be proved in Section 6), if R is large enough, $j\colon Z \looparrowright M_0$ is π_1-injective, and the convex core of $j_*(\pi_1(Z)) < \pi_1(M_0) < \mathrm{Isom}_+(\mathbb{H}^3)$ is homeomorphic to $\mathscr{X} \setminus \partial_p\mathscr{X}$.

So the covering space $\widetilde{M}_0$ of M_0 corresponding to $j_*(\pi_1(Z)) < \pi_1(M_0)$ contains a submanifold homeomorphic $\mathscr{X} \setminus \partial_p\mathscr{X}$, and $\partial_p\mathscr{X}$ corresponds to cusp ends of $\widetilde{M}_0$. Now we chop off cusp ends of M_0 and consider it as a compact 3-manifold with tori boundary, the corresponding $\widetilde{M}_0$ contains a compact submanifold homeomorphic to $\mathscr{X}$ such that $\partial\widetilde{M}_0 \cap \mathscr{X} = \partial_p\mathscr{X}$, and $\partial\widetilde{M}_0 \cap \mathscr{X}$ is mapped to $T_1 \cup T_2 \subset \partial M_0 \cap \partial M$.

Let $\widetilde{M}$ be the covering space of M corresponding to $j_*(\pi_1(Z)) < \pi_1(M_0) < \pi_1(M)$, then it contains a submanifold homeomorphic to $\widetilde{M}_0$. Since ∂M contains the tori boundary components T_1 and T_2 of M_0, $\mathscr{X} \subset \widetilde{M}_0$ is also a compact submanifold of $\widetilde{M}$ such that $\partial\widetilde{M} \cap \mathscr{X} = \partial_p\mathscr{X}$. Since $j: Z \to M$ maps into a JSJ hyperbolic piece $M_0 \subset M$, by [21] (which heavily relies on [2]), $j_*(\pi_1(Z))$ is a separable subgroup of $\pi_1(M)$. Then by [18], there is an intermediate finite cover M' of $\widetilde{M} \to M$, such that $\mathscr{X}$ is mapped into M' via embedding, and $\mathscr{X} \cap \partial M' = \partial_p\mathscr{X}$ holds.

Again, we identify N with a neighborhood $\mathscr{N}(N_0)$ of N_0, which has an induced handle structure. By Lemma 5.18, there is a proper map $g\colon (\mathscr{X}, \partial\mathscr{X}) \to (\mathscr{N}^{(2)}, \partial\mathscr{N}^{(2)})$ such that $g^{-1}(\partial_p\mathscr{N}^{(2)}) = \partial_p\mathscr{X}$ and $\deg(g) \in \{1, 2, 4\}$. Then we extend g to a proper map $f\colon (M', \partial M') \to (\mathscr{N}(N_0), \partial\mathscr{N}(N_0))$. For each component K of $M' \setminus \mathscr{X}$, the map f can be defined exactly as in the proof of Theorem 5.1, which maps each component of $K \cap \partial M'$ to a point in $\partial\mathscr{N}(N_0)$. Then we have $\deg(f) = \deg(g) \in \{1, 2, 4\}$.

6 Proof of π_1-injectivity

In this section, we devote to prove the π_1-injectivity result Theorem 5.17, and the proof is more difficult than the π_1-injectivity results in [19; 23]. One reason is that our constructions of Z and $j : Z \looparrowright M$ are more complicated. The more important reason is that the induced map $\tilde{j} : \widetilde{Z} \to \widetilde{M} = \mathbb{H}^3$ on universal covers is not a quasi-isometric embedding.

We sketch the structure of this section in the following. In Section 6.1, we define an (ideal) 3-complex Z^3 that contains Z as a deformation retract, then we define a family of representations

$$\rho_t : \pi_1(Z^3) \to \mathrm{Isom}(\mathbb{H}^3)$$

and a family of maps $\tilde{j}_t : \widetilde{Z}^3 \to \mathbb{H}^3$ that is ρ_t-equivariant, such that $\tilde{j}_1|_{\widetilde{Z}} = \tilde{j}$ and $\rho_1 = j_*$. In Section 6.2, we prove that $\tilde{j}_0 : \widetilde{Z}^3 \to \mathbb{H}^3$ is a quasi-isometric embedding and some further details, via a more complicated definition of *modified sequence* than in [19]. In Section 6.3, we prove that any $\tilde{j}_t$ is a quasi-isometric embedding, and finish the proof of Theorem 5.17.

6.1 An extension of Z and a family of maps

By Construction 5.15(2), for any $s, t \in \{1, 2\}$, each triangle $\Delta_{ijk} \subset \partial N_0$ gives a triangle $\Delta^{Z,t}_{ijk,s}$ in Z with $s, t \in \{1, 2\}$. If we fix s and t, then the union of all such $\Delta^{Z,t}_{ijk,s}$ gives a torus $T^t_s \subset Z$. Each $T^t_s \subset Z$ is combinatorially isomorphic to ∂N_0, and the restriction of j on T^t_s is homotopic to a covering map to $T_s \subset M$. Since both T_1 and T_2 are peripheral tori in M, the lifting of $j| : T^t_s \to M$ to the universal cover is not a quasi-isometric embedding, so neither is $\tilde{j} : \widetilde{Z} \to \widetilde{M} = \mathbb{H}^3$. To treat this undesired situation, we extend Z to an ideal 3-complex Z^3 as following.

Definition 6.1 For each triangulated torus $T^t_s \subset Z$ with $s, t = 1, 2$, we take a cone over T^t_s and delete the cone point, and denote the resulting ideal 3-complex by $C(T^t_s)$. We define Z^3 to be the union of Z and all these $C(T^t_s)$.

It is clear that Z^3 deformation retracts to Z. For any vertex $v^t_{i,s}$, edge $e^{Z,t}_{ij,s}$ and triangle $\Delta^{Z,t}_{ijk,s}$ (copies of v_i, e^Z_{ij}, Δ_{ijk} in T^t_s) contained in T^t_s, their cones give an edge, a triangle and a tetrahedron in Z^3 respectively, with the cone point deleted. We denote these cones by $e^{Z,t}_{i\infty,s}$, $\Delta^{Z,t}_{ij\infty,s}$ and $T^{Z,t}_{ijk\infty,s}$ respectively.

We define $j_1 : Z^3 \looparrowright M$ as following, which is an extension of $j : Z \looparrowright M$. We require that j_1 maps $e^{Z,t}_{i\infty,s}$ to the geodesic ray from $j(v^t_{i,s})$ to the ideal point corresponding to the cusp end T_s of ∂M, maps $\Delta^{Z,t}_{ij\infty,s}$ to the ideal triangle given by $j(e^{Z,t}_{ij,s})$ and the ideal point, and maps $T^{Z,t}_{ijk\infty,s}$ to the ideal tetrahedron given by $j(\Delta^{Z,t}_{ijk,s})$ and the ideal point. To prove Theorem 5.17, the main step is to prove that the lifting $\tilde{j}_1 : \widetilde{Z}^3 \to \widetilde{M} = \mathbb{H}^3$ of $j_1 : Z^3 \looparrowright M$ to the universal cover is a quasi-isometric embedding.

Now we define a family of maps $\{\tilde{j}_t : \widetilde{Z}^3 \to \mathbb{H}^3 \mid t \in [0, 1]\}$ (the map given by $t = 1$ is the above $\tilde{j}_1$) and a family of representations $\{\rho_t : \pi_1(Z^3) \to \mathrm{Isom}_+(\mathbb{H}^3) \mid t \in [0, 1]\}$, such that $\tilde{j}_t$ is ρ_t-equivariant. We also have the property that $\tilde{j}_0$ maps each component of the preimage of $Z \setminus Z^{(1)}$ in $\widetilde{Z}^3$ to a totally geodesic subsurface of $\mathbb{H}^3$.

The map $j_1\colon Z^3 \looparrowright M$ gives us the following parameters. All parameters are similar to the ones given in [23, Parameter 5.4], except items (3) and (4).

Parameter 6.2 (1) For each vertex $v_{i,s}^t \in Z^{(0)}$ and each edge $e_{ij,s}^{Z,t} \subset Z^{(1)} \setminus \partial_p Z^{(1)}$ (or $e_{ijk,s}^{Z,t}$) adjacent to $v_{i,s}$, the initial frame of $\mathfrak{s}_{ij,s}$ (or $\mathfrak{s}_{ijk,s}$) equals $F_{ijk}^M \cdot A_{ij,s}$ (or $F_{ijk,s}^M \cdot A_{ijk,s}$) for some $A_{ij,s} \in \mathrm{SO}(3)$ (or $A_{ijk,s}$) that is $\frac{\epsilon}{10}$-close to $\mathrm{id} \in \mathrm{SO}(3)$ (see Construction 5.12(2)(b) and (3)(b)).

(2) For each edge $e_{ij,s}^{Z,t} \in Z^{(1)} \setminus \partial_p Z^{(1)}$ (or $e_{ijk,s}^{Z,t}$), the complex length of the associated ∂-framed segment $\mathfrak{s}_{ij,s}$ (or $\mathfrak{s}_{ijk,s}$) equals $2R + \lambda_{ij,s}$ (or $2R + \lambda_{ijk,s}$) for some complex number $\lambda_{ij,s}$ (or $\lambda_{ijk,s}$) with modulus at most $\frac{\epsilon}{5}$ (see Construction 5.12(2)(a) and (3)(a)).

(3) For each edge $e_{ij,s}^{Z,t} \subset \partial_p Z^{(1)}$, $j(e_{ij,s}^{Z,t})$ is homotopic to an Euclidean geodesic segment $g_{ij,s}$ in the horotorus $T_s \subset M$ of length $(1 + \tau_{ij,s})R$ with $|\tau_{ij,s}| < \epsilon$ (see Construction 5.12(1) and Lemma 5.6(1)).

(4) For each vertex $v_{i,s}^t \in \partial_p Z^{(1)}$, we take a preferred edge $e_{ij,s}^{Z,t} \subset \partial_p Z^{(1)}$, and let n_k be the vertex of N such that n_i, n_j, n_k form a triangle in $N^{(2)}$ not contained in ∂N_0, and n_k lies to the left of the edge e_{ij} (as in Figure 2). Let $\boldsymbol{F}_{i,s} = (j(v_{i,s}^t), \vec{v}_{ij,s}, \vec{n}_{i,s})$ be the frame based at $j(v_{i,s}^t)$ such that $\vec{v}_{ij,s}$ is tangent to $g_{ij,s}$ (in item (3)) and $\vec{n}_{i,s}$ is tangent to $j_1(e_{i\infty,s}^{Z,t})$. Then we have

$$\boldsymbol{F}_{ijk,s}^M = \boldsymbol{F}_{i,s} \cdot X \cdot B_{i,s},$$

where $X \in \mathrm{SO}(3)$ satisfies

$$\begin{pmatrix} \vec{v}_2^{\,T} & \vec{v}_3^{\,T} & (\vec{v}_2 \times \vec{v}_3)^T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \cdot X$$

for vectors $\vec{v}_2$ and $\vec{v}_3$ in Remark 5.8 and $B_{i,s} \in \mathrm{SO}(3)$ is ϵ-close to $\mathrm{id} \in \mathrm{SO}(3)$.

(5) For each decomposition curve C of some surface $S_{ijk,s}$ (or S_{ijk}) and is not an inner cuff of any hamster wheel, the corresponding good curve has complex length $2R' + \xi_C$, for some complex number ξ_C with $|\xi_C| < 2\epsilon$ (the condition of (R', ϵ)-good curves).

(6) For each hamster wheel H in some $S_{ijk,s}$ or S_{ijk}, it has R' rungs (common perpendicular segments of its two outer cuffs) $r_{H,1}, \dots, r_{H,R'}$, and these rungs divide both outer cuffs c and c' to R' geodesic segments $s_{H,1}, \dots, s_{H,R'}$ and $s'_{H,1}, \dots, s'_{H,R'}$ respectively. Then for any $i = 1, \dots, R'$, the complex distance between c and c' along $r_{H,i}$ is $R' - 2 \log \sinh 1 + \mu_{H,i}$ with $|\mu_{H,i}| < \epsilon / R'$ [11, (2.9.1)]; and for any $i = 1, \dots, R' - 1$, the complex distance between $r_{H,i}$ and $r_{H,i+1}$ along $s_{H,i}$ and $s'_{H,i}$ are $2 + \nu_{H,i}$ and $2 + \nu'_{H,i}$ respectively, with $|\nu_{H,i}|, |\nu'_{H,i}| < \epsilon / R'$ [11, (2.9.3)].

(7) For each decomposition curve C of some surface $S_{ijk,s}$ (or S_{ijk}), the feet of its two adjacent good components differ by $1 + \pi i + \eta_C$, for some complex number η_C with $|\eta_C| < 100\epsilon$ (the (R, ϵ)-well-matched condition in [11, Section 2.10]) and $|\eta_C| < \epsilon / R'$ if formal feet are defined on both sides of C. Here if C is contained in $\partial S_{ijk,s}$, the foot from the three-cornered annulus (or two-cornered annulus) $A_{ijk,s}$ is the foot of the shortest geodesic segment from $\gamma_{ijk,s}$ to a preferred vertex $v_{i,s}$, as in Construction 5.15(3), (4) and (5).

So we have parameters

$$A_{ij,s}, A_{ijk,s}, B_{i,s} \in \mathrm{SO}(3), \quad \lambda_{ij,s}, \lambda_{ijk,s}, \xi_C, \eta_C, \mu_{H,i}, \nu_{H,i}, \nu'_{H,i} \in \mathbb{C}, \quad \tau_{ij,s} \in \mathbb{R}$$

associated to $j_1 \colon Z^3 \looparrowright M$, and these parameters are very small with respect to metrics of SO(3), $\mathbb{C}$ and $\mathbb{R}$, respectively.

Note that the data in Parameter 6.2(5) and (6) determine shapes of all (R', ϵ)-good components, as in the discussion after [23, Parameter 5.4].

For any $t \in [0,1]$, we take parameters

$$tA_{ij,s}, tA_{ijk,s}, tB_{i,s} \in \mathrm{SO}(3), \quad t\lambda_{ij,s}, t\lambda_{ijk,s}, t\xi_C, t\eta_C, t\mu_{H,i}, t\nu_{H,i}, t\nu'_{H,i} \in \mathbb{C}, \quad t\tau_{ij,s} \in \mathbb{R}.$$

Here for any $A \in \mathrm{SO}(3)$ close to id, tA denotes the image of $t \in [0,1]$ under the shortest geodesic $[0,1] \to \mathrm{SO}(3)$ from id to A. These parameters give rise to a map $\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3$ that is equivariant with respect to a representation $\rho_t \colon \pi_1(Z^3) \to \mathrm{Isom}_+(\mathbb{H}^3)$, and $\tilde{j}_t$ can be defined by a developing argument as following.

We use Z' to denote the subcomplex of Z^3 consisting of following pieces:

- $Z^{(1)} \subset Z$ (as in Construction 5.11),
- all triangles $\Delta^t_{ijk,s} \subset Z$ corresponding to triangles $\Delta_{ijk} \subset \partial N_0$,
- all ideal edges $e^{Z,t}_{i\infty,s}$, ideal triangles $\Delta^{Z,t}_{ij,\infty,s}$ and ideal tetrahedra $T^{Z,t}_{ijk\infty,s}$.

We define Z'' to be the union of Z' and all decomposition curves and boundary components of $S_{ijk,s}$ and S_{ijk}. The inclusion $Z'' \hookrightarrow Z^3$ is π_1-injective on each component. We further define Z''' to be the union of Z'' and following pieces:

- For each two-cornered or three-cornered annulus $A^t_{ijk,s}$ in $Z \subset Z^3$, take an edge from a vertex $v^t_{i,s}$ to the corresponding good curve $\gamma^t_{ijk,s}$.
- For each pair of pants in $Z \subset Z^3$, take its three seams.
- For each hamster wheel in $Z \subset Z^3$, take all of its short seams between adjacent inner cuffs, and $2R'$ seams from two outer cuffs to all inner cuffs.

Let $\pi \colon \widetilde{Z}^3 \to Z^3$ be the universal cover of Z^3, then each component of $\pi^{-1}(Z') = \widetilde{Z}' \subset \widetilde{Z}^3$ is the universal cover of a component of Z'. We also use $\pi_M \colon \mathbb{H}^3 \to M$ to denote the universal cover of M. The construction of $\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is given in the following.

Construction 6.3 We will first define $\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3$ on $\widetilde{Z}'' = \pi^{-1}(Z'') \subset \widetilde{Z}^3$, by the following steps.

(1) We start with a vertex $\tilde{v}^1_{i,1} \in \widetilde{Z}^3$ such that $\pi(\tilde{v}^1_{i,1}) = v^1_{i,1}$, and a point $p \in \mathbb{H}^3$ such that $\pi_M(p) = j(v^1_{i,1}) \in M$. Then we have an isometry of tangent spaces $(d\pi_M)_p \colon T_p\mathbb{H}^3 \to T_{\pi_M(p)}(M)$, and we define $\tilde{j}_t(\tilde{v}^1_{i,1}) = p$.

(2) Let $\tilde{e}^{Z,1}_{ij,1}$ be an edge from $\tilde{v}^1_{i,1}$ to another vertex $\tilde{v}^1_{j,1}$ such that it projects to $e^{Z,1}_{ij,1} \subset Z^1 \setminus \partial_p Z$. We map $\tilde{e}^{Z,1}_{ij,1}$ to a geodesic segment in $\mathbb{H}^3$ of length $R + t\,\mathrm{Re}(\lambda_{ij,1})$ from p to some $q \in \mathbb{H}^3$, such that its tangent vector at p is the tangent vector of

$$\widetilde{\boldsymbol{F}}^M_{ijk,1}(t) := (d\pi_M)^{-1}_p(\boldsymbol{F}^M_{ijk,1} \cdot tA_{ij,1}).$$

We parallel transport $-\widetilde{\boldsymbol{F}}^M_{ijk,1}(t)(t\,\mathrm{Im}\,\lambda_{ij,1}) \cdot (tA_{ji,1})^{-1}$ along this geodesic segment to get a frame $\widetilde{\boldsymbol{F}}^M_{jik,1}(t) \in \mathrm{SO}_q(\mathbb{H}^3)$. Here $\widetilde{\boldsymbol{F}}^M_{ijk,1}(t)(t\,\mathrm{Im}\,\lambda_{ij,1})$ denotes the frame rotation of $\widetilde{\boldsymbol{F}}^M_{ijk,1}(t)$ by angle $t\,\mathrm{Im}\,\lambda_{ij,1}$. Then we take an isometry

$$T_q\mathbb{H}^3 \to T_{j(v^1_{j,1})}M$$

that identifies $\widetilde{\boldsymbol{F}}^M_{jik,1}(t)$ with $\boldsymbol{F}^M_{jik,1} \in \mathrm{SO}_{j(v^1_{j,1})}(M)$. Under this identification, we can further define the map $\tilde{j}_t$ on edges adjacent to $\tilde{v}^1_{j,1}$ that do not project to $\partial_p Z$, as in item (1).

(3) If $\tilde{v}^1_{i,1}$ corresponds to a vertex $n_i \in \partial N_0$, then $v^1_{i,1}$ is contained in the torus $T^1_1 \subset \partial_p Z$ and let $\widetilde{T}^1_1$ be the component of $\pi^{-1}(T^1_1) \subset \widetilde{Z}^3$ containing $\tilde{v}^1_{i,1}$, then we do the following construction. Let $\tilde{e}^{Z,1}_{ij,1} \subset \widetilde{Z}^{(1)}$ be the edge of $\widetilde{Z}^{(1)}$ adjacent to $\tilde{v}^1_{i,1}$ corresponding to the preferred edge $e_{ij} \in \partial N_0$ as in Parameter 6.2(4), and let n_k be the vertex of $N^{(0)}$ as in Figure 2. The geodesic ray starting from $p = \tilde{j}_t(\tilde{v}^1_{i,1})$ and tangent with the normal vector of

$$(d\pi_M)^{-1}_p(\boldsymbol{F}^M_{ijk,1} \cdot (X \cdot tB_{i,1})^{-1})$$

gives an ideal point $b \in \partial\mathbb{H}^3$, and it determines a horoplane P going through p. Then we take an Euclidean geodesic segment in P tangent to the tangent vector of $(d\pi_M)^{-1}_p(\boldsymbol{F}^M_{ijk,1} \cdot (X \cdot tB_{i,1})^{-1})$ with length $(1 + t\tau_{ij,1})R$ and map $\tilde{v}^1_{j,1}$ to its endpoint. The parameters $\tau_{st,1}$ (in Parameter 6.2(3)) inductively give triangles in P with edge length $(1 + t\tau_{st,1})R$. In this way, we get a map $\tilde{i}_t \colon \widetilde{T}^1_1 \to P$.

(4) For $\tilde{i}_t \colon \widetilde{T}^1_1 \to P$ defined in (3), we homotopy each edge in $\widetilde{T}^1_1$ to a geodesic segment in $\mathbb{H}^3$ (relative to endpoints) and each triangle in $\widetilde{T}^1_1$ to the corresponding totally geodesic triangle, to get the desired map $\tilde{j}_t| \colon \widetilde{T}^1_1 \to \mathbb{H}^3$. For each ideal edge $e^{Z,1}_{i\infty,1}$, ideal triangle $\Delta^{Z,1}_{ij\infty,1}$, ideal tetrahedron $T^{Z,1}_{ij\infty,1}$ in $\widetilde{Z}^3$ adjacent to $\widetilde{T}^1_1$, we map them to the geodesic ray, the ideal geodesic triangle and the ideal hyperbolic tetrahedron determined by the ideal point $b \in \partial\mathbb{H}^3$ and $\tilde{j}_t \colon \widetilde{T}^1_1 \to \mathbb{H}^3$, respectively.

(5) For any vertex $\tilde{v}^1_{i',1} \in \widetilde{T}^1_1$ that corresponds to $v^1_{i',1} \in T^1_1$, there is an preferred edge $\tilde{e}^{Z,1}_{i'j',1}$ adjacent to $\tilde{v}^1_{i',1} \in \widetilde{T}^1_1$. Then we get a frame $\widetilde{\boldsymbol{F}}_{i',1}$ based at $\tilde{j}_t(\tilde{v}^1_{i',1})$ such that its tangent vector is tangent to $\tilde{i}_t(\tilde{e}^{Z,1}_{i'j',1})$ and its normal vector points to $b \in \partial\mathbb{H}^3$. We take the isometry $T_{\tilde{j}_t(\tilde{v}^1_{i',1})}\mathbb{H}^3 \to T_{j(v^1_{i',1})}M$ that maps $\widetilde{\boldsymbol{F}}_{i',1}$ to $\boldsymbol{F}^M_{i'j'k',1} \cdot X \cdot (tB_{i',1})$. Then we construct $\tilde{j}_t$ on edges of $\widetilde{Z}^{(1)}$ adjacent to $\tilde{v}^1_{i',1} \in \widetilde{T}^1_1$ as in item (2).

(6) By applying the construction in items (2), (3), (4) and (5) inductively, we define $\tilde{j}_t$ on the component $\widetilde{W}$ of $\widetilde{Z}' \subset \widetilde{Z}^3$ containing $\tilde{v}^1_{i,1}$, and let W be the image of $\widetilde{W}$ in Z'. Note that when $t = 0$, for any triangle Δ_{ijk} (or bigon B_{ijk}) in $N_0 \setminus \partial N_0$, the $\tilde{j}_0$-image of any component of $\pi^{-1}(e^{Z,1}_{ij,1} \cup e^{Z,1}_{jk,1} \cup e^{Z,1}_{ki,1})$ (or $\pi^{-1}(e^{Z,1}_{ijk,1} \cup e^{Z,1}_{ji,1})$) lies in a hyperbolic plane in $\mathbb{H}^3$. For Δ_{ijk}, this claim follows from item (2) since the normal vector of $\widetilde{F}^M_{ijk,1}(0)$ is parallel transported to the normal vector of $-\widetilde{F}^M_{jik,1}(0)$, and the same

holds if i, j and k are permuted. For B_{ijk}, this claim follows from Remark 5.8 that the $t=0$ case is modeled by an equilateral tessellation of the horoplane.

When $t=1$, $\tilde{j}_1$ is exactly the restriction of $\tilde{j}\colon \widetilde{Z} \to \mathbb{H}^3$ on $\widetilde{W}$. Since $\tilde{j}_t|_{\widetilde{W}}$ is defined by geometric parameters, it induces a representation $\rho_t^{\widetilde{W}}\colon \pi_1(W) \to \mathrm{Isom}_+(\mathbb{H}^3)$, such that $\tilde{j}_t|_{\widetilde{W}}$ is $\rho_t^{\widetilde{W}}$-equivariant.

(7) Now we work on $\widetilde{Z}'''$. For any line component $l \subset \widetilde{Z}'''$ of the preimage of some $C \subset \partial S_{ijk,1}$ (or ∂S_{ijk}) that is adjacent to $\widetilde{W}$, it corresponds to a bi-infinite concatenation of edges in $\widetilde{W}$, and $\tilde{j}_t$ maps this concatenation to a bi-infinite quasigeodesic in $\mathbb{H}^3$. We map l to the bi-infinite geodesic that share end points with the above quasigeodesic, equivariant under the $\pi_1(C)$-action (via $\rho_t^{\widetilde{W}}$). For each edge e of $\widetilde{Z}'''$ adjacent to l that is mapped to an edge contained in a three-cornered annulus (or two-cornered annulus) in Z, we map it to the shortest geodesic segment between the corresponding vertex in $\widetilde{W}$ and the bi-infinite geodesic $\tilde{j}_t(l)$ in $\mathbb{H}^3$.

(8) For each line component $l' \subset \widetilde{Z}''$ in the preimage of an (R',ϵ)-good curve C' in S_{ijk} and adjacent to l in $\widetilde{Z}''$, we map the seam s between l and l' to the geodesic segment whose feet are the $(1+\pi i+t\eta_C)$-shift of the closest feet on $\tilde{j}_t(l)$ (arising from an edge in a three-cornered or a two-cornered annulus in item (7)). The complex length of $\tilde{j}_t(s)$ is determined by the parameters of this good component, which determines the $\tilde{j}_t$-image of l', a bi-infinite geodesic in $\mathbb{H}^3$. Let $\pi_1(C')$ acts on $\mathbb{H}^3$ by translating along $\tilde{j}_t(l')$, with complex translation length $2R'+\xi_{C'}$, and we define $\tilde{j}_t$ on l' to be $\pi_1(C')$-equivariant.

Then we apply this process inductively to define $\tilde{j}_t$ on one component of $\widetilde{Z}^3 \setminus \widetilde{Z}'$. For a hamster wheel, the complex lengths of its seams are determined by parameters $\mu_{H,i}, \nu_{H,i}, \nu'_{H,i}$ in Parameter 6.2(6) and complex lengths of its outer cuffs.

(9) Then we define $\tilde{j}_t$ on $\widetilde{Z}'''$ inductively by the process in items (2)–(8). Again, since $\tilde{j}_t\colon \widetilde{Z}''' \to \mathbb{H}^3$ is defined by geometric parameters, it induces a representation $\rho_t\colon \pi_1(Z^3) \to \mathrm{Isom}(\mathbb{H}^3)$ and $\tilde{j}_t$ is ρ_t-equivariant.

At the end, since each component of $Z^3 \setminus Z'''$ is topologically a disc, we can further triangulate Z^3 and map each new triangle in Z^3 to a geodesic triangle in $\mathbb{H}^3$. Then the map $\tilde{j}_t\colon \widetilde{Z}''' \to \mathbb{H}^3$ extends to a ρ_t-equivariant map $\tilde{j}_t\colon \widetilde{Z}^3 \to \mathbb{H}^3$. Here $\tilde{j}_1\colon \widetilde{Z} \to \mathbb{H}^3$ is the lifting of $j\colon Z^3 \looparrowright M$ to universal covers, and $\tilde{j}_0$ maps each component of $\widetilde{Z}^3 \setminus \widetilde{Z}'$ into a hyperbolic plane in $\mathbb{H}^3$.

Note that $\tilde{j}_0\colon \widetilde{Z}^3 \to \mathbb{H}^3$ maps each component of $\widetilde{Z}^3 \setminus \widetilde{Z}'$ to a totally geodesic subsurface of $\mathbb{H}^3$, maps each ideal tetrahedron in $\widetilde{Z}^3$ to an ideal tetrahedron in $\mathbb{H}^3$, and the union of these pieces is $\widetilde{Z}^3$. Then $\tilde{j}_0$ pulls back the hyperbolic metric on $\mathbb{H}^3$ to metrics on aforementioned pieces of $\widetilde{Z}^3$, and further induces a path metric on $\widetilde{Z}^3$. Each component of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ is called a 2*-dimensional piece* or a 3*-dimensional piece* of $\widetilde{Z}^3$, according to its dimension.

Now we prove a lemma that describes the coarse geometry of $\widetilde{Z}^3$, with respect to the above metric. These estimates may not be optimal.

Lemma 6.4 *If $R > 0$ is large enough, the following estimates hold for $\widetilde{Z}^3$, with respect to the above metric.*

(1) *For any vertex $v \in \widetilde{Z}^3$ and any edge $e \in \widetilde{Z}^{(1)}$ not containing v, $d_{\widetilde{Z}^3}(v,e) \geq \log R$.*

(2) *For any two distinct 3-dimensional pieces of $\widetilde{Z}^3$, their distance is at least $\frac{9}{10}\log R$.*

(3) *For any two distinct edges $e_1, e_2 \in \widetilde{Z}^{(1)}$ that do not share a vertex and do not lie in the same 3-dimensional piece of $\widetilde{Z}^3$, we have $d_{\widetilde{Z}^3}(e_1,e_2) \geq \frac{1}{3}\log R$.*

(4) *For any two distinct vertices $v_1, v_2 \in \widetilde{Z}^3$, we have $d_{\widetilde{Z}^3}(v_1,v_2) \geq \log R$.*

Proof We take $R > 0$ large enough that

$$\log R \geq \max\{4, 2I(\pi - \phi_0) + 10, -10\log(\sin\phi_0)\}.$$

Note that all edges of $\widetilde{Z}^{(1)}$ that project to $\partial_p Z$ have length at least $2\log R$ (Construction 5.12(1)), and all other edges have length at least $2R - 1 > 2\log R$ (Construction 5.12(2) and (3)).

(1) Let e be an edge of $\widetilde{Z}^{(1)}$, let v be a vertex of $\widetilde{Z}^3$ not contained in e, and let γ be the shortest oriented path in $\widetilde{Z}^3$ from v to e.

If v and e do not lie in the closure of the same component of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$, then the interior of γ intersects with $\widetilde{Z}^{(1)}$ at finitely many edges, and let e' be the first such edge. Then e' and v lie in the same component of the closure of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ and v is not a vertex of e'. So it suffices to assume that v and e lie in the closure C of the same component of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$.

(a) We first suppose that C is a 3-dimensional piece. Let the ideal point corresponding to C be ∞, then the projection of C gives a tessellation of $\mathbb{R}^2$ consisting of equilateral triangles of length R. A computation in hyperbolic geometry gives $d_{\widetilde{Z}^3}(v,e) > \log R$. So we can assume that C is a 2-dimensional piece.

(b) If γ intersects with the preimage of some surface piece $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, Remark 2.12 implies that $d_{\widetilde{Z}^3}(v,e) = l(\gamma) \geq R > \log R$. So γ does not intersect with any $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, and is homotopic to a concatenation of subsegments of edges in $\widetilde{Z}^{(1)}$ relative to endpoints, as shown in Figure 4.

(c) There must be $k \geq 2$ edges in Figure 4, and all edges have length at least $2\log R$ except the last one. Since γ lies in a 2-dimensional piece, each inner angle between edges is at least ϕ_0. If the last edge has length at least $\frac{1}{4}\log R$, then Lemma 3.2(1) implies

$$\begin{aligned} d_{\widetilde{Z}^3}(v,e) = l(\gamma) &\geq (k-1)\cdot 2\log R + \tfrac{1}{4}\log R - (k-1)(I(\pi - \phi_0) + 1) \\ &\geq \log R + (k-1)\big(\tfrac{1}{2}\log R - I(\pi - \phi_0) - 1\big) \geq \log R. \end{aligned}$$

If the last edge has length less than $\frac{1}{4}\log R$, then Lemma 3.2(1) implies

$$\begin{aligned} d_{\widetilde{Z}^3}(v,e) = l(\gamma) &\geq (k-1)\cdot 2\log R - (k-2)(I(\pi - \phi_0) + 1) - \tfrac{1}{4}\log R \\ &\geq \log R + (k-2)(2\log R - I(\pi - \phi_0) - 1) \geq \log R. \end{aligned}$$

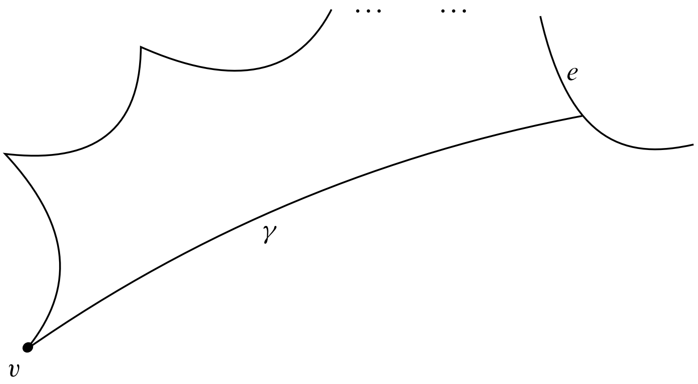


Figure 4: The shortest path in a piece of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ from v to e, without intersecting $\widetilde{S}_{ijk,s}$ of $\widetilde{S}_{ijk}$.

(2) Let C_1 and C_2 be two distinct 3-dimensional pieces of $\widetilde{Z}^3$; then $C_1 \cap C_2 = \varnothing$ holds. Let γ be the shortest path in $\widetilde{Z}^3$ between C_1 and C_2. We call intersections of γ with 3-dimensional and 2-dimensional pieces of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ as 3-dimensional subsegments and 2-dimensional subsegments of γ. We can assume that γ does not have any 3-dimensional subsegment, otherwise the proof can be reduced to a subsegment of γ. So we can assume that γ only has 2-dimensional subsegments.

(a) If γ intersects with some component of $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, then $d_{\widetilde{Z}^3}(v,e) = l(\gamma) \geq R > \log R$ as in (1)(b). So we assume that γ does not intersect with any such $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, and the intersection of γ with 2-dimensional pieces of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ are as shown in Figure 5.

(b) If a 2-dimensional subsegment of γ is homotopic to a concatenation of $k \geq 3$ subsegments of edges of $\widetilde{Z}^{(1)}$ relative to endpoints, as in Figure 5, left, then all such subsegments have lengths at least $2\log R$ except the first and last one. We divide into three cases. If both the first and last subsegments of edges have length at least $\frac{1}{4}\log R$, by Lemma 3.2(1), we have

$$\begin{aligned} l(\gamma) &\geq 2 \times \tfrac{1}{4}\log R + (k-2)2\log R - (k-1)(I(\pi-\phi_0)+1) \\ &= \tfrac{3}{2}\log R + (2k-5)(\log R - 2I(\pi-\phi_0)-2) + (3k-9)(I(\pi-\phi_0)+1) \\ &\geq \log R. \end{aligned}$$

If exactly one of the first and last subsegment of edges has length at least $\frac{1}{4}\log R$, we have

$$\begin{aligned} l(\gamma) &\geq \tfrac{1}{4}\log R + (k-2)2\log R - (k-2)(I(\pi-\phi_0)+1) - \tfrac{1}{4}\log R \\ &= \tfrac{3}{2}\log R + \left(2k-\tfrac{11}{2}\right)(\log R - 2I(\pi-\phi_0)-2) + (3k-9)(I(\pi-\phi_0)+1) \\ &\geq \log R. \end{aligned}$$

If both the first and last subsegments of edges have length at most $\frac{1}{4}\log R$, we have

$$\begin{aligned} l(\gamma) &\geq (k-2)2\log R - (k-3)(I(\pi-\phi_0)+1) - 2\cdot\tfrac{1}{4}\log R \\ &= \tfrac{5}{4}\log R + \left(2k-\tfrac{23}{4}\right)(\log R - 2I(\pi-\phi_0)-2) + \left(3k-\tfrac{17}{2}\right)(I(\pi-\phi_0)+1) \\ &\geq \log R. \end{aligned}$$

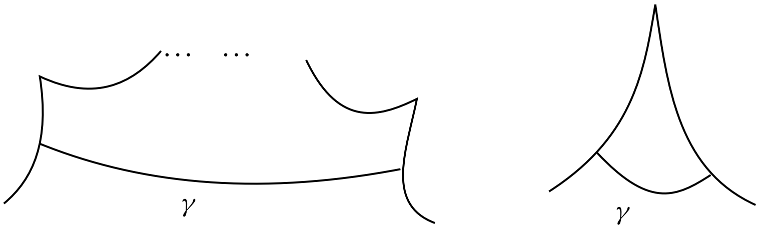


Figure 5: Intersection of γ and components of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$.

(c) By (2)(b) above, we can assume that only Figure 5, right, shows up. If all subsegments of edges in Figure 5, right, have length less than $\log R$, since all edges of $\widetilde{Z}^{(1)}$ have length at least $2\log R$, the vertices in each occurrence of Figure 5, right, must be the same vertex, contradicting with the fact that γ connects two distinct 3-dimensional pieces. So some subsegment of edge in Figure 5, right, must have length at least $\log R$. Since the angle in Figure 5, right, is at least ϕ_0, a computation in hyperbolic geometry gives $\sinh(l(\gamma)) \geq \sinh(\log R)\cdot \sin\phi_0$; thus

$$l(\gamma) \geq \log R + \log(\sin\phi_0) > \tfrac{9}{10}\log R.$$

(3) Let e_1 and e_2 be two distinct edges of $\widetilde{Z}^{(1)}$ that do not share a vertex and do not lie in the same 3-dimensional piece of $\widetilde{Z}^3$. Let γ be the shortest path in $\widetilde{Z}^3$ between e_1 and e_2. If γ intersects with a component of $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, then $l(\gamma) \geq \log R$ as in (1)(b) above. So we assume that γ does not intersect with any such $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, and we consider the intersection of γ with components of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$, as the following cases.

(a) If γ has a 2-dimensional subsegment as shown in Figure 5, left, with $k \geq 3$ edges in $\widetilde{Z}^{(1)}$ show up in the picture. Then (2)(b) implies $l(\gamma) \geq \log R$ holds. So we can assume that all 2-dimensional subsegments of γ are as shown in Figure 5, right.

(b) If some subsegment of edge in Figure 5, right, has length at least $\frac{2}{3}\log R$, then the argument as in (2)(c) implies $l(\gamma) > \frac{2}{3}\log R + \log(\sin\phi_0) > \frac{1}{2}\log R$. So we can assume that, for any 2-dimensional subsegment of γ as in Figure 5, right, all subsegments of edges have length smaller than $\frac{2}{3}\log R$.

(c) We call concatenations of adjacent 2-dimensional subsegments of γ as 2-dimensional pieces of γ, and we also call 3-dimensional subsegments of γ as 3-dimensional pieces. Note that by (3)(b), each 2-dimensional piece of γ lies in the link of a vertex of $\widetilde{Z}^3$. In particular, γ cannot be a single 2-dimensional piece, since e_1 and e_2 do not share a vertex of $\widetilde{Z}^3$, thus γ must contain at least one 3-dimensional piece.

(d) If γ contains two 3-dimensional pieces, since each 3-dimensional piece of $\widetilde{Z}^3$ is convex (as a subset of $\mathbb{H}^3$), these two 3-dimensional pieces of γ must be contained in two distinct 3-dimensional pieces of $\widetilde{Z}^3$. Then item (2) implies that $l(\gamma) \geq \frac{9}{10}\log R$ holds. So we can assume that γ has at most one 3-dimensional piece.

(e) So γ is a concatenation of one 2-dimensional piece, one 3-dimensional piece and one 2-dimensional piece. It is possible that one 2-dimensional piece of γ may degenerate, but these 2-dimensional pieces cannot both degenerate, since e_1 and e_2 do not lie in the same 3-dimensional piece of $\widetilde{Z}^3$. If exactly one of the two 2-dimensional pieces of γ degenerates, we assume that the degenerated edge contains the terminal point of γ. The unique 2-dimensional piece of γ lies in the $\left(\frac{2}{3}\log R\right)$-neighborhood of a vertex v. Then v is a vertex of e_1, and is contained in a 3-dimensional piece C of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$. Since e_1 and e_2 do not share a vertex, e_2 is contained in C and does not have v as its vertex. By (1), $d_{\widetilde{Z}^3}(e_2, v) \geq \log R$ holds. So we have $l(\gamma) \geq \log R - \frac{2}{3}\log R = \frac{1}{3}\log R$.
If neither two 2-dimensional pieces of γ degenerate, then there are two vertices v_1 and v_2 of the same 3-dimensional piece C of $\widetilde{Z}^3$, such that the two 2-dimensional pieces of γ lie in $\left(\frac{2}{3}\log R\right)$-neighborhoods of v_1 and v_2 respectively. Since e_1 and e_2 do not share a common vertex, v_1 and v_2 are two distinct vertices of C. Since $d_{\widetilde{Z}^3}(v_1, v_2) = d_C(v_1, v_2) \geq 2\log R$ holds, we have $l(\gamma) \geq 2\log R - 2\cdot\frac{2}{3}\log R = \frac{2}{3}\log R$.

(4) Let v_1 and v_2 be two distinct vertices of $\widetilde{Z}^3$, and let γ be the shortest path in $\widetilde{Z}^3$ between v_1 and v_2. If γ intersects with some component of $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, or if γ intersects with some 2-dimensional piece of $\widetilde{Z}^3$ as in Figure 5, left, with $k \geq 3$, then (1)(b) and (2)(b) imply $l(\gamma) \geq \log R$ hold, respectively. So we assume that each 2-dimensional subsegment of γ is as shown in Figure 5, right.

(a) If the initial point v_1 of γ is contained in a 2-dimensional subsegment of γ, as in Figure 5, right, then the subsegment of edge containing v_1 has length at least $2\log R$. So as in (2)(c), we have $l(\gamma) \geq \log R$.

(b) If the initial point v_1 of γ is contained in a 3-dimensional subsegment of γ, then the other end point of this 3-dimensional subsegment is contained in an edge not containing v_1. Then by (1), we have $l(\gamma) \geq \log R$. □

6.2 Estimation on the ideal model of Z

In this section, we prove the following result on the map $\tilde{j}_0\colon \widetilde{Z}^3 \to \mathbb{H}^3$ defined in the last section.

Proposition 6.5 *Given the metric on* $\widetilde{Z}^3$, *for large enough* $R > 0$, *the following statements hold.*

(1) *The map* $\tilde{j}_0\colon \widetilde{Z}^3 \to \mathbb{H}^3$ *is a quasi-isometric embedding.*

(2) *The representation* $\rho_0\colon \pi_1(Z^3) \to \mathrm{Isom}_+(\mathbb{H}^3)$ *is injective.*

(3) *The map* $\tilde{j}_0\colon \widetilde{Z}^3 \to \mathbb{H}^3$ *is an embedding.*

(4) *The convex core of* $\mathbb{H}^3/\rho_0(\pi_1(Z^3))$ *is homeomorphic to the* 3*-manifold* $\mathfrak{X} \setminus \partial_p\mathfrak{X}$ *in Theorem 5.17, as oriented manifolds.*

The idea of the proof of Proposition 6.5 is similar to proofs of corresponding results in [19; 23], but the actual proof is more complicated, since the construction of the 3-complex Z^3 is more involved. In particular, we will have a more complicated definition of the *modified sequence*.

We take large enough R so that there is an L satisfying the inequality

$$\max\{1000, 2I(\pi-\phi_0)\} \leq L \leq \tfrac{1}{320}\log R. \tag{6-1}$$

Here ϕ_0 is as defined in Notation 5.7(3).

For any two points $x, y \in \widetilde{Z}'' \subset \widetilde{Z}^3$, we will estimate $d_{\mathbb{H}^3}(\tilde{j}_0(x), \tilde{j}_0(y))$. Let γ be the shortest path in $\widetilde{Z}^3$ from x to y, and we will assume that γ intersects with $\widetilde{Z}^{(1)}$ nontrivially in its interior. Let $x_1, x_2, \dots, x_n$ be the intersection points of $\gamma \cap \widetilde{Z}^{(1)}$ that follow the orientation of γ. If γ contains a subsegment of an edge in $\widetilde{Z}^{(1)}$, we only record the endpoints of this subsegment. This sequence $x_1, x_2, \dots, x_n$ is called the *intersection sequence* of γ, and let $x_0 = x$ and $x_{n+1} = y$. We use $\gamma_i = \overline{x_i x_{i+1}}$ to denote the subsegment of γ (and the geodesic segment) from x_i to x_{i+1}. Such a γ_i is called a 3*-dimensional piece* of γ if it is contained in the union of ideal tetrahedra of $\widetilde{Z}^3$.

Now we make the following assumption on γ.

Assumption 6.6 For the segment $\overline{xx_1}$, we assume the following hold.

1. Either x is a vertex of $\widetilde{Z}^3$, or x lies on an edge of $\widetilde{Z}^{(1)}$ and its distance to any vertex of $\widetilde{Z}^3$ is at least L, or the distance between x and any vertex of $\widetilde{Z}^3$ is at least $L + \frac{1}{160}\log R$.
2. If $\overline{xx_1}$ is not a 3-dimensional piece, then $d(x, x_1) \geq \frac{1}{160}\log R$.

We also assume the same condition holds for $\overline{x_n y}$.

If γ satisfies Assumption 6.6, we construct the *modified sequence* of γ as follows.

Construction 6.7 For any $i = 1, \dots, n$, we do the following modification.

1. If neither γ_{i-1} nor γ_i are 3-dimensional pieces and both of following hold:
   - $d(x_i, v_i) < L$ for some vertex $v_i \neq x_i$ (then x_i and v_i lie on the same edge of $\widetilde{Z}^3$ and v_i is unique, by Lemma 6.4(1)(4)),
   - $d(x_{i-1}, x_i) < \frac{1}{160}\log R$ or $d(x_i, x_{i+1}) < \frac{1}{160}\log R$,

   then we replace x_i by v_i.
2. If γ_{i-1} or γ_i is a 3-dimensional piece, then exactly one of them is. Without loss of generality, we assume γ_i (from x_i to x_{i+1}) is a 3-dimensional piece, and γ_{i-1} is not. Then we do the following two steps.
   - (a) If $d(x_i, v_i) < L$ for some vertex $v_i \neq x_i$ and $d(x_{i-1}, x_i) < \frac{1}{160}\log R$, we replace x_i by v_i. Similarly, if $d(x_{i+1}, v_{i+1}) < L$ for some vertex $v_{i+1} \neq x_{i+1}$, then $x_{i+1} \neq y$ and x_{i+2} exists by Assumption 6.6(1). In this case, if $d(x_{i+1}, x_{i+2}) < \frac{1}{160}\log R$, we replace x_{i+1} by v_{i+1}.
   - (b) If the modification in step (1) is done for x_i but not for x_{i+1} and $d(x_{i+1}, v_i) < L$, then by Lemma 6.4(1), v_i is contained in an edge containing x_{i+1}, and we replace x_{i+1} by v_i. We do a similar process if the modification in step (1) is done for x_{i+1} but not for x_i.

Note that during the modification process, if some x_i is replaced by v_i, then they lie on the same edge of $\widetilde{Z}^{(1)}$ and we must have $d(x_i, v_i) < L$.

After doing the above process, by replacing certain x_i by corresponding vertices of $\widetilde{Z}^3$, we get the *modified sequence* $x = y_0, y_1, \dots, y_m, y_{m+1} = y$ of γ. Note that a few points in the intersection sequence might be replaced by the same point y_i. Then y_i and y_{i+1} lie in the same component of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$. Each component of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ is either a simply connected convex hyperbolic surface with piecewise geodesic boundary, or a convex hyperbolic 3-manifold obtained by an infinite union of ideal tetrahedra. Let $\gamma_i' = \overline{y_i y_{i+1}}$ be the shortest path in the piece of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ from y_i to y_{i+1}, and we say that γ_i' is a 2-dimensional or a 3-dimensional piece if it lies in a 2-dimensional or a 3-dimensional piece of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$, respectively. We define the *modified path* γ' of γ to be the concatenation of $\gamma_0', \gamma_1', \dots, \gamma_m'$.

Any y_i that also lies in the intersection sequence of γ is called *an unmodified point*, otherwise it is called *a modified point*. Any γ_i' that is also a piece of γ is called *an unmodified piece* of γ'.

We will prove a few properties of the modified path in the following.

Lemma 6.8 *If γ satisfies Assumption 6.6, then for any $i = 0, 1, \dots, m$, γ_i' is either an unmodified* 3*-dimensional piece or satisfies $l(\gamma_i') \geq L$.*

Proof **Case I** If both y_i and y_{i+1} are modified points, by Lemma 6.4(4), we have

$$l(\gamma_i') = d(y_i, y_{i+1}) \geq \log R \geq L.$$

Case II If both y_i and y_{i+1} are unmodified points, we divide into the following subcases.

- If γ_i' is a 3-dimensional piece, then it is unmodified and the result trivially holds. So we assume that γ_i' is not a 3-dimensional piece in the following.
- If $y_i = x$ of $y_{i+1} = y$, we assume $y_i = x$ without loss of generality. Then Assumption 6.6(2) implies $l(\gamma_i') = d(y_i, y_{i+1}) \geq \frac{1}{160} \log R > L$. So we assume that y_i and y_{i+1} are not x and y respectively.
- If γ_i' intersects with any component of $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, then $l(\gamma_i') \geq R > L$, by Remark 2.12.
- If γ_i' does not intersect with $\widetilde{S}_{ijk,s}$ or $\widetilde{S}_{ijk}$, we have two cases. If $l(\gamma_i') \geq \frac{1}{160} \log R$, then the result trivially holds. If $l(\gamma_i') < \frac{1}{160} \log R$, then by Lemma 6.4(3), y_i and y_{i+1} must lie on two adjacent edges of $\widetilde{Z}^{(1)}$, with a common vertex v. Then we must have $d(y_i, v), d(y_{i+1}, v) \geq L$, otherwise y_i or y_{i+1} should be modified in Construction 6.7(1) or (2)(a). Since we have $\angle y_i v y_{i+1} \geq \phi_0$ (Notation 5.7(3)), we get

$$d(y_i, y_{i+1}) \geq d(y_i, v) + d(y_{i+1}, v) - I(\pi - \phi_0) \geq 2L - I(\pi - \phi_0) \geq L.$$

Case III If exactly one of y_i and y_{i+1} is modified, we assume that y_i is unmodified and y_{i+1} is modified. Then there are x_i and x_{i+1} in the intersection sequence of γ, such that $x_i = y_i$ and x_{i+1} is modified to y_{i+1} (in Construction 6.7). If $y_i = x_i = x$, then since y_{i+1} is a vertex of $\widetilde{Z}^3$ and distinct from y_i, Assumption 6.6(1) implies $d(y_i, y_{i+1}) \geq L$. So we assume that $y_i \neq x$ in the following; thus it lies on an edge of $\widetilde{Z}^{(1)}$.

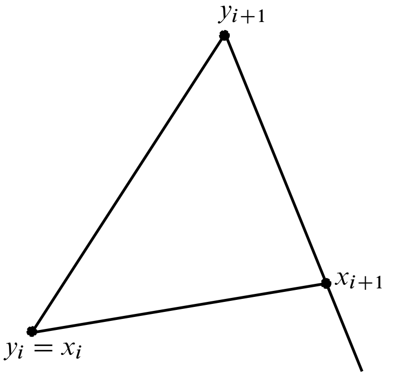


Figure 6: The position of $x_i = y_i, x_{i+1}$ and y_{i+1}.

If y_i and y_{i+1} do not lie on the same edge of $\widetilde{Z}^{(1)}$, since y_{i+1} is a vertex of $\widetilde{Z}^{(1)}$, Lemma 6.4(1) implies $d(y_i, y_{i+1}) \geq \log R > L$.

So y_i and y_{i+1} lie on the same edge, and the picture is shown in Figure 6. Suppose that $d(y_i, y_{i+1}) < L$ holds. Since x_{i+1} is modified to y_{i+1}, we have $d(x_{i+1}, y_{i+1}) < L$. Since $x_i = y_i$, we have

$$d(x_i, x_{i+1}) < 2L < \frac{1}{160} \log R.$$

If $\overline{x_i x_{i+1}}$ is not a 3-dimensional piece, then x_i should be modified, according to Construction 6.7(1) or (2)(a), and we get a contradiction. If $\overline{x_i x_{i+1}}$ is a 3-dimensional piece, then the fact that x_{i+1} is modified implies that x_i should be modified, according to Construction 6.7(2)(b), which is impossible. So we must have $d(y_i, y_{i+1}) \geq L$. □

The next job is to estimate the angle $\angle y_{i-1} y_i y_{i+1}$. To obtain this estimate, we first prove the following lemma.

Lemma 6.9 *We suppose that γ satisfies Assumption 6.6. Let x_i and x_{i+1} be two consecutive points in the intersection sequence of γ with $x_i \neq x$, such that when producing the modified sequence, x_i is not modified (thus $y_i = x_i$) and x_{i+1} is replaced by y_{i+1}, then the following hold.*

(1) *If $\overline{x_i x_{i+1}}$ is not a 3-dimensional piece, then $\angle x_{i+1} y_i y_{i+1} \leq 2e^{-L/2}$.*

(2) *If $\overline{x_i x_{i+1}}$ is a 3-dimensional piece, while x_i and y_{i+1} do not lie in the same edge of $\widetilde{Z}^{(1)}$, then $\angle x_{i+1} y_i y_{i+1} \leq 2e^{-L/2}$.*

(3) *If $\overline{x_i x_{i+1}}$ is a 3-dimensional piece, while x_i and y_{i+1} lie in the same edge of $\widetilde{Z}^{(1)}$, then $\angle x_{i+1} y_i y_{i+1} < \frac{\pi}{2}$.*

Proof Since x_{i+1} is replaced by y_{i+1}, by Construction 6.7, we have $d(x_{i+1}, y_{i+1}) < L$. Since $\overline{y_i y_{i+1}}$ is a piece of the modified sequence and is not an unmodified 3-dimensional piece, Lemma 6.8 implies $d(x_i, y_{i+1}) = d(y_i, y_{i+1}) \geq L$.

Case I Suppose that $y_i = x_i$ does not lie in any edge of $\widetilde{Z}^{(1)}$ containing y_{i+1}. Since y_{i+1} is a vertex of $\widetilde{Z}^{(1)}$, by Lemma 6.4(1), we have $d(y_i, y_{i+1}) \geq \log R$. Then we get

$$\angle x_{i+1} y_i y_{i+1} \leq 2 \sin \angle x_{i+1} y_i y_{i+1} \leq 2 \frac{\sinh d(y_{i+1}, x_{i+1})}{\sinh d(y_i, y_{i+1})} \leq 2e^{-L/2}. \tag{6-2}$$

So the proof of (2) is done.

Case II Suppose that $y_i = x_i$ lies in an edge of $\widetilde{Z}^{(1)}$ containing y_{i+1}, the picture is shown in Figure 6.

(a) Suppose that $\overline{x_i x_{i+1}}$ is not a 3-dimensional piece. Then the triangle in Figure 6 does not lie in a 3-dimensional piece of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$, we have $\angle y_i y_{i+1} x_{i+1} \geq \phi_0$, and

$$\begin{aligned} d(y_i, x_{i+1}) &\geq d(y_i, y_{i+1}) + d(y_{i+1}, x_{i+1}) - I(\pi - \phi_0) \\ &\geq d(y_{i+1}, x_{i+1}) + (L - I(\pi - \phi_0)) \geq d(y_{i+1}, x_{i+1}) + \tfrac{1}{2} L. \end{aligned}$$

So we have

$$\angle x_{i+1} y_i y_{i+1} \leq 2 \sin \angle x_{i+1} y_i y_{i+1} \leq 2 \frac{\sinh d(x_{i+1}, y_{i+1})}{\sinh d(y_i, x_{i+1})} \leq 2e^{-L/2}. \tag{6-3}$$

Then (6-2) and (6-3) together imply (1).

(b) Suppose that $\overline{x_i x_{i+1}}$ is a 3-dimensional piece. Since x_i is not modified, by Construction 6.7(2)(b), we have $d(y_i, y_{i+1}) \geq L > d(x_{i+1}, y_{i+1})$. So $\angle x_{i+1} x_i y_{i+1} < \frac{\pi}{2}$, and the proof of (3) is done. □

The next technical lemma estimates the angle $\angle y_{i-1} y_i y_{i+1}$ in the modified path.

Lemma 6.10 *There exists $\eta_0 > 0$ (only depend on the geometry of the triangulation of N) such that for large enough $L > 0$ (depending on η_0) and large enough $R > 0$ (depending on L), the following statements hold for any γ satisfying Assumption 6.6.*

(1) *If neither γ'_{i-1} nor γ'_i are unmodified 3-dimensional pieces, then $\angle y_{i-1} y_i y_{i+1} \geq \eta_0$.*

(2) *If either γ'_{i-1} or γ'_i is an unmodified 3-dimensional piece, then $\angle y_{i-1} y_i y_{i+1} \geq \frac{\pi}{2} + \eta_0$.*

Note that it is impossible that both γ'_{i-1} and γ'_i are unmodified 3-dimensional pieces.

Proof Recall that the constant $\phi_0 > 0$ defined in Notation 5.7(3) is a lower bound of all inner angles of triangles in N and all dihedral angles between intersecting totally geodesic triangles in N. We take a smaller ϕ_0 if necessary, such that $\phi_0 \in \left(0, \frac{\pi}{20}\right)$.

For any vertex v^Z of $\widetilde{Z}^3$, we have a subspace $S_{v^Z} \subset T^1_{\tilde{j}_0(v^Z)} = S^2$, consisting of all unit tangent vectors at v^Z tangent to $\tilde{j}_0$-images of pieces of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ that are adjacent to v^Z. If v^Z corresponds to a vertex v in $N_0 \setminus \partial N_0$, S_{v^Z} is a union of finitely many geodesic arcs in S^2, and it is determined by the geometry of N near v. If v^Z corresponds to a vertex in ∂N_0, S_{v^Z} is a union of finitely many geodesic arcs and one hexagon in S^2 (corresponding to a 3-dimensional component of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$; see Figure 1). Moreover, in the second case, S_{v^Z} also depends on R. The hexagon degenerates to a point when R goes to infinity, and the limit geometry only depends on the geometry of N. The metric on S^2 induces a path metric on each S_{v^Z}. For fixed R, we only have finitely many isometric classes of S_{v^Z}, since the vertex set of $\widetilde{Z}^{(1)}$

is a finite union of $\pi_1(Z)$-orbits. So there exists $R_0>0$ and a constant $\theta_0>0$, such that for any $R>R_0$, any vertex v^Z of $\widetilde{Z}^3$ and any two vectors $v_1,v_2\in S_{v^Z}\subset S^2$, if their distance under the path metric of S_{v^Z} is at least ϕ_0, then the angle between them is at least θ_0.

We take $\eta_0=\min\{\frac{1}{2}\phi_0,\frac{1}{2}\theta_0\}$. We also take large L and R such that $L>2\log(8/\eta_0)$, $R>(8/\eta_0)^{640}$ and (6-1) holds. Since the definition of the modified sequence is complicated, we need to run a case-by-case argument.

Case I We first assume that y_i is an unmodified point.

(1) Both y_{i-1} and y_{i+1} are unmodified points. Then the concatenation of $\overline{y_{i-1}y_i}$ and $\overline{y_iy_{i+1}}$ is the shortest path in $\widetilde{Z}^3$ from y_{i-1} to y_{i+1}.

(a) We first suppose that neither of $\overline{y_{i-1}y_i}$ or $\overline{y_iy_{i+1}}$ are 3-dimensional pieces. Since the dihedral angle between corresponding totally geodesic subsurfaces in $\mathbb{H}^3$ is at least ϕ_0 (Notation 5.7(3)),

$$\angle x_{i-1}x_ix_{i+1}=\angle y_{i-1}y_iy_{i+1}\geq\phi_0>\eta_0. \tag{6-4}$$

(b) We suppose that one of $\overline{y_{i-1}y_i}$ and $\overline{y_iy_{i+1}}$ is a 3-dimensional piece. By the explicit vectors in Remark 5.13(3), the dihedral angle between the boundary of an ideal tetrahedron and an adjacent geodesic subsurface (not the boundary of an ideal tetrahedron) in $\widetilde{Z}^3$ is at least $\arccos\bigl(-\frac{\sqrt{3}}{2\sqrt{7}}\bigr)>\frac{3\pi}{5}$. So we have

$$\angle x_{i-1}x_ix_{i+1}=\angle y_{i-1}y_iy_{i+1}>\tfrac{3\pi}{5}>\tfrac{\pi}{2}+\eta_0. \tag{6-5}$$

Note that (6-4) and (6-5) will be repeatedly used in the remainder of this proof.

(2) Exactly one of y_{i-1} and y_{i+1} is a modified point, and we assume that y_{i-1} is unmodified and y_{i+1} is modified. Then we have $y_{i-1}=x_{i-1}$, $y_i=x_i$, and y_{i+1} is replaced by x_{i+1} as in Construction 6.7. In particular, we have $d(x_{i+1},y_{i+1})<L$ holds.

(a) We first suppose that $\overline{x_ix_{i+1}}$ is not a 3-dimensional piece. Then Lemma 6.9(1) implies that $\angle x_{i+1}y_iy_{i+1}\leq 2e^{-L/2}<\frac{1}{2}\phi_0$. By (6-4) and (6-5) respectively, $\angle y_{i-1}y_ix_{i+1}$ is at least ϕ_0 or $\frac{3\pi}{5}$ in the two cases of this lemma. So we get that $\angle y_{i-1}y_iy_{i+1}$ is at least η_0 or $\frac{\pi}{2}+\eta_0$ in these two cases.

(b) Now we suppose that $\overline{x_ix_{i+1}}$ is a 3-dimensional piece, then $\overline{x_{i-1}x_i}$ is not a 3-dimensional piece. So neither $\overline{y_{i-1}y_i}$ nor $\overline{y_iy_{i+1}}$ are unmodified 3-dimensional pieces, and $d(y_{i-1},y_i),d(y_i,y_{i+1})\geq L$ by Lemma 6.8.

(i) If y_i and y_{i+1} do not lie on the same edge of $\widetilde{Z}^{(1)}$, then Lemma 6.9(2) implies

$$\angle x_{i+1}y_iy_{i+1}\leq 2e^{-L/2},$$

and (6-4) implies $\angle y_{i-1}y_iy_{i+1}\geq\eta_0$.

(ii) If y_i and y_{i+1} lie on the same edge of $\widetilde{Z}^{(1)}$, then Lemma 6.9(3) implies $\angle x_{i+1}y_iy_{i+1}\leq\frac{\pi}{2}$. Since γ is the shortest path in $\widetilde{Z}^3$, (6-5) implies

$$\angle y_{i-1}y_iy_{i+1}=\angle x_{i-1}x_iy_{i+1}\geq\angle x_{i-1}x_ix_{i+1}-\angle x_{i+1}y_iy_{i+1}\geq\tfrac{3\pi}{5}-\tfrac{\pi}{2}>\eta_0.$$

(3) Both y_{i-1} and y_{i+1} are modified points. So $y_i = x_i$, while y_{i-1} and y_{i+1} are modified from x_{i-1} and x_{i+1} respectively. In this case, neither $\overline{y_{i-1}y_i}$ nor $\overline{y_i y_{i+1}}$ are unmodified 3-dimensional pieces, and $d(y_{i-1}, y_i), d(y_i, y_{i+1}) \geq L$ by Lemma 6.8.

(a) Neither $\overline{x_{i-1}x_i}$ nor $\overline{x_i x_{i+1}}$ are 3-dimensional pieces. By Lemma 6.9(1), we have

$$\angle x_{i-1} y_i y_{i-1}, \angle x_{i+1} y_i y_{i+1} \leq 2e^{-L/2}.$$

Then by (6-4), we have $\angle x_{i-1} y_i x_{i+1} \geq \phi_0$, so

$$\angle y_{i-1} y_i y_{i+1} \geq \phi_0 - 4e^{-L/2} \geq \tfrac{1}{2}\phi_0 \geq \eta_0.$$

(b) Both $\overline{x_{i-1}x_i}$ and $\overline{x_i x_{i+1}}$ are 3-dimensional pieces. It is impossible.

(c) Exactly one of $\overline{x_{i-1}x_i}$ and $\overline{x_i x_{i+1}}$ is a 3-dimensional piece. We assume that $\overline{x_{i-1}x_i}$ is not a 3-dimensional piece and $\overline{x_i x_{i+1}}$ is a 3-dimensional piece. By Lemma 6.9(1), $\angle x_{i-1} y_i y_{i-1} \leq 2e^{-L/2}$.

- If y_i and y_{i+1} do not lie on the same edge of $\widetilde{Z}^{(1)}$, then Lemma 6.9(2) implies

$$\angle x_{i+1} y_i y_{i+1} \leq 2e^{-L/2}.$$

Since $\angle x_{i-1} y_i x_{i+1} \geq \phi_0$ (equation (6-4)), we have $\angle y_{i-1} y_i y_{i+1} \geq \phi_0 - 4e^{-L/2} \geq \frac{1}{2}\phi_0 \geq \eta_0$.

- If y_i and y_{i+1} lie on the same edge of $\widetilde{Z}^{(1)}$, then Lemma 6.9(3) implies $\angle x_{i+1} y_i y_{i+1} \leq \frac{\pi}{2}$. Since $\angle x_{i-1} y_i x_{i+1} \geq \frac{3\pi}{5}$ (equation (6-5)), we get

$$\begin{aligned}\angle y_{i-1} y_i y_{i+1} &\geq \angle x_{i-1} y_i x_{i+1} - \angle x_{i-1} y_i y_{i-1} - \angle x_{i+1} y_i y_{i+1} \\ &\geq \tfrac{3\pi}{5} - 2e^{-L/2} - \tfrac{1}{2}\pi \geq \eta_0.\end{aligned}$$

So the proof of Case I is done.

Case II Now we assume that y_i is a modified point. In this case, there might be several consecutive points $x_i, \dots, x_{i+k}$ in the intersection sequence of γ that are modified to y_i; then we have $d(y_i, x_i), \dots, d(y_i, x_{i+k}) < L$. Note that neither $\overline{y_{i-1}y_i}$ nor $\overline{y_i y_{i+1}}$ are unmodified 3-dimensional pieces in this case.

(1) Both y_{i-1} and y_{i+1} are unmodified points.

(a) We assume that $k = 0$ holds; then the picture is shown as in Figure 7(a). This figure shows a flattened picture of $\widetilde{Z}^3$ in a hyperbolic plane, while the actual picture is bended in $\mathbb{H}^3$. By Construction 6.7, at least one of $d(x_{i-1}, x_i)$ and $d(x_i, x_{i+1})$ is smaller than $\frac{1}{160}\log R$ and we assume $d(x_{i-1}, x_i) < \frac{1}{160}\log R$. Then we have

$$d(x_{i-1}, y_i) \leq d(x_{i-1}, x_i) + d(x_i, y_i) < \frac{1}{160}\log R + L.$$

By Assumption 6.6(1), even if $x_{i-1} = x$, we know that x_{i-1} lies on an edge of $\widetilde{Z}^{(1)}$. Since

$$d(x_{i-1}, y_i) < \frac{1}{160}\log R + L < \frac{1}{80}\log R,$$

by Lemma 6.4(1), x_{i-1} and y_i must lie on the same edge of $\widetilde{Z}^{(1)}$.

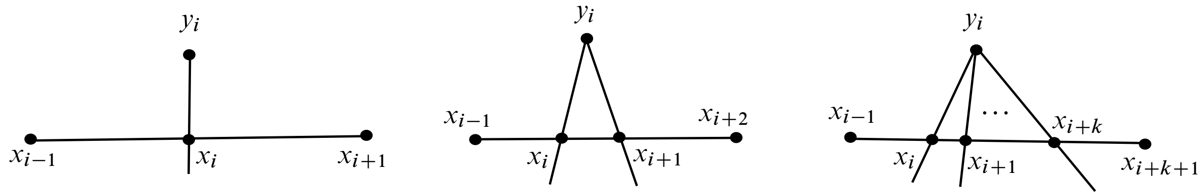


Figure 7: y_i is a modified point, while $y_{i-1} = x_{i-1}$ and $y_{i+1} = x_{i+k+1}$ are not.

- If $\overline{x_{i-1}x_i}$ is not a 3-dimensional piece, then $\angle x_{i-1} y_i x_i \ge \phi_0$ holds by the definition of ϕ_0 in Notation 5.7(3).
- If $\overline{x_{i-1}x_i}$ is a 3-dimensional piece, then $\overline{x_i x_{i+1}}$ is not a 3-dimensional piece. Since x_i is a modified point and x_{i-1} is not, by Construction 6.7(2), we have $d(x_i, x_{i+1}) < \frac{1}{160}\log R$. The same argument as above implies $\angle x_i y_i x_{i+1} \ge \phi_0$.

Let S be the subset of S^2 corresponding to vertex y_i given in the beginning of this proof, and let $\vec{v}_{i-1}$ and $\vec{v}_{i+1}$ be points in S given by tangent vectors of $\overline{y_i x_{i-1}}$ and $\overline{y_i x_{i+1}}$ respectively. The above inequalities on $\angle x_{i-1} y_i x_i$ and $\angle x_i y_i x_{i+1}$ imply $d_S(\vec{v}_{i-1}, \vec{v}_{i+1}) \ge \phi_0$, otherwise the concatenation $\overline{x_{i-1}x_i} \cdot \overline{x_i x_{i+1}}$ is not the shortest path in $\widetilde{Z}^3$ from x_{i-1} to x_{i+1}. The choice of θ_0 implies $\angle y_{i-1} y_i y_{i+1} = \angle x_{i-1} y_i x_{i+1} \ge \theta_0 \ge \eta_0$ holds. This argument will be used repeatedly in the following part of this proof, referred as "the argument in Case II(1)(a)".

(b) We assume that $k = 1$ holds; then the picture is shown in Figure 7(b).

- If $\overline{x_i x_{i+1}}$ is not a 3-dimensional piece, then we have $\angle x_i y_i x_{i+1} \ge \phi_0$. The argument in Case II(1)(a) implies $\angle y_{i-1} y_i y_{i+1} = \angle x_{i-1} y_i x_{i+1} \ge \theta_0 \ge \eta_0$.
- If $\overline{x_i x_{i+1}}$ is a 3-dimensional piece, then by Construction 6.7(2), either $d(x_{i-1}, x_i) < \frac{1}{160}\log R$ or $d(x_{i+1}, x_{i+2}) < \frac{1}{160}\log R$ holds. We assume that $d(x_{i-1}, x_i) < \frac{1}{160}\log R$ holds. By Assumption 6.6(1) and Lemma 6.4(1) again, even if $x_{i-1} = x$ holds, x_{i-1} lies on an edge of $\widetilde{Z}^{(1)}$ containing y_i. Since $\overline{x_{i-1}x_i}$ is not a 3-dimensional piece, we have $\angle x_{i-1} y_i x_i \ge \phi_0$. The argument in Case II(1)(a) implies $\angle y_{i-1} y_i y_{i+1} = \angle x_{i-1} y_i x_{i+2} \ge \theta_0 \ge \eta_0$.

(c) We assume that $k \ge 2$ holds; then the picture is shown in Figure 7(c). Here either $\overline{x_i x_{i+1}}$ or $\overline{x_{i+1} x_{i+2}}$ is not a 3-dimensional piece; thus either $\angle x_i y_i x_{i+1} \ge \phi_0$ or $\angle x_{i+1} y_i x_{i+2} \ge \phi_0$ holds. Again, the argument in Case II(1)(a) implies $\angle y_{i-1} y_i y_{i+1} = \angle x_{i-1} y_i x_{i+k+1} \ge \theta_0 \ge \eta_0$.

(2) Exactly one of y_{i-1} and y_{i+1} is a modified point, and we assume y_{i-1} is modified (from x_{i-1}) and y_{i+1} is unmodified (equals x_{i+k+1}). Since y_{i-1} and y_i are distinct vertices of $\widetilde{Z}^{(1)}$, by Lemma 6.4(4), we have $d(y_{i-1}, y_i) \ge \log R$. Since x_{i-1} and x_i are modified to y_{i-1} and y_i respectively, we have $d(x_{i-1}, y_{i-1}), d(x_i, y_i) < L$. So we have

$$\angle x_{i-1} y_i y_{i-1} \le 2 \sin \angle x_{i-1} y_i y_{-1} \le 2 \frac{\sinh d(x_{i-1}, y_{i-1})}{\sinh d(y_{i-1}, y_i)} \le 2e^{-R/2}.$$

Moreover,

$$d(x_{i-1}, x_i) \ge d(y_{i-1}, y_i) - d(x_{i-1}, y_{i-1}) - d(x_i, y_i) \ge \log R - 2L \ge \tfrac{1}{2}\log R.$$

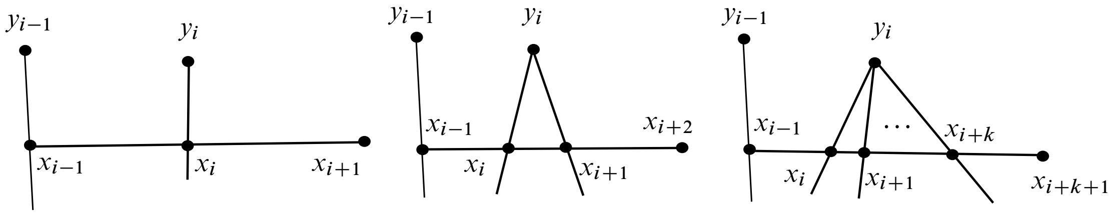


Figure 8: y_{i-1} and y_i are modified points, while $y_{i+1} = x_{i+k+1}$ is not.

Now we claim that $\angle x_{i-1} y_i x_{i+k+1} \geq \theta_0$ holds, and the proof divides into following cases.

(a) We first assume that $k = 0$ holds; then the picture is shown in Figure 8, left. Since x_i is modified to y_i, by Construction 6.7, $\overline{x_i x_{i+1}}$ cannot be a 3-dimensional piece, and we must have $d(x_i, x_{i+1}) < \frac{1}{160} \log R$. By Assumption 6.6(1) and Lemma 6.4(1), even if $x_{i+1} = y$ holds, x_{i+1} lies on an edge of $\widetilde{Z}^{(1)}$ containing y_i, and we have $\angle x_i y_i x_{i+1} \geq \phi_0$. By the argument in Case II(1)(a), we have $\angle x_{i-1} y_i x_{i+1} \geq \theta_0$.

(b) We assume that $k = 1$ holds, then the picture is shown in Figure 8, middle.

- If $\overline{x_i x_{i+1}}$ is not a 3-dimensional piece, then we have $\angle x_i y_i x_{i+1} \geq \phi_0$. The argument in Case II(1)(a) implies $\angle x_{i-1} y_i x_{i+2} \geq \theta_0$.
- If $\overline{x_i x_{i+1}}$ is a 3-dimensional piece, then $\overline{x_{i+1} x_{i+2}}$ is not a 3-dimensional piece. Since $d(x_{i-1}, x_i) \geq \frac{1}{2} \log R$, by Construction 6.7(2), we must have $d(x_{i+1}, x_{i+2}) < \frac{1}{160} \log R$. By Assumption 6.6(1) and Lemma 6.4(1), even if $x_{i+2} = y$ holds, x_{i+2} lies on an edge of $\widetilde{Z}^{(1)}$ containing y_i, and $\angle x_{i+1} y_i x_{i+2} \geq \phi_0$ holds. Again, the argument in Case II(1)(a) implies $\angle x_{i-1} y_i x_{i+2} \geq \theta_0$.

(c) We assume that $k \geq 2$ holds; then the picture is shown in Figure 8, right. Then either $\overline{x_i x_{i+1}}$ or $\overline{x_{i+1} x_{i+2}}$ is not a 3-dimensional piece; thus either $\angle x_i y_i x_{i+1} \geq \phi_0$ or $\angle x_{i+1} y_i x_{i+2} \geq \phi_0$ holds. Again, the argument in Case II(1)(a) implies $\angle x_{i-1} y_i x_{i+k+1} \geq \theta_0$.

So the claim is established, and we have

$$\angle y_{i-1} y_i y_{i+1} = \angle y_{i-1} y_i x_{i+k+1} \geq \angle x_{i-1} y_i x_{i+k+1} - \angle x_{i-1} y_i y_{i-1} \geq \theta_0 - 2e^{-R/2} \geq \eta_0.$$

(3) Both y_{i-1} and y_{i+1} are modified points. Then y_{i-1} and y_{i+1} are obtained by modifying x_{i-1} and x_{i+k+1} respectively. Since y_{i-1}, y_i and y_{i+1} are distinct vertices of $\widetilde{Z}$, by Lemma 6.4(4), we have $d(y_{i-1}, y_i), d(y_i, y_{i+1}) \geq \log R$. By the modification process in Construction 6.7,

$$d(x_{i-1}, y_{i-1}), d(x_i, y_i), \dots, d(x_{i+k}, y_i), d(x_{i+k+1}, y_{i+1}) < L.$$

By the computation at the beginning of Case II(2), we have

$$\angle x_{i-1} y_i y_{i-1}, \angle x_{i+k+1} y_i y_{i+1} \leq 2e^{-R/2}, \quad d(x_{i-1}, x_i), d(x_{i+k}, x_{i+k+1}) > \tfrac{1}{2} \log R.$$

As in Case II(2), we claim that $\angle x_{i-1} y_i x_{i+k+1} \geq \theta_0$, and the proof divides into following cases.

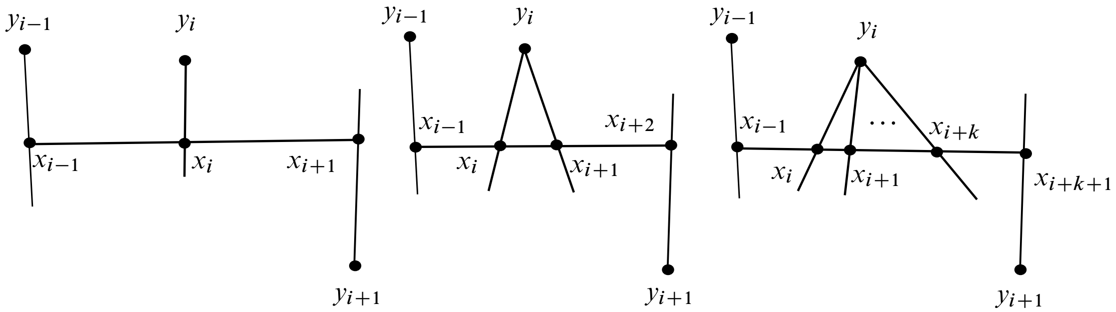


Figure 9: All of y_{i-1}, y_i and y_{i+1} are modified points.

(a) We first assume that $k=0$ holds; then the picture is shown in Figure 9, left. Since

$$d(x_{i-1},x_i), d(x_i,x_{i+1}) > \tfrac{1}{2}\log R,$$

Construction 6.7 implies that x_i should not be modified. This case is impossible.

(b) We assume that $k=1$ holds; then the picture is shown in Figure 9, middle.

- If $\overline{x_i x_{i+1}}$ is not a 3-dimensional piece, then we have $\angle x_i y_i x_{i+1} \geq \phi_0$. The argument in Case II(1)(a) implies $\angle x_{i-1} y_i x_{i+2} \geq \theta_0$.
- If $\overline{x_i x_{i+1}}$ is a 3-dimensional piece, then since

$$d(x_{i-1},x_i), d(x_{i+k},x_{i+k+1}) > \tfrac{1}{2}\log R,$$

Construction 6.7(2) implies that x_i and x_{i+1} should not be modified. This case is impossible.

(c) We assume that $k \geq 2$ holds; then the picture is shown in Figure 9, right. Then either $\overline{x_i x_{i+1}}$ or $\overline{x_{i+1} x_{i+2}}$ is not a 3-dimensional piece; thus either $\angle x_i y_i x_{i+1} \geq \phi_0$ or $\angle x_{i+1} y_i x_{i+2} \geq \phi_0$ holds. Again, the argument in Case II(1)(a) implies $\angle x_{i-1} y_i x_{i+k+1} \geq \theta_0$.

So the claim is established, and we have

$$\angle y_{i-1} y_i y_{i+1} \geq \angle x_{i-1} y_i x_{i+k+1} - \angle x_{i-1} y_i y_{i-1} - \angle x_{i+k+1} y_i y_{i+1} \geq \theta_0 - 4e^{-R/2} \geq \eta_0.$$

The proof of Case II is done and the proof of this lemma is finished. □

Now we are ready to prove Proposition 6.5.

Proof of Proposition 6.5 We first take $\eta_0 > 0$ in Lemma 6.10, and take $L > 0$ such that $\frac{1}{2}L$ satisfies the assumption of Proposition 3.5 with respect to $\frac{1}{2}\eta_0$. Then we enlarge L and take large $R > 0$ so that (6-1) and Lemma 6.10 hold.

(1) $\tilde{j}_0\colon \tilde{Z}^3 \to \mathbb{H}^3$ is a quasi-isometric embedding Since $\tilde{Z}''$ (defined before Construction 6.3) is 2-dense in $\tilde{Z}^3$, we only need to prove that the restriction $\tilde{j}_0|\colon \tilde{Z}'' \to \mathbb{H}^3$ is a quasi-isometric embedding. More precisely, for any $x, y \in \tilde{Z}''$, we will prove that

$$\tfrac{1}{2} d_{\tilde{Z}^3}(x,y) - \tfrac{3}{40}\log R - 4L \leq d_{\mathbb{H}^3}(\tilde{j}_0(x), \tilde{j}_0(y)) \leq d_{\tilde{Z}^3}(x,y) + \tfrac{3}{40}\log R + 4L. \tag{6-6}$$

We first do the following two-step modification on x and y.

Modification I If x or y lie in the $\bigl(L+\frac{1}{80}\log R\bigr)$-neighborhood of some vertex of $\widetilde{Z}^3$ (which is unique by Lemma 6.4(4)), we replace it by the corresponding vertex. So we can assume that x and y are either vertices of $\widetilde{Z}^3$ or do not lie in the $\bigl(L+\frac{1}{80}\log R\bigr)$-neighborhood of any vertex of $\widetilde{Z}^3$. Under this assumption, we only need to prove the following estimate, which implies (6-6):

$$\tfrac{1}{2}d_{\widetilde{Z}^3}(x,y)-\tfrac{1}{40}\log R\le d_{\mathbb{H}^3}(\tilde{j}_0(x),\tilde{j}_0(y))\le d_{\widetilde{Z}^3}(x,y)+\tfrac{1}{40}\log R. \tag{6-7}$$

Let γ be the shortest path in $\widetilde{Z}^3$ from x to y. If the interior of γ is contained in $\widetilde{Z}^3\setminus\widetilde{Z}^{(1)}$, then $\tilde{j}_0(\gamma)$ is a geodesic segment in $\mathbb{H}^3$; thus $d_{\widetilde{Z}^3}(x,y)=d_{\mathbb{H}^3}(\tilde{j}_0(x),\tilde{j}_0(y))$ holds and (6-7) holds. So we can assume that γ is not contained in $\widetilde{Z}^3\setminus\widetilde{Z}^{(1)}$.

Modification II We take the intersection sequence $x_1,\dots,x_n$. If

$$d(x,x_1)<\frac{1}{160}\log R\quad\text{or}\quad d(x_n,y)<\frac{1}{160}\log R,$$

we replace x or y by x_1 or x_n respectively. We still denote the new initial and terminal points by x and y, and still denote the new shortest path between x and y by γ. Then we only need to prove the following estimate, which implies (6-7):

$$\tfrac{1}{2}d_{\widetilde{Z}^3}(x,y)\le d_{\mathbb{H}^3}(\tilde{j}_0(x),\tilde{j}_0(y))\le d_{\widetilde{Z}^3}(x,y). \tag{6-8}$$

We claim that the path γ obtained by Modifications I and II satisfies Assumption 6.6. We will only argue for the initial point x, and the proof for the terminal point y is the same.

- If we did Modification I for x, then the new x is a vertex of $\widetilde{Z}^3$ and its distance to any edge of $\widetilde{Z}^{(1)}$ not containing x is at least $\log R$ (by Lemma 6.4(1)). So Modification II is not applied to x, and Assumption 6.6 holds.
- If we did not do Modification I but did Modification II for x, then the new x lies on an edge of $\widetilde{Z}^{(1)}$ and its distance to any vertex of $\widetilde{Z}^3$ is at least $\bigl(L+\frac{1}{80}\log R\bigr)-\frac{1}{160}\log R=L+\frac{1}{160}\log R$. After Modification II, if $\overline{xx_1}$ is not a 3-dimensional piece and $d(x,x_1)<\frac{1}{160}\log R$, then the edges of $\widetilde{Z}^{(1)}$ containing x and x_1 share a vertex v (by Lemma 6.4(3)) and $\angle xvx_1\ge\phi_0$ holds. So we have
$$d(x,x_1)\ge d(x,v)+d(v,x_1)-I(\pi-\phi_0)\ge\Bigl(L+\frac{1}{160}\log R\Bigr)-I(\pi-\phi_0)>\frac{1}{160}\log R,$$
which is impossible. So Assumption 6.6 holds in this case. If $\overline{xx_1}$ is a 3-dimensional piece, Assumption 6.6 trivially holds.
- If we do neither Modification I nor Modification II for x, then the distance between x and any vertex of $\widetilde{Z}^3$ is at least $L+\frac{1}{80}\log R>L+\frac{1}{160}\log R$, and we have $d(x,x_1)\ge\frac{1}{160}\log R$. So Assumption 6.6 holds.

Now we take the modified sequence $y_1,\dots,y_m$ of γ, and let γ_i' be the shortest path (in a piece of $\widetilde{Z}^3\setminus\widetilde{Z}^{(1)}$) from y_i to y_{i+1}. The modified path γ' is the concatenation of $\gamma_0',\gamma_1',\dots,\gamma_m'$.

For each consecutive γ_i' and γ_{i+1}' in the modified path γ, we have the following possibilities.

- If neither of them are unmodified 3-dimensional pieces, then Lemma 6.8 implies $l(\gamma_i'), l(\gamma_{i+1}') \geq L$, and Lemma 6.10(1) implies the bending angle is at most $\pi - \eta_0$.
- If either γ_i' or γ_{i+1}' is an unmodified 3-dimensional piece, exactly one of them is. Then Lemma 6.8 implies that one of $l(\gamma_i')$ or $l(\gamma_{i+1}')$ is at least L, and Lemma 6.10(2) implies the bending angle is at most $\frac{\pi}{2} - \eta_0$.

Then Proposition 3.5 implies

$$d_{\mathbb{H}^3}(\tilde{j}_0(x), \tilde{j}_0(y)) = l(\gamma_0'\gamma_1'\cdots\gamma_m') \geq \frac{1}{2}\sum_{i=0}^{m} l(\gamma_i') = \frac{1}{2}\sum_{i=0}^{m} d_{\widetilde{Z}^3}(y_i, y_{i+1}) \geq \frac{1}{2} d_{\widetilde{Z}^3}(x,y). \tag{6-9}$$

On the other hand, since the metric on $\widetilde{Z}^3$ is a path metric induced by the metric of $\mathbb{H}^3$, we always have $d_{\mathbb{H}^3}(\tilde{j}_0(x), \tilde{j}_0(y)) \leq d_{\widetilde{Z}^3}(x,y)$. So (6-8) holds for the path γ obtained after Modifications I and II; thus (6-6) holds for any $x, y \in \widetilde{Z}''$. This implies that

$$\tfrac{1}{2} d_{\widetilde{Z}^3}(x,y) - \tfrac{3}{40}\log R - 4L - 8 \leq d_{\mathbb{H}^3}(\tilde{j}_0(x), \tilde{j}_0(y)) \leq d_{\widetilde{Z}^3}(x,y) + \tfrac{3}{40}\log R + 4L + 8 \tag{6-10}$$

holds for any $x, y \in \widetilde{Z}^3$, by the 2-denseness of $\widetilde{Z}'' \subset \widetilde{Z}^3$. So $\tilde{j}_0 \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is a quasi-isometric embedding.

(2) π_1-injectivity of ρ_0 Since $\pi_1(Z) \cong \pi_1(Z^3)$ is torsion-free and $\tilde{j}_0 \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is ρ_0-equivariant, the fact that $\tilde{j}_0$ is a quasi-isometric embedding implies that ρ_0 is injective.

Moreover, since $\pi_1(Z^3)$ is neither a surface group nor a free group, the covering theorem [6] implies that $\rho_0(\pi_1(Z^3)) < \mathrm{Isom}_+(\mathbb{H}^3)$ is a geometrically finite subgroup.

(3) Injectivity of $\tilde{j}_0$ Now we prove that $\tilde{j}_0 \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is injective. For $x, y \in \widetilde{Z}^3$ such that

$$d_{\widetilde{Z}^3}(x,y) > \tfrac{1}{5}\log R > \tfrac{3}{20}\log R + 8L + 16,$$

the left hand side of (6-10) implies $\tilde{j}_0(x) \neq \tilde{j}_0(y)$. Moreover, if x and y lie in the same component of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$, then since $\tilde{j}_0$ restricts to an embedding on this component, we have $\tilde{j}_0(x) \neq \tilde{j}_0(y)$.

So we can assume that $d_{\widetilde{Z}^3}(x,y) \leq \frac{1}{5}\log R$ holds, while x and y lie in different components of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$. We take the shortest path γ in $\widetilde{Z}^3$ from x to y and take the intersection sequence $x_1, \dots, x_n$. Let $x = x_0$ and $y = x_{n+1}$, and we denote the subpath of γ from x_i to x_{i+1} by γ_i. Then we have $l(\gamma_i) \leq \frac{1}{5}\log R$ for all $i = 0, \dots, n$. So for any γ_i with $i = 1, \dots, n-1$, one of the following hold:

(i) either γ_i is a 3-dimensional piece,

(ii) or γ_i is a 2-dimensional piece, and by Lemma 6.4(3), the two edges of $\widetilde{Z}^{(1)}$ containing x_i and x_{i+1} share a vertex v_i; moreover, since $\angle x_i v_i x_{i+1} \geq \phi_0$, we have $d(x_i, v_i), d(x_{i+1}, v_i) < \frac{2}{5}\log R$.

Moreover, γ contains at most one 3-dimensional piece, since by Lemma 6.4(2), any two different 3-dimensional components of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ have distance at least $\frac{9}{10}\log R$.

Now we prove $\tilde{j}_0(x) \neq \tilde{j}_0(y)$ by dividing into the following cases.

(a) If γ contains no 3-dimensional pieces, then all vertices v_i in item (ii) above must be the same vertex, thus x and y lie in two 2-dimensional pieces of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ that share a vertex. Since $\tilde{j}_0$ maps these two pieces to two totally geodesic subsurfaces in $\mathbb{H}^3$ that are disjoint except at the common edge or vertex, we have $\tilde{j}_0(x) \neq \tilde{j}_0(y)$. So we can assume that γ contains exactly one 3-dimensional piece in the following.

(b) If γ_0 or γ_n is a 3-dimensional piece, we assume that γ_0 is. Then by item (ii) again, the 3-dimensional piece of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ containing x and the 2-dimensional piece of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ containing y share a vertex. Then as in case (a), the geometry of $\tilde{j}_0$ implies $\tilde{j}_0(x) \neq \tilde{j}_0(y)$. So we further assume that neither γ_0 nor γ_n are 3-dimensional pieces, and $n \geq 2$ holds in the following.

(c) If $n = 2$, then γ_1 is a 3-dimensional piece. By the proof of Lemma 6.10, Case I (1)(b), the bending angle at y_1 and y_2 are both at most $\frac{2}{5}\pi < \frac{\pi}{2}$, so we have $\tilde{j}_0(x) \neq \tilde{j}_0(y)$. So we can further assume that $n \geq 3$ in the following.

(d) So $n \geq 3$. For some $1 \leq i \leq n-1$, γ_i is the unique 3-dimensional piece of γ. Then x_i is $\left(\frac{2}{5}\log R\right)$-close to a vertex v_i of $\widetilde{Z}^{(1)}$, and x_{i+1} is $\left(\frac{2}{5}\log R\right)$-close to a vertex v_{i+1}. We must have $v_i = v_{i+1}$, otherwise $d_{\widetilde{Z}^3}(v_i, v_{i+1}) \geq \log R$ by Lemma 6.4(4) and $d_{\widetilde{Z}^3}(x_i, x_{i+1}) \geq \frac{1}{5}\log R$, which is impossible. So both x_i and x_{i+1} lie on edges of $\widetilde{Z}^{(1)}$ containing v_i, and by item (ii), the 2-dimensional pieces of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ containing x and y share the vertex v_i. So we obtain $\tilde{j}_0(x) \neq \tilde{j}_0(y)$ as in case (a).

The proof of the injectivity of $\tilde{j}_0 \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is finished.

(4) Homeomorphic type of the convex core Since $\tilde{j}_0 \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is injective and is ρ_0-equivariant, the image $\tilde{j}_0(\widetilde{Z}^3)$ has a closed $\rho_0(\pi_1(Z^3))$-equivariant neighborhood $\mathcal{N}(\widetilde{Z}^3)$ in $\mathbb{H}^3$. By the construction of $\mathscr{Z}$, we can see that $\mathcal{N}(\widetilde{Z}^3)/\rho_0(\pi_1(Z^3))$ is homeomorphic to $\mathscr{Z} \setminus \partial_p\mathscr{Z}$, as oriented manifolds.

Also note that $\mathcal{N}(\widetilde{Z}^3)/\rho_0(\pi_1(Z^3))$ is a finite volume submanifold of $\mathbb{H}^3/\rho_0(\pi_1(Z^3))$ such that the inclusion induces an isomorphism on π_1. Since all boundary components of $\mathscr{Z}$ are incompressible, by tameness of open hyperbolic 3-manifolds [1; 5], each component of

$$\big(\mathbb{H}^3/\rho_0(\pi_1(Z^3))\big) \setminus \big(\mathcal{N}(\widetilde{Z}^3)/\rho_0(\pi_1(Z^3))\big)$$

is homeomorphic to the product of a surface and $(0, \infty)$. So $\mathbb{H}^3/\rho_0(\pi_1(Z^3))$ is homeomorphic to $\mathscr{Z} \setminus \partial_p\mathscr{Z}$.

Since $\rho_0(\pi_1(Z^3)) < \mathrm{Isom}_+(\mathbb{H}^3)$ is a geometrically finite subgroup and $\partial_p\mathscr{Z}$ corresponds to cusp ends of $\mathbb{H}^3/\rho_0(\pi_1(Z^3))$, the convex core of $\mathbb{H}^3/\rho_0(\pi_1(Z^3))$ is homeomorphic to $\mathscr{Z} \setminus \partial_p\mathscr{Z}$. Note that all above homeomorphisms preserve natural orientations on involved manifolds. □

6.3 Proof of quasi-isometric embedding

The following proposition is the main result of this subsection, which is the last technical piece of the proof of Theorem 5.17.

Proposition 6.11 *For any $t \in [0,1]$, $\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is a quasi-isometric embedding.*

To prove Proposition 6.11, we need the following two lemmas. The first lemma appeared as Lemma 5.7 of [23], which estimates the geometry of $\tilde{j}_t$ on 2-dimensional pieces of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$.

Lemma 6.12 *For any $\delta \in (0, 10^{-6})$, there exists $\epsilon_0 > 0$ and $R_0 > 0$, such that for any positive numbers $\epsilon \in (0, \epsilon_0)$, $R > R_0$ and any positive integer R' greater than all of the R_{ij} and R_{ijk} in Lemma 5.14, the following statement holds.*

If $\{\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3 \mid t \in [0,1]\}$ is constructed with respect to ϵ, R and R', then for any $t \in [0,1]$ and any x and y lying in the closure of a 2-dimensional piece $C \subset \widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$ such that $x \in \partial C$, we have

$$\tfrac{1}{2} d_{\mathbb{H}^3}(\tilde{j}_t(x), \tilde{j}_t(y)) \le d_{\widetilde{Z}^3}(x, y) \le 2 d_{\mathbb{H}^3}(\tilde{j}_t(x), \tilde{j}_t(y)). \tag{6-11}$$

Moreover, let e be an edge in $\widetilde{Z}^{(1)}$ containing x (with a preferred orientation). If $d(x,y) \ge 100$, then

$$d_{S^2}\big(\Theta(x, y, e), \Theta(\tilde{j}_t(x), \tilde{j}_t(y), \tilde{j}_t(e))\big) < 10\delta. \tag{6-12}$$

Here $\Theta(x,y,e)$ denotes the point in S^2 determined by the tangent vector of $\overline{xy}$ in $\mathbb{H}^3$, with respect to a coordinate of $T_x(\mathbb{H}^3)$ given by a frame $\boldsymbol{p} = (x, \vec{v}, \vec{n})$, where $\vec{v}$ is tangent to e and $\vec{n}$ is tangent to C (points inward). Similarly, $\Theta(\tilde{j}_t(x), \tilde{j}_t(y), \tilde{j}_t(e))$ is defined with respect to a frame based at $\tilde{j}_t(x)$, with the first vector tangent to $\tilde{j}_t(e)$, and the second vector is ϵ-close to be tangent to $\tilde{j}_t(C)$ (points inward).

The second lemma estimates the geometry of $\tilde{j}_t$ on 3-dimensional pieces of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$. Since this lemma only concerns 3-dimensional pieces of $\widetilde{Z}^3$, we do not need the R_{ij}, R_{ijk} and R' part of $\widetilde{Z}^3$, but we still state them in the following lemma, so that its statement is parallel with the statement of Lemma 6.12. We will only give a sketch of the proof of this lemma and some computations are skipped.

Lemma 6.13 *For any $\delta \in (0, 10^{-6})$, there exists $\epsilon_0 > 0$ and $R_0 > 0$, such that for any positive numbers $\epsilon \in (0, \epsilon_0)$, $R > R_0$ and any positive integer R' greater than all of the R_{ij} and R_{ijk} in Lemma 5.14, the following statement holds.*

If $\{\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3 \mid t \in [0,1]\}$ is constructed with respect to ϵ, R and R', then for any $t \in [0,1]$, the following hold. For any x and y lying in the closure of a 3-dimensional piece $C \subset \widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$, we have

$$(1-\delta) d_{\mathbb{H}^3}(\tilde{j}_t(x), \tilde{j}_t(y)) \le d_{\widetilde{Z}^3}(x, y) \le (1+\delta) d_{\mathbb{H}^3}(\tilde{j}_t(x), \tilde{j}_t(y)). \tag{6-13}$$

Moreover, if x belongs to an oriented edge $e \subset \widetilde{Z}^{(1)}$ contained in the boundary of C, then

$$d_{S^2}\big(\Theta(x, y, e), \Theta(\tilde{j}_t(x), \tilde{j}_t(y), \tilde{j}_t(e))\big) < 10\delta. \tag{6-14}$$

Here $\Theta(x,y,e)$ denotes the point in S^2 determined by the tangent vector of $\overline{xy}$ in $\mathbb{H}^3$, with respect to a coordinate of $T_x(\mathbb{H}^3)$ given by a frame $\boldsymbol{p} = (x, \vec{v}, \vec{n})$, where $\vec{v}$ is tangent to e and $\vec{n}$ is tangent to a face of C (points inward). $\Theta(\tilde{j}_t(x), \tilde{j}_t(y), \tilde{j}_t(e))$ is defined by a similar frame based at $\tilde{j}_t(x)$, given by $\tilde{j}_t(e)$ and $\tilde{j}_t(C)$.

Proof Note that we did not give a precise definition of $\tilde{j}_t$ on ideal tetrahedra of $\widetilde{Z}^3$ in Construction 6.3(4), so it suffices to prove this lemma for some choice of $\tilde{j}_t$.

We can use the Klein model of the hyperbolic space to (noncanonically) identify each ideal tetrahedron in $\widetilde{Z}^3$ with a 3-simplex in the 3-ball (with one vertex on the boundary). Then we use the linear structure of the 3-simplex to define the $\tilde{j}_t$ map on each tetrahedron, which is piecewise smooth on C. Moreover, if $\epsilon > 0$ is small enough, the restriction of $\tilde{j}_t$ on each ideal tetrahedron is very close to an isometry, up to the second derivative, in the following sense.

(1) For any two unit tangent vectors $\vec{v}_1$ and $\vec{v}_2$ based at the same point $z \in C$, we have

$$|\langle \vec{v}_1, \vec{v}_2\rangle - \langle D\tilde{j}_t(\vec{v}_1), D\tilde{j}_t(\vec{v}_1)\rangle| < \delta^3. \tag{6-15}$$

Here if z lies in the boundary of an ideal tetrahedron, then $\vec{v}_1$ and $\vec{v}_2$ point toward the same ideal tetrahedron.

(2) For any geodesic γ contained in an ideal tetrahedron contained in C, the geodesic curvature of $\tilde{j}_t \circ \gamma$ is always bounded above by δ^3.

Equation (6-15) implies that $\tilde{j}_t|_C$ is a $(1+\delta)$-bi-Lipschitz map, so (6-13) holds.

Now we work on the angle estimate, and we identify both C and $\tilde{j}_t(C)$ with convex subsets of the upper half space model of $\mathbb{H}^3$, such that all vertices have z-coordinate 1. We take projections of x and y to $\mathbb{R}^2$, and let the Euclidean distance between these projections by d.

Case I We first suppose that $d \geq 4R/\delta$. Since the z-coordinate of x is at most $\sqrt{(R/2)^2+1}$ (by Construction 6.3(3)), an elementary computation implies that the tangent vector of $\overline{xy}$ at x is at most δ away from $(0,0,1)$. Since $\tilde{j}_t|_C$ is induced by an almost isometry of the equilateral tessellation of $\mathbb{R}^2$, the Euclidean distance between the projections of $\tilde{j}_t(x)$ and $\tilde{j}_t(y)$ to $\mathbb{R}^2$ is at least $2R/\delta$. So the tangent vector of $\overline{\tilde{j}_t(x)\tilde{j}_t(y)}$ at $\tilde{j}_t(x)$ is at most 2δ away from $(0,0,1)$. Since the geometry of C and $\tilde{j}_t(C)$ are close on their boundaries, (6-14) holds in this case.

Case II Now we suppose that $d \leq 4R/\delta$. Since the equilateral tessellation of $\mathbb{R}^2$ has side length R, an elementary area estimate implies that $\overline{xy}$ intersects with $m \leq 20/\delta$ ideal tetrahedra of C. Let $\gamma_1, \gamma_2, \dots, \gamma_m$ be the intersections of $\overline{xy}$ with these tetrahedra such that their concatenation gives $\overline{xy}$. Then $\tilde{j}_t \circ \gamma_1, \tilde{j}_t \circ \gamma_2, \dots, \tilde{j}_t \circ \gamma_m$ are smooth curves in tetrahedra of $\tilde{j}_t(C)$ such that their concatenation is homotopic to $\overline{\tilde{j}_t(x)\tilde{j}_t(y)}$, and their geodesic curvatures are always bounded above by δ^3.

For each $i = 1, 2, \dots, m-1$, by (6-15), the angle between the terminal tangent vector of $\tilde{j}_t(\gamma_i)$ and the initial tangent vector of $\tilde{j}_t(\gamma_{i+1})$ is at most $4\delta^3$.

Let γ_i' be the geodesic segment that share endpoints with $\tilde{j}_t(\gamma_i)$. The condition on geodesic curvatures implies that the initial tangent vectors of γ_i' and $\tilde{j}_t(\gamma_i)$ differ by at most $2\delta^3$, and the same holds for their terminal tangent vectors. So the angle between the terminal tangent vector of γ_i' and the initial tangent vector of γ_{i+1}' is at most $10\delta^3$.

Since we have $m \le 20/\delta$ geodesic segments γ_i', then initial tangent vectors of γ_1' and $\overline{\tilde{j}_t(x)\tilde{j}_t(y)}$ differ by at most $10\delta^3 \cdot 20/\delta = 200\delta^2 \le \delta$. Since the tangent vector of γ_1' is $2\delta^3$-close to the tangent vector of $\tilde{j}_t \circ \gamma_1$, (6-14) holds in this case. □

Given these two lemmas, we are ready to prove Proposition 6.11.

Proof of Proposition 6.11 For η_0 given in Lemma 6.10, we take $\delta \in (0, \eta_0/40)$. Then we take small $\epsilon > 0$ and large $R > 0$ satisfying Lemmas 6.12 and 6.13.

To prove $\tilde{j}_t$ is a quasi-isometric embedding, for any two points $x, y \in \widetilde{Z}^3$, we want to prove the following inequality:

$$\text{(6-16)} \qquad \frac{1}{4} d_{\widetilde{Z}}(x, y) - 5\Big(L + \frac{3}{160}\log R + 2\Big) \le d_{\mathbb{H}^3}(\tilde{j}_t(x), \tilde{j}_t(y)) \\ \le 4 d_{\widetilde{Z}^3}(x, y) + 12\Big(L + \frac{3}{160}\log R + 2\Big).$$

As in the proof of Proposition 6.5, we replace x and y by two 2-close points in $\widetilde{Z}''$ and then do Modifications I and II. After this modification process, the shortest path γ in $\widetilde{Z}^3$ from x to y satisfies Assumption 6.6.

Recall that the modification process moves both x and y by distance at most $L + \frac{3}{160}\log R + 2$. Lemmas 6.12 and 6.13 imply that $\tilde{j}_t(x)$ and $\tilde{j}_t(y)$ are moved by distance at most $2\big(L + \frac{3}{160}\log R + 2\big)$. So to prove (6-16), it suffices to prove the following inequality for points $x, y \in \widetilde{Z}''$ satisfying Assumption 6.6:

$$\text{(6-17)} \qquad \tfrac{1}{4} d_{\widetilde{Z}^3}(x, y) \le d_{\mathbb{H}^3}(\tilde{j}_t(x), \tilde{j}_t(y)) \le 4 d_{\widetilde{Z}^3}(x, y).$$

For the new x and y, we take the shortest path γ in $\widetilde{Z}^3$ from x to y, and take the modified sequence $y_1, \dots, y_n$. Let γ_i' be the shortest path in $\widetilde{Z}^3$ from y_i to y_{i+1} (in the closure of a component of $\widetilde{Z}^3 \setminus \widetilde{Z}^{(1)}$), and let γ' be the concatenation of γ_i' for $i = 0, 1, \dots, n$. Since γ satisfies Assumption 6.6, Lemmas 6.8 and 6.10 imply the following:

- Each γ_i' is either an unmodified 3-dimensional piece, or $l(\gamma_i') \ge L$ holds.
- If neither γ_{i-1}' nor γ_i' are unmodified 3-dimensional pieces, then $\angle y_{i-1} y_i y_{i+1} \ge \eta_0$.
- If γ_{i-1}' or γ_i' is an unmodified 3-dimensional piece, then $\angle y_{i-1} y_i y_{i+1} \ge \frac{\pi}{2} + \eta_0$.

Let $z_i = \tilde{j}_t(y_i)$, and let δ_i be the geodesic segment in $\mathbb{H}^3$ from z_i to z_{i+1}. Then we have $z_0 = \tilde{j}_t(x)$ and $z_{n+1} = \tilde{j}_t(y)$. By Lemmas 6.12 and 6.13, since $\delta < \eta_0/40$, the following conditions hold for δ_i.

(1) For any $i = 0, \dots, n$, we have $\frac{1}{2} l(\gamma_i') \le l(\delta_i) \le 2 l(\gamma_i')$. Moreover, if γ_i' is not an unmodified 3-dimensional piece, $l(\delta_i) \ge \frac{1}{2} L$ holds.

(2) If neither γ_{i-1}' nor γ_i' are unmodified 3-dimensional pieces, then $\angle z_{i-1} z_i z_{i+1} \ge \frac{1}{2}\eta_0$.

(3) If γ_{i-1}' or γ_i' is an unmodified 3-dimensional piece, then $\angle z_{i-1} z_i z_{i+1} \ge \frac{\pi}{2} + \frac{1}{2}\eta_0$.

On one hand, we have

$$\begin{aligned} d_{\mathbb{H}^3}(\tilde{j}_t(x), \tilde{j}_t(y)) &\le \sum_{i=0}^{n} l(\delta_i) \\ &\le 2\sum_{i=0}^{n} l(\gamma_i') && \text{(by item (1))} \\ &\le 4d_{\mathbb{H}^3}(\tilde{j}_0(x), \tilde{j}_0(y)) && \text{(by (6-9))} \\ &\le 4d_{\widetilde{Z}^3}(x, y). \end{aligned}$$

On the other hand, since L is large with respect to η_0, items (2) and (3) and Proposition 3.5 imply that

$$\begin{aligned} d_{\mathbb{H}^3}(\tilde{j}_t(x), \tilde{j}_t(y)) &\ge \frac{1}{2}\sum_{i=0}^{n} l(\delta_i) && \text{(by Proposition 3.5)} \\ &\ge \frac{1}{4}\sum_{i=0}^{n} l(\gamma_i') && \text{(by item (1))} \\ &\ge \frac{1}{4}d_{\widetilde{Z}^3}(x, y). \end{aligned}$$

We have proved (6-17) holds for the modified endpoints x and y; thus (6-16) holds for any $x, y \in \widetilde{Z}^3$. So $\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is a quasi-isometric embedding. □

Now we finish the proof of Theorem 5.17.

Proof of Theorem 5.17 By Proposition 6.11, each $\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is a quasi-isometric embedding. Moreover, since $\pi_1(Z^3)$ is torsion free and $\tilde{j}_t$ is ρ_t-equivariant, each representation

$$\rho_t \colon \pi_1(Z^3) \to \mathrm{Isom}_+(\mathbb{H}^3)$$

is injective. Again, since $\tilde{j}_t \colon \widetilde{Z}^3 \to \mathbb{H}^3$ is a quasi-isometric embedding and $Z^3 = \widetilde{Z}^3/\pi_1(Z^3)$ is compact after truncating cusp ends, $\rho_t(\pi_1(Z^3)) < \mathrm{Isom}_+(\mathbb{H}^3)$ is a geometrically finite subgroup.

So $\{\rho_t(\pi_1(Z^3)) \mid t \in [0,1]\}$ forms a continuous family of geometrically finite subgroups of $\mathrm{Isom}_+(\mathbb{H}^3)$. Then the convex core of $\mathbb{H}^3/j_*(\pi_1(Z^3)) = \mathbb{H}^3/\rho_1(\pi_1(Z^3))$ is homeomorphic to the convex core of $\mathbb{H}^3/\rho_0(\pi_1(Z^3))$, which is homeomorphic to $\mathscr{X} \setminus \partial_p\mathscr{X}$ (as oriented manifolds) by Proposition 6.5(4). □

References

[1] **I Agol**, *Tameness of hyperbolic* 3*-manifolds*, preprint (2004) arXiv math/0405568

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

[3] **I Berstein**, **A L Edmonds**, *On the construction of branched coverings of low-dimensional manifolds*, Trans. Amer. Math. Soc. 247 (1979) 87–124 MR Zbl

[4] **M Boileau**, **S Wang**, *Non-zero degree maps and surface bundles over S^1*, J. Differential Geom. 43 (1996) 789–806 MR Zbl

[5] **D Calegari**, **D Gabai**, *Shrinkwrapping and the taming of hyperbolic 3-manifolds*, J. Amer. Math. Soc. 19 (2006) 385–446 MR Zbl

[6] **R D Canary**, *A covering theorem for hyperbolic 3-manifolds and its applications*, Topology 35 (1996) 751–778 MR Zbl

[7] **J A Carlson**, **D Toledo**, *Harmonic mappings of Kähler manifolds to locally symmetric spaces*, Inst. Hautes Études Sci. Publ. Math. 69 (1989) 173–201 MR Zbl

[8] **A L Edmonds**, *Deformation of maps to branched coverings in dimension three*, Math. Ann. 245 (1979) 273–279 MR Zbl

[9] **A Gaifullin**, *Universal realisators for homology classes*, Geom. Topol. 17 (2013) 1745–1772 MR Zbl

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

[11] **J Kahn**, **A Wright**, *Nearly Fuchsian surface subgroups of finite covolume Kleinian groups*, Duke Math. J. 170 (2021) 503–573 MR Zbl

[12] **Y Liu**, *Immersing quasi-Fuchsian surfaces of odd Euler characteristic in closed hyperbolic 3-manifolds*, J. Differential Geom. 111 (2019) 457–493 MR Zbl

[13] **Y Liu**, **V Markovic**, *Homology of curves and surfaces in closed hyperbolic 3-manifolds*, Duke Math. J. 164 (2015) 2723–2808 MR Zbl

[14] **Y Liu**, **H Sun**, *Virtual 1-domination of 3-manifolds*, Compos. Math. 154 (2018) 621–639 MR Zbl

[15] **E Martínez-Pedroza**, *Combination of quasiconvex subgroups of relatively hyperbolic groups*, Groups Geom. Dyn. 3 (2009) 317–342 MR Zbl

[16] **R Myers**, *Homology cobordisms, link concordances, and hyperbolic 3-manifolds*, Trans. Amer. Math. Soc. 278 (1983) 271–288 MR Zbl

[17] **P Przytycki**, **D T Wise**, *Mixed 3-manifolds are virtually special*, J. Amer. Math. Soc. 31 (2018) 319–347 MR Zbl

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

[19] **H Sun**, *Virtual domination of 3-manifolds*, Geom. Topol. 19 (2015) 2277–2328 MR Zbl

[20] **H Sun**, *Virtual homological torsion of closed hyperbolic 3-manifolds*, J. Differential Geom. 100 (2015) 547–583 MR Zbl

[21] **H Sun**, *A characterization on separable subgroups of 3-manifold groups*, J. Topol. 13 (2020) 187–236 MR Zbl

[22] **H Sun**, *The panted cobordism groups of cusped hyperbolic 3-manifolds*, J. Topol. 15 (2022) 1580–1634 MR Zbl

[23] **H Sun**, *Virtual domination of 3-manifolds, II*, J. Lond. Math. Soc. 108 (2023) 869–915 MR Zbl

Department of Mathematics, Rutgers University
Piscataway, NJ, United States

hongbin.sun@rutgers.edu

Received: 21 April 2023 Revised: 31 October 2023

Algebraic & Geometric Topology **25**:3 (2025) 1667–1730
DOI: 10.2140/agt.2025.25.1667
Published: 20 June 2025

The Kakimizu complex for genus one hyperbolic knots in the 3-sphere

LUIS G VALDEZ-SÁNCHEZ

The Kakimizu complex $\mathrm{MS}(K)$ for a knot $K \subset \mathbb{S}^3$ is the simplicial complex with vertices the isotopy classes of minimal genus Seifert surfaces in the exterior of K and simplices any set of vertices with mutually disjoint representative surfaces. We determine the structure of the Kakimizu complex $\mathrm{MS}(K)$ of genus one hyperbolic knots $K \subset \mathbb{S}^3$. In contrast with the case of hyperbolic knots of higher genus, it is known that the dimension d of $\mathrm{MS}(K)$ is universally bounded by 4, and we show that $\mathrm{MS}(K)$ consists of a single d-simplex for $d = 0, 4$ and otherwise of at most two d-simplices which intersect in a common $(d-1)$-face. For the cases $1 \leq d \leq 3$ we also construct infinitely many examples of such knots where $\mathrm{MS}(K)$ consists of two d-simplices.

57K10; 57K30

1 Introduction

Let K be a knot in the 3-sphere $\mathbb{S}^3$ with exterior $X_K = \mathbb{S}^3 \setminus \operatorname{int} N(K)$, where $N(K) \subset \mathbb{S}^3$ is regular neighborhood of K. The knot K is the boundary of orientable, compact and connected surfaces embedded in $\mathbb{S}^3$, called Seifert surfaces for the knot. Equivalently, there is a unique slope $J \subset \partial X_K$, the longitude of K, that bounds orientable compact surfaces in X_K which correspond in a natural way to the Seifert surfaces in $\mathbb{S}^3$ bounded by the knot. The genus of the knot K, a topological invariant of K, is then defined as the smallest genus of the Seifert surfaces bounded by the knot.

The Kakimizu complex $\mathrm{MS}(K)$ was defined in [12] for knots (and links) $K \subset \mathbb{S}^3$ as the simplicial complex with vertices the equivalence isotopy classes of minimal genus Seifert surfaces for K in X_K, such that any set of vertices with mutually disjoint representative surfaces comprise a simplex. For instance, it is well known that the figure-eight knot bounds a unique Seifert torus and so its Kakimizu complex consists of a single 0-simplex.

For hyperbolic knots $K \subset \mathbb{S}^3$ much is known about the complex $\mathrm{MS}(K)$. It is a consequence of Eisner [4] that $\mathrm{MS}(K)$ is a finite complex. It is also known that $\mathrm{MS}(K)$ is a flag simplicial complex, by Schultens [19], which is connected, by Scharlemann and Thompson [18], and contractible, by Przytycki and Schultens [15].

For the family of hyperbolic knots $K \subset \mathbb{S}^3$ of a fixed genus $g \geq 2$ no universal bound on the dimension of $\mathrm{MS}(K)$ is known. Moreover, Y Tsutsumi [22] shows that for each genus $g \geq 2$ there are hyperbolic knots $K \subset \mathbb{S}^3$ of genus g such that the number of vertices of $\mathrm{MS}(K)$, and hence of simplices, is arbitrarily large.

In [17] M Sakuma and K J Shackleton provide a bound for the diameter of the (1-skeleton of) $\mathrm{MS}(K)$, quadratic in the genus of the knot, and show that the diameter of $\mathrm{MS}(K)$ is 1 or 2 for genus one knots, with an example that realizes the diameter of 2. These bounds on the diameter of $\mathrm{MS}(K)$ do not however bound the number of top-dimensional simplices present in $\mathrm{MS}(K)$.

In this paper we determine the structure of the Kakimizu complex for genus one hyperbolic knots $K \subset \mathbb{S}^3$ and obtain a picture opposite to that of the case of genus $g \geq 2$. Our main result is the following.

Theorem 1 *If $K \subset \mathbb{S}^3$ is a genus one hyperbolic knot and $d = \dim \mathrm{MS}(K)$ then*

(1) $0 \leq d \leq 4$;

(2) $\mathrm{MS}(K)$ *consists of at most two d-dimensional simplices which intersect in a common $(d-1)$-face, and exactly one d-simplex if $d = 0, 4$;*

(3) *for each integer $1 \leq d \leq 3$ there are infinitely many genus one hyperbolic knots $K \subset \mathbb{S}^3$ such that* $\mathrm{MS}(K)$ *consists of two d-simplices.*

A simplex of $\mathrm{MS}(K)$ corresponds to a collection of mutually disjoint and nonparallel Seifert tori

$$\mathbb{T} = T_1 \sqcup \cdots \sqcup T_N \subset X_K.$$

We refer to $\mathbb{T}$ as a *simplicial collection* of Seifert tori for short and assume that its components are labeled consecutively as they appear around X_K, as shown in Figure 4. The collection $\mathbb{T}$ is maximal if its number of components $|\mathbb{T}| = N$ is largest among all possible simplicial collections of Seifert tori in X_K, in which case $\dim \mathrm{MS}(K) = N - 1$. Components T_i and T_j of $\mathbb{T}$ then cobound a region $R_{i,j}$ in X_K with boundary a surface of genus two that contains the longitude slope J of K.

It was proved in [24] that the exterior X_K of a genus one hyperbolic knot in $\mathbb{S}^3$ contains at most 5 mutually disjoint and nonparallel Seifert tori, which gives the bound $\dim \mathrm{MS}(K) \leq 4$ in Theorem 1(1). This was achieved by studying the properties of maximal simplicial collections $\mathbb{T}$ of Seifert tori in X_K with at least 5 components; all but at most one of the complementary regions $R_{i,i+1} \subset X_K$ of $\mathbb{T}$ are then genus two handlebodies whose structure was determined in [24].

It is the low genus of the complementary handlebody regions of $\mathbb{T}$ that makes it possible to establish Theorem 1(2) and construct the examples in Theorem 1(3). One of the difficulties here, which need not be handled in detail in [24], is that some region $R_{i,i+1} \subset X_K$ in a maximal simplicial collection may not be a handlebody.

The paper is organized as follows. Most of Sections 2 and 3 contain background material and extension of definitions from [24] adapted to the needs of the present paper. Specifically, in Section 2 we introduce the notation and some basic results that are used throughout the paper. The definition of a pair (H, J) given in [24], that H is a genus two handlebody and $J \subset \partial H$ a separating circle which is nontrivial in H, is updated to include 3-manifolds H other than handlebodies. Pairs of the form $(R_{i,j}, J)$, where $J \subset \partial R_{i,j}$ is the longitudinal slope in ∂X_K, are then naturally produced by any maximal simplicial collection of Seifert tori $\mathbb{T} \subset X_K$. The structure of several types of handlebody pairs is also described in some detail.

In Section 4 (see Lemma 4.2.1) the structure of $\mathrm{MS}(K)$ is shown to depend on the presence of *annular pairs* $(R_{i,j}, J)$: pairs such that $R_{i,j}$ contains an incompressible spanning annulus with one boundary component in T_i and the other in T_j.

The properties of general annular pairs are established in Section 3, while the properties of the annular pairs in $X_K \subset \mathbb{S}^3$ for K a hyperbolic knot are developed further in Section 5. Two major restrictions arise once we restrict our attention to annular pairs $(R_{i,j}, J)$ in $\mathbb{S}^3$. The first one concerns the case where $R_{i,j}$ is not a genus two handlebody. In this case $R_{i,j}$ must be an atoroidal, irreducible and boundary irreducible manifold which is the complement in $\mathbb{S}^3$ of a genus two handlebody $V \subset \mathbb{S}^3$. In the context of Koda and Ozawa [13], V is a nontrivial genus two handlebody knot in $\mathbb{S}^3$ and as such it has a restricted structure outlined in Lemma 5.1.1. The second one, the content of Proposition 5.2.3, restricts even further the structure of the annular pair $(R_{i,j}, J)$ so that $R_{i,j} = R_{i,i+1}$ or $R_{i,j} = R_{i,i+2}$; that is, $R_{i,j}$ may contain at most one Seifert torus not parallel to T_i or T_j. It is also established in Proposition 5.2.3 that $\mathrm{MS}(K)$ has more than one top-dimensional simplex if and only if for each maximal simplicial collection $\mathbb{T} \subset X_K$ there is at least one annular pair of the form $(R_{i,i+2}, J)$.

We call an annular pair of the form $(R_{i,i+2}, J)$ an *exchange pair*, given that there is a Seifert torus T'_{i+1} in $R_{i,i+2}$ that can be exchanged with $T_{i+1} \subset R_{i,i+2}$ to construct a new simplicial collection $\mathbb{T}'$ not isotopic to $\mathbb{T}$ in X_K. The properties of this *exchange trick* are established in Sections 5.2 and 5.3.

In Section 6 we show that any maximal simplicial collection $\mathbb{T} \subset X_K$ of size 5 does not produce exchange pairs and hence that such a collection $\mathbb{T}$ is unique up to isotopy. Maximal simplicial collections $\mathbb{T} \subset X_K$ of size $2 \leq |\mathbb{T}| \leq 4$ are handled in Section 7, where it is shown in Proposition 7.0.2 that $\mathbb{T}$ produces at most one exchange pair and hence that, up to isotopy, there are at most two maximal simplicial collections of Seifert tori in X_K. The results so far make possible the proof of parts (1) and (2) of Theorem 1 by the end of Section 7.

Section 8 is devoted to the construction of examples of genus one hyperbolic knots in $\mathbb{S}^3$ with maximal simplicial collections satisfying various conditions. The extreme examples in Section 8.2 where $|\mathbb{T}| = 4$ and $\mathrm{MS}(K)$ consists of two 3-simplices are particularly challenging to construct since each of the four complementary regions $R_{i,i+1}$ of $\mathbb{T}$ must be genus two handlebodies. These examples establish Theorem 1(3) in the case $\dim \mathrm{MS}(K) = 3$.

For the cases $\dim \mathrm{MS}(K) = 1, 2$ in Theorem 1(3) there are two possible types of examples, depending on whether the exchange region $R_{i,i+2}$ is a handlebody or not. These examples are constructed in Sections 8.3 and 8.5.

For the examples in Section 8.5 where the exchange region $R_{i,i+2}$ is not a handlebody we use the construction in [13] of genus two nontrivial handlebody knots in $\mathbb{S}^3$ whose exteriors contain essential annuli and of excellent 1-submanifolds of 3-manifolds of Myers [14]. Separating the cases $\dim \mathrm{MS}(K) = 1$ and $\dim \mathrm{MS}(K) = 2$ requires the use of another special type of handlebody pair, the general *basic pair*,

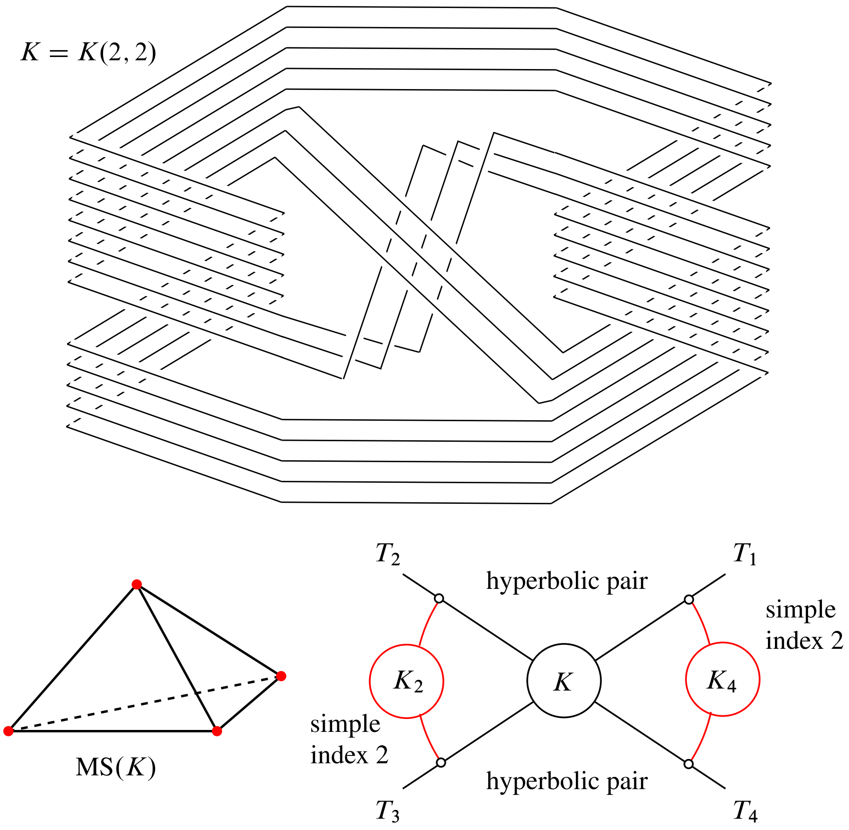


Figure 1: The genus one hyperbolic knot $K = K(2,2) \subset \mathbb{S}^3$ with $\mathrm{MS}(K)$ a single 3-simplex.

whose classification is discussed in Lemma 2.4.1. The proof of Theorem 1(3) is given at the end of Section 8.5.

Examples of knots $K \subset \mathbb{S}^3$ with a maximal simplicial collection of size 4 as above are not easy to render in a regular projection. However, making use of basic hyperbolic pairs, in Section 8.4 we construct an infinite family of genus one hyperbolic knots $K(p,q) \subset \mathbb{S}^3$ with $\mathrm{MS}(K)$ consisting of a single 3-simplex, all of which have the simple projections shown in Figure 25. The smallest member $K(2,2)$ of this family is represented in Figure 1 with a crossing number of 141 along with the structure of the pairs in their exteriors X_K (see Section 2.3 for definitions).

In Section 8.3 an infinite family of genus one hyperbolic knots $K = K(-1,n,2)$, $|n| \geq 2$, with at most $14+6|n|$ crossings is constructed such that $\mathrm{MS}(K)$ consists of two 2-dimensional simplices. The structure of their exteriors is represented in Figure 2 (where $\boxed{n}$ stands for n full twists).

Examples of genus one satellite knots $K \subset \mathbb{S}^3$ for which the dimension of $\mathrm{MS}(K)$ is arbitrarily large, showing that the restriction to hyperbolic knots is necessary, can be explicitly constructed as follows: Let $A_1, A_2 \subset \mathbb{S}^3$ be trivial unlinked and untwisted annuli connected by a band B whose core follows the pattern of a connected sum of nontrivial knots $K_1 \# \cdots \# K_n$, $n \geq 2$, with K the boundary component of

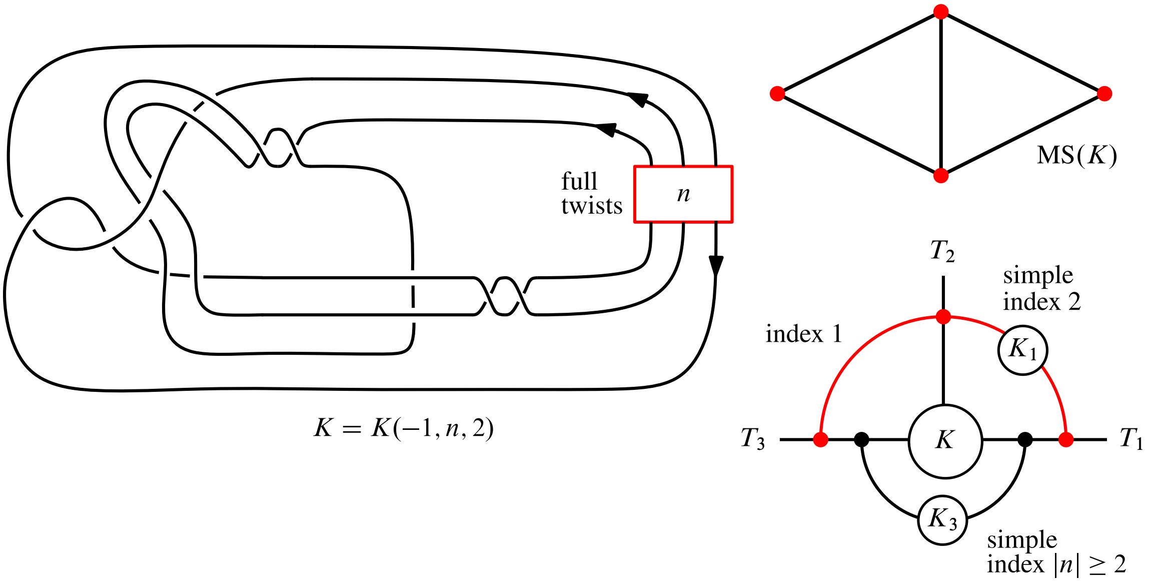


Figure 2: The genus one hyperbolic knot $K = K(-1, n, 2) \subset \mathbb{S}^3$ with $\mathrm{MS}(K)$ consisting of two 2-simplices.

the resulting pants indicated in Figure 3. The knot K is then a nontrivial zero winding number satellite of each knot K_i. Attaching an annulus to the free boundary components of A_1 and A_2 that swallows the factors $K_1 \# \cdots \# K_s$ and follows the factors $K_{s+1} \# \cdots \# K_n$ produces a Seifert torus $T_s \subset X_K$. It is not hard to see that the Seifert tori $T_1, \ldots, T_n \subset X_K$ can be constructed so as to be mutually disjoint and hence nonparallel in X_K.

On the structure of maximal simplicial collections $\mathbb{T} \subset X_K$ some questions remain unresolved. For instance, an infinite family of genus one hyperbolic knots $K \subset \mathbb{S}^3$ with $|\mathbb{T}| = 5$ was constructed in [24], all of whose pairs $(R_{i,i+1}, J)$ are of a type called *simple* (see Section 2.3). It is not known if hyperbolic knots with a simplicial collection of size $|\mathbb{T}| = 5$ can be constructed where at least one pair $(R_{i,i+1}, J)$ is not simple.

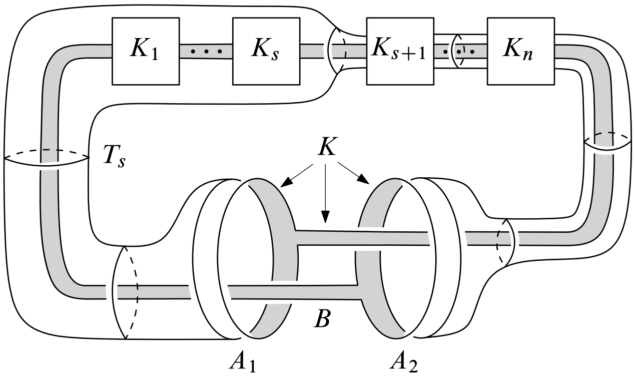


Figure 3: The satellite knot $K \subset \mathbb{S}^3$ and the swallow-follow Seifert torus $T_s \subset X_K$.

More specifically, in the case $|\mathbb{T}| = 5$ there are two options for a nonsimple pair $(R_{i,i+1}, J)$: a *primitive pair* or a *hyperbolic pair* (see Lemma 3.1.1). Realizing the case $|\mathbb{T}| = 5$ where one of the pairs $(R_{i,i+1}, J)$ is *primitive* could produce an example of a hyperbolic knot in $\mathbb{S}^3$ with a nonintegral Seifert surgery. One example realizing such a primitive pair is constructed in Proposition 8.2.1 (see Figure 21, bottom) but with a maximal simplicial collection of size $|\mathbb{T}| = 3$.

Acknowledgements

We thank M Eudave-Muñoz and J Schultens, organizers of the 2021 BIRS-CMO virtual workshop *Knots, surfaces and* 3*-manifolds*, and K Baker, organizer of the 2023 conference *Low-dimensional topology and circle-valued Morse functions* at UMiami-IMSAC, where preliminary versions of some of the results in this paper were presented. We also thank the reviewers for their careful reading of the paper and useful suggestions. This research was partially funded by a generous gift from Microsoft Co (2022–2023).

2 Preliminaries

We work in the PL category. For definitions of basic concepts see [10] or [11]. We will make use of many of the definitions and results in [24], some of which are reproduced throughout the paper.

Unless otherwise stated, submanifolds are assumed to be properly embedded. For A a submanifold of B, $\mathrm{cl}(A)$, $\mathrm{int}(A)$ and $\mathrm{fr}(A) = \mathrm{cl}(\partial A \setminus \partial B)$ denote the closure, interior and frontier of A in B, respectively.

If A is a finite set or a manifold then $|A|$ denotes the cardinality or the number of components of A, respectively.

The isotopy class of a circle in a surface is the *slope* of the circle. The circle is *nontrivial* if it does not bound a disk in the surface.

Any two circles α and β in a surface can be isotoped so as to intersect transversely and minimally, in which case $|\alpha \cap \beta|_{\min}$ denotes their minimal number of intersections.

The algebraic intersection number between 1-submanifolds α and β of a surface will be denoted by $\alpha \cdot \beta \in \mathbb{Z}$.

Let M be a 3 manifold and $F \subset M$ a surface. The components of $\partial F \subset \partial M$ are sometimes indexed as $\partial F = \partial_1 F \sqcup \partial_2 F \sqcup \cdots$.

The manifold obtained by cutting M along the surface $F \subset M$ is denoted by $M|F = \mathrm{cl}[M \setminus N(F)]$. Two surfaces F and G in M are *parallel* if they cobound a product region in M of the form $F \times I$, where $F \times \{0\}$ corresponds to F and $F \times \{1\}$ to G.

Observe that if two properly embedded surfaces F and G in M intersect transversely and $|\partial F \cap \partial G|$ is not as small as possible then it is possible to isotope F and G near ∂M to reduce $|\partial F \cap \partial G|$ without increasing $|F \cap G|$. Hence we will say that F and G intersect *minimally* if they intersect transversely so that the pair $(|\partial F \cap \partial G|, |F \cap G|)$ is smallest in the lexicographic order, in which case $|F \cap G|_{\min}$ denotes the number of components in the minimal intersection.

For a 1-submanifold $\Gamma \subset \partial M$ we let $M(\Gamma)$ denote the 3-manifold obtained by adding 2-handles to ∂M along the components of Γ and capping off any resulting 2-sphere boundary components with 3-balls. For a surface $F \subset M$ we denote by $\widehat{F} \subset M(\partial F)$ the closed surface obtained by capping off ∂F with disks.

Let $\gamma \subset \partial M$ be a circle which is nontrivial in M. An annulus $A \subset M$ is a *companion annulus* for γ if the circles ∂A cobound an annular regular neighborhood of γ in ∂M and A is not parallel to ∂M. The following result on the properties of companion annuli is established in [23, Lemma 5.1].

Lemma 2.0.1 [23] *Let M be an irreducible and atoroidal* 3*-manifold with boundary. If a circle $\gamma \subset \partial M$ has a companion annulus $A \subset M$ then*

(1) *A is unique up to isotopy*;

(2) *A cobounds with ∂M a companion solid torus V around which A runs $p \geq 2$ times.* □

We denote by $F(a, b, \dots)$ a Seifert fiber space over the surface F with singular fibers of indices $a, b, \dots \geq 1$. Typically the surface F will be a disk $\mathbb{D}^2$, an annulus $\mathbb{A}^2$ or a 2-sphere $\mathbb{S}^2$. $L_p \neq \mathbb{S}^3, \mathbb{S}^1 \times \mathbb{S}^2$ stands for a lens space with finite fundamental group of order $p \geq 2$

2.1 Genus one hyperbolic knots

With very few exceptions, for the rest of this paper we restrict our attention to hyperbolic knots in $\mathbb{S}^3$. For notation, let $K \subset \mathbb{S}^3$ be a genus one hyperbolic knot and $J \subset \partial X_K = \partial N(K)$ the longitudinal slope of K.

Recall that by a *simplicial collection of Seifert tori in X_K* we mean a collection $\mathbb{T} \subset X_K$ of mutually disjoint and nonparallel Seifert tori in X_K. The collection $\mathbb{T}$ is *maximal* if its number of components $|\mathbb{T}|$ is as large as possible. By [24] we have that $|\mathbb{T}| \leq 5$, with $|\mathbb{T}| = 5$ being the largest attainable bound.

The components of $\mathbb{T}$ are labeled consecutively following their order of appearance around ∂X_K as $T_1, T_2, \dots, T_N$, $N = |\mathbb{T}|$. For $|\mathbb{T}| > 2$ we denote by $R_{i,i+1}$ the region in X_K cobounded by T_i and T_{i+1} which contains no components of $\mathbb{T}$ other than T_i, T_{i+1}; if $|\mathbb{T}| = 2$ then a region cobounded by $T_1 \sqcup T_2$ is chosen as $R_{1,2}$, and if $\mathbb{T} = T_1$ then we define $R_{1,1}$ as the complement of a product region $\mathrm{cl}(X_K \setminus T_1 \times [-1, 1])$. Here we interpret a label i modulo $N = |\mathbb{T}|$, so $N + 1 = 1$ etc. Notice that ∂T_i and ∂T_{i+1} cobound the annulus $\partial R_{i,i+1} \cap \partial X_K$ whose core has slope J in ∂X_K.

More generally we set $R_{i,i} = \mathrm{cl}(X_K \setminus T_i \times [-1, 1])$ and $R_{i,j} = R_{i,i+1} \cup \dots \cup R_{j-1,j}$. A simplicial collection $\mathbb{T} \subset X_K$ of size $|\mathbb{T}| = 5$ is represented in Figure 4.

The next result summarizes the general properties of the regions $R_{i,j} \subset X_K$, which follow from [24, Lemmas 3.7 and 4.1].

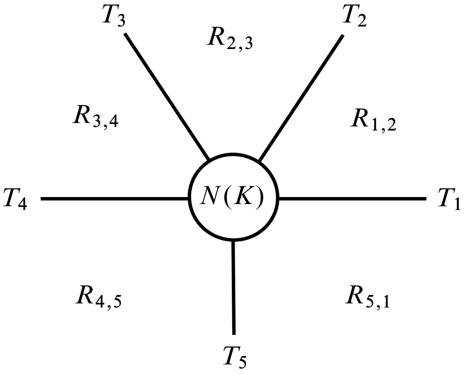


Figure 4: The knot $K \subset \mathbb{S}^3$ and a simplicial collection $\mathbb{T} = T_1 \sqcup \cdots \sqcup T_5 \subset X_K$.

Lemma 2.1.1 [24] *Let $K \subset \mathbb{S}^3$ be a hyperbolic knot with a simplicial collection of Seifert tori $\mathbb{T} \subset X_K$.*

(P1) *The manifold $R_{i,j}$ is either a genus two handlebody or an irreducible, boundary irreducible, atoroidal 3-manifold.*

(P2) *If $R_{i,j}$ is not a handlebody then $R_{j,i}$ is a handlebody.*

(P3) *If $R_{k,\ell} \subseteq R_{i,j}$ and $R_{i,j}$ is a handlebody then $R_{k,\ell}$ is a handlebody.*

(P4) *At most one region $R_{i,i+i}$ is not a handlebody, and if such a region is present then $|\mathbb{T}| \leq 4$.* □

In [24] a pair (H, J) consists of a genus two handlebody H and a separating circle $J \subset \partial H$ which is nontrivial in H; they were used to model the handlebody regions $R_{i,j}$ produced by a simplicial collection $\mathbb{T} \subset X_K$. By [24, Lemma 4.3], if $|\mathbb{T}| = 5$ then all regions $R_{i,i+1}$ are genus two handlebodies, but in the cases $|\mathbb{T}| \leq 4$ some region $R_{i,i+1}$ may not be a handlebody. In the next section we update this definition of a pair appropriately so as to be able to deal with nonhandlebody regions $R_{i,j}$.

2.2 Pairs

A *pair* (H, J) consists of an irreducible, atoroidal, connected 3-manifold H with boundary a genus two surface and $J \subset \partial H$ a separating circle which in H is nontrivial and has no companion annulus.

In the pair (H, J) the circle J separates ∂H into two once-punctured tori, T_1 and T_2, such that $\partial H = T_1 \cup_J T_2$, each of which is necessarily incompressible in H.

For convenience, a once-punctured torus in H with boundary slope J will be called a *J-torus*.

A pair (H, J) is *minimal* if any J-torus in H is parallel into T_1 or T_2.

The next result shows that handlebodies and atoroidal regions $R_{i,j} \subset X_K$ for arbitrary genus one knots $K \subset \mathbb{S}^3$ satisfy this more general definition of pair.

Lemma 2.2.1 (1) *If (H, J) is a pair then*

(a) *H is either boundary irreducible or a genus two handlebody;*

(b) *if H is a handlebody then (H, J) is a pair for any nontrivial separating circle $J \subset \partial H$.*

(2) *If $K \subset \mathbb{S}^3$ is a genus one knot and $\mathbb{T} = T_1 \sqcup \cdots \sqcup T_N \subset X_K$ is a simplicial collection of Seifert tori such that the region $R_{i,j}$ is atoroidal, as is the case when K is hyperbolic, then $(R_{i,j}, J)$ is a pair.*

Proof Part (1)(a) follows as in the proof of [24, Lemma 4.1] and (1)(b) from [24, Lemma 3.3].

For (2), if $R_{i,j}$ is atoroidal and the circle $J \subset \partial R_{i,j}$ has a companion annulus then by Lemma 2.0.1 J has a companion solid torus $V \subset R_{i,j}$ around which it runs $p \geq 2$ times, which implies that $R_{i,j}(J)$ has a lens space connected summand L_p. However, by [7, Corollary 8.3] the manifold $X_K(J)$ is irreducible and each torus $\widehat{T}_i \subset X_K(J)$ is incompressible; hence the manifold $R_{i,j}(J) \subset X_K(J)$ is irreducible. This contradiction shows that J has no companion annuli in $R_{i,j}$ and hence that $(R_{i,j}, J)$ is a pair. □

2.3 Handlebody pairs

If H is a genus two handlebody then we call (H, J) a *handlebody pair*. Handlebody pairs were introduced in [24] and play a prominent role in the structure of a genus one knot exterior. In this section we gather the main examples and properties of handlebody pairs.

As usual we write $\partial H = T_1 \cup_J T_2$. Two nonseparating circles in T_1 and T_2 are *coannular* if they cobound an annulus in H.

2.3.1 Primitive and power circles The fundamental group of handlebody H (of any genus) is a free group. We say that a circle in ∂H is *primitive* or a *power* in H if it represents a primitive or a power $p \geq 2$ of a nontrivial element in the fundamental group $\pi_1(H)$ of H, in which case the circle must be nonseparating in ∂H (see [24]). By [24, Lemma 3.3], a circle $\omega \subset \partial H$ is primitive or a power in H if and only if the surface $\partial H \setminus \omega$ compresses in H.

By [1], if $\omega \subset \partial H$ is a power circle then ω is a power of a primitive element in $\pi_1(H)$. Equivalently, the circle $\omega \subset \partial H$ is a power circle if and only if it has a companion annulus in H, in which case ω is a power of primitive element of $\pi_1(H)$ represented by the core of its companion solid torus.

In the special case that H is a genus two handlebody, if $w(x, y)$ is a cyclically reduced word representing a primitive element of the free group $\pi_1(H) = \langle x, y \mid - \rangle$ other than x, x^{-1}, y, or y^{-1} then by [2] there is an $\varepsilon \in \{\pm 1\}$ and an integer $n \in \mathbb{Z}$ such that in $w(x, y)$ all exponents of x (resp. y) are ε and all exponents of y (resp. x) are n or $n+1$. The same conclusion holds when $w(x, y)$ is a power of some primitive word $w'(x, y)$, as $w'(x, y)$ must then be cyclically reduced.

2.3.2 Circles with companion annuli in general pairs By [24, Lemma 3.1], if (H, J) is a general pair with $\partial H = T_1 \cup_J T_2$ and $i \in \{1, 2\}$ then up to isotopy there is at most one circle $\omega_i \subset T_i$ which has a companion annulus and companion solid torus in H, and these companion objects are unique in H up to isotopy.

Handlebody pairs (H, J) include the following types. Here we write $\partial H = T_1 \cup_J T_2$.

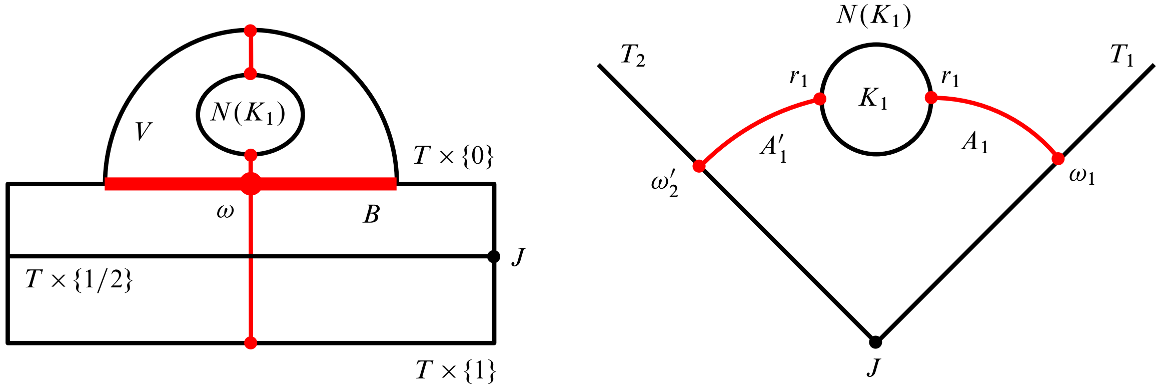


Figure 5: The core knot K_1 and the annuli A_1 and A_1' in a simple pair (H,J).

2.3.3 Trivial pair $H=T\times I$ for T a once-punctured torus and J the slope of the core of the annulus $(\partial T)\times I$. By [24, Lemma 3.7(4)] a handlebody pair (H,J) with $\partial H=T_1\cup_J T_2$ is trivial if and only if $H(J)\approx \widehat{T}_1\times I$.

2.3.4 Simple pair $H=(T\times I)\cup_B V$ where V is a solid torus, $B=(T\times\{0\})\cap\partial V$ is a closed annular neighborhood of a nonseparating circle $\omega\times\{0\}$ in $T\times\{0\}$, and B runs $p_1\geq 2$ times around V. The separating circle J corresponds to $\partial T\times\{1/2\}$. The core of the annulus $\partial V\setminus B\subset T_2$ is then coannular in H to $\omega\times\{0\}\subset T_1$.

Simple pairs are discussed in detail in [24, Sections 3.2 and 6.1]. In this case, for $\partial H=T_1\cup_J T_2$, there are coannular p-power circles $\omega_1=\omega\times\{0\}\subset T_1$ and $\omega_2\subset T_2$ which cobound a nonseparating annulus A in H and there is a nonseparating disk $D\subset H\setminus A$, all unique under isotopy.

The minimal intersection of D and J satisfies $|D\cap J|=2$. In fact, by [24, Lemma 3.11] a handlebody pair (H,J) is trivial or simple if and only if there is a disk in H which minimally intersects J in two points.

The core K_1 (defined up to isotopy) of the solid torus $H|D$ obtained by cutting H along D is called the *core of the pair* (H,J).

Thus K_1 is isotopic in $H=(T\times I)\cup_B V$ to the core of the solid torus V. It follows that, in the exterior $XH_{K_1}=H\setminus \operatorname{int} N(K_1)$ of K_1 in H, there are disjoint annuli A_1 and A_1' such that

$$\partial_1 A_1=\omega_1\subset T_1,\quad \partial_1 A_1'=\omega_1'\subset T_2,\quad \partial_2 A_1\sqcup\partial_2 A_1'\subset\partial N(K_1),$$

and each circle $\partial_2 A_1,\partial_2 A_1'$ has nonintegral boundary slope in $\partial N(K_1)$ of the form $r_1=a_1/p_1$ with $\gcd(a_1,p_1)=1$. These objects are represented in Figure 5. Figure 6, top left, shows an actual simple pair (H,J) with the disk $D\subset J$ that intersects J in two points.

The *index of the simple pair* (H,J) is defined as the integer $p_1\geq 2$; we also say that its core K_1 has index p_1.

2.3.5 Operations with simple pairs Let (H, J) be a pair with $\partial H = T_1 \cup_J T_2$ and $T \subset H$ a J-torus such that $H|T$ consists of two components, H_1 and H_2, with $H = H_1 \cup_T H_2$ and $\partial H_i = T \cup_J T_i$. Suppose that (H_1, J) is a simple pair of index $p \geq 2$ and $\omega_1 \subset T_1$ and $\omega \subset T$ are the coannular p-power circles in H_1 in Section 2.3.4. The following observations will be useful in the analysis of general pairs.

(1) *If $V_1 \subset H_1$ is the companion solid torus of $\omega_1 \subset T_1$ then $H_2 \approx \mathrm{cl}[H \setminus V_1]$.* Equivalently, if $A_1 \subset H_1$ is the companion annulus of $\omega_1 \subset T_1$ then the components of $H|A_1$ are homeomorphic to H_2 and V_1.

(2) *If $V \subset H_1$ is the companion solid torus of $\omega \subset T$ then $H \approx H_2 \cup V$.*

(3) *H is a handlebody if and only if H_2 is a handlebody and $\omega \subset T$ is a primitive circle in H_2.*

(4) *If H is a handlebody and T is not parallel into ∂H then at least one of the pairs (H_1, J) and (H_2, J) is simple; hence there is a circle in T_1 or T_2 which is a power in H.*

Items (1) and (2) follow by construction of the simple pair (H_1, J) and Figure 5. Item (3) is the content of [24, Lemma 6.3], and (4) follows from [24, Lemma 3.7(3)].

2.3.6 Primitive pair A nontrivial pair (H, J) such that there is a nonseparating annulus $A \subset H$ with each boundary component $\partial_1 A \subset T_1$ and $\partial_2 A \subset T_2$ a primitive circle in H. By [24, Lemma 6.9] the circle $\partial_i A \subset T_i$ is the unique circle in T_i which is primitive in H, and A is also unique up to isotopy.

2.3.7 Basic pair We say that a pair of circles $\omega_1 \subset T_1$ and $\omega_2 \subset T_2$ are basic in H if, relative to some base point $*$, the circles represent a basis for the free group $\pi_1(H, *)$. By [24, Section 3] this is equivalent to saying that in H the circles $\omega_1 \subset T_1$ and $\omega_2 \subset T_2$ are primitive and separated by a disk.

A pair (H, J) is *basic* if it contains a pair of basic circles as above.

2.3.8 Double pair This is a variation of a simple pair, essentially the union of two simple pairs. Let T be a once-punctured torus and $\omega_0, \omega_1 \subset T$ two circles which intersect minimally in one point. Then the circles $\omega_0 \times \{0\} \subset T \times \{0\}$ and $\omega_1 \times \{1\} \subset T \times \{1\}$ are basic in $T \times [0, 1]$. Attaching solid tori V_0 and V_1 to $T \times I$ along annular neighborhoods $B_0 \subset T \times \{0\}$ and $B_1 \subset T \times \{1\}$ of $\omega_0 \times \{0\}$ and $\omega_1 \times \{1\}$, with the circles $\omega_0 \times \{0\}$ and $\omega_1 \times \{1\}$ running $p_0, p_1 \geq 2$ times around V_0 and V_1, respectively, produces a handlebody $H = T \times I \cup_{B_0} V_0 \cup_{B_1} V_1$, with J corresponding to the circle $(\partial T) \times \{1/2\}$.

The J-torus $T \times \{1/2\} \subset H$ separates H into two handlebodies $H_0 \supset T \times \{0\}$ and $H_1 \supset T \times \{1\}$ such that (H_0, J) and (H_1, J) are simple pairs of indices p_0 and p_1, respectively, with the power circles $\omega_0 \times \{1/2\} \subset T \times \{1/2\} \subset H_0$ and $\omega_1 \times \{1/2\} \subset T \times \{1/2\} \subset H_1$ intersecting minimally in one point in $T \times \{1/2\}$.

By [24, Lemma 6.8(2)(a)], any J-torus in the double pair (H, J) is parallel to a boundary J-torus or to the J-torus T that splits it into the simple subpairs (H_0, J) and (H_1, J).

Figure 6, top right, shows a double pair (H, J) that splits into simple subpairs of index 2.

2.3.9 Maximal pair If (H, J) is a genus two handlebody pair then there are at most two J-tori in H that are mutually disjoint and nonparallel, and not parallel to ∂H; this follows from Lemma 2.0.1 and the fact that any J-torus in H is boundary compressible (see [21]). The pair (H, J) is *maximal* if it contains two such J-tori.

By [24, Lemma 6.8], a maximal pair (H, J) contains two disjoint J-tori $T_1', T_2' \subset H$ such that, in H, the J-tori T_i and T_i' cobound a simple pair (H_i, J) for $i = 1, 2$ and T_1' and T_2' cobound a nontrivial basic pair (H_0, J). Specifically, the circles $\omega_1' \subset T_1'$ and $\omega_2' \subset T_2'$ that are power circles in H_1 and H_2, respectively, are basic circles in H_0. The situation is represented in Figure 6, bottom left.

Moreover no circle in T_1 or T_2 is primitive in H, as otherwise by Section 2.3.5(2) and (3) it would be possible to construct a handlebody pair (H', J) that contains 3 J-tori that are mutually disjoint and nonparallel, and not parallel to $\partial H'$, which is impossible.

2.3.10 Induced simple pair and induced J-torus Suppose that (H, J) is a general pair and $\omega_1 \subset T_1$ a circle with companion annulus $A \subset H$. Let $V \subset H$ be the companion solid torus cobounded by A and an annulus neighborhood $B \subset T_1$ of ω_1 (Lemma 2.0.1), such that ω_1 runs $p \geq 2$ times around V.

Then a small regular neighborhood $H' = N(T_1 \cup V) \subset H$ is a genus two handlebody and (H', J) is a simple pair of index p with ω_1 a p-power circle in H'.

We say that the pair (H', J) and J-torus $T_1' = \mathrm{fr}(H') \subset H$ are the simple pair and J-torus *induced* by T_1, and more specifically by the power circle $\omega_1 \subset T_1$. Since by Section 2.3.2 the J-torus T_1 contains at most one circle with a companion annulus and solid torus, all of which are unique up to isotopy, it follows that the J-torus T_1' and simple pair (H', J) induced by T_1 are unique in H up to isotopy. The situation is represented in Figure 6, bottom right.

Minimal handlebody pairs are characterized as follows.

Lemma 2.3.11 *If (H, J) is a minimal nontrivial handlebody pair then*

$$H(J) = \begin{cases} \mathbb{A}^2(p) & \text{if } (H, J) \text{ is a simple pair of index } p \geq 2, \\ \textit{toroidal} & \text{if } (H, J) \text{ is a primitive pair}, \\ \textit{hyperbolic} & \text{otherwise}. \end{cases}$$

Proof By [24, Lemma 3.7(1)] the manifold $H(J)$ is irreducible and boundary irreducible, so if $H(J)$ is anannular and atoroidal then it is hyperbolic by [20].

Since H is a handlebody, by [5, Theorems 1 and 2] if $H(J)$ contains an incompressible annulus or torus which is not boundary parallel then H contains an incompressible annulus A disjoint from J which is not boundary parallel.

Let $\partial H = T_1 \cup_J T_2$. If $\partial A \subset T_i$ then by [24, Lemma 3.3(2)] A is a companion annulus for a power circle in T_i and hence the pair (H, J) is simple by [24, Lemma 6.2]. In this case the manifold $H(J)$ is a cable space of the form $\mathbb{A}^2(*)$.

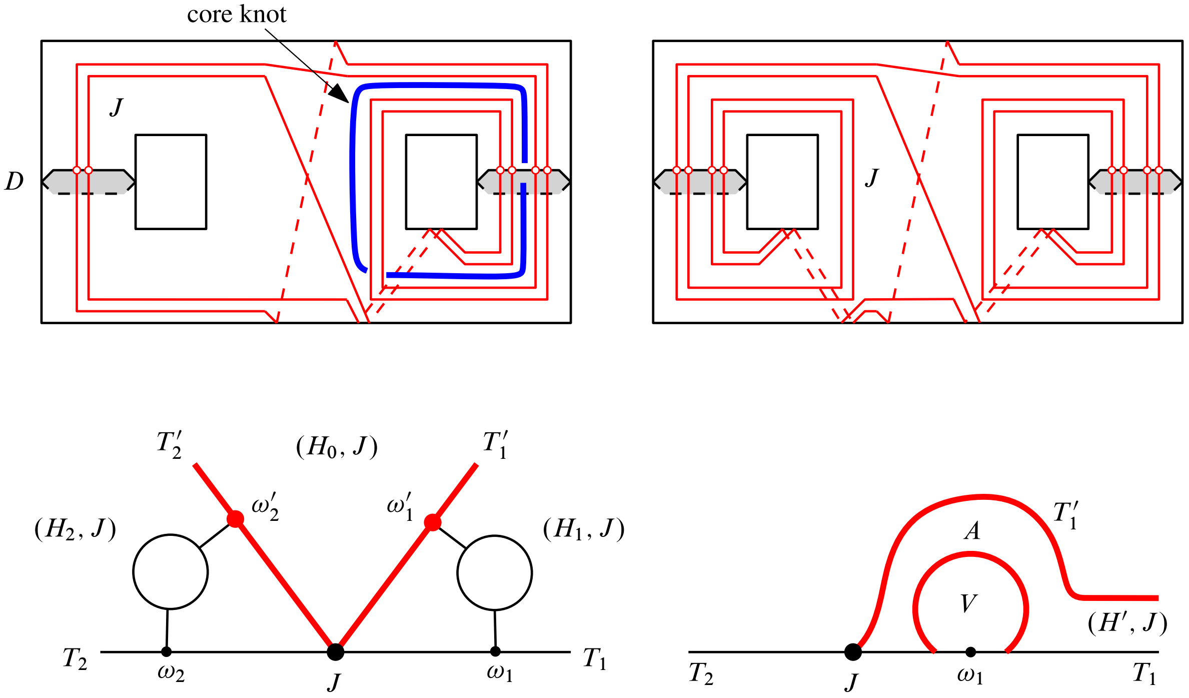


Figure 6: Examples of simple, double, maximal and induced pairs.

Suppose now that $\partial_1 A \subset T_1$ and $\partial_2 A \subset T_2$. By [24, Lemma 3.4] the components of ∂A are both power circles or both primitive circles in H. In the first case the pair (H, J) is simple by [24, Lemma 6.2].

In the second case the pair (H, J) is primitive. Let W be a regular neighborhood of $\widehat{T}_1 \cup \widehat{T}_2 \cup A$ in $H(J)$. Then W is a composing space of the form $P \times \mathbb{S}^1$ for some pants P and $\partial W = \widehat{T}_1 \sqcup \widehat{T}_2 \sqcup T$ where T is a torus. Since $H(J)$ is irreducible and boundary irreducible, if T compresses in $H(J)$ then it bounds a solid torus $V \subset H(J)$ and hence $H(J) = W \cup_T V$ is a Seifert fiber space of the form $\mathbb{A}^2(*)$. Thus $H(J)(\beta)$ is an atoroidal Seifert fiber space of the form $\mathbb{D}^2(*, *)$ for each circle $\beta \subset T_1$ with $\Delta(\beta, \partial_1 A) \geq 2$, contradicting [24, Lemma 6.9] that the manifold $H(\beta) = H(J)(\beta)$ is toroidal. Therefore $H(J)$ is a toroidal manifold. □

In light of Lemma 2.3.11, we will say that a handlebody pair (H, J) is *hyperbolic* if the manifold $H(J)$ is hyperbolic.

Examples of hyperbolic pairs are provided by some basic pairs as established in the next result.

Lemma 2.3.12 (1) *Primitive*, *simple*, *basic and hyperbolic handlebody pairs are minimal.*

(2) *A primitive pair is neither basic nor simple.*

(3) *A basic pair is trivial*, *simple or hyperbolic.*

Proof Simple pairs are minimal by [24, Lemma 3.9(2)].

If (H, J) is a handlebody pair and $T \subset H$ is a J-torus not parallel into $\partial H = T_1 \cup_J T_2$ then

$$H_1(J) \neq \widehat{T} \times I \neq H_2(J)$$

by [24, Lemma 3.7(4)] and so $\widehat{T}$ is an incompressible torus in $H(J) = H_1(J) \cup_{\widehat{T}} H_2(J)$ that is not boundary parallel. It follows that any hyperbolic handlebody pair is minimal.

We claim that if a handlebody pair (H, J) is simple or nonminimal then there is a circle $\beta \subset \partial H \setminus J$ which is a p-power circle in H for some $p \geq 2$. In such case $H(\beta)$ is a reducible manifold of the form $H(\beta) = \mathbb{S}^1 \times \mathbb{D}^2 \# L_p$ for some lens space L_p with finite fundamental group of order p.

In the case where (H, J) is a simple pair a p-power circle β exists in each component of $\partial H \setminus J$ by definition. If there is a J-torus $T \subset H$ which is not parallel to T_1 or T_2 then by [24, Lemma 3.7(3)] T separates H into two genus two handlebodies H_1 and H_2 with $\partial H_i = T \cup T_i$, where we may assume that (H_1, J) is a simple pair. Thus there is a circle $\beta \subset T_1$ which is a power circle in H_1 and hence in H.

Now, if (H, J) is a primitive pair then by Section 2.3.6 the circles $\alpha_1 \subset T_1$ and $\alpha_2 \subset T_2$ which are primitive in H are unique and coannular, hence not basic in H, so (H, J) is not a basic pair, and by [24, Lemma 6.9] the manifold $H(\beta)$ is irreducible for each circle $\beta \subset T_1$ other than the primitive circle α_1, so (H, J) is minimal and not a simple pair by Section 2.3.5(4).

Suppose now that the pair (H, J) is basic, with basic circles $\alpha_1 \subset T_1$ and $\alpha_2 \subset T_2$ that are separated by a disk $D \subset H$ (see Section 2.3.7), and $T \subset H$ is a J-torus which is not parallel to T_1 or T_2. We may assume that D intersects T minimally, so that $D \cap T$ consists of a nonempty collection of arcs. Let $E \subset D$ be a subdisk cut off by an arc in $D \cap T$ which is outermost in D, with $(\operatorname{int} E) \cap T = \varnothing$, and suppose that $E \subset H_1$. Then $|E \cap J| = |E \cap \partial T| = 2$ and so the pair (H_1, J) is simple by [24, Remark 3.8 and Lemma 3.11]. As the circle $\alpha_1 \subset T_1$ is primitive in H, it must be primitive in H_1 and hence it must intersect E minimally in one point by [24, Lemma 6.2(5)]. Thus α_1 intersects D, which is not the case. This contradiction shows that the basic pair (H, J) is minimal. Therefore (1) and (2) hold, and (3) follows now by Lemma 2.3.11 □

2.4 Construction of basic pairs

Recall from Lemma 2.3.12 that any basic pair is trivial, simple or hyperbolic. In this section we construct all basic pairs and give simple conditions to determine their nature.

In preparation for this we set up the following items:

(i) A genus two handlebody H.

(ii) Circles $\omega_1, \omega_2 \subset \partial H$ that are basic in H and separated by a disk $D \subset H$.

(iii) A decomposition $H = V_1 \cup (D \times I) \cup V_2$ where V_1 and V_2 are solid tori with meridian disks $D_1 \subset V_1$ and $D_2 \subset V_2$, such that $|D_1 \cap \omega_1|_{\min} = 1 = |D_2 \cap \omega_2|_{\min}$.

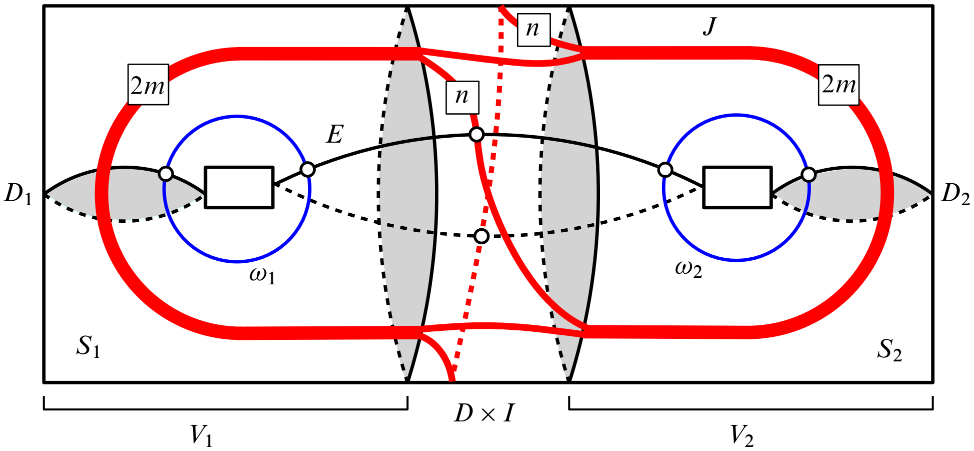


Figure 7: Construction of the basic pair (H, J), with $\boxed{k}$ representing k parallel strands.

(iv) A decomposition $\partial H = S_1 \cup A \cup S_2$ with the once-punctured tori and annulus $S_1 = \partial V_1 \cap \partial H$, $S_2 = \partial V_2 \cap \partial H$, and $A = (\partial D) \times I$.

(v) We remark that D_1 and D_2 are up to isotopy the only disks in H that satisfy the conditions

$$|D_i \cap \omega_j| = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \qquad \text{for all } \{i, j\} = \{1, 2\}.$$

If (H, J) is a basic pair with basic circles ω_1 and ω_2 then J can be isotoped in $\partial H \setminus (\omega_1 \sqcup \omega_2)$ to intersect the circles $\partial D_1 \sqcup \partial D_2 \sqcup \partial A$ minimally, in which case:

(vi) For $i = 1, 2$, $J \cap S_i$ is a collection of $2m \geq 2$ parallel arcs disjoint from ω_i.

(vii) The arcs $J \cap A$ split into 4 collections of parallel arcs each of size n or $2m - n$, where n is an integer such that $1 \leq n \leq m$ and $\gcd(n, 2m) = 1$.

The situation is represented in Figure 7, where each arc represents one of the collections of parallel arcs in (vi)-(vii) of the size indicated by the number in the box of top of the arc.

It is then always possible to construct a nonseparating disk $E \subset H$ which satisfies the following properties:

(E1) $E \cap D_1 = \varnothing = E \cap D_2$,

(E2) E intersects each circle ω_1 and ω_2 minimally in one point,

(E3) E intersects J minimally in $2n$ points.

The boundary of one such disk E is shown in Figure 7.

Lemma 2.4.1 *Any disk in H which satisfies* (E2) *is isotopic to a disk obtained by performing some number of half-Dehn twists to E along the separating disk D and hence intersects J in at least $2n$ points.*

Proof It suffices to show that any disk in H which satisfies (E2) can be isotoped to satisfy (E2) and be disjoint from $D_1 \sqcup D_2$.

So let $F \subset H$ be a disk that satisfies (E2). It is then possible to isotope F so that it satisfies (E2) and intersects $D_1 \sqcup D_2$ minimally; in particular $F \cap (D_1 \sqcup D_2)$ has no circle components.

If $F \cap (D_1 \sqcup D_2) \neq \varnothing$, say $F \cap D_1 \neq \varnothing$ for definiteness, then there is an outermost arc c of the graph $F \cap D_1 \subset D_1$ which cobounds a disk face $D_0 \subset D_1$ disjoint from ω_1. The frontier of $N(F \cup D_0)$ then consists of three disks, F_0, F_1 and F_2, properly embedded in H, say with F_0 parallel to F.

If the arc $c \subset F$ separates the points $F \cap \omega_1$ and $F \cap \omega_2$ then may assume that $|F_i \cap \omega_j| = \delta_{i,j}$ for all $\{i, j\} = \{1, 2\}$. By (v) it follows that in H the disks F_1 and F_2 are isotopic to D_1 and D_2, respectively, and hence that F can be isotoped to be disjoint from $D_1 \sqcup D_2$, contradicting our hypothesis.

If the arc $c \subset F$ does not separate the points $F \cap \omega_1$ and $F \cap \omega_2$ then one of the disks, say F_1, is disjoint from the basic circles $\omega_1 \sqcup \omega_2$ and hence must be parallel into ∂H. This implies that it is possible to reduce $|F \cap (D_1 \sqcup D_2)|$ by an isotopy that replaces a component of $F \setminus c$ with the disk D_0, again contradicting our hypothesis.

Therefore F may be assumed to be disjoint from $D_1 \sqcup D_2$ and hence isotopic to a disk obtained by performing some number of half-Dehn twists to the disk E along D. Since $1 \leq n \leq m$ holds by (vii), and hence $n \leq 2m - n$, it is then not hard to see that F must intersect J in at least $2n$ points. □

The next result classifies the basic pair (H, J) in terms of the numbers m and n in (vi) and (vii).

Lemma 2.4.2 *The basic pair (H, J) is trivial for $m = 1$, simple of index $m \geq 2$ if $n = 1$, and otherwise hyperbolic.*

Proof Let $\partial H = T_1 \cup_J T_2$. For $m = 1$ it is not hard to see that $H \approx T_1 \times I$ and hence that (H, J) is a trivial pair.

If $n = 1$ and $m \geq 2$ then it is not hard to find a circle in, say, T_1, which represents the power $(D_1 D_2)^m$ in $\pi_1(H) = \langle D_1, D_2 \mid - \rangle$. Since by Lemma 2.3.12(1) the pair (H, J) is minimal, by Section 2.3.10 it must be simple.

Conversely, if (H, J) is a simple pair then by Section 2.3.4 and [24, Lemma 6.2(5)] there is a disk $F \subset H$ that intersects J minimally in two points and satisfies (E2). Since by Lemma 2.4.1 the disk F must intersect J in at least $2n$ points it follows that $n = 1$.

Finally, if $n \geq 2$ then by the above argument the pair (H, J) is neither primitive nor simple and hence must be hyperbolic by Lemma 2.3.12(3). □

Remark 2.4.3 The simplest example of a basic hyperbolic pair (H, J) is constructed in Section 8.4 and represented in Figure 24, top. In Proposition 8.4.1 this hyperbolic pair is used to construct an example of a genus one hyperbolic knot $K \subset \mathbb{S}^3$ with a simplicial collection $\mathbb{T} \subset X_K$ which decomposes X_K into two simple pairs and two hyperbolic pairs homeomorphic to (H, J).

3 Annular pairs

Here we generalize the notions of simple and primitive handlebody pairs to arbitrary pairs.

3.1 Spanning annuli

Let (H, J) be a pair with $\partial H = T_1 \cup_J T_2$. A *spanning annulus* for (H, J) is an annulus $A \subset H$ with $\partial_1 A \subset T_1$ and $\partial_2 A \subset T_2$ nonseparating circles, in which case we say that the circles $\partial A \subset \partial H$ are *coannular* in H.

Any spanning annulus A for a pair (H, J) is nonseparating and incompressible. If H is a genus two handlebody then by [24, Lemma 3.4] the boundary circles $\partial_1 A = A \cap T_1$ and $\partial_2 A = A \cap T_2$ are both primitive or both p-power circles in H for some $p \geq 2$ and cobound at most two nonisotopic spanning annuli in H. The next result generalizes these facts to arbitrary pairs.

Lemma 3.1.1 *Let (H, J) be a pair with $\partial H = T_1 \cup_J T_2$.*

(1) *Any two spanning annuli A_1 and A_2 for the pair (H, J) which intersect transversely with*

$$\partial A_1 \cap \partial A_2 \cap T_i = \varnothing$$

for some $i = 1, 2$ can be isotoped so as to be mutually disjoint.

(2) *If (H, J) contains two spanning annuli with nonempty minimal intersection then (H, J) is a trivial pair.*

(3) *A nontrivial pair (H, J) admits at most two isotopy classes of spanning annuli. Specifically, any two nonisotopic spanning annuli A_1 and A_2 for the pair (H, J) that intersect minimally are mutually disjoint and cobound a solid torus region $V \subset H$ such that A_1 and A_2 each run $p \geq 2$ times around V.*

In particular, for any nontrivial pair (H, J) with a spanning annulus,

(4) *the boundary slopes $\omega_1 \subset T_1$ and $\omega_2 \subset T_2$ of spanning annuli in H are unique up to isotopy,*

(5) *(H, J) admits two nonisotopic spanning annuli if and only if $\omega_1 \subset T_1$ or $\omega_2 \subset T_2$ has a companion annulus in H, in which case*

   (a) *in H each slope $\omega_1 \subset T_1$ and $\omega_2 \subset T_2$ has a companion annulus and a companion solid torus around which the slope runs $p \geq 2$ times,*

   (b) *if the pair (H, J) is minimal then it is simple.*

Proof For part (1), let $A_1, A_2 \subset H$ be any two spanning annuli for the pair (H, J) which intersect transversely. We assume that $\partial_i A_j \subset T_i$ for all $i, j \in \{1, 2\}$, and that $\partial_2 A_1 \cap \partial_2 A_2 = \varnothing$ for definiteness. It follows that any arc component of $A_1 \cap A_2$ is parallel to $\partial_1 A_1$ and $\partial_1 A_2$ in A_1 and A_2, respectively, and hence, by a standard outermost arc/innermost circle argument (using the fact that H is irreducible and T_1 is incompressible in H), that A_1 and A_2 can be isotoped to intersect minimally so that each component of $A_1 \cap A_2$, if any, is a nontrivial circle in A_1 and A_2.

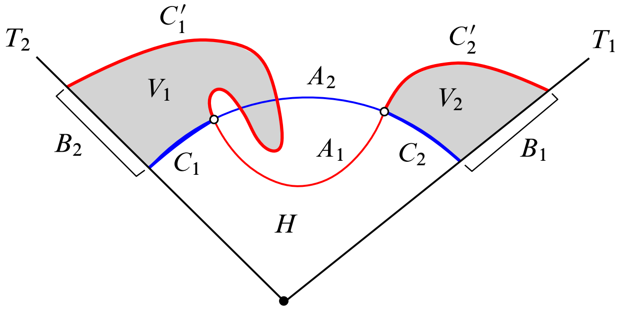


Figure 8: The spanning annuli A_1 and A_2 in H.

Now, each pair of circles $\partial_1 A_1 \sqcup \partial_1 A_2 \subset T_1$ and $\partial_2 A_1 \sqcup \partial_2 A_2 \subset T_2$ cobound annuli $B_1 \subset T_1$ and $B_2 \subset T_2$, respectively. Let $C_1, C_2 \subset A_2$ be the annular closures of the components of $A_2 \setminus A_1$ that contain the circles $\partial_1 A_2$ and $\partial_2 A_2$, respectively, and let $C'_1, C'_2 \subset A_1$ be the annuli cobounded by the pairs of circles $(C_1 \cap A_1) \sqcup \partial_1 A_1$ and $(C_2 \cap A_1) \sqcup \partial_2 A_1$, respectively. By the minimality of $A_1 \cap A_2$, for $i = 1, 2$ the annulus $C_i \cup C'_i$ is a companion annulus in H for the core circle of B_i, and hence B_i and $C_i \cup C'_i$ cobound a solid torus $V_i \subset H$, with B_i running at least twice around V_i. The situation is represented in Figure 8.

It follows that the manifold $M = N(A_1 \cup V_1 \cup V_2) \subset H$ is a Seifert fiber space over the disk with two singular fibers, contradicting Lemma 2.0.1 applied to the torus $\partial M \subset H$. Therefore we must have $A_1 \cap A_2 = \varnothing$ and so (1) holds.

For (2), let $A_1, A_2 \subset H$ be any two spanning annuli for the pair (H, J) which intersect minimally with $A_1 \cap A_2 \neq \varnothing$. By (1), for $i = 1, 2$ the circles $\partial_i A_1, \partial_i A_2 \subset T_i$ intersect minimally and, as T_i is a once-punctured torus, coherently in T_i, and so $A_1 \cap A_2$ consists of a nonempty disjoint collection of mutually parallel spanning arcs in A_1 and A_2. If $A_1 \cap A_2$ has at least two arc components then, from the closure of a rectangular component of $A_2 \setminus A_1$, it is possible to construct a spanning annulus A'_2 for (H, J) which intersects A_1 minimally in one arc. We may therefore assume that $A_1 \cap A_2$ is a single spanning arc, in which case $S_1 = N(\partial_1 A_1 \cup \partial_1 A_2) \subset T_1$ is a once-punctured torus with ∂S_1 parallel in T_1 to $J = \partial T_1$, while $N(A_1 \cup A_2) \subset H$ is homeomorphic to the genus two handlebody $S_1 \times I$, with $A_1 \cup A_2$ corresponding to $(\partial_1 A_1 \cup \partial_1 A_2) \times \{1/2\}$.

It follows that the frontier A of $N(A_1 \cup A_2) \subset H$ is an annulus with boundary circles ∂A parallel to J in H. Since the circle J has no companion annuli in H, the annulus A must be parallel to ∂H in H, which implies that H is homeomorphic to $S_1 \times I$ and hence that the pair (H, J) is trivial, so (2) holds.

Suppose now that (H, J) is a nontrivial pair. By (1) any two spanning annuli A_1 and A_2 for (H, J) can be isotoped so as to be disjoint, whence the circles $\partial_1 A_1, \partial_1 A_2 \subset T_1$ and $\partial_2 A_1, \partial_2 A_2 \subset T_2$ cobound annuli $B_1 \subset T_1, B_2 \subset T_2$, respectively. By Lemma 2.0.1, the torus $A_1 \cup A_2 \cup B_1 \cup B_2$ bounds a solid torus $V(A_1, A_2) \subset H$, with each annulus $A_1, A_2 \subset \partial V(A_1, A_2)$ running $n(A_1, A_2) \geq 1$ times around $V(A_1, A_2)$.

If $n(A_1, A_2) = 1$ for all spanning annuli A_1, A_2 of (H, J) then any two such spanning annuli are mutually isotopic, so (H, J) contains a unique spanning annulus.

Otherwise, suppose that $p = n(A_1, A_2) \geq 2$ for some mutually disjoint spanning annuli $A_1, A_2 \subset H$, and let $A \subset H$ be any spanning annulus for (H, J). Isotope A so as to be disjoint from A_1, whence A and A_1 have the same boundary slope, and then isotope A so as to intersect A_2 minimally subject to $A \cap A_1 = \varnothing$. An argument similar to the one used in the proof of part (1) then shows that we must have $A \cap A_2 = \varnothing$ too, whence A can be isotoped so as to be disjoint from $A_1 \sqcup A_2$. It follows that either $A \subset V(A_1, A_2)$ or $A_i \subset V(A, A_j)$ for some $\{i, j\} = \{1, 2\}$, which implies that A is parallel to A_1 or A_2.

Therefore A_1 and A_2 are up to isotopy the only spanning annuli in H, hence their boundary slopes are the only slopes in T_1 and T_2 that cobound a spanning annulus in H; and as $p \geq 2$ the solid torus $V(A_1, A_2)$ is a companion solid torus for each slope $\omega_1 = \partial_1 A_1 \subset T_1$ and $\omega_2 = \partial_2 A_1 \subset T_2$.

Conversely, if A is a spanning annulus for (H, J) and $V \subset H$ is a companion solid torus for, say, the circle $\omega_1 = A \cap T_1 \subset T_1$, so that ω_1 runs $p \geq 2$ times around V, then A can be isotoped so that $A \cap \operatorname{int} V = \varnothing$ and $A \cap V = A \cap T_1$, in which case $N(A \cup V) \subset H$ is a solid torus whose frontier consists of two disjoint spanning annuli for (H, J), each of which runs p times around $N(A \cup V)$. Therefore the first part of (5) holds.

Finally, let V_1 be the solid torus obtained by pushing $V(A_1, A_2)$ slightly off T_2. Then V_1 is a companion solid torus for $\omega_1 = \partial_1 A_1 \subset T_1$ which by Section 2.3.10 induces a J-torus $T \subset H$ that splits H into genus two handlebodies $H_1 = N(T_1 \cup V_1)$ and $H_2 \subset H$, with $\partial H_1 = T_1 \cup_J T$ and $\partial H_2 = T_2 \cup_J T$, such that (H_1, J) is a simple pair of index $p \geq 2$. Thus if the pair (H, J) is minimal then T is parallel to T_2 and so the pair (H, J) is simple. A similar conclusion holds if we push the solid torus $V(A_1, A_2)$ slightly off T_1 to obtain a companion solid torus for $\omega_2 = \partial_2 A_1 \subset T_2$. Therefore (4) and (5) hold. □

3.2 The index of an annular pair

We make the following definitions and observations based on the properties obtained in Lemma 3.1.1.

(A1) A pair (H, J) is said to be *annular* if it is nontrivial and contains a spanning annulus.

(A2) If (H, J) is an annular pair and $A \subset H$ is a spanning annulus then the *index* of (H, J) and A is the number $p = n(A_1, A_2) \geq 2$ given in Lemma 3.1.1(3) if A is not unique, and otherwise it is 1.

(A3) By [24, Lemma 3.4(4)(a)], a handlebody pair (H, J) is annular of index 1 if and only if it is a primitive pair.

(A4) For an annular pair (H, J) with a spanning annulus A of index $p \geq 2$, the solid torus

$$V = V(A_1, A_2) \subset H$$

in Lemma 3.1.1(3) is unique up to isotopy and its core is called the *core knot of* (H, J).

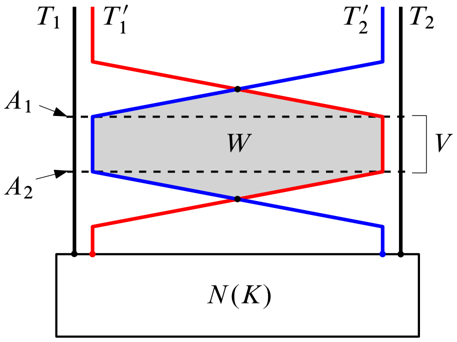


Figure 9: The J-tori $T_1', T_2' \subset R_{1,2}$ induced along T_1 and T_2 in the annular pair $(R_{1,2}, J)$.

Also, following the notation in the proof of Lemma 3.1.1(3), for each $\{i, j\} = \{1, 2\}$ the manifold $W_i = N(T_i \cup V) \subset H$ is a handlebody which, after being pushed slightly off from T_j, produces a simple pair (W_i, J) of index $p \geq 2$ cobounded by its two frontier J-tori T_i and $T_i' \subset H$. As in Section 2.3.10 the pair (W_i, J) and T_i' are the simple pair and J-torus *induced by the annular pair* (H, J) *along* T_i, unique up to isotopy in H. The situation is represented in Figure 9.

4 Seifert tori in X_K

In this section we establish several properties of Seifert tori in the exterior of a hyperbolic knot $K \subset \mathbb{S}^3$. In particular we determine the structure of the pairs generated by a simplicial collection of Seifert tori in X_K in the presence of a Seifert torus not isotopic to any of those in the collection.

4.1 General properties

We use the following notation. Let $T \subset X_K$ be a J-torus and $T \times [-1, 1]$ a product neighborhood of T in X_K with T corresponding to $T \times \{0\}$. For a surface $F \subset X_K$, not necessarily properly embedded, such that $T \cap \operatorname{int} F = \varnothing$ and $T \cap \partial F \neq \varnothing$, we say that F *locally lies on one side of* T if $F \cap (T \times [-1, 0]) = \varnothing$ or $F \cap (T \times [0, 1]) = \varnothing$, and otherwise that F *locally lies on both sides of* T. For instance, a companion annulus for a slope in T locally lies on one side of T, while ∂X_K locally lies on both sides of T.

Lemma 4.1.1 *Let $T_1, T_2, T_3 \subset X_K$ be J-tori in the exterior of a hyperbolic knot $K \subset \mathbb{S}^3$.*

(1) *If $F \subset X_K$ is a properly embedded surface which intersects T_1 transversely with $(\partial T_1) \cap (\partial F) = \varnothing$ then the number of circle components of $T_1 \cap F$ that are nonseparating in T_1 is even.*

(2) *If A is an annulus in X_K with $A \cap T_1 = (\partial A) \cap T_1$ such that each of the circles ∂A is nonseparating in T_1 and $\Delta(\partial_1 A, \partial_2 A) = 0$ then A is a companion annulus that locally lies on one side of T_1.*

(3) *Any two companion annuli for a circle in T_1 locally lie on the same side of T_1 and are mutually isotopic.*

(4) *If T_1 and T_2 are mutually disjoint and A is a spanning annulus for the pair $(R_{1,2}, J)$ then the circles ∂A are not coannular in $R_{2,1}$ and not both have companion annuli in $R_{2,1}$. Moreover, if a component of ∂A has a companion annulus in $R_{2,1}$ then $A \subset R_{1,2}$ has index 1.*

(5) *If $B \subset R_{1,2}$ is an annulus with $\partial_1 B$ a nonseparating circle in T_1 and $\partial_2 B$ a circle in T_i for some $i = 1, 2$ then $\partial_2 B$ is also a nonseparating circle in T_i.*

(6) *Suppose that $T_1, T_2, T_3 \subset X_K$ are mutually disjoint and nonparallel Seifert tori with $T_2 \subset R_{1,3}$. If $A \subset R_{1,3}$ is a spanning annulus which intersects T_2 minimally then $A_1 = A \cap R_{1,2}$ and $A_2 = A \cap R_{2,3}$ are spanning annuli in $R_{1,2}$ and $R_{2,3}$, respectively, and $(R_{1,3}, J)$ has index $p \geq 1$ if and only if one of the pairs $(R_{1,2}, J)$ or $(R_{2,3}, J)$ has index 1 and the other index $p \geq 1$.*

(7) *If the J-tori T_1, T_2 and T_3 are mutually disjoint and nonparallel in X_K, T_2 lies in the region $R_{1,3}$, and the pair $(R_{1,2}, J)$ is simple with $\omega \subset T_2$ a $p \geq 2$ power circle in $R_{1,2}$, then $R_{1,3}$ is a handlebody if and only if $R_{2,3}$ is a handlebody and ω is a primitive circle in $R_{2,3}$.*

In particular, if the pair $(R_{2,3}, J)$ is primitive then $R_{1,3}$ is a handlebody if and only if the slopes of the spanning annuli in $R_{1,2}$ and $R_{2,3}$ agree on T_2.

Proof For part (1), observe that if $\partial F \neq \varnothing$ then, after suitably capping off with disks any boundary components of ∂F that are trivial in X_K, we may assume that each component of ∂F is a nontrivial circle in X_K of slope J.

$T_1 \cap F$ has no arc components since $(\partial T_1) \cap (\partial F) = \varnothing$ and so each component of $T_1 \cap F$ is a circle which either is parallel to ∂T_1, bounds a disk in T_1, or does not separate T_1. Let $\mathcal{N} \subseteq T_1 \cap F$ be the collection of circles that are nonseparating in T_1 and assume that $\mathcal{N} \neq \varnothing$. As the circles in $\mathcal{N}$ are mutually parallel in T, there is a circle $\beta \subset T_1$ which intersects each component of $\mathcal{N}$ transversely in one point and is disjoint from $T_1 \cap F \setminus \mathcal{N}$. After pushing β slightly away from T_1, we may assume that β is disjoint from T_1 and intersects F transversely in $|\mathcal{N}|$ points. If F separates X_K then $|\mathcal{N}|$ is even, so we may further assume that F does not separate X_K. Since $H_2(X_K(J); \mathbb{Z}_2) = \mathbb{Z}_2$, the nonseparating closed surfaces $\widehat{T}_1, \widehat{F} \subset X_K(J)$ belong to the only nontrivial homology class of $H_2(X_K(J); \mathbb{Z}_2)$; hence $\widehat{T}_1$ and $\widehat{F}$ must have the same homological intersection number mod 2 with β, ie

$$0 \equiv \beta \cdot \widehat{T}_1 \equiv \beta \cdot \widehat{F} \equiv |\mathcal{N}| \mod 2$$

and so $|\mathcal{N}|$ is even.

For part (2) suppose that A does not locally lie on one side of T_1. Since $\Delta(\partial_1 A, \partial_2 A) = 0$, the circles ∂A have the same slope in T_1 and hence A can be isotoped in X_K so that $A \cap T_1 = (\partial A) \cap T_1$ and $\partial_1 A = \partial_2 A$, in which case the resulting closed surface A contradicts the conclusion of part (1). Therefore A locally lies on one side of T_1.

Part (3) is the content of [24, Lemmas 3.1 and 5.1].

For part (4), if B is an annulus in $R_{2,1}$ with $\partial B = \partial A$ then $A \cup B \subset X_K$ is a closed surface in X_K that intersects T_1 minimally in one circle, contradicting (1); thus the circles ∂A are not coannular in $R_{2,1}$.

Suppose now that $B_1, B_2 \subset R_{2,1}$ are companion annuli for the circles $\partial_1 A$ and $\partial_2 A$, respectively, isotoped so as to intersect minimally, and let $V_1, V_2 \subset R_{2,1}$ be the companion solid tori cobounded by B_1, T_1 and B_2, T_2, respectively. Let A_1 and A_2 be mutually disjoint spanning annuli for the pair $(R_{1,2}, J)$ that are parallel to A with $\partial A_1 \sqcup \partial A_2 = \partial B_1 \sqcup \partial B_2$.

If $B_1 \cap B_2 \neq \varnothing$ then each component of $B_1 \cap B_2$ is a core circle of B_1 and B_2, so it is possible to construct a spanning annulus B for the pair $(R_{2,1}, J)$ with $\partial B = \partial A$, contradicting the argument above. And if $B_1 \cap B_2 = \varnothing$ then B_1 and $A_1 \cup B_2 \cup A_2$ are companion annuli for the circle $T_1 \cap \partial A$ that lie on opposite sides of T_1, contradicting (3).

Finally, if A has index $p \geq 2$ then each circle $\partial_i A \subset T_i$ has a companion annulus in $R_{1,2}$ by Lemma 3.1.1(5), and hence by (3) cannot have a companion annulus in $R_{2,1}$. Therefore (4) holds.

For part (5) if in $\widehat{T}_i \subset X_K(J)$ the circle $\partial_2 B \subset T_i$ bounds a disk then the disk $\widehat{B}_1$ compresses the nonseparating torus $\widehat{T}_1$ in $X_K(J)$ into a nonseparating 2-sphere, contradicting [7, Corollary 8.3] that the manifold $X_K(J)$ is irreducible. Therefore $\partial_2 B$ is nonseparating in $\widehat{T}_i$ and hence in T_i.

For part (6), each component of $A \cap T_2$ is a nontrivial circle in A and in T_2 and so each component of $A \cap R_{1,2}$ and $A \cap R_{2,3}$ is an annulus.

Let A_1 be the component of $A \cap R_{1,2}$ with $A \cap T_1 \subseteq A_1 \cap T_1$. Then necessarily $A_1 \cap T_2 \neq \varnothing$ and so by (5) the circle $\alpha_1 = A_1 \cap T_2$ is nonseparating in T_2, hence A_1 is a spanning annulus in $R_{1,2}$. Similarly the component A_2 of $A \cap R_{2,3}$ with $A \cap T_3 \subseteq A_2 \cap T_3$ is a spanning annulus in $R_{2,3}$ with $\alpha_2 = A_2 \cap T_2$ a nonseparating circle in T_2. In particular either $\alpha_1 = \alpha_2$ or α_1 and α_2 are disjoint and mutually parallel circles in T_2.

If $\alpha_1 \neq \alpha_2$ then by (5) the component of $A \cap R_{2,3}$ which contains α_1 is a companion annulus for α_1 in $R_{2,3}$, and similarly α_1 has a companion annulus in $R_{1,2}$, contradicting (3). Therefore $\alpha_1 = \alpha_2$ and so $A = A_1 \cup A_2$.

Suppose now that $B \subset R_{1,3}$ is a spanning annulus disjoint from A. By the argument above we may assume that $B_1 = B \cap R_{1,2} \subset R_{1,2}$ and $B_2 = B \cap R_{2,3} \subset R_{2,3}$ are spanning annuli. Let V be the region in $R_{1,3}$ cobounded by A and B, and let $C \subset V$ be the annulus cobounded by the circles $(A \sqcup B) \cap T_2$. By Lemma 3.1.1(3) the region V is a solid torus and so C separates V into two solid tori $V_1 = V \cap R_{1,2}$ and $V_2 = V \cap R_{2,3}$. Necessarily C runs once around one of the solid tori V_1 or V_2.

If the pair $(R_{1,3}, J)$ has index $p \geq 2$ and A runs p times around V then necessarily A_1, say, runs p times around V_1. Therefore the pair $(R_{1,2}, J)$ has index $p \geq 2$, while by Lemma 3.1.1(5) the core of $C \subset T_2$ has a companion annulus in $R_{1,2}$ and so by (3) the core of C cannot have a companion annulus in $R_{2,3}$, which implies that the pair $(R_{2,3}, J)$ has index 1.

Conversely, if $(R_{1,2}, J)$, say, has index $p \geq 2$ then there is a spanning annulus and $A_1' \subset R_{1,2}$ disjoint from A_1 that cobounds with A_1 a solid torus $V' \subset R_{1,2}$ around which each annulus runs p times. Thus

$W = N(A_1 \cup V')$ is a solid torus in $R_{1,3}$ and its frontier consists of two spanning annuli in $R_{1,3}$ each of which runs p times around W, so the pair $(R_{1,3}, J)$ has index p. Therefore (6) holds.

The first part of (7) follows from Section 2.3.5(2) and (3). If the pair $(R_{2,3}, J)$ is primitive then by [24, Lemma 6.9(2)] the slope $\omega' \subset T_2$ of the spanning annulus in $R_{2,3}$ is the unique circle which is primitive in $R_{2,3}$, while $\omega \subset T_2$ is the slope of the spanning annulus in $R_{1,2}$. Therefore $R_{1,3}$ is a handlebody if and only if ω and ω' have the same slope in T_2. □

4.2 Intersections of Seifert tori

By [17, Lemma 5.2] the minimal intersection between two nonisotopic Seifert tori in $X_K \subset \mathbb{S}^3$ (which is assumed to be only atoroidal) consists of two circles which are nonseparating in each of the surfaces. Here we extend this result to give a more detailed picture of the minimal intersection between a Seifert torus S and a simplicial collection of Seifert tori $\mathbb{T} \subset X_K$. In particular we will see that a nontrivial such intersection produces an annular pair of index 1 within a complementary region of $\mathbb{T}$.

The next result is the first approximation to the main classification given in Proposition 7.0.2.

Lemma 4.2.1 *Let $\mathbb{T} = T_1 \sqcup \cdots \sqcup T_N$ be a simplicial collection of Seifert tori in X_K and let $S \subset X_K$ be a Seifert torus which is not isotopic in X_K to any component of $\mathbb{T}$, such that either*

(i) *S intersects the collection $\mathbb{T}$ minimally, or*

(ii) *$N \geq 2$, $S \subset R_{i,j}$ ($i = j$ allowed), and S intersects the collection $\mathbb{T} \cap R_{i,j}$ minimally.*

Then:

(1) *For each j, $S \cap T_j$ is either empty or consists of two circle components that are nonseparating in S and T_j; in particular, S and T_j intersect minimally in X_K.*

(2) *The closure of each component of $S \setminus \mathbb{T}$ is either a pants P or an annulus.*

(3) *$P \cap \mathbb{T} = P \cap T_i$ for some $1 \leq i \leq N$.*

(4) *There is a Seifert torus $T \subset X_K \setminus (P \cup \mathbb{T})$ such that if $R, R' \subset X_K$ are the regions cobounded by $T \sqcup T_i$ and $P \subset R$ then*

   (a) *$R \cap \mathbb{T} = T_i$,*

   (b) *the pair (R, J) is annular of index 1 with spanning annulus $A_R \subset R$,*

   (c) *the pair (R', J) is nontrivial and $A = S \cap R'$ is a companion annulus; moreover, P and A lie on opposite sides of T_i and the annuli A and A_R have the same boundary slope in T_i.*

Proof (I) Since S and $\mathbb{T}$ have the same boundary slopes, by conditions (i) and (ii) we have that $\partial S \cap \partial \mathbb{T} = \varnothing$ and so $S \cap \mathbb{T}$ is a nonempty collection of circles that are nontrivial in S and $\mathbb{T}$.

(II) *For each j, $S \cap T_j$ consists of a collection circles that are nonseparating in S and T_j, hence mutually disjoint and parallel in S and T_j. Thus* (1) *holds.*

For let c be a circle component of $S \cap \mathbb{T}$; by (I) c is nontrivial in S and $\mathbb{T}$. Consider the case where c is parallel to ∂S in S; the case where c is parallel to $\partial \mathbb{T}$ can be dealt with in a similar way. We may assume that c is outermost in S, that is, c cobounds with ∂S an annulus $A_c \subset S$ with interior disjoint from $\mathbb{T}$.

Let $T_j \subset \mathbb{T}$ be the component containing c. By Lemma 4.1.1(5) the circle c separates T_j and so it cobounds an annulus $A'_c \subset T_j$ with ∂T_j. The annulus $A_c \cup_c A'_c$ is then properly embedded in X_K with J as boundary slope, and as the knot K is hyperbolic this annulus must be parallel in X_K into an annulus $B \subset \partial X_K$. Thus the annuli $A_c \cup_c A'_c$ and B cobound a solid torus $V \subset X_K$ around which each annulus runs once and such that $V \cap \mathbb{T} = A'_c$. It is then possible to reduce $|S \cap \mathbb{T}|$ by an isotopy of $\mathbb{T}$ that exchanges the annulus A'_c with A_c within the solid torus V and pushes the resulting surface slightly off S, contradicting the minimality of $|S \cap \mathbb{T}|$.

(III) *$S \cap T_j$ has an even number of components* by Lemma 4.1.1(1).

(IV) *If $S \cap T_j \neq \varnothing$ then $|S \cap T_j| = 2$ and the closures of the components of $S \setminus T_j$ are a pants P_j and a companion annulus A_j that locally lie on opposite sides of T_j, with $P_j \cap T_j = A_j \cap T_j = \partial A_j$.*

By (II) and (III) the closures of the components of $S \setminus T_j$ consist of a pants component P_j and an odd number of annuli. By Lemma 4.1.1(3), each such annulus component is a companion annulus for the slope of the circles $S \cap T_j \subset S$, and all such annular components lie on the same side of T_j. Therefore there can be only one such annular component A_j, so $|S \cap T_j| = 2$ and the rest of the properties of P_j and A_j follow.

(V) Similarly, by (II) and (III) $S \cap \mathbb{T}$ consists of an even number of circle components which are nonseparating in S and so the closures of the components of $S \setminus \mathbb{T}$ consist of a pants component P and an odd number of annuli. If $\partial P = \partial S \sqcup \alpha_1 \sqcup \alpha_2$ then by (IV) $P \cap T_i = S \cap T_i = \alpha_1 \sqcup \alpha_2$ for some $T_i \subset \mathbb{T}$ and the annulus $A = \mathrm{cl}[S \setminus P]$ is a companion annulus of the slope of the circles $S \cap T_i$, with P and A lying on opposite sides of T_i and $P \cap \mathbb{T} = P \cap T_j$. Thus (2) and (3) hold.

(VI) Let P, A and T_i be as in (V) so that $S = P \cup A$. Since P and $\mathbb{T} \setminus T_i$ are disjoint, there is regular neighborhood $N(P \cup T_i) \subset X_K$ which is disjoint from $\mathbb{T} \setminus T_i$. The frontier of $N(P \cup T_i) \subset X_K$ contains two J-tori, T_P and T'_P, with T_P on the same side of T_i as P and T'_P on the opposite side and parallel to T_i.

Let $R, R' \subset X_K$ be the two regions cobounded by T_P and T_i, with $P = S \cap R$ and $A = S \cap R'$. If T_P and T_i are parallel in R or R' then by [25, Corollary 3.2] P or A is parallel in R or R' into T_i, respectively, and so S can be isotoped by pushing P or A across and to the other side of T_i, thus reducing $|S \cap \mathbb{T}|$, which is not possible. Therefore T_P and T_i are not parallel in X_K and so the pairs (R, J) and (R', J) are nontrivial.

Let $B \subset T_i$ be the annulus cobounded by the circles $\alpha_1 \sqcup \alpha_2 = P \cap T_i$. Then the J-torus $P \cup B \subset R$ is parallel in R to T_P, that is, the region in R between T_P and $P \cup B$ is a product of the form $(P \cup B) \times [0,1]$, with $T_P = (P \cup B) \times \{0\}$ and $P \cup B = (P \cup B) \times \{1\}$. It follows that $A_R = \alpha_1 \times [0,1]$ is a spanning annulus for the region R. Since the annulus $A = S \cap R'$ is a companion for the slope $\alpha_1 \subset T_i$ outside R, A_R has index 1 by Lemma 4.1.1(4) and so the pair (R, J) is annular of index 1. Therefore (4) holds. □

5 Minimality of index 1 annular pairs in X_K

Suitably gluing together two annular pairs of index 1 results in a new annular pair of index 1 which is not minimal. In this section we show that an annular pair of index 1 produced by two Seifert tori in the exterior of a hyperbolic knot in $\mathbb{S}^3$ must be minimal.

5.1 Annular pairs in X_K

Let $K \subset \mathbb{S}^3$ be a genus one hyperbolic knot and let (R, J) and (R', J) be pairs cobounded by two mutually disjoint and nonparallel Seifert tori in X_K, so that $X_K = R \cup R'$.

If (R', J) is an annular pair and the region $R' \subset X_K$ is boundary irreducible then by Lemma 2.1.1(P2) the region $R \subset \mathbb{S}^3$ is a genus two handlebody, an example of a *nontrivial handlebody knot in* $\mathbb{S}^3$. In [13, Lemma 3.8] Y Koda and M Ozawa classify R as a certain type of handlebody knot using a result of C Gordon [13, Lemma 3.6], along with that any 4-punctured sphere with integral boundary slope in a knot exterior in $\mathbb{S}^3$ is compressible. We remark that the compressibility of many-punctured spheres with nonintegral and nonmeridional boundary slope follows from the results in [3, Sections 2.5 and 2.6], in particular Proposition 2.5.6.

We use a similar strategy to impose restrictions on the pairs (R, J) or (R', J) in X_K whenever one of them is an annular pair. A classification of handlebody annular pairs is obtained which will be extended and refined in Proposition 5.2.3 to arbitrary annular pairs in a knot exterior X_K. We will see in Section 7 that the properties of this type of pair are the key to bound the number of maximal simplicial collections of Seifert tori in X_K.

Lemma 5.1.1 *Let $K \subset \mathbb{S}^3$ be a genus one hyperbolic knot and $T_1 \sqcup T_2 \subset X_K$ a simplicial collection of Seifert tori such that the pair $(R_{1,2}, J)$ is annular with spanning annulus $A \subset R_{1,2}$. Then one of the following holds*:

(1) *The region $R_{1,2}$ is a genus two handlebody and the pair $(R_{1,2}, J)$ is either primitive*, *simple*, *or splits along some J-torus in $R_{1,2}$ into a simple and a primitive pair.*

(2) *The region $R_{2,1}$ is a genus two handlebody*, *the circles $\partial_1 A \subset T_1$ and $\partial_2 A \subset T_2$ are separated in $R_{2,1}$ and one of the following holds*:

   (a) *The pair $(R_{2,1}, J)$ is basic*, *with the components of ∂A as basic circles in $R_{2,1}$.*

   (b) *One of the two components of ∂A*, *say $\partial_2 A \subset T_2$*, *is a power circle in $R_{2,1}$ which induces a J-torus $T_3 \subset R_{2,1}$*, *such that the pair $(R_{2,3}, J)$ is simple with spanning annulus B of index $p \geq 2$ and the pair $(R_{3,1}, J)$ is basic with $A \cap T_1$ and $B \cap T$ basic circles in $R_{3,1}$*; *in particular*, *any J-torus in $R_{2,1}$ is isotopic in $R_{2,1}$ to T_1*, *T_2 or T_3.*

Proof Let the knot $L \subset \mathbb{S}^3$ be the core of A. Then $N(A) \subset R_{1,2}$ is a solid torus neighborhood of L and so X_L can be identified with $\mathbb{S}^3 \setminus \operatorname{int} N(A)$. Extending T_1 and T_2 radially in $N(K)$ so that $\partial T_1 = K = \partial T_2$ yields the genus two surface $F = T_1 \cup T_2$ such that $P = \operatorname{cl}[F \setminus N(A)] \subset F$ is a 4-punctured 2-sphere in X_L. Notice that P has integral boundary slope in ∂X_L and separates X_L into two components, $W_1 = \operatorname{cl}[R_{1,2} \setminus N(A)] \subset R_{1,2}$ and $W_2 = R_{2,1} \cup N(A) \supset R_{2,1}$.

By [13, Lemma 3.6] the 4-punctured sphere P compresses in X_L along some disk E. We consider two cases.

Case 1 $E \subset W_1$.

Then $E \subset R_{1,2}$, so $\partial R_{1,2}$ compresses in $R_{1,2}$ and so the region $R_{1,2}$ is a genus two handlebody by Lemma 2.1.1(P1); hence the pair $(R_{1,2}, J)$ is primitive if A has index 1.

Suppose now that A has index $p \geq 2$, so that $\partial_1 A = A \cap T_1$ is a p-power circle in $R_{1,2}$. If the pair $(R_{1,2}, J)$ is minimal then it is simple by Lemma 3.1.1(5). If the pair $(R_{1,2}, J)$ is not minimal and $T_a \subset R_{1,2}$ is some J-torus not parallel to T_1 or T_2 then by [24, Lemma 3.7(2)(3)] each region $R_{1,a}, R_{a,2} \subset R_{1,2}$ is a handlebody and we may assume that $(R_{1,a}, J)$ is simple, hence annular of index p, and hence by Lemma 4.1.1(6) that the pair $(R_{a,2}, J)$ is annular of index 1, primitive. Therefore (1) holds.

Case 2 $E \subset W_2$.

The disk E is disjoint from $N(A)$ and so it is properly embedded in $R_{2,1}$; therefore $R_{2,1}$ is a genus two handlebody by Lemma 2.1.1(P1). We also have that the circles $\partial A, \partial E$ are mutually disjoint and, as T_1 and T_2 are incompressible in $R_{2,1}$, ∂E is not parallel to any component of ∂A.

If the disk $E \subset R_{2,1}$ is nonseparating then by [24, Lemma 3.4] the circles ∂A are coannular in $R_{2,1}$, contradicting Lemma 4.1.1(4). Therefore E must be a separating disk in $R_{2,1}$ and so by [24, Lemma 3.4] each circle $\partial_1 A$ and $\partial_2 A$ is a primitive or power circle in $R_{2,1}$, and by Lemma 4.1.1(4) not both can be power circles.

In $R_{2,1}$, if the separated circles $\omega_1 = \partial_1 A \subset T_1$ and $\omega_2 = \partial_2 A \subset T_2$ are both primitive then by Section 2.3.7 they are basic circles in $R_{2,1}$ and so the pair $(R_{2,1}, J)$ is basic.

Suppose now for definiteness that, in $R_{2,1}$, $\omega_1 \subset T_1$ is a primitive circle and $\omega_2 \subset T_2$ is a $p \geq 2$ power circle. Since ω_2 is disjoint from the separating disk $E \subset R_{2,1}$ a companion solid torus of $V_2 \subset R_{2,1}$ of ω_2 can be isotoped so as to be disjoint from E. Therefore we may assume that the J-torus $T_3 \subset R_{2,1}$ induced by ω_2 as in Section 2.3.10 is disjoint from E, the pair $(R_{2,3}, J)$ is simple of index p, and $E \subset R_{3,1}$.

Let $\omega_3 \subset T_3$ be the power circle of the simple pair $(R_{2,3}, J)$. As $R_{2,1}$ is a handlebody, by Section 2.3.5(3) the circle $\omega_3 \subset T_3$ is primitive in $R_{3,1}$. The circle $\omega_1 \subset T_1$ is primitive in $R_{2,1}$ and hence it must be primitive in $R_{3,1}$. Since $E \subset R_{3,1}$ separates ω_3 and ω_1 it follows from Section 2.3.7 that the circles ω_3 and ω_1 are basic in $R_{3,1}$ and hence that the pair $(R_{3,1}, J)$ is basic.

Suppose now that S is a J-torus in $R_{2,1}$. If S is not isotopic to T_1, T_2 or T_3 in $R_{2,1}$ then by Lemma 4.2.1(3) applied to S and the collection $T_3 \subset R_{2,1}$ we have that one of the pairs $(R_{2,3}, J)$ or $(R_{3,1}, J)$ must be primitive, which is not the case by Lemma 2.3.12(2). Therefore S is isotopic in $R_{2,1}$ to T_1, T_2 or T_3 and hence (2) holds. □

Remark 5.1.2 (1) The handlebody region $R_{1,2}$ in conclusion (1) of Lemma 5.1.1 which splits into a primitive and a simple pair is an example of an *exchange region*. These regions are classified in general in Section 5.3 and their properties will be used in Sections 6 and 7 to obtain the restricted structure of the complex $\mathrm{MS}(K)$ in Theorem 1.

(2) The 4-punctured sphere $P \subset X_L$ constructed in the proof of Lemma 5.1.1 compresses in X_L on one of its sides along a disk which is also a compression disk of either $\partial R_{1,2}$ in $R_{1,2}$ or $\partial R_{2,1}$ in $R_{2,1}$, corresponding to conclusions (1) and (2) of the lemma. If P compresses on both sides then conclusions (1) and (2) hold simultaneously and hence a maximal simplicial collection of Seifert tori in X_K has at most 4 components.

An example where the surface $P \subset X_L$ compresses only on its side contained in the annular pair $(R_{1,2}, J)$ is provided by the family of knots constructed in Proposition 8.4.1(1) and represented in Figure 27, right. In these examples the pair $(R_{2,3}, J)$ is simple, hence annular, but $R_{3,2}$ does not satisfy conclusion (2) of Lemma 5.1.1.

5.2 Index 1 annular pairs in X_K

By Lemma 2.1.1, for an index 1 annular pair $(R_{i,j}, J)$ in X_K the region $R_{i,j}$ may be a handlebody or a boundary irreducible manifold, and in the latter case the complementary region $R_{j,i}$ must be a handlebody. In this section we will see that this relationship between the regions $R_{i,j}$ and $R_{j,i}$, whose union is the exterior X_K of the knot $K \subset \mathbb{S}^3$, greatly limits the topology of the annular pair $(R_{i,j}, J)$.

We first establish a technical result that applies to manifolds like the regions $R_{i,j} \subset X_K$.

Lemma 5.2.1 *Let H be an irreducible manifold with ∂H a surface of genus two, and let $\alpha, \beta, \gamma \subset \partial H$ be nonseparating circles such that*

1. *α is disjoint from $\beta \cup \gamma$,*
2. *β and γ intersect minimally in one point,*
3. *α and β cobound an annulus $A \subset H$,*
4. *$H(\alpha)$ and $H(\gamma)$ are solid tori.*

Then $H(\alpha \sqcup \gamma) = \mathbb{S}^3$ and H is either a genus two handlebody or a toroidal irreducible manifold with irreducible boundary.

Proof Conditions (1)–(4) imply that the circles α, β and γ are nontrivial in H and the circle α is not parallel in ∂H to β or γ. Also the nonseparating annulus $A \subset H$ in (3) turns into the meridian disk $\widehat{A}$

of the solid torus $H(\alpha)$ with $\beta = \partial\widehat{A}$ intersecting $\gamma \subset \partial H(\alpha)$ minimally in one point. Therefore γ is a longitude of $H(\alpha)$ and so

$$\mathbb{S}^3 = H(\alpha)(\gamma) = H(\alpha \sqcup \gamma) = H(\gamma)(\alpha),$$

which implies that α is a longitude of the solid torus $H(\gamma)$.

Let τ be the core of the 2-handle $\mathbb{D}^2 \times I$ used in the construction of $H(\gamma)$. Then τ is an arc properly embedded in the solid torus $H(\gamma)$ with regular neighborhood $N(\tau) = \mathbb{D}^2 \times I \subset H(\gamma)$, such that

(i) $H \subset H(\gamma)$ is the closure of $H(\gamma) \setminus N(\tau)$;

(ii) $N(\gamma) = (\partial\mathbb{D}^2) \times I \subset \partial H$ is an annular neighborhood of γ in ∂H;

(iii) $S_0 = \mathrm{cl}[\partial H(\gamma) \setminus N(\tau)] = \mathrm{cl}[\partial H \setminus N(\gamma)]$ is a twice punctured torus such that $\partial H = S_0 \cup \partial N(\gamma)$;

(iv) $\beta = \beta_1 \cup \beta_2$, where
- $\beta_1 = \beta \cap S_0$ is an arc properly embedded in S_0 connecting the boundary components of S_0,
- $\beta_2 = \beta \cap N(\gamma)$ is a spanning arc of the annulus $N(\gamma)$ which intersects γ minimally in one point.

The situation is represented in Figure 10, top.

Since the circle α and the arc β_1 are disjoint and properly embedded in the twice punctured torus $S_0 \subset \partial H(\gamma)$ and the circle α is a longitude of the solid torus $H(\gamma)$, there is a meridian disk $E \subset H(\gamma)$ such that ∂E lies in $S_0 \subset \partial H(\gamma)$, intersects α minimally in one point, is disjoint from β_1, and intersects the arc $\tau \subset H(\gamma)$ minimally among all meridian disks of $H(\gamma)$ satisfying the previous conditions.

We may therefore assume that $F = E \cap H$ is a punctured disk with boundary the circle $\partial E \subset S_0 \subset \partial H$ and some circle components parallel to γ in the annulus $N(\gamma) \subset \partial H$. Keeping ∂F fixed, we further isotope F in H so as to intersect the annulus $A \subset H$ minimally.

Since $\alpha \subset \partial A$ intersects $\partial E \subset \partial F$ in one point, and each component of ∂F in $N(\gamma) \subset \partial H$ intersects $\beta \subset \partial A$ in one point, it follows that the graph $G_A = A \cap F \subset A$ consists of one spanning arc $a_0 \subset A$ and perhaps some arcs b_i, $1 \le i \le n$, each with both boundary points on the subarc β_2 of the circle $\beta \subset \partial A$.

The presence of the spanning arc a_0 implies that any circle component of $A \cap F$ is trivial in A. If an innermost such circle component c is nontrivial in F then surgery of E along the disk bounded in A by c produces a meridian disk for $H(\gamma)$ satisfying all conditions above but having fewer intersections with A, contradicting the minimality of $A \cap F$. Therefore $A \cap F$ has no circle components and so the graph G_A consists only of arc components, as represented in Figure 10, bottom.

If the arcs b_i are present then the graph G_A has a disk face D_j cobounded by a subarc of β and an outermost arc b_j; but then the disk D_j may be used to boundary compress F in H and reduce by 2 the number of intersections in $H(\gamma)$ between E and τ, contradicting the minimality of $E \cap \tau$.

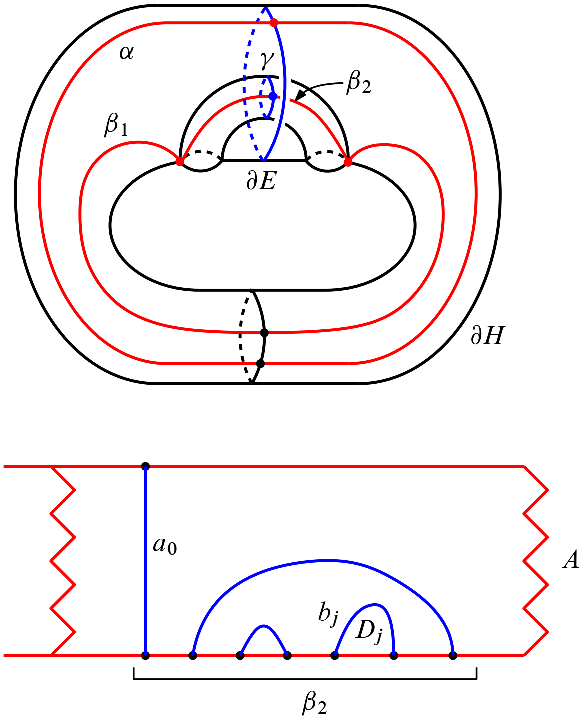


Figure 10: The circles α, $\beta = \beta_1 \cup \beta_2$, γ and ∂E in ∂H (top) and the graph $G_A = A \cap E \subset A$ (bottom).

Therefore $E \cap \tau$ consists of a single point, so $F \subset H$ is an annulus and $A \cap F$ consists of the single arc a_0. This final situation is represented in Figure 10, top.

It follows that $W = N(A \cup F) \subset H$ is a product of the form $T_0 \times I$, where $T_0 = T_0 \times \{0\}$ is the once-punctured torus $N(\alpha \cup \partial E) \subset \partial H$ and $T_0 \times \{1\}$ the once-punctured torus $N(\beta \cup \gamma) \subset \partial H$. The circles ∂T_0 and ∂T_1 are then separating and disjoint in ∂H and hence cobound an annulus $B_0 \subset \partial H$; moreover $B = \operatorname{fr} W = (\partial T_0) \times I$ is a separating annulus properly embedded in H with $\partial B = \partial B_0$.

Let $V = \operatorname{cl}[H \setminus W] \subset H$ be the region in H cobounded by B_0 and B, so that $H = W \cup_B V$. Since $\partial V = B_0 \cup B$ is a torus and $V \subset H \subset \mathbb{S}^3$, V is either a solid torus or the exterior of a nontrivial knot in $\mathbb{S}^3$. Since the circle $\partial T_0 \subset \partial V$ bounds the surface T_0 outside V, if $V \subset \mathbb{S}^3$ is a solid torus then ∂T_0 runs once around V. Therefore $\partial T_0 \subset \partial B_0$ is a longitude of V and so V is a parallelism between the annuli B_0 and B in $H = W \cup_B V$. It follows that $H \approx W = T_0 \times I$ is a genus two handlebody.

Suppose now that V is the exterior of a nontrivial knot in $\mathbb{S}^3$. Then the torus ∂V is incompressible in V and the annulus $B = (\partial T_0) \times I$, which is incompressible in $W = T_0 \times I$, is therefore incompressible in $H = W \cup_B V$. It follows that H is an irreducible and boundary irreducible manifold and that ∂V, when pushed slightly into the interior of H, is an incompressible torus in H. □

With the help of Lemmas 5.1.1 and 5.2.1 we now obtain more information about the topology of an index 1 annular pair $(R_{i,j}, J)$ in X_K and spanning annulus $A \subset R_{i,j}$ by analyzing the manifold $R_{i,j}(\partial A)$.

Lemma 5.2.2 *Let $T_1 \sqcup T_2 \subset X_K$ be a simplicial collection of Seifert tori that cobound an annular pair $(R_{1,2}, J)$ of index 1 with spanning annulus $A \subset R_{1,2}$, such that $\partial_1 A \subset T_1$ and $\partial_2 A \subset T_2$. Then*

(1) *there is a closed 3-manifold M such that, for $\{i,j\} = \{1,2\}$, $R_{1,2}(\partial_i A) = \mathbb{S}^1 \times \mathbb{D}^2 \# M$ with $\partial_j A$ the slope of the meridian of the solid torus summand $\mathbb{S}^1 \times \mathbb{D}^2$;*

(2) *if $R_{1,2}(\partial_1 A)$ is a solid torus then $R_{1,2}$ is a handlebody and so the pair $(R_{1,2}, J)$ is primitive, hence minimal.*

Proof Observe that for $\{i,j\} = \{1,2\}$ the boundary of the manifold $R_{1,2}(\partial_i A)$ is a torus. Since the spanning annulus $A \subset R_{1,2}$ turns into a disk in $R_{1,2}(\partial_i A)$ with boundary the circle $\partial_j A \subset \partial R_{1,2}(\partial_i A)$, it follows that $R_{1,2}(\partial_i A) = \mathbb{S}^1 \times \mathbb{D}^2 \# M_i$ for some closed 3-manifold M_i, where the meridian slope of the solid torus factor $\mathbb{S}^1 \times \mathbb{D}^2$ is $\partial_j A$. Therefore,

$$R_{1,2}(\partial A) = R_{1,2}(\partial_1 A)(\partial_2 A) = \mathbb{S}^1 \times \mathbb{S}^2 \# M_1 = R_{1,2}(\partial_2 A)(\partial_1 A) = \mathbb{S}^1 \times \mathbb{S}^2 \# M_2,$$

whence $M_1 \approx M_2$ by uniqueness of prime factorization. Thus (1) holds.

For part (2) suppose that $R_{1,2}(\partial_1 A)$ is a solid torus and $R_{1,2}$ is not a handlebody. By Lemma 5.1.1 the region $R_{2,1}$ is a handlebody, the circles $\partial_1 A \subset T_1$ and $\partial_2 A \subset T_2$ are separated by a disk in $R_{2,1}$, and we may assume that $\partial_1 A \subset T_1$ is a primitive circle in $R_{2,1}$. Therefore there is a disk $D \subset R_{2,1}$ which intersects $\partial_1 A$ transversely in one point and is disjoint from $\partial_2 A$.

Let $V \subset R_{2,1}$ be the solid torus $R_{2,1}|D$ and denote its core by $L \subset V$. The exterior of the knot $L \subset \mathbb{S}^3 = R_{1,2} \cup R_{2,1}$ is then the manifold $X_L = \mathbb{S}^3 \setminus \operatorname{int} V \approx R_{1,2}(\partial D)$

By (1) the manifold $R_{1,2}(\partial_2 A)$ is a solid torus with meridian slope the circle $\partial_1 A$, and since $\Delta(\partial D, \partial_1 A) = 1$ and $\partial D \cap \partial_2 A = \varnothing$ it follows that

$$\mathbb{S}^3 = R_{1,2}(\partial_2 A)(\partial D) = R_{1,2}(\partial D)(\partial_2 A) = X_L(\partial_2 A).$$

Since $\partial_2 A$ does not bound a disk in $R_{2,1}$, hence neither in the solid torus V, by [8] the knot $L \subset \mathbb{S}^3$ is trivial and hence $R_{1,2}(\partial D) \approx X_L$ is a solid torus. But then by Lemma 5.2.1 applied to the 4-tuple $(H, \alpha, \beta, \gamma) = (R_{1,2}, \partial_1 A, \partial_2 A, \partial D)$ the manifold $R_{1,2}$ must be toroidal, contradicting Lemma 2.1.1(P1). Therefore $R_{1,2}$ is a handlebody and so the pair $(R_{1,2}, J)$ is primitive, hence minimal by Lemma 2.3.12. □

The above results can now be combined to obtain the minimality of any index 1 annular pair in X_K. As a consequence we extend the classification of handlebody annular pairs in X_K given in Lemma 5.1.1(1) to include nonhandlebody such pairs.

For convenience we may denote by $R_{S,S'}$ and $R_{S',S}$ the regions in X_K cobounded by two disjoint Seifert tori $S, S' \subset X_K$, so that $X_K = R_{S,S'} \cup R_{S',S}$.

Proposition 5.2.3 *Let $\mathbb{T} = T_1 \sqcup \cdots \sqcup T_N$, $N \geq 1$, be a simplicial collection of Seifert tori in X_K with minimal pairs. Then:*

(1) *Any index* 1 *annular pair* $(R_{i,j}, J)$ $(i = j$ *allowed) in* X_K *is minimal, and if* $R_{i,j}$ *is not a handlebody then* $|\mathbb{T}| \leq 3$.

(2) *If* $|\mathbb{T}| \geq 2$ *and for some* $k \geq 2$ *the nonminimal pair* $(R_{1,1+k}, J)$ *is annular of index* $p \geq 2$ *then* $k = 2$ *and* $R_{1,1+k} = R_{1,3}$, *and if* $T_1', T_3' \subset R_{1,3}$ *are the* J*-tori induced by* T_1 *and* T_3, *respectively, then*

 (a) T_1' *and* T_3' *are not isotopic in* X_K,

 (b) *any* J*-torus in* $R_{1,3}$ $(T_2$ *for instance) is isotopic to* T_1, T_3, T_1' *or* T_3',

 (c) *the pair* $(R_{T_1,T_1'}, J)$ *is simple of index* p *and* (R_{T_1',T_3}, J) *is annular of index* 1,

 (d) *the pair* $(R_{T_1,T_3'}, J)$ *is annular of index* 1 *and* (R_{T_3',T_3}, J) *is simple of index* p.

Proof If the annular pair $(R_{i,j}, J)$ is minimal and $R_{i,j}$ is not a handlebody then $R_{j,i}$ is a handlebody by Lemma 2.1.1(P2), in which case by Lemma 5.1.1(2) either:

- The pair $(R_{j,i}, J)$ is basic, hence minimal by Lemma 2.3.12(1); in this case we obtain $|\mathbb{T}| \leq 2$, where $|\mathbb{T}| = 1$ if the basic pair $(R_{j,i}, J)$ is trivial.
- $R_{j,i}$ contains exactly one J-torus not parallel to T_i or T_j, in which case $|\mathbb{T}| = 3$.

Therefore the second part of (1) follows from the first part. For the first part of (1) we argue by contradiction. For definiteness suppose that the pair $(R_{1,j}, J)$ is annular of index 1 for some $j \neq 1, 2$, so that $R_{1,3} \subseteq R_{1,j}$ with $(R_{1,3}, J)$ a nonminimal pair. By Lemma 4.1.1(6) we may therefore assume that $j = 3$. Let $A \subset R_{1,3}$ be a spanning annulus.

(I) By Lemma 4.1.1(6) we may assume that $A_1 = A \cap R_{1,2}$ and $A_2 = A \cap R_{2,3}$ are spanning annuli of index 1 in $R_{1,2}$ and $R_{2,3}$, respectively. Therefore each of the pairs $(R_{1,2}, J)$ and $(R_{2,3}, J)$ is annular of index 1 and so the region $R_{1,3}$ is not a handlebody by [24, Lemma 3.7(3)].

(II) $R_{1,2}$ and $R_{2,3}$ are handlebodies.

If $R_{1,2}$ is not a handlebody then $R_{2,3} \subseteq R_{2,1}$ is a handlebody by Lemma 2.1.1(P2) and (P3), and hence the pair $(R_{2,3}, J)$ is primitive by (I). This contradicts Lemma 5.1.1(2)(b) since a primitive pair is neither basic nor simple by Lemma 2.3.12. Therefore $R_{1,2}$ is a handlebody, and by a similar argument $R_{2,3}$ is also a handlebody.

(III) $R_{1,3}(\partial_1 A)$ is a solid torus.

By (I) and (II) the pairs $(R_{1,2}, J)$ and $(R_{2,3}, J)$ are primitive with spanning annuli $A_1 \subset R_{1,2}$ and $R_{2,3}$. Denote the boundary components of A_1 by $\omega_1 = A_1 \cap T_1$ and $\omega_2 = A_1 \cap T_2$; these are primitive circles in $R_{1,2}$, and $\omega_2 = A_2 \cap T_2$ is primitive in $R_{2,3}$. Therefore $R_{1,2}(\omega_1)$ is a solid torus with meridian disk $\widehat{A}_1$ such that $\partial \widehat{A}_1 = \omega_2 \subset \partial R_{1,2}(\omega_1)$, and $R_{2,3}(\omega_2)$ is also a solid torus.

For $i = 1, 2$ each manifold $R_{i,i+1}(J)$ has boundary the tori $\widehat{T}_i$ and $\widehat{T}_{i+1}$, and

$$R_{1,3}(J) = R_{1,2}(J) \cup_{\widehat{T}_2} R_{2,3}(J).$$

Therefore,

$$\begin{aligned} R_{1,3}(\partial_1 A) = R_{1,3}(\omega_1) &= R_{1,3}(J)(\omega_1) \\ &= [R_{2,3}(J) \cup_{\widehat{T}_2} R_{1,2}(J)](\omega_1) \\ &= R_{2,3}(J) \cup_{\widehat{T}_2} [R_{1,2}(J)(\omega_1)] = R_{2,3}(J)(\partial \widehat{A}_1) = R_{2,3}(\omega_2) = \text{solid torus}. \end{aligned}$$

This contradicts Lemma 5.2.2(2) since by (I) the region $R_{1,3}$ is not a handlebody. Therefore (1) holds.

Part (2)(a) will be established in the next section in Lemma 5.3.2. Parts (2)(c) and (2)(d) follow from Lemma 4.1.1(6) and the properties of induced tori in Section 2.3.10.

For part (2)(b) suppose that $S \subset R_{1,3}$ is a J-torus that is not parallel to T_1 or T_3. By Lemma 4.1.1(6) we may assume that the pair $(R_{T_1,S}, J)$ is annular of index $p \geq 2$ while (R_{S,T_3}, J) is annular of index 1.

By Section 3.2(A4) the J-torus T_1' induced by T_1 in $R_{T_1,S}$ is isotopic to the J-torus induced by T_1 in $R_{1,3}$, so it is not parallel to T_3 in $R_{1,3}$, and cobounds a region $R_{T_1,T_1'} \subset R_{T_1,S}$ such that the pair $(R_{T_1,T_1'}, J)$ is simple of index $p \geq 2$.

By Lemma 4.1.1(6) the pair (R_{T_1',T_3}, J) is then annular of index 1 and hence minimal by (1). As $R_{T_1',T_3} = R_{T_1',S} \cup_S R_{S,T_3}$ and the pair (R_{S,T_3}, J) is nontrivial, it follows that the pair $(R_{T_1',S}, J)$ must be trivial and hence that S is parallel to T_1'. Therefore (2)(b) holds. □

Examples of genus one hyperbolic knots in $\mathbb{S}^3$ realizing the conditions of Proposition 5.2.3(1), with $R_{i,j}$ a handlebody or not, can be found in Sections 8.2, 8.3 and 8.5.

5.3 Exchange regions and the exchange trick

By Proposition 5.2.3(2)(c) and (d), given an annular pair $(R_{1,3}, J)$ of index $p \geq 2$ of a simplicial collection of Seifert tori $\mathbb{T} \subset X_K$ with minimal pairs, the region $R_{1,3} \subset X_K$ contains two nontrivial minimal subpairs $(R_{T_1,T}, J)$ and $R_{T,T_3}, J)$ where the nature of each subpair alternates between being simple of index p or annular of index 1 depending on the choice of splitting J-torus $T \subset R_{1,3}$.

We will refer to any region in X_K with properties similar to those of the region $R_{1,3} \subset X_K$ as an *exchange region*, to the pair $(R_{1,3}, J)$ as an *exchange pair*, and to the switch of type of subpair in $R_{1,3}$ adjacent to T_1 or T_3 between a simple and an index 1 pair as the *exchange trick*.

In this section we prove Lemma 5.3.2 which states that the two induced J-tori in an exchange region are not isotopic in X_K, thus completing the proof of Proposition 5.2.3(2)(a). We first review the construction of the induced J-tori $T_1', T_3' \subset R_{1,3}$ given in Section 2.3.10 and Section 3.2(A4).

By hypothesis the pair $(R_{1,3}, J)$ is annular of index $p \geq 2$ and so by Lemma 3.1.1(5) there are disjoint spanning annuli $A, A' \subset R_{1,3}$ which cobound a solid torus $V \subset R_{1,3}$ around which each spanning annulus runs p times.

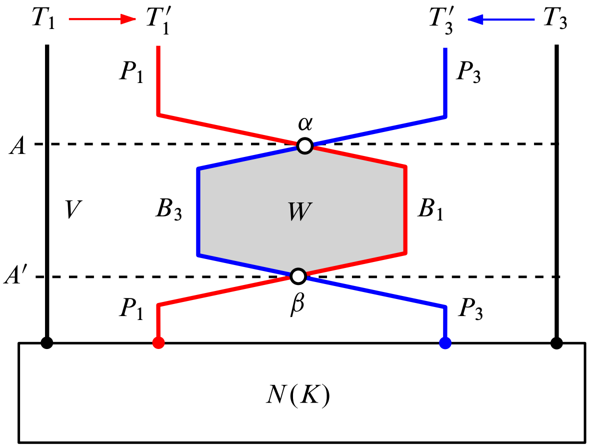


Figure 11: The induced J-tori T'_1 and T'_3 in the exchange region $R_{1,3} \subset X_K$.

Push V off T_3 and into $R_{1,3}$ to obtain a companion solid torus $V_1 \subset R_{1,3}$ of the circle $\omega_1 = A \cap T_1$. Similarly the circle $\omega_3 \subset T_3$ has a companion solid torus $V_3 \subset R_{1,3}$. For $i = 1, 3$ the frontier of a thin regular neighborhood $N(T_i \cup V_i) \subset R_{1,3}$ then consists of T_i and the J-torus T'_i induced by T_i. The situation is represented in Figure 11.

Notice that T'_1 and T'_3 intersect transversely in two circles $T'_1 \cap T'_3 = \alpha \sqcup \beta$ that are nonseparating in T'_1 and T'_3. The closures of the components of $T'_1 \setminus T'_3$ consist of an annulus B_1 and a pants P_1, and those of $T'_3 \setminus T'_1$ of an annulus B_3 and a pants P_3, with $\partial B_1 = \alpha \sqcup \beta = \partial B_3$ as shown in Figure 11.

By the construction of the induced tori T'_1 and T'_3, the torus $B_1 \cup B_3 \subset R_{1,3}$ bounds a solid torus $W \subset R_{1,3}$ obtained by pushing the solid torus $V \subset R_{1,3}$ off from T_1 and T_3, as represented in Figure 11. Since the index of $(R_{1,3}, J)$ is $p \geq 2$, each spanning annulus runs p times around V and hence each circle α and β runs p times around W.

In order to establish the isotopy properties of the induced J-tori $T'_1, T'_3 \subset R_{1,3}$ we use the following result of [16, Proposition 4.8], an elaboration of the results in [25].

Let P and Q be surfaces properly embedded in a 3-manifold M and which intersect transversely. A *product region* between P and Q is an embedded copy of a manifold of the form $\tilde{\Sigma} = \Sigma \times I/\!\sim$ in M, where Σ is a surface, b is a compact 1-submanifold of $\partial\Sigma$, and

(i) for each $x \in b$ the relation $\sim$ collapses the arc $\{x\} \times I$ to a point,

(ii) $\Sigma \times \{0\} \subset P$, $\Sigma \times \{1\} \subset Q$, and $\mathrm{cl}[(\partial\Sigma - b)] \times I \subset \partial M$,

(iii) $P \cap \operatorname{int} \tilde{\Sigma} = \varnothing$, and $Q \cap \operatorname{int} \tilde{\Sigma}$ may be nonempty only when Σ is a disk and $P \cap \partial M$ is connected.

Lemma 5.3.1 [16] *Let M be a Haken 3-manifold with incompressible boundary, and let $P, Q \subset M$ be properly embedded incompressible and boundary incompressible surfaces which intersect transversely with $\partial P \cap \partial Q = \varnothing$. If in M the surfaces P and Q are isotopic or P is isotopic to a surface disjoint from Q then there is a product region between P and Q.* □

In the presence of a product region $\widetilde{\Sigma}$ between P and Q it is possible to reduce $|P \cap Q|$ by an isotopy that exchanges $\Sigma \times \{1\} \subset Q$ with $\Sigma \times \{0\}$ and pushes the resulting new surface Q slightly off P.

We will apply the lemma to surfaces with disjoint boundary, in which case the 1-submanifold $b \subset \partial\Sigma$ is simply a union of components of $\partial\Sigma \subset \partial P \sqcup \partial Q \sqcup (P \cap Q)$.

Lemma 5.3.2 *The Seifert tori $T_1', T_3' \subset X_K$ are nonisotopic in X_K.*

Proof By Lemma 5.3.1 it suffices to show that there are no product regions in X_K between T_1' and T_3'.

By Proposition 5.2.3(2) for $i = 1, 3$ the pairs $(R_{T_1,T_i'}, J)$ and (R_{T_i',T_3}, J) are nontrivial and so the induced J-tori T_1' and T_3' are not parallel to T_1 or T_3 in $R_{1,3}$. Therefore T_1' and T_3' intersect minimally in $R_{1,3}$ and so by Lemma 5.3.1 there are no product regions between T_1' and T_3' contained in the region $R_{1,3}$.

By the above construction of the induced tori T_1' and T_3' any product region $\widetilde{\Sigma}$ in X_K between T_1' and T_3' must run between the following subsurfaces of T_1' and T_3':

(a) B_1 and B_3: Here the only possible product region $\widetilde{\Sigma} \subset X_K$ must be constructed from $\Sigma = B_1$ and $b = \partial B_1 = \partial B_3$ with

$$\widetilde{\Sigma} \cap T_1' = \Sigma \times \{0\} = B_1, \quad \widetilde{\Sigma} \cap T_3' = \Sigma \times \{1\} = B_3, \quad \widetilde{\Sigma} \cap \partial X_K = \varnothing,$$

whence necessarily $\widetilde{\Sigma} = W \subset R_{1,3}$, contradicting the argument above that $\widetilde{\Sigma} \not\subset R_{1,3}$.

(b) P_1 and P_3: The surface $P_1 \cup P_3$ is a separating twice punctured torus properly embedded in X_K and so any product region $\widetilde{\Sigma}$ between P_1 and P_3 must be constructed from $\Sigma = P_1$ and $b = \partial P_1 \cap \partial P_3 = \alpha \sqcup \beta$ with

$$\widetilde{\Sigma} \cap T_1' = \Sigma \times \{0\} = P_1, \quad \widetilde{\Sigma} \cap T_3 = \Sigma \times \{1\} = P_3, \quad \widetilde{\Sigma} \cap \partial X_K = (\partial T_1') \times I.$$

Thus $\widetilde{\Sigma} \approx P_1 \times I$ is a genus two handlebody such that each component of $\partial P_1 \supset \alpha \sqcup \beta$ is a primitive circle in $\widetilde{\Sigma}$. However, as $\widetilde{\Sigma} \not\subset R_{1,3}$, we must have $W \subset \widetilde{\Sigma}$ (see Figure 11) which implies that α and β are $p \geq 2$ power circles in $\widetilde{\Sigma}$.

This last contradiction shows that there are no product regions in X_K between T_1' and T_3'. □

6 No exchange regions for $|\mathbb{T}| = 5$

In this section we assume that $K \subset \mathbb{S}^3$ is a genus one hyperbolic knot which bounds a maximal simplicial collection of five Seifert tori $\mathbb{T} = T_1 \sqcup \cdots \sqcup T_5 \subset X_K$. Our goal is to prove the following result:

Proposition 6.0.1 *If $|\mathbb{T}| = 5$ then no pair $(R_{i,i+2}, J)$ is an exchange pair and the maximal simplicial collection $\mathbb{T} \subset X_K$ is unique up to isotopy.*

A sketch of the proof goes like this. Both an exchange region, say $R_{1,3}$, and its complementary region $R_{3,1}$ must be genus two handlebodies; thus the pair $(R_{3,1}, J)$ is maximal. At this point we use a method developed in [24, Section 7.3] to construct a Heegaard diagram from the Heegaard decomposition $R_{1,3} \cup R_{3,1}$ which applies whenever one of the pairs $(R_{1,3}, J)$ or $(R_{3,1}, J)$ is maximal. That the Heegaard decomposition $R_{1,3} \cup R_{3,1}$ cannot correspond to $\mathbb{S}^3$ can then be detected from the fact that otherwise the core knot of a simple pair in $R_{1,3}$ should be a trivial or torus knot in $\mathbb{S}^3$, which we show cannot be the case.

6.1 The regions $R_{i,i+1}$

The following general result restricts the types of pairs $(R_{i,i+1}, J)$ produced by the simplicial collection $\mathbb{T} \subset X_K$. Its proof relies on an analysis of the essential graphs of intersection between $\mathbb{T}$ and a *Gabai* meridional planar surface for $\mathbb{T}$ from [7], along with some basic results from [24, Sections 2.1 and 2.2].

Lemma 6.1.1 *For each $1 \le i \le 5$ the region $R_{i,i+1}$ is a handlebody, the pair $(R_{i,i+1}, J)$ is minimal, and at least one of the pairs $(R_{i,i+1}, J)$ or $(R_{i+1,i+2}, J)$ is simple.*

Proof Each pair $(R_{i,i+1}, J)$ is minimal since the simplicial collection $\mathbb{T}$ is maximal, and each region $R_{i,i+1}$ is a handlebody by [24, Lemma 4.1(3)]. We show that of any two consecutive pairs, say $(R_{1,2}, J)$ and $(R_{2,3}, J)$, at least one of them is simple.

Let $\mathbb{T}' = T_1 \sqcup T_2 \sqcup T_3$. By [7] there is a planar surface $Q \subset X_K$ with meridional boundary slope which intersects $\mathbb{T}'$ transversely in essential graphs $G_Q = Q \cap \mathbb{T}' \subset Q$ and $G' = Q \cap \mathbb{T}' \subset \mathbb{T}'$. Necessarily each cycle around a face of G_Q has an even number of edges.

If the graph G_Q has no parallel edges then it is a reduced graph and the degree of each of its vertices is 3; therefore by [24, Lemma 2.3(2)] G_Q has a disk face D_4 with 4 edges around its boundary. Otherwise G_Q has parallel edges, that is, G_Q has a disk face D_2 with 2 edges around its boundary.

The disk face $E \in \{D_2, D_4\}$ of G_Q lies in one of the regions $R_{1,2}$, $R_{2,3}$ or $R_{3,1}$, and by [24, Lemma 2.1(3)] intersects J minimally in 2 or 4 points.

If $E \subset R_{i,j}$ then $R_{i,j}$ is a handlebody by Lemma 2.1.1(P1) and so, by [24, Lemma 6.1], $(R_{i,j}, J)$ is either a simple pair, which is minimal, or a nonminimal double pair. However, by Section 2.3.8, in a double pair any J-torus is parallel to a boundary J-torus or to the J-torus separating the double pair into simple subpairs. Since $R_{3,1}$ contains the two J-tori T_4 and T_5, which are not parallel to the boundary or to each other, it follows that $R_{i,j} \neq R_{3,1}$.

Therefore $E \subset R_{1,2}$ or $E \subset R_{2,3}$, in which case, respectively, the minimal pair $(R_{1,2}, J)$ or $(R_{2,3}, J)$ is simple by [24, Lemma 6.1]. □

For the rest of Section 6 we assume that $R_{1,3}$ is the exchange region for the simplicial collection $\mathbb{T}$ with exchange J-tori T_2 and $T_{2'}$ as shown in Figure 12. In the figure the elements of each simple pair will be represented using the notation set up in Section 2.3.4 and Figure 5. Circles of distinct slope, like $\omega_1, \omega_5' \subset T_1$, are represented as nonoverlapping.

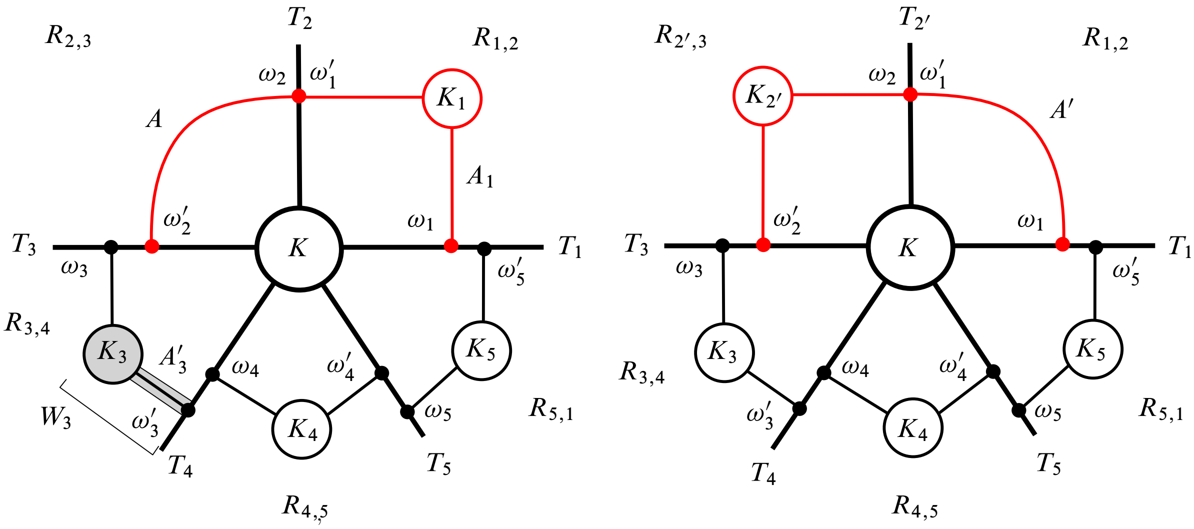


Figure 12: The knot $K \subset \mathbb{S}^3$ with the exchange pair $(R_{1,3}, J)$.

In the exchange region $R_{1,3}$ the pairs $(R_{1,2}, J)$ and $(R_{2',3}, J)$ in Figure 12 are simple with core knots K_1 and $K_{2'}$, respectively.

6.2 The regions $R_{i,j}$

In this section we establish some of the general properties of the pairs $(R_{i,j}, J)$. We will see that indeed each of the pairs $(R_{i,i+1}, J)$ is simple for $i = 3, 4, 5$ as represented in Figure 12.

We will use the following notation: a circle γ in the boundary of a genus two handlebody H is a *Seifert circle* if $H(\omega) = \mathbb{D}^2(p,q)$ for some integers $p, q \geq 2$.

(E1) By Lemma 6.1.1 applied to the regions $R_{2,4}$ and $R_{5,2'}$, each of the pairs $(R_{3,4}, J)$ and $(R_{5,1}, J)$ is simple, with cores the knots K_3 and K_5, respectively.

In Figure 12, left, $\omega_2' \neq \omega_3$ by Lemma 4.1.1(4) and so the region $R_{2,4}$ is not a handlebody by Lemma 4.1.1(7). Therefore the region $R_{4,2}$ is a handlebody by Lemma 2.1.1(P2) and so the pair $(R_{4,2}, J)$ is maximal. By [24, Lemma 6.8] it follows that

(a) $(R_{4,5}, J)$ is a simple pair with core the knot K_4,

(b) the simple pair $(R_{5,1}, J)$ is a basic pair with basic circles $\omega_4' \subset T_5$ and $\omega_1 \subset T_1$,

(c) $\Delta(\omega_4', \omega_5) = 1 = \Delta(\omega_5', \omega_1)$.

A similar argument using Figure 12, right, shows that

(d) the simple pair $(R_{3,4}, J)$ is a basic pair with basic circles $\omega_2' \subset T_3$ and $\omega_4 \subset T_4$,

(e) $\Delta(\omega_2', \omega_3) = 1 = \Delta(\omega_3', \omega_4)$.

Let $p_1, p_2, p_3, p_4, p_5 \geq 2$ be the indices of the simple pairs $(R_{1,2}, J)$, $(R_{2',3}, J)$, $(R_{3,4}, J)$, $(R_{4,5}, J)$ and $(R_{5,1}, J)$, respectively.

(E2) The region $R_{3,1}$ is a handlebody and the core knots K_3 and K_5 are hyperbolic Eudave-Muñoz knots of indices $p_3 = 2 = p_5$.

Let W_3 be the solid torus neighborhood of $K_3 \sqcup A'_3$ in $R_{3,4}$ indicated in Figure 12, left, where A'_3 is the annulus constructed in Section 2.3.4 (see also Figure 5) such that the slope r_3 of the boundary circle $A'_3 \cap \partial N(K_3)$ is nonintegral (relative to $N(K_3)$) of the form $*/p_3$. We identify the exterior $X_{K_3} \subset \mathbb{S}^3$ of the core knot K_3 with $\mathbb{S}^3 \setminus \operatorname{int} W_3$. Back in Figure 12, left, observe that

- by (E1), $(R_{3,4}, J)$ is a simple pair and $R_{2,4}$ is not a handlebody,
- $R_{3,2}$ is not a handlebody by Lemma 2.1.1 and Section 2.3.9.

Therefore [24, Lemma 7.1(1)] applies to the simple pair $(R_{3,4}, J)$ (denoted by $(R_{3,4}, K)$ in [24]) and the J-tori T_2, T_3 and T_4 to conclude that the separating two-punctured torus $F = \operatorname{cl}[T_2 \cup T_4 \setminus W_3]$ is incompressible in X_{K_3} and the torus $\widehat{F}$ is incompressible in $X_{K_3}(r_3)$. Moreover, the closures F^+ and F^- of the components of $X_{K_3} \setminus F$ can be identified with the handlebodies $F^+ = R_{4,2}$ and $F^- = \operatorname{cl}[R_{2,4} \setminus W_3] \approx R_{2,3}$ (by Section 2.3.5(1)), and since the slope $r_3 \subset \partial X_{K_3}$ is nonintegral (with denominator $p_3 \geq 2$) it follows from [13, Lemma 3.14] that K_3 is a hyperbolic Eudave-Muñoz knot.

By [9] r_3 is the unique nonintegral toroidal slope for K_3 and we must have $p_3 = 2$. By [6, Theorem 2.1 and Proposition 2.2] the torus $\widehat{F}$ is the unique essential torus in $X_{K_3}(r_3)$ and it decomposes $X_{K_3}(r_3)$ into a union of two Seifert fiber spaces of the form $\mathbb{D}^2(*, *)$ for $* \geq 2$. As

$$X_{K_3}(r_3) \approx R_{2,3}(\omega'_2) \cup_{\widehat{F}} R_{4,2}(\omega'_3),$$

it follows that $R_{4,2}(\omega'_3)$ is a space of the form $\mathbb{D}^2(*, *)$ for $* \geq 2$ and hence that $\omega'_3 \subset T_4$ is a Seifert circle in the handlebody $R_{4,2}$. As $\partial R_{4,2} \setminus \omega'_3$ contains the power circle $\omega'_1 \subset T_2$, [24, Lemma 6.10(2)(b)] applied to the pair $(R_{4,2}, K)$ and the J-torus $T_1 \subset R_{4,2}$ yields that ω'_3 is a primitive circle in the handlebody $R_{4,1}$, which by [24, Lemma 7.2(5)] implies that $R_{3,1}$ is a handlebody.

Using Figure 12, right, a symmetric argument shows that K_5 is also a hyperbolic Eudave-Muñoz knot and $p_5 = 2$.

(E3) The core knot K_1 (Figure 12, left) is a trivial or torus knot.

Since the pair $(R_{2,3}, J)$ is primitive with spanning annulus A, $R_{2,3}(\omega'_1)$ is a solid torus with meridian disk $\widehat{A}$ and hence meridian slope $\partial\widehat{A} = \omega'_2 \subset T_3$. As the circles ω'_2 and ω_4 are basic in $R_{3,4}$, it follows that $V_3 = R_{2,4}(\omega'_1) = R_{2,4}(J)(\omega'_1)$ is a solid torus with meridian slope that intersects ω_4 in one point. Similarly, as the circles ω_1 and ω'_4 are basic in $R_{5,1}$, $V_5 = R_{5,1}(\omega_1) = R_{5,1}(J)(\omega_1)$ is a solid torus with meridian slope that intersects ω'_4 in one point. Therefore, if $r_1 \subset \partial N(K_1)$ is the nonintegral slope $*/p_1$ of the boundary circle $A_1 \cap \partial N(K_1)$ (see Figure 12, left) constructed in Section 2.3.4, then we have

$$\begin{aligned} X_{K_1}(r_1) \approx R_{2,1}(\omega_1 \sqcup \omega'_1) &= R_{2,1}(J)(\omega_1 \sqcup \omega'_1) \\ &= (R_{2,4} \cup R_{4,5} \cup R_{5,1})(J)(\omega_1 \sqcup \omega'_1) \approx R_{4,5}(J) \cup V_3 \cup V_5 = \mathbb{S}^2(p_4, 1, 1) \end{aligned}$$

and so $X_{K_1}(r_1)$ is either $\mathbb{S}^3$, $\mathbb{S}^1\times\mathbb{S}^2$ or a lens space. In fact, as r_1 is a nonintegral slope, for homological reasons $X_{K_1}(r_1)$ cannot be $\mathbb{S}^1\times\mathbb{S}^2$. Therefore by [3] and [8] K_1 is a trivial or torus knot.

(E4) The circles ω_1 and ω_2' are Seifert circles in $R_{3,1}$.

Since the region $R_{4,2}$ in Figure 12, left, is a handlebody, the circle ω_1 is primitive in $R_{4,1}$ by [24, Lemma 6.8(1)(b)]; hence ω_1 is a Seifert circle in $R_{3,1}$ by [24, Lemma 6.8(1)(d)].

A similar argument applied to the handlebody region $R_{2',5}$ of Figure 12, right, shows that ω_2' is also a Seifert circle in $R_{3,1}$.

6.3 The maximal pair $(R_{3,1}, K)$

For the rest of this section we assume that the regular neighborhood $N(K)\subset\mathbb{S}^3$ has been retracted radially onto K, so that the circles J and ∂T_i become identified with K. Thus we use the notation $(R_{i,i+1}, K)$ for the pairs $(R_{i,i+1}, J)$

We construct a complete disk system for the maximal pair $(R_{3,1}, K)$ as follows. First observe that by Lemma 4.1.1(6) the maximal pair $(R_{3,1}, K)$ is not annular; hence by [24, Lemma 3.4] there is a disk $E\subset R_{3,1}$, unique up to isotopy, which separates the power circles $\omega_3, \omega_5'\subset R_{3,1}$. Thus $R_{3,1}|E$ consists of two solid tori V_3 and V_5 with the power circles $\omega_3\subset\partial V_3$ and $\omega_5'\subset\partial V_3$ intersecting meridian disks $D_3\subset V_3$ and $D_5\subset V_5$ in $p_3=2$ and $p_5=2$ points, respectively. Using the method outlined in [24, Section 7.3 and Lemma 7.6(2)], the 6-tuple $(\partial E, K, \omega_1, \omega_2', \omega_3, \omega_5')$ is homeomorphic to the 6-tuple in Figure 13, top, or to the 6-tuple obtained by reflecting $\partial R_{3,1}$ across the plane of the page. The two 6-tuples are then homeomorphic and hence we only consider the case of Figure 13, top.

The meridian circles $\partial D_3, \partial D_5\subset\partial R_{3,1}$ can then be constructed as homological sums of the form

$$\partial D_3 = q_3\omega_3 + a_3\beta_3 \quad\text{and}\quad \partial D_5 = q_5\omega_5' + a_5\beta_5,$$

where $\gcd(a_3,q_3)=1=\gcd(a_5,q_5)$ and $\beta_3\subset\partial V_3$ and $\beta_5\subset\partial V_5$ are the two circles indicated in Figure 13, bottom, such that

$$\Delta(\beta_3,\omega_3)=1=\Delta(\beta_5,\omega_5').$$

The values of a_3, a_5, q_3 and q_5 can be found by performing some computations in the fundamental group of $R_{3,1}$. To this end we set $x=\partial D_3$, $y=\partial D_5$, and $\pi_1(R_{3,1})=\langle x,y\mid -\rangle$ relative to some base point. Thus a circle $c\subset\partial R_{3,1}$ which intersects $x\sqcup y$ transversely is represented by an unreduced word $c(x,y)\in\pi_1(R_{3,1})=\langle x,y\mid -\rangle$ obtained by reading the consecutive signed intersections of c with x and y without introducing any cancellations, relative to a base point in $c\setminus(x\sqcup y)$. Notice that if the word $c(x,y)$ is cyclically reduced then c intersects $x\sqcup y$ minimally, but not conversely.

For convenience we use the notation $X=x^{-1}$ and $Y=y^{-1}$ to denote the inverses of x and y in $\langle x,y\mid -\rangle$.

The following relations now follow from Figure 13, bottom (relative to some base point and intersection signs scheme):

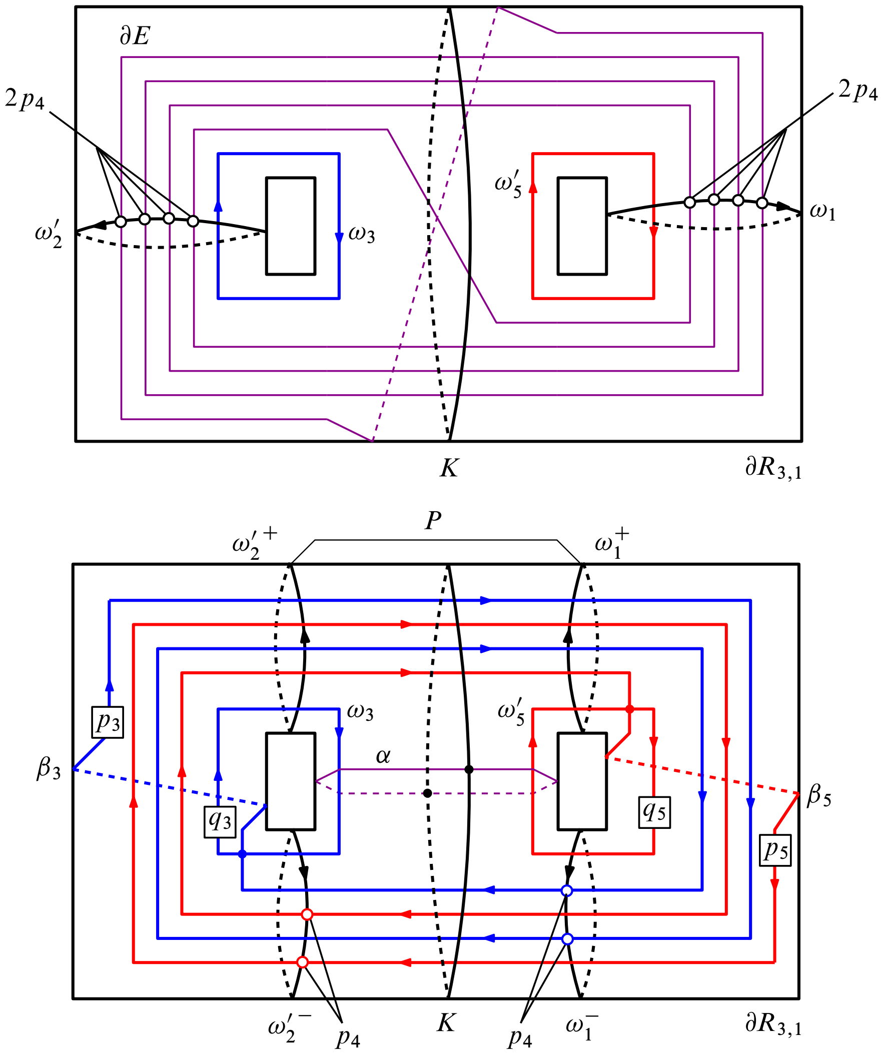


Figure 13: Construction of the circles $\partial D_3 = q_3\omega_3 + p_3\beta_3$ and $\partial D_5 = q_5\omega_5' + p_5\beta_5$ in $\partial R_{3,1}$ (with $p_4 = 2$).

(E5) (a) $\omega_3(x,y) = x^{a_3}$ and $\omega_5'(x,y) = y^{a_5}$. Since ω_3 and ω_5' are p_3 and p_4 power circles in $R_{3,1}$, respectively, we may choose $a_3 = p_3 = 2$ and $a_5 = p_5 = 2$;

(b) $\omega_1(x,y) = (y^{p_5}x^{p_3})^{p_4}y^{q_5}$ and $\omega_2'(x,y) = (y^{p_5}x^{p_3})^{p_4}x^{q_3}$. Observe that $\omega_1(x,y) = W(x^{p_3},y)$ where $W(x,y) = (y^{p_5}x)^{p_4}y^{q_5}$. As ω_1 is a Seifert circle in $R_{3,1}$, by the argument of [24, Lemma 7.11] we must have that $W(x,y)$ is a primitive word in the free group $\langle x,y \mid - \rangle$ and hence that $q_5 = \pm 1$. In a similar way we must have that $q_3 = \pm 1$.

For each circle ω_3, ω_5', β_3 and β_5 in Figure 13, bottom, the coefficient in the box on top of the circle represents the number of copies of that circle used in the homological sum construction of a given meridian circle $x = \partial D_3$ and $y = \partial D_5$.

We summarize the above facts in the next result.

Lemma 6.3.1 *The 7-tuple $(\partial R_{3,1}, \partial E, K, \omega_1, \omega'_2, \omega_3, \omega'_5)$ is homeomorphic to the 7-tuple in Figure 13, top (where we use $p_4 = 2$ for simplicity). Moreover, the circles $\partial D_3, \partial D_5 \subset \partial R_{3,1}$ can be represented as the homological sums*

$$\partial D_3 = q_3\omega_3 + p_3\beta_3 \quad \text{and} \quad \partial D_5 = q_5\omega'_5 + p_5\beta_5$$

where β_3 and β_5 are the circles indicated in Figure 13, bottom, with $p_3 = 2 = p_5$ and $q_3, q_5 = \pm 1$. □

Since the circles $\omega_1, \omega'_2 \subset \partial R_{1,3}$ cobound an annulus in $R_{1,3}$, by [24, Lemma 3.4] the surface

$$\partial R_{1,3} \setminus (\omega_1 \sqcup \omega'_2)$$

compresses along a nonseparating disk $D \subset R_{1,3}$ (unique up to isotopy), and necessarily $R_{1,3}|D$ is a solid torus neighborhood of the knot K_1. Therefore we may set

$$X_{K_1} = R_{3,1}(\partial D)$$

By (E3) the core knot $K_1 \subset \mathbb{S}^3$ of the simple pair $(R_{1,2}, K)$ is either trivial or a torus knot. Therefore $X_{K_1} = R_{3,1}(\partial D)$ is either a solid torus or a Seifert fiber space of the form $\mathbb{D}^2(*,*)$ for $* \geq 2$, or, equivalently:

(E6) The circle ∂D is either primitive or Seifert in $R_{3,1}$.

The next two results will be useful in restricting the possible embeddings of the circle ∂D in $\partial R_{3,1}$.

Lemma 6.3.2 *Let $\omega \subset \partial R_{3,1}$ be any circle that intersects $x \sqcup y \subset \partial R_{3,1}$ minimally. If some cyclic reorderings of the word $\omega(x,y) \in \pi_1(R_{3,1})$ contain strings of the form x^a and y^b for some integers $|a|, |b| \geq 2$ then $\omega(x,y)$ is cyclically reduced and ω is neither a primitive nor a power circle in $R_{3,1}$.*

Proof Let Q be the 4-punctured 2-sphere $\partial R_{3,1} \setminus \operatorname{int} N(x \sqcup y)$ with boundary components the circles x^+, x^- and y^+, y^- corresponding to the two sides of x and y in $\partial R_{3,1}$, respectively. Since ω intersects $x \sqcup y$ minimally, $\omega \cap Q$ consists of a collection of properly embedded arcs none of which is parallel into ∂Q.

By hypothesis, some cyclic reordering of the word $\omega(x,y)$ contains a string of the form x^a for some integer $|a| \geq 2$ and so there is an arc component $c_x \subset Q$ with one endpoint in x^+ and the other in x^-. Thus the circle $\gamma = \operatorname{fr} N(x^+ \cup c_x \cup x^-) \subset Q$ separates $x^+ \sqcup x^-$ from $y^+ \sqcup y^-$.

Suppose that some cyclic reordering of the word $\omega(x,y)$ has a canceling string of the form yY or Yy. Then there is an arc component $c_y \subset \omega \cap Q$ with both endpoints on, say, the boundary component $y^+ \subset \partial Q$. As c_y is disjoint from c_x, it is also disjoint from γ and so c_y separates $x^+ \sqcup x^-$ from y^- in Q (see Figure 14), which is impossible since ω is a closed circle in $\partial R_{3,1}$.

Therefore, no cyclic reordering of the word $\omega(x,y)$ has canceling strings of the form yY or Yy, and in a similar way neither of the form xX or Xx; hence it is a cyclically reduced word. Since the word $\omega(x,y)$

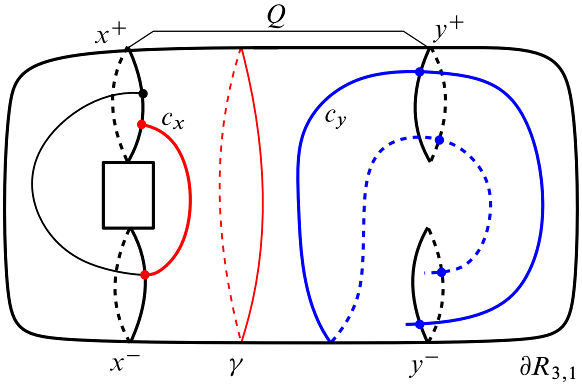


Figure 14: The arc components $c_x, c_y \subset \omega \cap Q$.

contains strings of the form x^a and y^b for some $|a|, |b| \geq 2$, by Section 2.3.1 the circle $\omega \subset \partial R_{3,1}$ is neither primitive nor a power in $R_{3,1}$. □

We now take parallel copies ω_1^+, ω_1^- of ω_1 and $\omega_2'^+, \omega_2'^-$ of ω_2' in $\partial R_{3,1}$ as shown in Figure 13, bottom, and let P be the 4-punctured 2-sphere in $\partial R_{3,1}$ cobounded by the 4 circles ω_1^+, ω_1^-, $\omega_2'^+$ and $\omega_2'^-$.

For each pair of values of $q_3, q_5 = \pm 1$ let $\Gamma \subset P$ denote the graph $P \cap (\partial D_3 \sqcup \partial D_5)$ and $\overline{\Gamma} \subset P$ the reduced graph obtained by amalgamating each collection of parallel edges of Γ into a single edge. By minimality of $|(x \sqcup y) \cap \partial D_3|$ and $|(x \sqcup y) \cap \partial D_5|$, no edge of Γ or $\overline{\Gamma}$ is parallel into ∂P, that is, the graphs Γ and $\overline{\Gamma}$ are essential.

Lemma 6.3.3 *If a circle $c \subset P$ intersects the reduced graph $\overline{\Gamma}$ minimally then the word $c(x, y)$ is cyclically reduced. In particular, any circle in P which is primitive in $R_{3,1}$ is isotopic to the circle $\alpha \subset P$ in Figure 13, bottom.*

Proof We consider the case $q_3 = +1$ and $q_5 = -1$; the other cases being similar. Figure 15, top, shows the graph $\Gamma \subset P$ where the thicker lines represent 2 amalgamated parallel edges of one of the circles ∂D_3 or ∂D_5 (corresponding to the values $p_3 = 2 = p_5$), while the thinner lines represent single arcs.

The reduced graph $\overline{\Gamma} \subset P$ is shown in Figure 15, middle, where each amalgamated edge shows the common orientation of its components. The set of faces of $\overline{\Gamma}$ consists of the two 4-sided disk faces R_1 and R_2 in Figure 15, bottom, where each edge of R_i is labeled and oriented as the corresponding edge in the unreduced graph $\Gamma \subset P$.

Let $c \subset P$ be any circle which intersects $\overline{\Gamma} \subset P$ minimally. Then the sink/source pattern of the oriented edges around the faces R_1 and R_2 guarantee that the word $c(x, y)$ does not contain any of the canceling pairs xX, Xx, yY or Yy, and hence that it is cyclically reduced.

Notice that if c intersects any of the horizontal edges of $\overline{\Gamma}$ then the word $c(x, y)$ contains one of the strings x^2y^2, y^2x^2 or their inverses, and hence by Section 2.3.1 the word $c(x, y)$ cannot be primitive

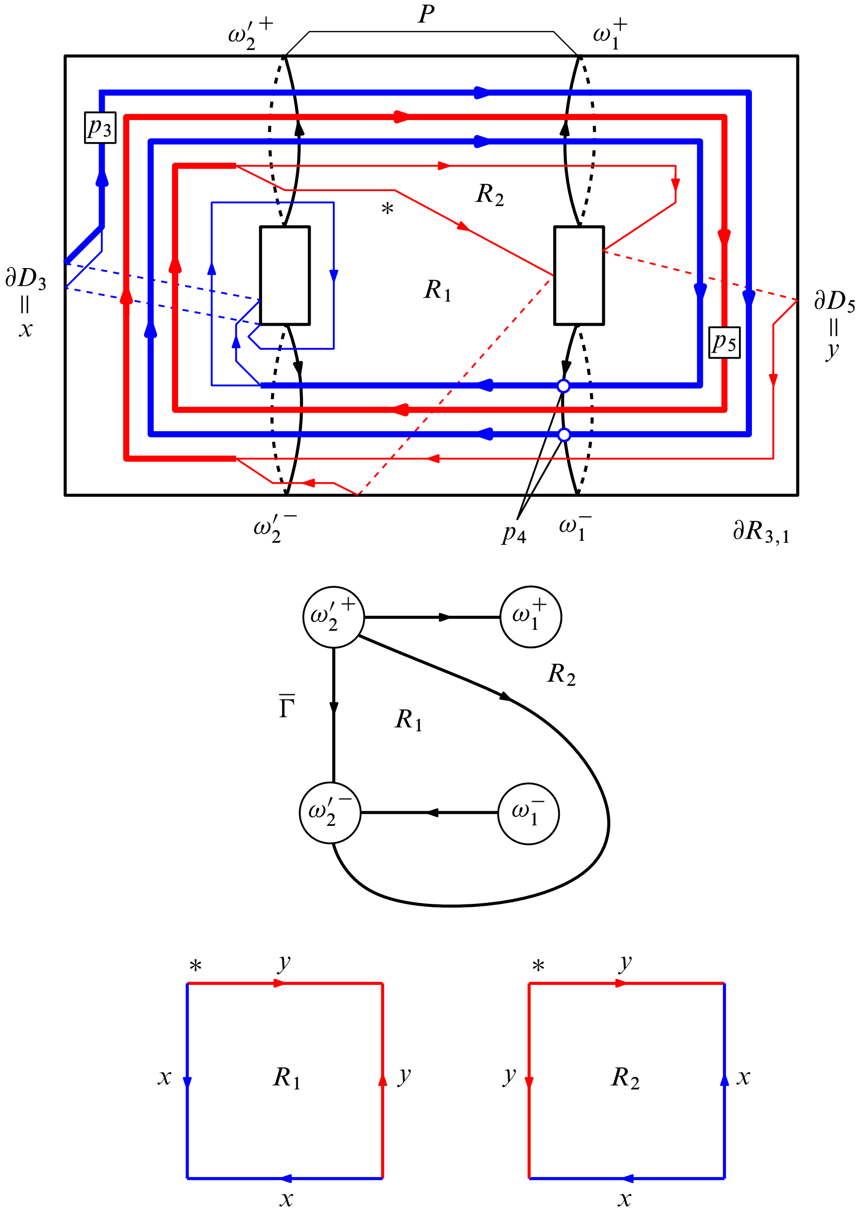


Figure 15: The circles $\partial D_3, \partial D_5 \subset \partial R_{3,1}$ for $q_3 = +1$ and $q_5 = -1$.

in the free group $\pi_1(R_{3,1}) = \langle x, y \mid - \rangle$. Therefore if $c \subset P$ is a primitive circle in $R_{3,1}$ then c can be isotoped in P so as to be disjoint from the horizontal edges of the graph $\overline{\Gamma} \subset P$. As the horizontal edges of $\overline{\Gamma}$ cut P into an annulus with core the circle $\alpha \subset P$, it follows that c must be isotopic to α in P, hence in $\partial R_{3,1}$. □

Proof of Proposition 6.0.1 By (E6) the circle $\partial D \subset \partial R_{3,1}$ is either a primitive or a Seifert circle in $R_{3,1}$. We consider two cases and arrive at a contradiction in each.

Case 1 $\partial D \subset \partial R_{3,1}$ is a primitive circle in $R_{3,1}$ ($K_1 \subset \mathbb{S}^3$ is a trivial knot).

As the circle $\partial D \subset \partial R_{3,1}$ is disjoint from the circles $\omega_1 \sqcup \omega_2' \subset \partial R_{3,1}$ it can be isotoped so as to lie in P and so by Lemma 6.3.3 it must be isotopic in P to the circle α in Figure 13, top. Since $|\partial D \cap K| = |\alpha \cap K| = 2$ and $D \subset R_{1,3}$ by Section 2.3.4 the pair $(R_{1,3}, K)$ is simple, hence minimal, which is not the case. Therefore this case does not occur.

Case 2 $\partial D \subset P$ is a Seifert circle in $R_{3,1}$ ($K_1 \subset \mathbb{S}^3$ is a nontrivial torus knot).

By [24, Lemma 6.7] there is a circle $h \subset \partial R_{3,1} \setminus \partial D$ which is a power circle in $R_{3,1}$.

By Lemma 6.3.3 isotopying ∂D in P so as intersect $\overline{\Gamma}$ minimally yields a cyclically reduced word $\partial D(x,y)$. Once ∂D has been isotoped, isotopying h in $\partial R_{3,1} \setminus \partial D$ so as to intersect $x \sqcup y = \partial D_3 \sqcup \partial D_5$ minimally produces the minimal intersection in $\partial R_{3,1}$ between h and $x \sqcup y$.

Now, by [24, Lemma 6.7] the circle $h \subset R_{3,1}(\partial D) = X_{K_1}$ is a fiber of the Seifert fiber space $X_{K_1} = \mathbb{D}^2(*,*)$. Since by (E3) $X_{K_1}(r_1) \approx R_{3,1}(\partial D)(\omega_1) \approx R_{3,1}(\partial D)(\omega_2')$ is either $\mathbb{S}^3$ or a lens space, it follows that $\Delta(h,\omega_1) = 1 = \Delta(h,\omega_2')$ in ∂X_{K_1} and hence that h intersects each circle ω_1 and ω_2' nontrivially in $\partial R_{3,1}$.

Therefore there is an arc component h' of $h \cap P$ with one endpoint in $\omega_1^+ \sqcup \omega_1^-$ and the other endpoint in $\omega_2'^+ \sqcup \omega_2'^-$.

If h' intersects transversely at least one of the horizontal edges in the reduced graph $\overline{\Gamma} \subset P$ then the word $h'(x,y)$, and hence $h(x,y)$, contains one of the strings x^2y^2, y^2x^2 or their inverses, contradicting Lemma 6.3.2 since h is a power circle in $R_{3,1}$. So if h' has endpoints on $\omega_1^+ \sqcup \omega_2'^+$ or $\omega_1^- \sqcup \omega_2'^-$ then h' can be isotoped so as to be parallel to one of the horizontal edges of $\overline{\Gamma} \subset P$, which implies that ∂D, being disjoint from h', is isotopic in P to the primitive circle $\alpha \subset P$, contradicting the hypothesis that ∂D is a Seifert circle in $R_{3,1}$.

Therefore the arc $h' \subset P$ must have endpoints on, say, ω_1^+ and $\omega_2'^-$. As ∂D and h' are disjoint, in the first integral homology group

$$H_1(R_{3,1}) = H_1(R_{3,1};\mathbb{Z}) = x\mathbb{Z} \oplus y\mathbb{Z},$$

the circle ∂D is the homological sum $\omega_1^+ +_{h'} \omega_2'^-$ of ω_1^+ and $\omega_2'^-$ along the arc $h' \subset P$. Using the orientations for ω_1^+ and $\omega_2'^-$ in Figure 13, top, and the relations $\omega_1^+(x,y) = (y^2x^2)^{p_4}y^{q_5}$ and $\omega_2'^-(x,y) = (x^2y^2)^{p_4}x^{q_3}$ (up to conjugation) found in (E5)(b) we obtain, in $H_1(R_{3,1})$,

$$\omega_1^+ = 2p_4x + (2p_4+q_5)y \quad \text{and} \quad \omega_2'^- = (2p_4+q_3)x + 2p_4y$$
$$\implies \partial D = \omega_1^+ + \omega_2'^- = (4p_4+q_3)x + (4p_4+q_5)y.$$

On the other hand, as $R_{3,1}(\partial D)$ is a knot exterior in $\mathbb{S}^3$ and hence a homology solid torus, the circle ∂D must be primitive in the abelian group $H_1(R_{3,1}) = x\mathbb{Z} \oplus y\mathbb{Z}$, so we must have

$$1 = \gcd(4p_4+q_3, 4p_4+q_5) = \gcd(4p_4+q_3, q_3-q_5) = \gcd(4p_4+q_3, 1-q_3q_5).$$

As $q_3, q_5 \in \{\pm 1\}$ and $p_4 \geq 2$ this implies that

$$q_3 q_5 = -1,$$

and hence that

$$\begin{aligned} H_1(X_{K_1}(r_1)) &= H_1\big(R_{3,1}(\omega_1^+ \sqcup \omega'^+_2)\big) \\ &= x\mathbb{Z} \oplus y\mathbb{Z}/\langle 2p_4 x + (2p_4 + q_5)y, (2p_4 + q_3)x + 2p_4 y\rangle \\ &= \{0\} \quad \text{since } \det\begin{bmatrix} 2p_4 & 2p_4 + q_5 \\ 2p_4 + q_3 & 2p_4 \end{bmatrix} = 1. \end{aligned}$$

Since by (E3) the manifold $X_{K_1}(r_1)$ is $\mathbb{S}^3$ or a lens space, it follows that $X_{K_1}(r_1) = \mathbb{S}^3$. But then, as r_1 is a nonintegral slope of the form a_1/p_1, $p_1 \geq 2$, by [3] K_1 is a trivial knot, contradicting the fact that K_1 is a nontrivial torus knot.

Therefore Case 2 does not occur and so the simplicial collection $\mathbb{T} \subset X_K$ does not produce any exchange regions. By Lemma 4.2.1(4) any Seifert torus in X_K is then isotopic to some component of $\mathbb{T}$ (see the proof of Proposition 7.0.2 for more details); hence the collection $\mathbb{T} \subset X_K$ is unique up to isotopy. □

7 Simplicial collections $\mathbb{T} \subset X_K$ with minimal pairs $(R_{i,i+1}, J)$

In this section we show that for a genus one hyperbolic knot $K \subset \mathbb{S}^3$ there are, up to isotopy, at most two maximal simplicial collections of Seifert tori in X_K. This restricted number of such collections is the result of the interplay between the restrictions on the complementary regions of a maximal simplicial collection $\mathbb{T} \subset X_K$ in Lemma 4.1.1, the small size of an annular pair $(R_{i,j}, J)$ of index ≥ 2 found in Proposition 5.2.3, and the bound $|\mathbb{T}| \leq 5$.

Lemma 7.0.1 *Any simplicial collection $\mathbb{T}$ of Seifert tori in X_K with minimal pairs $(R_{i,i+1}, J)$ has at most one exchange region, and if so then $2 \leq |\mathbb{T}| \leq 4$.*

Proof Set $|\mathbb{T}| = N$ where $1 \leq N \leq 5$ by [24]. Clearly there are no exchange regions when $N = 1$. If $N = 2$ then by Lemma 4.1.1(4) only one of $R_{1,1}$ or $R_{2,2}$ can be an exchange region, while in the case $N = 5$ there are no exchange regions by Proposition 6.0.1.

Therefore we may assume that $N = 3, 4$ in which case any two exchange regions must intersect. Arguing by contradiction, we only need to consider the following two cases.

Case 1 $R_{1,3}$ and $R_{2,4}$ are exchange regions (with $T_1 = T_4$ allowed).

The situation is represented in Figure 16, left. By the exchange trick of Section 5.3 we may assume that the pair $(R_{2,3}, J)$ is annular of index 1 while the pairs $(R_{1,2}, J)$ and $(R_{3,4}, J)$ are simple of indices ≥ 2. So if A and B are spanning annuli for $R_{1,3}$ and $R_{2,4}$, respectively, then by Lemma 4.1.1(6) we may assume that $A \cap R_{2,3}$ and $B \cap R_{2,3}$ are spanning annuli of index 1 in $R_{2,3}$, hence isotopic by Lemma 3.1.1. Thus the circles $A \cap T_2$ and $B \cap T_2$ have the same slope on T_2, and similarly the circles $A \cap T_3$ and

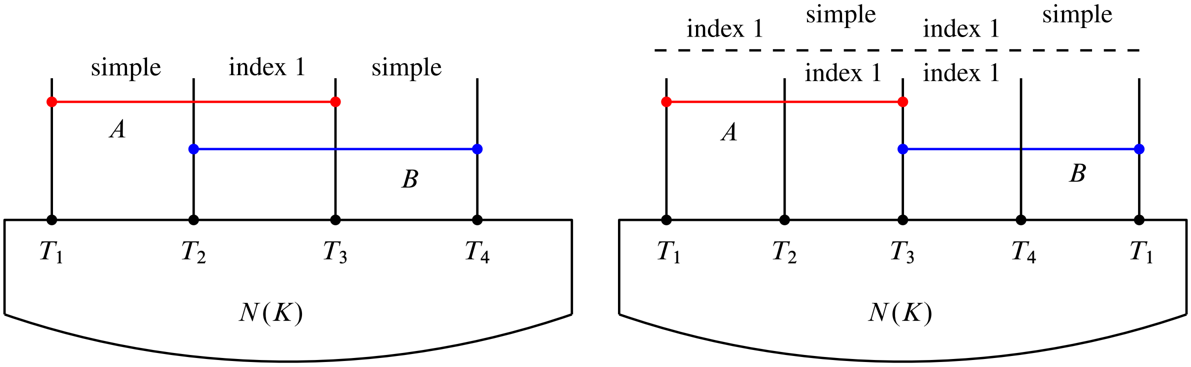


Figure 16: Intersecting exchange regions in X_K.

$B \cap T_3$ have the same slope on T_3. This implies that the boundary components of the spanning annulus $A \cap R_{2,3} \subset R_{2,3}$ have companion annuli in $R_{1,2}$ and $R_{3,4}$, contradicting Lemma 4.1.1(4).

Case 2 $N = 4$ and $R_{1,3}$ and $R_{3,1}$ are exchange regions.

Let A and B be spanning annuli for $R_{1,3}$ and $R_{3,1}$, respectively, and let $\Delta_3 = \Delta(A \cap T_3, B \cap T_3)$.

Suppose that $\Delta_3 = 0$. By the exchange trick and Lemma 4.1.1(6) we may assume that $A \cap R_{2,3}$ and $B \cap R_{3,4}$ are spanning annuli of index 1 in $R_{2,3}$ and $R_{3,4}$, respectively, as shown in Figure 16, right, below the dashed line.

Therefore the annulus $A \cap R_{2,3}$ can be isotoped in $R_{2,3}$ so that $A \cap R_{2,3} = B \cap R_{3,4}$, in which case their union becomes an index 1 spanning annulus for the region $R_{2,4}$, contradicting Proposition 5.2.3(1).

Therefore $\Delta_3 \neq 0$ and, by the exchange trick, this time we may assume that the pairs $(R_{1,2}, J)$ and $(R_{3,4}, J)$ are annular of index 1 while the pairs $(R_{2,3}, J)$ and $(R_{4,5}, J)$ are simple of index ≥ 2, as shown in Figure 16, right, above the dashed line.

If $R_{2,4}$ is a handlebody then $R_{3,4}$ is a handlebody by Lemma 2.1.1(3) and so, being of index 1, $(R_{3,4}, J)$ is a primitive pair with spanning annulus $B \cap R_{3,4}$. On the other hand $A \cap R_{2,3}$ is a spanning annulus for the simple pair $(R_{2,3}, J)$. As $\Delta_3 \neq 0$, the slopes of the spanning annuli $A \cap R_{2,3}$ and $B \cap R_{3,4}$ disagree on T_3, contradicting Lemma 4.1.1(7). Therefore $R_{2,4}$ is not a handlebody and hence the region $R_{4,2}$ is a handlebody by Lemma 2.1.1(P2).

However, as $N = 4$, by the argument above we also have that $\Delta_1 = \Delta(A \cap T_1, B \cap T_1) \neq 0$ and hence that the region $R_{4,2}$ is not a handlebody, a contradiction.

Therefore there cannot be two exchange regions for the collection $\mathbb{T}$ in X_K. □

Proposition 7.0.2 *Let K be a hyperbolic knot in $\mathbb{S}^3$.*

(1) *A simplicial collection $\mathbb{T} = \bigsqcup_i T_i \subset X_K$ of Seifert tori is maximal if and only if its complementary pairs $(R_{i,i+1}, J)$ are all minimal. In particular, any simplicial collection of Seifert tori in X_K can be extended to a maximal such collection by suitably adding J-tori to each nonminimal pair of the collection.*

(2) *Up to isotopy, there are at most two maximal simplicial collections of Seifert tori in X_K. Specifically, if $\mathbb{T} \subset X_K$ is a maximal such collection then either:*

   (a) *$\mathbb{T}$ has no exchange region and $\mathbb{T}$ is the unique maximal simplicial collection of Seifert tori in X_K; any Seifert torus in X_K is isotopic to some component of $\mathbb{T}$.*

   (b) *$2 \leq |\mathbb{T}| \leq 4$ and $\mathbb{T}$ has a unique exchange region $R_{i-1,i+1}$ with induced tori*

$$T_{i+1}, T'_{i+1} \subset R_{i-1,i+1},$$

   and $\mathbb{T}$ and $(\mathbb{T} \setminus T_i) \sqcup T'_{i+1}$ are the unique maximal simplicial collections of Seifert tori in X_K; any Seifert torus in X_K is isotopic to some component of $\mathbb{T}$ or to T'_{i+1}.

Proof Let $\mathbb{T} = T_1 \sqcup \cdots \sqcup T_N \subset X_K$ be a simplicial collection of Seifert tori such that each pair $(R_{j,j+1}, J)$ is minimal, and let $\mathbb{S} \subset X_K$ be any simplicial collection of Seifert tori.

Isotope $\mathbb{S}$ in X_K so as to intersect $\mathbb{T}$ minimally with $\partial\mathbb{S} \cap \partial\mathbb{T} = \varnothing$. By the argument in Lemma 4.2.1(1) it follows that each component of $\mathbb{S}$ is either disjoint from $\mathbb{T}$ and hence parallel to some component of $\mathbb{T}$, or intersects $\mathbb{T}$ minimally in X_K.

Suppose that $S \subset \mathbb{S}$ is a Seifert torus which is not isotopic to any component of $\mathbb{T}$. By Lemma 4.2.1 there is a component $T_j \subset \mathbb{T}$ such that

(i) $|S \cap T_j| = 2$;

(ii) the closures of the components of $S \setminus \mathbb{T}$ consist of a pants P and a companion annulus A with $P \cap T_j = P \cap \mathbb{T} = A \cap T_j$ that lie on opposite sides of T_j;

(iii) there is a Seifert torus $T \subset X_K \setminus (P \cup \mathbb{T})$ which is not parallel to T_j in X_K;

(iv) if $R \subset X_K$ is the region cobounded by T and T_j that contains P then the pair (R, J) is annular of index 1 with spanning annulus $A_R \subset R$ having the same boundary slope on T_j as A, and $R \subset R_{j-1,j}$ or $R \subset R_{j,j+1}$.

Since each pair $(R_{k,k+1}, J)$ is minimal, by (iii) we must have $N \geq 2$ and by (iv) and Proposition 5.2.3(1) we may assume that $R = R_{j-1,j}$, in which case by (ii) we have $A \subset R_{j,j-1}$.

If $A \cap T_{j+1} \neq \varnothing$ then some component of $A \cap R_{j,j+1}$ is a spanning annulus of $R_{j,j+1}$, of the same boundary slope on T_j as A_R by (iv), and some component of $A \cap R_{j+1,j-1}$ is a companion annulus. Therefore by Lemma 4.1.1(6) the pair $(R_{j,j+1}, J)$ is annular of index 1 and so the pair $(R_{j-1,j+1}, J)$ is also annular of index 1, contradicting Proposition 5.2.3(1).

It follows that the companion annulus A lies in $R_{j,j+1}$ and hence that the minimal pair $(R_{j,j+1}, J)$ is simple by Lemma 3.1.1(5)(b). Therefore the pair $(R_{j-1,j+1}, J)$ is annular of index ≥ 2 and hence $R_{j-1,j+1}$ is the unique exchange region of $\mathbb{T}$. Moreover, since $S \subset R_{j-1,j+1}$ and $T_j \subset R_{j-1,j+1}$ is the J-torus induced by T_{j+1}, by Proposition 5.2.3(2) S is isotopic in $R_{j-1,j+1}$ to the J-torus $T'_j \subset R_{j-1,j+1}$ induced by T_{j-1} and $2 \leq |\mathbb{T}| \leq 4$.

Therefore the collection $\mathbb{S}$ is isotopic to some subset of one of the collections $\mathbb{T}$ or $(\mathbb{T} \setminus T_j) \sqcup T_j'$, both of which have size $|\mathbb{T}|$, and so the collection $\mathbb{T}$ is maximal.

That a maximal simplicial collection produces minimal pairs follows by definition of maximality. Therefore (1) holds, and now (2) holds by the above argument. □

Proof of Theorem 1(1)–(2) Set $d = \dim \mathrm{MS}(K)$. That $0 \le d \le 4$ follows from the bound $|\mathbb{T}| \le 5$ given in [24] for any maximal simplicial collection of Seifert tori $\mathbb{T} \subset X_K$. Hence part (1) holds.

Each d-dimensional simplex of $\mathrm{MS}(K)$ corresponds to the isotopy class of some such maximal collection $\mathbb{T} \subset X_K$, and by Proposition 7.0.2(2) any two such maximal collections differ up to isotopy by at most one component. Therefore $\mathrm{MS}(K)$ consists of at most two d-simplices, and two d-dimensional simplices in $\mathrm{MS}(K)$ intersect in a common $(d-1)$-face. Hence part (2) holds. □

8 Examples of hyperbolic knots in $\mathbb{S}^3$

By Propositions 6.0.1 and 7.0.2, a maximal simplicial collection $\mathbb{T} \subset X_K$ of size $|\mathbb{T}| = 5$ produces no exchange regions and is therefore unique up to isotopy. In this section we construct examples of hyperbolic knots $K \subset \mathbb{S}^3$ with maximal simplicial collections of Seifert tori $\mathbb{T} \subset X_K$ of sizes $2 \le |\mathbb{T}| \le 4$ that produce exchange regions and hence $\mathrm{MS}(K)$ consists of two top-dimensional simplices.

One example of such a knot K with a collection $\mathbb{T} \subset X_K$ having an exchange region was constructed in [17, Section 6]. In that example it is proved that there are nonisotopic Seifert tori in X_K that intersect nontrivially and hence the diameter of $\mathrm{MS}(K)$ must be 2; thus the presence of an exchange region for $\mathbb{T}$ is inferred from Proposition 7.0.2(2). We follow a different strategy in the construction of examples along with the results obtained so far which allows us to determine both the size of their maximal simplicial collection of Seifert tori and the Kakimizu complex of the constructed knots.

8.1 Detecting primitive pairs and exchange pairs

In the case of handlebody pairs by Section 2.3.5(3) and (4) an exchange pair can be thought of as an extension of a primitive pair by a simple pair. Both simple and exchange pairs are annular pairs of index ≥ 2, and the next result will be useful in distinguishing these types of pairs form each other.

Lemma 8.1.1 *Let (H, J) be a handlebody pair with $\partial H = T_1 \cup_J T_2$, $\omega_1 \subset T_1$ and $\omega_2 \subset T_2$ coannular circles in H, and $\gamma \subset T_1$ a circle with $\Delta(\omega_1, \gamma) = 1$. Then the surface $\partial H \setminus (\omega_1 \sqcup \omega_2)$ compresses in H along a nonseparating disk $D \subset H$, unique up to isotopy, and the following hold:*

1. *ω_1 and ω_2 are both primitive in H or both p-power circles for some $p \ge 2$;*
2. *$\gamma \cdot \partial D = \pm 1$;*
3. *$|\gamma \cap \partial D|_{\min} = 1$ if and only if (H, J) is a trivial or simple pair;*
4. *if $|\gamma \cap \partial D|_{\min} > 1$ then (H, J) is a primitive or an exchange pair if ω_1 is a primitive or a power circle in H, respectively.*

Proof That the surface $\partial H \setminus (\omega_1 \sqcup \omega_2)$ compresses in H along a nonseparating disk $D \subset H$ which unique up to isotopy and part (1) follow from [24, Lemma 3.4].

Isotope ∂D in $\partial H \setminus (\omega_1 \sqcup \omega_2)$ so as to intersect γ minimally. As the circles $\omega_1 \sqcup \omega_2 \sqcup \partial D$ separate ∂H into two pants the circle ∂D is homologous in ∂H to $\omega_1 \sqcup \omega_2$, and since $\partial D \cap \omega_2 = \varnothing$ we must have (up to some orientation scheme)

$$\gamma \cdot \partial D = \gamma \cdot \omega_1 + \gamma \cdot \omega_2 = \gamma \cdot \omega_1 = \pm 1,$$

so (2) holds.

Since $|\omega_1 \cap \gamma|_{\min} = 1$ and $\omega_1 \cup \gamma \subset T_1$, it follows that the circles J and $\partial N(\omega_1 \cup \gamma) \subset T_1$ are parallel in T_1 and hence that

$$|J \cap \partial D|_{\min} = 2 \cdot |\gamma \cap \partial D|_{\min}.$$

By Section 2.3.4 the pair (H, J) is trivial or simple if and only if the disk D intersects J minimally in two points; hence if and only if D intersects γ minimally in one point, so (3) holds.

For the case $|\gamma \cap \partial D|_{\min} > 1$ by (3) the pair (H, J) is nontrivial and not simple, so if ω_1 is a primitive circle in H then (H, J) is a primitive pair. Otherwise by (1) ω_1 is a $p \geq 2$ power circle in H and so (H, J) is an exchange pair by Lemma 5.1.1(1) and Remark 5.1.2(1); hence (4) holds. □

In the construction of examples of knots $K \subset \mathbb{S}^3$ we will make use of Lemma 2.2.1 to justify that the regions $R_{i,j}$ involved form pairs $(R_{i,j}, J)$ or $(R_{i,j}, K)$ before knowing that the knot $K \subset \mathbb{S}^3$ is hyperbolic. As this is automatically the case whenever the region $R_{i,j}$ is a handlebody, we will only invoke Lemma 2.2.1 when $R_{i,j}$ is not a handlebody.

8.2 Hyperbolic knots with $|\mathbb{T}| = 4$ and one exchange region

Suppose that $\mathbb{T} \subset X_K$ is a maximal simplicial collection of size $|\mathbb{T}| = 4$ such that $R_{1,3}$ is an exchange region with $(R_{1,2}, J)$ an index one annular pair. By Proposition 5.2.3(1) the region $R_{1,2}$ must then be a handlebody, in which case the pair $(R_{1,2}, J)$ is primitive and $R_{1,3}$ is a handlebody by Section 2.3.5(3). Arguments similar to those in Sections 6.1 and 6.2 can be used to prove that the region $R_{3,1}$ must also be a handlebody and at least one of the pairs $(R_{3,4}, J)$ or $(R_{4,1}, J)$ be simple.

In this section we construct a family of knots $K = K(q, k, \varepsilon) \subset \mathbb{S}^3$ for integers $q \geq 1$, $k \in \mathbb{Z}$, and $\varepsilon = \pm 1$, each of which bounds a maximal simplicial collection of 4 Seifert tori $\mathbb{T} = T_1 \sqcup T_2 \sqcup T_3 \sqcup T_4$ with an exchange region $R_{1,3}$ such that the regions $R_{1,3}$ and $R_{3,1}$ are handlebodies and both pairs $(R_{3,4}, J)$ and $(R_{4,1}, J)$ are simple. The collection $\mathbb{T} \subset X_K$ is represented in Figure 17.

(I) Construction of the circles $\omega_1, \omega_4' \subset T_1$ and $\omega_2', \omega_3 \subset T_3$ relative to the handlebody $R_{3,1}$.

Figure 18 shows the genus two handlebody $R_{3,1}$ with complete meridian system given by the disks $x \sqcup y \subset R_{3,1}$ and a disk $E \subset R_{3,1}$ separating x and y. We identify $\pi_1(R_{3,1})$ with the free group

$$\pi_1(R_{3,1}) = \langle x, y \mid - \rangle$$

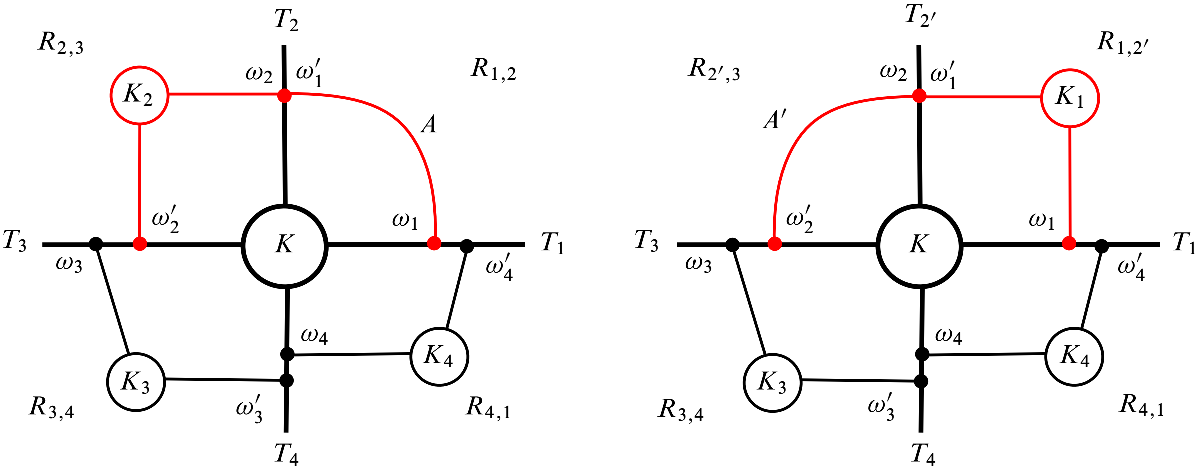


Figure 17: The knot $K = K(q,k,\varepsilon) \subset \mathbb{S}^3$ with exchange pair $(R_{1,3}, J)$.

relative to some base point. For $w_1, w_2 \in \langle x, y \mid - \rangle$ we write $w_1 \equiv w_2$ to indicate that the words differ by a cyclic permutation.

The circles ω_1, ω_4' and ω_2' in $\partial R_{3,1}$ are constructed as indicated in Figure 18, along with an extra circle u. Notice that to the left of the separating disk E all arcs in the figure are mutually parallel and intersect the disk x minimally in one point. To the right of E there are 3 disjoint arcs which intersect y minimally in q, $q+\varepsilon$ and $2q+\varepsilon$ points, as well as the circle ω_4' which intersects y minimally in $2q+\varepsilon$ points (since it is disjoint from the arc that intersects y in $2q+\varepsilon$ points). Figure 18 corresponds to the case $(q,\varepsilon) = (1,-1)$.

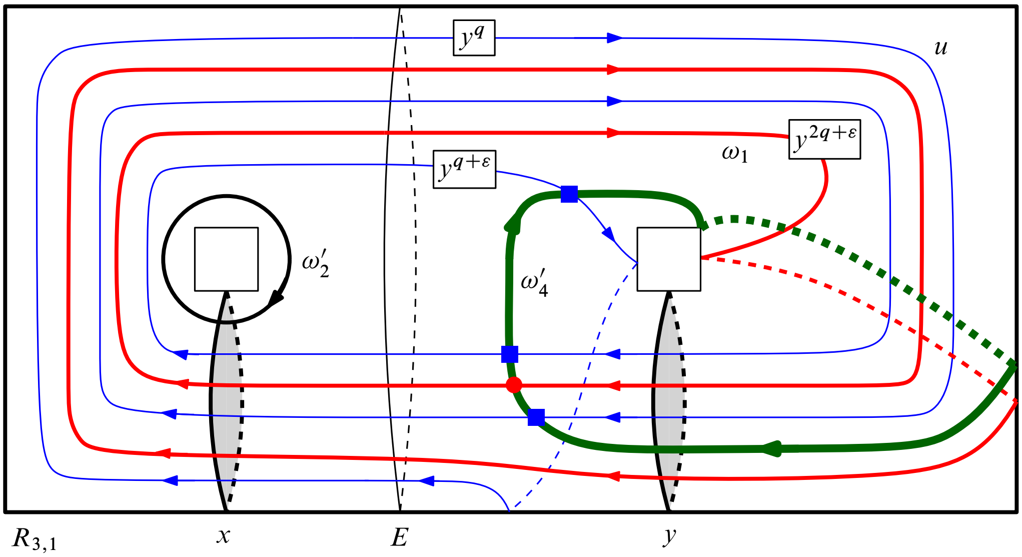


Figure 18: The circles ω_1, ω_2', ω_4' and u in $\partial R_{3,1}$ for $(q,\varepsilon) = (1,-1)$.

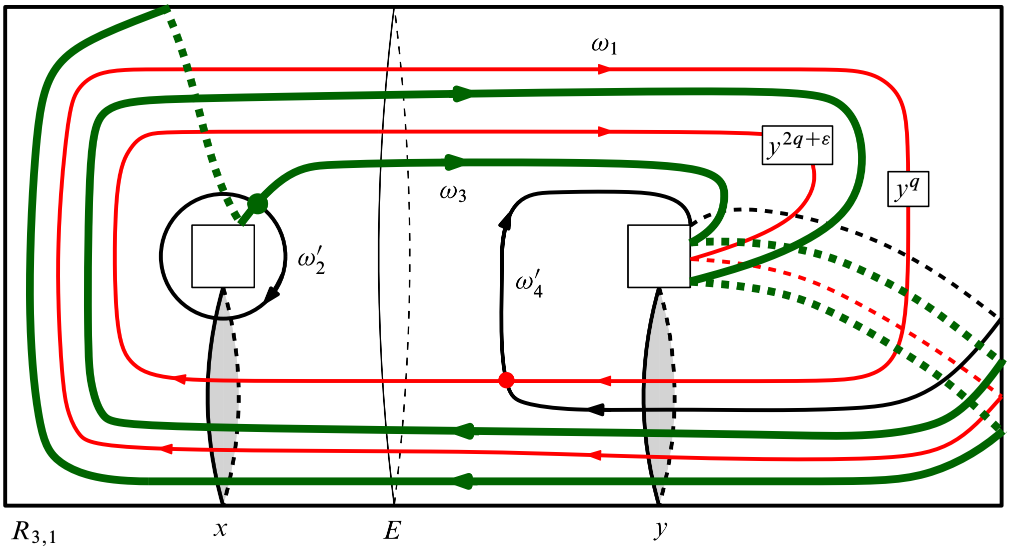


Figure 19: The circles ω_1, ω_2', ω_3 and ω_4' in $\partial R_{3,1}$.

The circle $\omega_3 \subset T_2$ is constructed in Figure 19. With their given orientations these circles satisfy the relations

$$(\omega_1 \cup \omega_4') \cap (\omega_2' \cup \omega_3) = \varnothing, \quad \Delta(\omega_1, \omega_4') = 1 = \Delta(\omega_2', \omega_3),$$

$$\omega_1(x,y) \equiv xy^q xy^{2q+\varepsilon}, \quad \omega_4'(x,y) \equiv y^{2q+\varepsilon}, \quad \omega_2'(x,y) = x, \quad \omega_3(x,y) \equiv (xy^{2q+\varepsilon})^2.$$

Therefore we define

$$K = \partial N(\omega_1 \sqcup \omega_4') \subset \partial R_{3,1}.$$

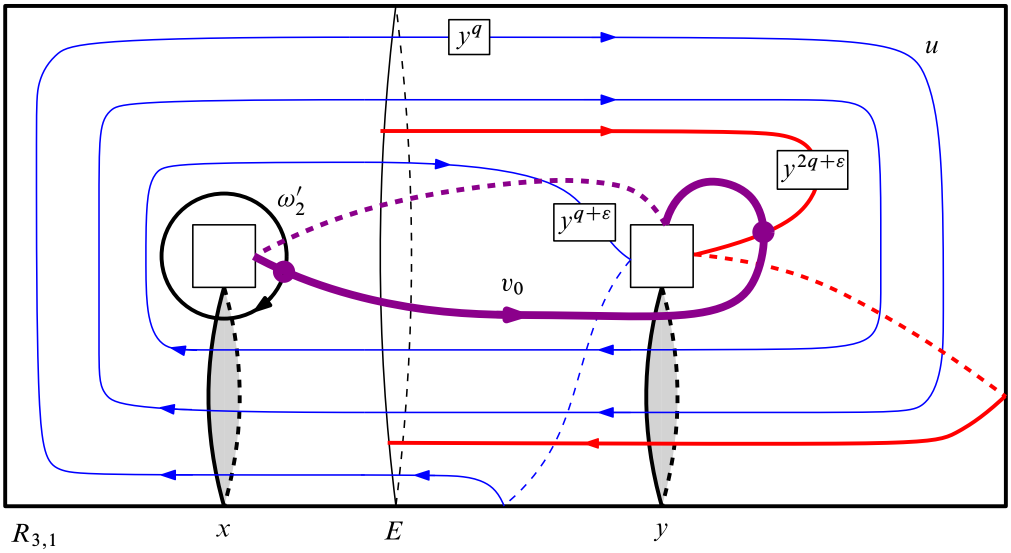


Figure 20: The circles ω_2', u and v_0 in $\partial R_{3,1}$.

Notice that for $(q,\varepsilon)=(1,-1)$ we have $\omega_1(x,y)\equiv(xy)^2\equiv\omega_3(x,y)$, and in fact from Figure 19 it follows directly that in this case the power circles ω_1 and ω_3' are coannular in $R_{3,1}$.

(II) Construction of the handlebody $R_{1,3}$.

We construct two disjoint and nonseparating circles $u,v\subset\partial R_{3,1}=\partial R_{1,3}$ representing the boundary of the complete system of disks for $R_{1,3}$.

The circle $u=\partial D\subset\partial R_{3,1}=\partial R_{1,3}$ is given in Figure 18 and satisfies the relations

$$u\cap(\omega_1\sqcup\omega_2')=\varnothing\quad\text{and}\quad u(x,y)=(xy^q)^3y^{\varepsilon}=\text{primitive in } R_{3,1}.$$

To obtain the circle $v\subset\partial R_{1,3}\setminus u$ we first construct the auxiliary circle $v_0\subset\partial R_{1,3}$ in Figure 20 such that

$$v_0\cap u=\varnothing,\quad |v_0\cap\omega_2'|=1,\quad v_0(x,y)=y^{\varepsilon}\quad\text{in } \pi_1(R_{3,1}).$$

The circle $v\subset\partial R_{1,3}\setminus u$ is then constructed in $\partial R_{1,3}$ as the homological sum

$$v=(1+3k)\cdot\omega_2'+[q+k(3q+\varepsilon)]\varepsilon\cdot v_0$$

where $k\in\mathbb{Z}$.

(III) The Heegaard decomposition $R_{1,3}\cup_\partial R_{3,1}\approx\mathbb{S}^3$.

In the first integral homology group $H_1(R_{3,1})=H_1(R_{3,1};\mathbb{Z})=x\mathbb{Z}\oplus y\mathbb{Z}$ we have

$$u=3x+(3q+\varepsilon)y\quad\text{and}\quad v=(1+3k)x+[q+k(3q+\varepsilon)]y$$

where

$$\det\begin{bmatrix}3 & 3q+\varepsilon\\ 1+3k & q+k(3q+\varepsilon)\end{bmatrix}=-\varepsilon=\pm1.$$

Therefore,

$$H_1(R_{3,1}(u\sqcup v))=x\mathbb{Z}\oplus y\mathbb{Z}/\langle 3x+(3q+\varepsilon)y,\ (1+3k)x+[q+k(3q+\varepsilon)]y\rangle=0,$$

and since

$$R_{1,3}\cup_\partial R_{3,1}\approx R_{3,1}(u\sqcup v)=R_{3,1}(u)(v)$$

and $R_{3,1}(u)$ is a solid torus it follows that $R_{1,3}\cup_\partial R_{3,1}\approx\mathbb{S}^3$.

(IV) The exchange region $R_{1,3}$.

By construction, in $\pi_1(R_{1,3})=\langle u,v\mid -\rangle$ we have $\omega_2'\equiv v^p$ for $p=|q+k(3q+\varepsilon)|$.

Since $u\cap(\omega_1\sqcup\omega_2')=\varnothing$ by (II), the nonseparating disk $u\subset R_{1,3}$ is a compression disk for $\partial R_{1,3}\setminus(\omega_1\sqcup\omega_2')$ and so the circles ω_1 and ω_2' are coannular in $R_{1,3}$ by [24, Lemma 3.4(2)]. From Figure 18 we can see that $|u\cap\omega_4'|_{\min}=3$ in $\partial R_{1,3}$ and so, by Lemma 8.1.1, $(R_{1,3},J)$ is an exchange pair if and only if $p=|q+k(3q+\varepsilon)|\geq2$.

We summarize the information above in the next result.

Proposition 8.2.1 *For integers $q \geq 1$, k, and $\varepsilon = \pm 1$, except for $(q,k)=(1,0)$ and $(q,\varepsilon)=(1,-1)$, each of the knots K in the family $K(q,k,\varepsilon) \subset \mathbb{S}^3$ is hyperbolic and bounds a maximal simplicial collection $\mathbb{T} = T_1 \sqcup T_2 \sqcup T_3 \sqcup T_4 \subset X_K$ of 4 Seifert tori such that*

(1) *the regions $R_{1,3}$ and $R_{3,1}$ are handlebodies;*

(2) *$(R_{1,3}, J)$ is an exchange pair of index $p = |q + k(3q+\varepsilon)| \geq 2$;*

(3) *$(R_{3,4}, J)$ is a simple pair of index 2 and $(R_{4,1}, J)$ is a simple pair of index $2q+\varepsilon \geq 3$;*

(4) *$\Delta(\omega_2', \omega_3) = \Delta(\omega_3', \omega_4) = \Delta(\omega_4', \omega_1) = 1$;*

(5) *the Kakimizu complex $\mathrm{MS}(K)$ is a union of two 3-simplices intersecting in a common 2-face, and each surgery manifold $X_K(r)$ is hyperbolic whenever $\Delta(r, J) \geq 2$.*

In the two exceptional cases $(q,k,\varepsilon) = (1,0,-1), (1,-1,-1)$ the knot K is hyperbolic and bounds a maximal collection of 2 Seifert tori $T_1 \sqcup T_3$ with $(R_{3,3}, J)$ an exchange pair (where $R_{3,3} = \mathrm{cl}[X_K \setminus T_3 \times I]$), and $\mathrm{MS}(K)$ the union of two 1-simplices along a common vertex (Figure 21, top left).

In the exceptional case $(q,k,\varepsilon) = (1,0,1)$ the knot K is hyperbolic and bounds a unique maximal collection of 3 Seifert tori $T_1 \sqcup T_3 \sqcup T_4$ with no exchange region (Figure 21, bottom).

In the remaining exceptional cases $(q,k,\varepsilon) = (1,k,-1)$ with $k \neq 0,-1$ the knot K is a satellite of a $(2,2k+1)$ torus knot (Figure 21, top right).

Proof Observe that

$$q \geq 1 \implies 0 < \frac{q}{3q+\varepsilon} < 1 \quad \text{and} \quad q \geq 2 \implies 0 < \frac{q \pm 1}{3q+\varepsilon} \leq \frac{q+1}{3q-1} < 1$$

and so, for $q \geq 1$,

$$\begin{aligned} |q+k(3q+\varepsilon)| \leq 1 \iff |q+k(3q+\varepsilon)| = 1 \iff -k = \frac{q\pm 1}{3q+\varepsilon} \in \mathbb{Z} \\ \iff q = 1 \text{ and } \begin{cases} k = 0 \text{ or} \\ k = -1 \text{ and } q+\varepsilon = 0. \end{cases} \end{aligned}$$

Therefore for integers $q \geq 1$, k, and $\varepsilon = \pm 1$ with $(q,k) \neq (1,0)$ and $(q,\varepsilon) \neq (1,-1)$ we have

$$p = |q+k(3q+\varepsilon)| \geq 2 \quad \text{and} \quad 2q+\varepsilon \geq 3,$$

and so by (IV) $(R_{1,3}, J)$ is an exchange pair. Moreover the circles $\omega_1 \subset T_1$ and $\omega_2' \subset T_3$ are coannular in $R_{1,3}$ and $\omega_1 \equiv v^p$ in $\pi_1(R_{1,3})$, therefore the index of each core knot $K_1, K_2 \subset R_{1,3}$ in Figure 17 is $p = |q+k(3q+\varepsilon)| \geq 2$.

By (I) $\omega_3(x,y) \equiv (xy^{2q+\varepsilon})^2$ and $\omega_4'(x,y) \equiv y^{2q+\varepsilon}$ in $\pi_1(R_{3,1})$ and so $\omega_3 \subset T_3$ and $\omega_4' \subset T_4$ are noncoannular power circles in $R_{3,1}$. By the argument in the proof of [24, Lemma 3.4(3)] it follows that there is a properly embedded disk $E \subset R_{3,1}$ that separates ω_3 and ω_4'. So if A_3 and A_4 are companion annuli in $R_{3,1}$ for the circles ω_3 and ω_4', respectively, then A_3 and A_4 can be isotoped to be disjoint from E and hence from each other. Therefore the J-tori T_3' and T_4 induced by ω_3 and ω_4' in $R_{3,1}$, respectively, are disjoint in $R_{3,1}$.

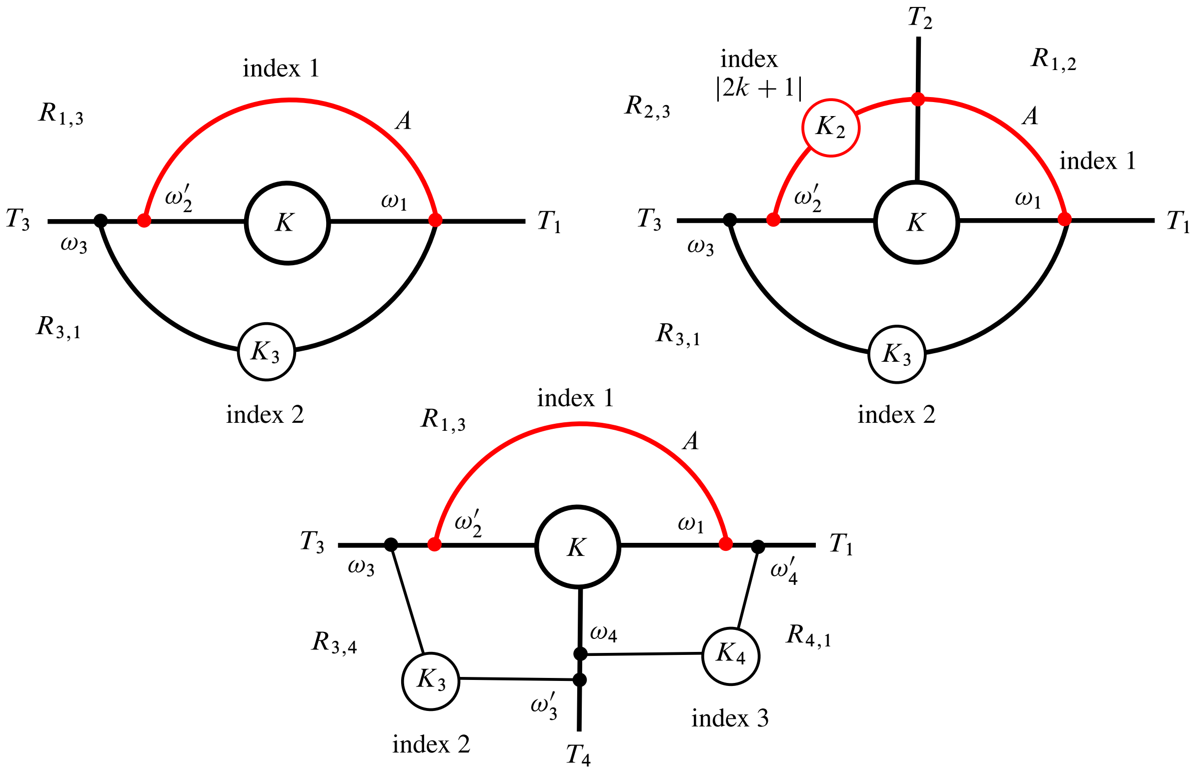


Figure 21: The Seifert tori bounded by the knot $K(q,k,\varepsilon)$ in the exceptional cases.

As we also have $\omega_2'(x,y)=x$ in $\pi_1(R_{3,1})$, the circle $\omega_2'\subset T_3$ is primitive in $R_{3,1}$ and so the pair $(R_{3,1},J)$ is not maximal by Section 2.3.9. This implies that the J-tori T_3' and T_4 are mutually parallel in $R_{3,1}$, and since by construction they cobound simple pairs with T_3 and T_1, respectively, by [24, Lemma 6.8(2)] $(R_{3,1},J)$ is a double pair with $T_4\subset R_{3,1}$ the unique J-torus not parallel into T_3 or T_1 and $\Delta(\omega_3',\omega_4)=1$ in T_4.

It now follows that the indices of the core knots K_3 and K_4 of the simple pairs $(R_{3,4},J)$ and $(R_{4,1},J)$ are 2 and $2q+\varepsilon\geq 3$, respectively.

Finally, in the case of T_1, the circle $\omega_1\subset T_1$ has a companion annulus in $R_{1,3}$ since it is the boundary of a spanning annulus in $R_{1,3}$ of index $p\geq 2$, while $\omega_4'\subset T_1$ has a companion annulus in $R_{3,1}$ since it is a power circle in $R_{3,1}$. As $\Delta(\omega_1,\omega_4')=1$ and by construction each region $R_{1,3}$ and $R_{3,1}$ is a handlebody, hence atoroidal, it follows from Lemma 2.0.1 that no circle on T_1 has a companion annulus on either side of T_1. A similar conclusion holds for the Seifert torus $T_3\subset X_K$ using the circles $\omega_2'\sqcup\omega_3$.

Therefore, by [24, Lemma 8.1] applied to the simplicial collection $T_1\sqcup T_3\subset X_K$, the knot

$$K=K(q,k,\varepsilon)\subset\mathbb{S}^3$$

is hyperbolic and each surgery manifold $X_K(r)$ is hyperbolic for any slope $r\subset\partial X_K$ with $\Delta(r,J)\geq 2$.

Since each of the pairs $(R_{i,i+1}, J)$ is minimal and $R_{1,3}$ is an exchange region, by Proposition 7.0.2 the collection $\mathbb{T} = T_1 \sqcup T_2 \sqcup T_3 \sqcup T_4$ is maximal and $\mathrm{MS}(K)$ consists of two 3 simplices intersecting in a common 2-face. Therefore (1)–(5) hold.

In the cases $(q, \varepsilon) = (1, -1)$ we have $2q + \varepsilon = 1$ and by definition the simple pair $(R_{4,1}, J)$ degenerates into a trivial pair. Moreover, by (I) the 2-power circles $\omega_3 \subset T_3$ and $\omega_1 \subset T_1$ become coannular in $R_{3,1}$, so $(R_{3,1}, J)$ becomes a simple pair of index 2.

If $k = 0, -1$ then $p = 1$ and so $(R_{2,3}, J)$ becomes a trivial pair, whence $(R_{1,3}, J)$ becomes a primitive pair with primitive circles ω_1 and ω_2'. It follows that the region $R_{3,3}$ is an exchange region for the collection $T_1 \sqcup T_3$. Since $\Delta(\omega_2', \omega_3) = 1$, by the above general argument the knot K is hyperbolic and bounds the maximal simplicial collection $T_1 \sqcup T_3$ with exchange region $R_{3,3}$; hence $\mathrm{MS}(K)$ is the union of two 1-simplices along a vertex (see Figure 21, top right).

If $k \neq 0, -1$ then $p = |2k + 1| \geq 3$ and so the spanning annulus A in $R_{1,2}$ has companion annuli on either boundary circle, so K is not a hyperbolic knot by Lemma 4.1.1(4); more precisely, by [24, Lemma 5.1] the knot K is a satellite of a $(2, 2k + 1)$ torus knot (see Figure 21, top left).

In the last exceptional case $(q, k, \varepsilon) = (1, 0, 1)$ we have $p = 1$ and $2q + \varepsilon = 3$. By a similar argument it follows that K is a hyperbolic knot that bounds the unique maximal simplicial collection $T_1 \sqcup T_3 \sqcup T_4$ of 3 Seifert tori with no exchange region (see Figure 21, bottom) and so $\mathrm{MS}(K)$ is a single 2-simplex. □

8.3 Hyperbolic knots with $|\mathbb{T}| = 2, 3$ and one exchange handlebody region

We construct a family of hyperbolic knots $K = K(-1, n, 2) \subset \mathbb{S}^3$, $n \in \mathbb{Z}$, each of which bounds a maximal simplicial collection $\mathbb{T} \subset X_K$ of $|\mathbb{T}| = 2, 3$ Seifert tori with a handlebody exchange region $R_{1,3}$. A projection of the knot $K(-1, n, 2)$ with at most $14 + 6|n|$ crossings is shown in Figure 2.

Unlike the case $|\mathbb{T}| = 4$ of Section 8.2, for $|\mathbb{T}| = 2, 3$ the exchange region $R_{1,3}$ need not be a handlebody; examples where $R_{1,3}$ is not a handlebody will be constructed in Section 8.5.

(1) Construction of the primitive pair $(R_{1,2}, J)$. Figure 22 shows a genus two handlebody $R_{1,2}$ standardly embedded in $\mathbb{S}^3$ with the following features:

(I) The disks $x \sqcup y \subset R_{1,2}$ form a complete disk system.

(II) The complementary handlebody $R_{2,1} = \mathbb{S}^3 \setminus \operatorname{int} R_{1,2}$ has complete disk system $a \sqcup b \subset R_{2,1}$.

(III) The circles $\omega_1, \omega_1' \subset \partial R_{1,2}$ are disjoint from each other and from the disk $x \subset R_{1,2}$, and each intersects the disk y minimally in one point. Thus ω_1 and ω_1' cobound an annulus $A \subset R_{1,2} \setminus x$, and by [24, Lemma 3.4] x is the unique compression disk of the surface $\partial R_{1,2} \setminus (\omega_1 \sqcup \omega_1')$.

(IV) The circles ω_1' and γ_0 intersect minimally in one point labeled $*$ in the figure.

(V) γ_0 and ∂x intersect minimally in 3 points.

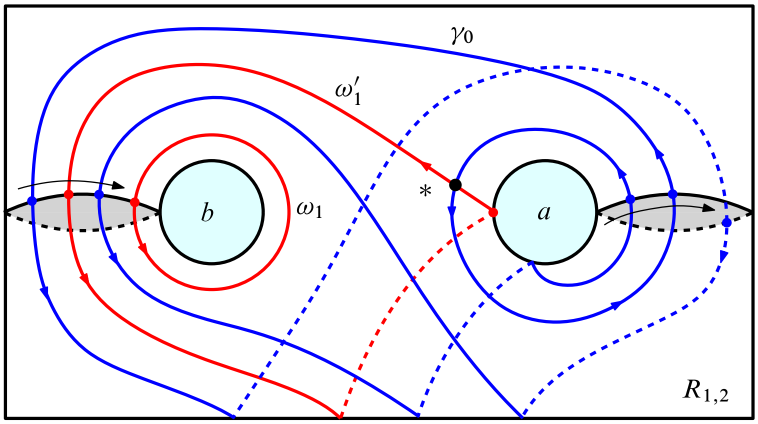


Figure 22: The complete disk systems $x \sqcup y \subset R_{1,2}$ and $a \sqcup b \subset R_{2,1}$, and the circles $\omega_1, \omega_1', \gamma_0 \subset \partial R_{1,2}$.

For any circle $\delta \subset \partial R_{1,2}$ and integers $k, n \in \mathbb{Z}$ we denote by $\delta(k,n) \subset \partial R_{1,2}$ the circle obtained by performing k full Dehn twists on δ around ∂x and n full Dehn twists around ∂y, where the Dehn twists are performed by cutting $\partial R_{1,2}$ along the circles $\partial x \sqcup \partial y$ and twisting on the side of these circles in the direction indicated by the arrows for positive twists.

For integers $k, n, p \in \mathbb{Z}$ we construct the following circles in $\partial R_{1,2}$:

(VI) The homological sum $\gamma_p = \gamma_0 + p\omega_1' \subset \partial R_{1,2}$, constructed so that it intersects ω_1' minimally in one point.

(VII) The circles $\omega_1(k,n)$, $\omega_1'(k,n)$ and $\gamma_p(k,n) \subset \partial R_{1,2}$.
By (4) the circles $\omega_1'(k,n)$ and $\gamma_p(k,n)$ intersect minimally in one point, and $\gamma_p(k,n))$ and ∂x intersect minimally in 3 points by (5).

(VIII) The separating circle $J = J(k,n,p) = \partial N(\omega_1'(k,n) \cup \gamma_p(k,n)) \subset \partial R_{1,2}$.

(2) Fundamental groups.

The fundamental groups of $R_{1,2}$ and $R_{2,1}$ have the presentations

$$\pi_1(R_{1,2}) = \langle x, y \mid - \rangle \quad \text{and} \quad \pi_1(R_{2,1}) = \langle a, b \mid - \rangle$$

relative to the base point $* = \omega_1' \cap \gamma_0$ in Figure 22. Therefore in $\pi_1(R_{1,2})$ we have

$$\omega_1(k,n)(x,y) = \omega_1'(k,n)(x,y) = y \quad \text{and} \quad \gamma_p(k,n)(x,y) = xyXyxy^p,$$

while in $\pi_1(R_{2,1})$ we compute

$$\omega_1(k,n)(a,b) = b^n, \quad \omega_1'(k,n)(a,b) = b^n a, \quad \gamma_p(k,n)(a,b) = a^k b^n A^k b^n a^{k+1}(b^n a)^p$$

where $X = x^{-1}$ and $A = a^{-1}$ as usual. Thus $\omega_1'(k,n)$ is a primitive circle in $R_{2,1}$.

(3) The knot $K = K(k,n,p) \subset \mathbb{S}^3$.

The circle $J = J(k,n,p)$ separates $\partial R_{1,2}$ into two once-punctured tori $T_1, T_2 \subset \partial R_{1,2}$ with $\omega_1(k,n) \subset T_1$ and $\omega_1'(k,n) \cup \gamma_p(k,n) \subset T_2$. We let $K = K(k,n,p) \subset \mathbb{S}^3$ be the knot represented by $J(k,n,p)$ and consider $T_1, T_2 \subset X_K$ as Seifert tori for K.

By (I)(3), (I)(7) and Lemma 8.1.1 it follows that the pair $(R_{1,2}, J)$ is primitive.

Since by (II) the circle $\omega_1'(k,n) \subset T_2$ is primitive in $R_{2,1}$, the pair $(R_{2,1}, J)$ is not maximal by Section 2.3.9.

(4) The power circles $\omega_1(k,n)$ and $\gamma_p(k,n)$ in $R_{2,1}$.

By [1], $\gamma_p(k,n) \subset T_2$ is a power circle in $R_{2,1}$ if and only if the word $\gamma_p(k,n)(a,b)$ is a power of some primitive word $w(a,b)$ in $\pi_1(R_{2,1})$, where by Section 2.3.1 in the cyclic reduction of $w(a,b)$ all exponents of a (b) are 1 or all -1 while all the exponents of b (a, respectively) are of the form $\ell, \ell+1$ for some integer ℓ.

Now, the following words are powers in $\pi_1(R_{2,1})$:

$$\begin{aligned} \omega_1(k,n) &= b^n && \text{for } |n| \geq 2, \\ \gamma_2(-1,n) &\equiv (b^{2n}a)^2 && \text{for all } n, \\ \gamma_{-1}(0,n) &= b^n && \text{for } |n| \geq 2, \end{aligned}$$

and we claim that these are the only cases when both $\omega_1(k,n)$ and $\gamma_p(k,n)$ are power circles in $R_{2,1}$. Indeed, suppose that $|n| \geq 2$ so that $\omega_1(k,n)$ is a power circle.

If $k \neq 0, -1$ then the cyclic reduction of the word $\gamma_p(k,n)(a,b)$ contains both a and A factors and hence it is not a power. If $k = 0$ then $\gamma_p(0,n) = b^n(b^n a)^{p+1}$ is a power if and only if $p = -1$, while if $k = -1$ then

$$\gamma_p(-1,n)(a,b) = Ab^n ab^n (b^n a)^p = AB^n \cdot (b^{2n}a)^2 \cdot (b^n a)^{p-2} \cdot b^n a \equiv (b^{2n}a)^2 \cdot (b^n a)^{p-2}$$

is a power if and only if $p = 2$.

The next result summarizes the information above.

Proposition 8.3.1 *For $|n| \geq 2$,*

(1) *the knot $K = K(-1,n,2) \subset \mathbb{S}^3$ is hyperbolic and bounds a maximal collection of 3 Seifert tori $\mathbb{T} = T_1 \sqcup T_2 \sqcup T_3 \subset X_K$ with one exchange pair $(R_{3,2}, J)$ of index $|n|$ and a simple pair $(R_{2,3}, J)$ of index 2, and $\mathrm{MS}(K)$ is the union of two 2-simplices along a common 1-subsimplex (see Figure 2 and Figure 23, left);*

(2) *the knot $K = K(0,n,-1) \subset \mathbb{S}^3$ is hyperbolic and bounds a maximal collection of 2 Seifert tori $\mathbb{T} = T_1 \sqcup T_2 \subset X_K$ with one exchange pair $(R_{1,1}, J)$ of index $|n|$, and $\mathrm{MS}(K)$ is the union of two 1-simplices along a common vertex (see Figure 23, right).*

Proof We sketch the proof following the argument in Proposition 8.2.1 closely.

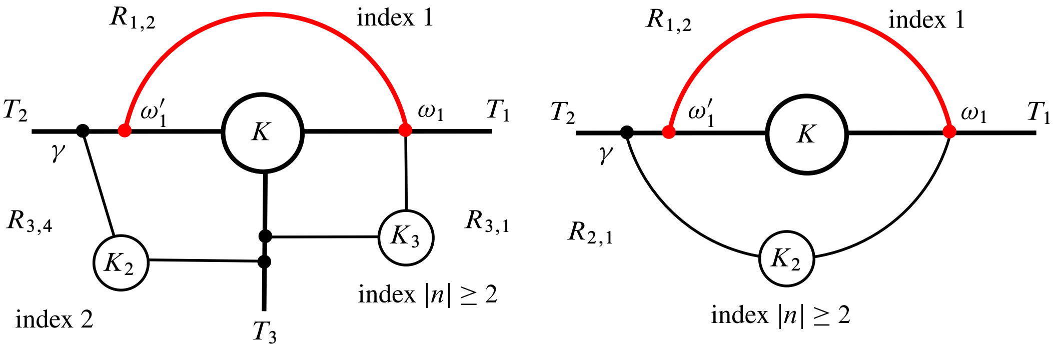


Figure 23: The Seifert tori bounded by the knots $K(-1,n,2)$ and $K(0,n,-1)$.

By (III), for all k, n and p the pair $(R_{1,2}, J)$ with $J = J(k,n,p)$ is primitive with primitive circles $\omega_1(k,n) \subset T_1$ and $\omega_1'(k,n) \subset T_2$. Let $|n| \geq 2$.

For the knot $K = K(-1,n,2) \subset \mathbb{S}^3$, by (IV) the circles $\omega_1(-1,n) \subset T_1$ and $\gamma_2(-1,n) \subset T_2$ are power circles in $R_{2,1}$ with words of the form b^n and $(b^{2n}a)^2$, respectively, and hence are not coannular in $R_{2,1}$. Since by (III) the pair $(R_{2,1}, J)$ is not maximal, either power circle $\omega_1(-1,n)$ or $\gamma_2(-1,n)$ induces a J-torus $T_3 \subset R_{2,1}$ which splits the pair $(R_{2,1}, J)$ into two simple pairs $(R_{2,3}, J)$ and $(R_{3,1}, J)$ of indices 2 and $|n|$, respectively. Therefore $R_{3,2} = R_{3,1} \cup_{T_1} R_{1,2}$ is an exchange region of index $|n|$.

For the knot $K = K(0,n,-1) \subset \mathbb{S}^3$, by (IV) we have $\omega_1(0,n) = b^n \equiv \gamma_{-1}(0,n)$ in $\pi_1(R_{2,1})$, and it can be seen directly from the corresponding diagram in Figure 22 that $\omega_1(0,n)$ and $\gamma_{-1}(0,n)$ are coannular $|n|$-power circles in $R_{2,1}$. It is also not hard to see that $a \subset R_{2,1}$ is the compression disk of the surface $\partial R_{2,1} \setminus [\omega_1(0,n) \sqcup \gamma_{-1}(0,n)]$. Since the disk $a \subset R_{2,1}$ intersects $\omega_1'(0,n) \subset T_2$ minimally in one point, by the definition of J and Lemma 8.1.1 it follows that $(R_{2,1}, J)$ is a simple pair of index $|n|$. Therefore $R_{1,1} = \mathrm{cl}[X_K \setminus N(T_1)]$ is an exchange region of index $|n|$.

As in the proof of Proposition 8.2.1, that the knots $K(-1,n,2)$ and $K(0,n,-1)$ are hyperbolic now follows from [24, Lemma 8.1]. Therefore (1) and (2) hold. □

8.4 Hyperbolic knots with $|\mathbb{T}| = 4$, two hyperbolic pairs, and no exchange region

We will use the notation set up in Lemma 2.4.1 in the classification of basic pairs.

Figure 24, top, shows a basic pair (H, J) constructed as in Lemma 2.4.1 using basic circles $\alpha \sqcup \beta \subset \partial H$ separated by the disk $D \subset H$, with parameters set for this example as $m = 4$ and $n = 3$. By Lemma 2.4.2 the pair (H, J) is therefore hyperbolic.

Cutting H along D produces two solid tori, V_1 and V_2, with $H = V_1 \cup (D \times [-1,1]) \cup V_2$, $\alpha \subset \partial V_1$ and $\beta \subset \partial V_2$, as shown in Figure 24.

We assume that the handlebody H is standardly embedded in $\mathbb{S}^3$ with complete disk system $D_1 \sqcup D_2$. Its complementary handlebody $H' = \mathbb{S}^3 \setminus \mathrm{int}\, H$ has complete disk system $D_1' \sqcup D_2'$ such that $|\partial D_i \cap \partial D_j'| = 0$ for $i \neq j$ and 1 for $i = j$.

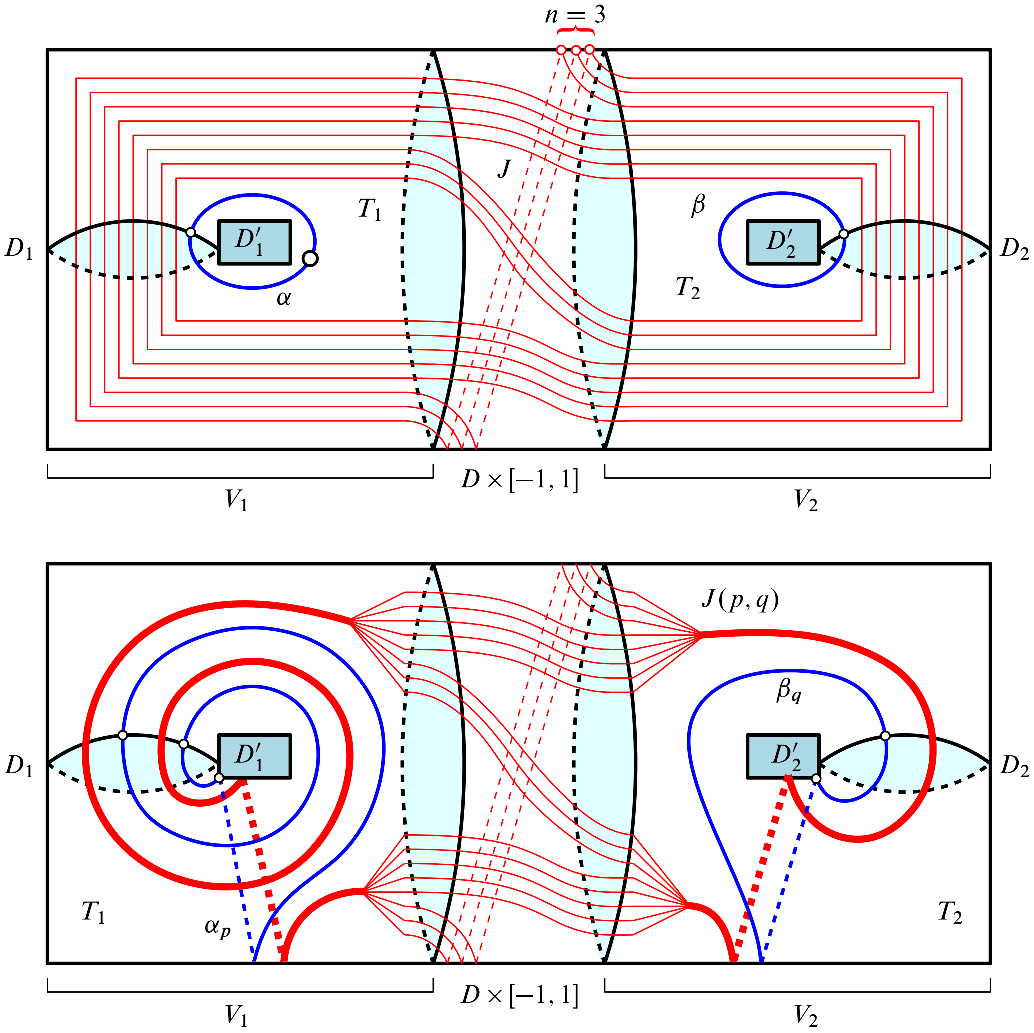


Figure 24: The hyperbolic basic pair (H, J) (top) and the associated hyperbolic pair $(H, J(p,q))$ for $p=2$ and $q=1$ (bottom).

Specifically the following circles are constructed on $\partial H = \partial H'$.

(H1) Basic circles $\alpha \sqcup \beta \subset \partial H$ parallel to $\partial D'_1$ and $\partial D'_2$, respectively.

(H2) The circle $\partial D \subset \partial H = \partial H'$ bounds a nontrivial separating disk $D' \subset H'$.

(H3) We write $\partial H = T_1 \cup_J T_2$ with $\alpha \subset T_1$ and $\beta \subset T_2$.

(H4) For integers $p, q \in \mathbb{Z}$ (and some orientation scheme) let $\alpha_p \subset T_1$ be a circle homologous to $\partial D_1 + p\alpha$, $\beta_q \subset T_2$ a circle homologous to $\partial D_2 + q\alpha$, and construct the separating circle $J(p,q) \subset \partial H$ by matching the endpoints of 8 arcs in $\partial V_1 \setminus \alpha_p$ with those of 8 arcs in $\partial V_2 \setminus \beta_q$ using the same pattern as for J.

The circle $J(p,q)$ is represented in Figure 24, bottom, for $|p|=2$ and $|q|=1$.

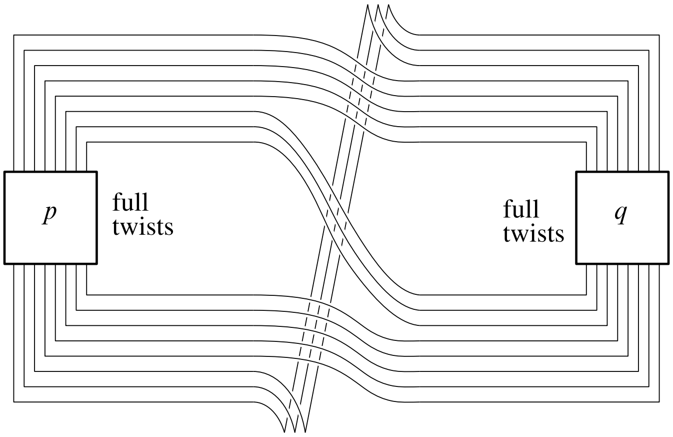


Figure 25: The knot $K(p,q) \subset \mathbb{S}^3$.

We denote by $K = K(p,q)$ the knot in $\mathbb{S}^3$ corresponding to $J(p,q)$. Thus K bounds two simplicial Seifert tori $T_1 \sqcup T_2 \subset X_K$ with $R_{1,2} = H$ and $R_{2,1} = H'$.

Equivalently $K(p,q)$ is the knot obtained by performing p and q full twists on the indicated strands of the trivial knot $K(0,0)$ shown in Figure 25.

If $p = 0$ then T_1 compresses in H' along the disk D'_1 an so the knots $K(0,q)$ are trivial. Similarly the knots $K(p,0)$ are trivial. The knot $K(2,2)$ is represented in Figure 1.

(H5) By (H2) and (H4) the circles $\alpha_p, \beta_q \subset \partial H'$ are basic circles in H' separated by the disk $D' \subset H'$ with $\partial D' = \partial D$.

Recall from Lemma 2.4.1 that the construction parameters m and n of a circle like $J(p,q) \subset \partial H'$ depend only on the distribution of parallel arcs in the intersection of $J(p,q)$ with the annulus $(\partial D') \times I \subset \partial H'$. As $\partial D' = \partial D$, the circles J and $J(p,q)$ share the same parameters, $m = 4$ and $n = 3$, and so the basic pair $(H', J(p,q))$ is hyperbolic and homeomorphic to (H, J).

Similarly, for $|p| = 1 = |q|$ the pair $(H, J(p,q))$ is hyperbolic and homeomorphic to (H, J).

Proposition 8.4.1 *For integers $p, q \neq 0$ the knot $K = K(p,q) \subset \mathbb{S}^3$ is hyperbolic. Specifically, setting $J^* = J(p,q)$:*

(1) *For $|p|, |q| \geq 2$ the knot K bounds a unique maximal simplicial collection of four Seifert tori $\mathbb{T} = T_1 \sqcup T_2 \sqcup T_3 \sqcup T_4 \subset X_K$, such that the pairs $(R_{1,2}, J^*)$ and $(R_{3,4}, J^*)$ are homeomorphic to the hyperbolic pair (H, J), while the pairs $(R_{2,3}, J^*)$ and $(R_{4,1}, J^*)$ are simple of index $|q|$ and $|p|$, respectively.*

(2) *If $|p| \geq 2$ and $|q| = 1$ then in (1) the pair $(R_{4,1}, J^*)$ becomes a trivial pair and so K bounds the unique maximal simplicial collection of three Seifert tori $T_1 \sqcup T_2 \sqcup T_3$ with the pairs $(R_{1,2}, J^*)$ and $(R_{3,1}, J^*)$ homeomorphic to the hyperbolic pair (H, J), and $(R_{2,3}, J^*)$ a pair of index $|p|$. A similar conclusion holds when $|p| = 1$ and $|q| \geq 2$.*

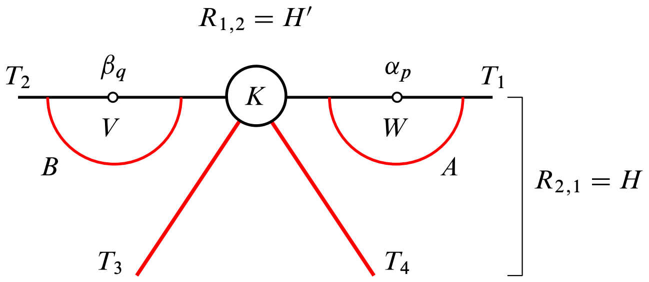


Figure 26: Construction of the Seifert tori $T_3, T_4 \subset X_K$.

(3) *If $|p| = 1 = |q|$ then K bounds the unique maximal simplicial collection of two Seifert tori $T_1 \sqcup T_2 \subset X_K$ with $(R_{1,2}, J^*)$ and $(R_{2,1}, J^*)$ homeomorphic to the hyperbolic pair (H, J).*

Proof Set $R_{1,2} = H'$, $R_{2,1} = H$ and $J^* = J(p,q)$.

Suppose that $|p|, |q| \geq 2$. Then in $R_{2,1} = H$ the circle $\alpha_p \subset T_1$ is a $|p|$-power circle and $\beta_q \subset T_2$ is a $|q|$-power circle. Let $T_3, T_4 \subset R_{2,1}$ be the J^*-tori induced by β_q and α_p, respectively, so that $(R_{2,3}, J^*)$ and $(R_{4,1}, J^*)$ are simple pairs of indices $|q|$ and $|p|$, respectively (see Section 2.3.10).

Let $B, V \subset R_{2,3}$ be the companion annulus and solid torus of the power circle $\beta_q \subset T_2$, with $V \cap T_2$ an annular regular neighborhood of β_q in T_2. Similarly let $A, W \subset R_{2,3}$ be the companion annulus and solid torus of the power circle $\alpha_p \subset T_1$, with $W \cap T_1$ an annular regular neighborhood of α_p. The situation is represented in Figure 26.

Then J^* is disjoint in $\partial R_{1,2}$ of the annuli $W \cap T_1$ and $V \cap T_2$, and we may assume that V and W are disjoint from the separating disk $D \subset H$.

Let $H^* \subset R_{2,1} = H$ be the genus two handlebody component of $R_{2,1}$ cut along the companion annuli $A \sqcup B$. By Section 2.3.5(1) the pair (H^*, J^*) is homeomorphic to $(R_{3,4}, J^*)$, and by [24, Lemma 6.8] the core circles α'_p and β'_q of the companion annuli A and B are basic circles in H^* disjoint from J^*.

The disk $D \subset H$ separates the solid tori V and W and hence it lies in H^*, which by the argument in (H5) implies that the pairs $(H^*, J^*) \approx (R_{3,4}, J^*)$ are homeomorphic to (H, J), hence hyperbolic. Thus (1) holds, and (2) and (3) follow by a similar argument. □

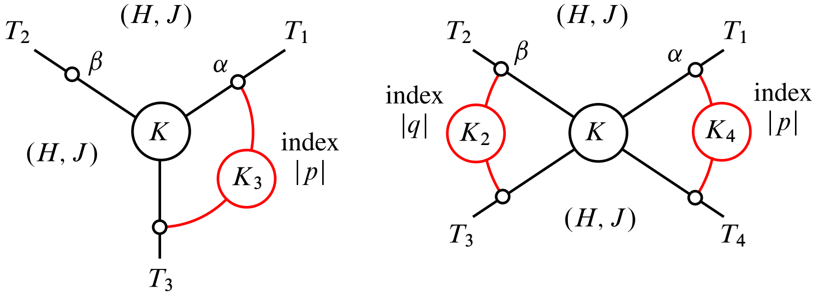


Figure 27: The Seifert tori in X_K for $|p| \geq 2$ and $|q| = 1$ (left) and $|p|, |q| \geq 2$ (right).

Remark 8.4.2 (1) In Figure 27, right, $K_2 \sqcup K_4 \subset \mathbb{S}^3$ is the unlink formed by the cores of the solid tori components of $H|D_0$ in Figure 24, bottom. This is one of the reasons why it is easy to render a projection of the knots $K(p,q)$ as in Figure 25. Moreover the regions $R_{1,3}$ and $R_{3,1}$ are handlebodies but $R_{2,1}$ and $R_{4,3}$ are not, making the latter the exteriors of nontrivial handlebody knots in $\mathbb{S}^3$ with an interesting internal structure.

(2) Similarly in Figure 27, left, the knot $K_3 \subset \mathbb{S}^3$ is trivial and the region $R_{1,3}$ is not a handlebody.

(3) We have used the values $m = 4$ and $n = 3$ for the parameters defined in Lemma 2.4.1; conclusions similar to those reached in Proposition 8.4.1 can be reached for any other suitable values of m and n.

8.5 Hyperbolic knots with $|\mathbb{T}| = 2\,, 3$ and one exchange nonhandlebody region

In this section we construct hyperbolic knots that bound a maximal simplicial collection $\mathbb{T} \subset X_K$ of $|\mathbb{T}| = 2, 3$ Seifert tori with a nonhandlebody exchange region $R_{1,3}$.

The pair $(R_{3,1}, J)$ is necessarily basic by Lemma 5.1.1(2) and in the context of [13] represents a handlebody knot in $\mathbb{S}^3$ whose exterior $R_{1,3}$ contains an essential annulus. Such handlebody knots are completely classified in [13] and the spanning annulus in $R_{1,3}$ corresponds to a type 3-3 annulus as constructed in [13, Section 3, Figure 8].

As in Section 6.3 we retract the regular neighborhood $N(K) \subset \mathbb{S}^3$ radially onto K so that the circles J and ∂T_i become identified with K and use the notation $(R_{i,i+1}, K)$ for the pairs $(R_{i,i+1}, J)$

(I) Construction of handlebody knots $V \subset \mathbb{S}^3$ with an incompressible type 3-3 annulus in their exterior.

(i) Let $L \subset \mathbb{S}^3$ be a fixed knot with exterior $X_L = \mathbb{S}^3 \setminus \operatorname{int} N(L) \subset \mathbb{S}^3$ and let $\alpha \subset \partial X_L = \partial N(L)$ be a circle of nonintegral slope of the form $r = a/p$ for some $p \geq 2$. If L is the trivial knot we choose α to be a nontrivial torus knot.

Let $L' \subset \operatorname{int} X_L$ be a knot which cobounds an annulus with α; thus $L' \subset \mathbb{S}^3$ is a nontrivial knot which is a cable of L. The exterior $X = \mathbb{S}^3 \setminus \operatorname{int}[N(L) \cup N(L')]$ of the link $L \sqcup L' \subset \mathbb{S}^3$ then contains a properly embedded annulus A with the circle $\alpha = \partial A \cap \partial N(L)$ of nonintegral slope in $N(L)$, and the circle $\partial A \cap \partial N(L')$ of integral slope r' in $\partial N(L')$. The situation is represented in Figure 28, top.

Since at least one component of the link $L \sqcup L'$ is a nontrivial knot, the boundary tori ∂X and the annulus A are incompressible in X.

(ii) Let $N(A)$ be a thin regular neighborhood of A in X and $M = \operatorname{cl}[X \setminus N(A)]$ the exterior of A in X; thus ∂M is a torus. By [14, Theorem 1.1] there is a properly embedded arc $e \subset M$ such that

(M1) e has one endpoint on each of the annuli $\partial N(L) \cap \partial M$ and $\partial N(L') \cap \partial M$;

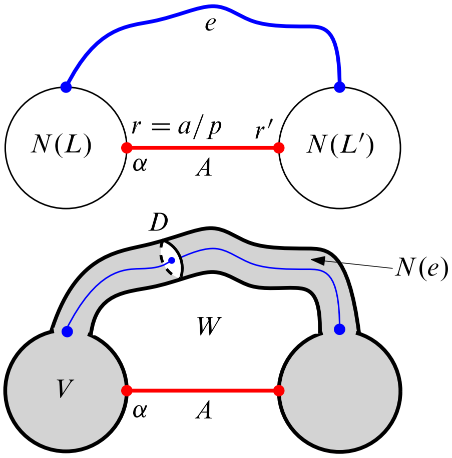


Figure 28: Construction of the genus two handlebody knot $V \subset \mathbb{S}^3$ with exterior W.

(M2) e is an excellent 1-submanifold of M; that is, the exterior $M_e = \mathrm{cl}[M \setminus N(e)]$ of e in M is irreducible, boundary irreducible, atoroidal and *anannular* (any incompressible annulus in M_e is boundary parallel).

(iii) We set

$$V = N(L) \cup N(L') \cup N(e) \subset \mathbb{S}^3 \quad \text{and} \quad W = \mathbb{S}^3 \setminus \operatorname{int} V = \mathrm{cl}[X \setminus N(e)] = M_e \cup N(A).$$

Thus V is a genus two handlebody knot in $\mathbb{S}^3$ whose exterior W contains the nonseparating annulus A. Being a subset of $X \subset \mathbb{S}^3$, the manifold W is irreducible and the annulus $A \subset W$ is incompressible in W. It is then not hard to see from (M2) that W is boundary irreducible and atoroidal, and that any annulus in W is isotopic to A.

(iv) The cocore disk D of the 1-handle $N(e)$ attached to $N(L) \sqcup N(L')$ is a nontrivial disk in V which separates the components of $\partial A \subset V$. By (i), in V the circle $\alpha = \partial A \cap \partial N(L)$ is a $p \geq 2$ power circle and $\partial A \cap \partial N(L')$ is a primitive circle. The situation is represented in Figure 28, bottom.

(II) The genus one hyperbolic knots $K \subset \mathbb{S}^3$.

We consider the family of knots $K \subset \mathbb{S}^3$ consisting of separating circles in $\partial V = \partial W$ that are disjoint from ∂A and nontrivial in V, or, equivalently, such that K and $\partial D \subset V$ intersect minimally in $|K \cap \partial D| \geq 4$ points.

Since the circle $K \subset \partial V = \partial W$ is nontrivial in V and W, K is a genus one knot in $\mathbb{S}^3$ that bounds two Seifert tori from the decomposition $\partial V = T_1 \cup_K T_2$. We set $R_{1,2} = W$ and $R_{2,1} = V$, so that $N(L) \sqcup D \subset R_{2,1}$, and choose the notation $\partial_1 A \subset T_1$ and $\partial_2 A = \alpha \subset T_2$.

We summarize the above conclusions in the next result.

Proposition 8.5.1 *There exist infinitely many genus one hyperbolic knots $K \subset \mathbb{S}^3$ which bound a maximal simplicial collection of* 2 *or* 3 *Seifert tori with one nonhandlebody exchange region.* □

Proof By (iii) and Lemma 2.2.1(2) $(R_{1,2}, K)$ and $(R_{2,1}, K)$ are pairs. Since the region $R_{1,2}$ contains the spanning annulus A, which is unique by (M2), the pair $(R_{1,2}, K)$ is annular of index 1.

By (I)(iv) the circle $\partial_2 A = \alpha \subset T_2$ is a power p-circle in $R_{2,1}$ with companion solid torus constructed out of $N(L) \subset R_{2,1}$, and we set $T_3 \subset R_{2,1}$ as the K-torus induced by α so that $(R_{2,3}, K)$ is a simple pair of index $p \geq 2$. Therefore,

- the pair $(R_{1,3}, K)$ is an exchange pair,
- $R_{1,3}$ is not a handlebody by Section 2.3.5(3),
- the knot $K \subset \mathbb{S}^3$ is hyperbolic by the argument used in Proposition 8.2.1,
- $(R_{3,1}, K)$ is a basic pair by Lemma 5.1.1.

If the pair $(R_{3,1}, K)$ is trivial then T_1 and T_3 are parallel in X_K and so, by Proposition 7.0.2(1), $\mathbb{T} = T_1 \sqcup T_2 \subset X_K$ is a maximal simplicial collection of Seifert tori bounded by K. Otherwise $\mathbb{T} = T_1 \sqcup T_2 \sqcup T_3 \subset X_K$ is such a maximal simplicial collection.

Knots with a maximal simplicial collection of size $|\mathbb{T}| = 3$ can be constructed by choosing $K \subset \partial V$ to follow the pattern of the knot $J(p,q) \subset \partial H$ with $q = 1$ in Figure 24, bottom, so that

$$(V, K, \partial_1 A, \partial_2 A, \partial D) = (H, J(p,q), \beta_q, \alpha_p, D).$$

By the argument in the proof of Proposition 8.4.1 the pair $(R_{3,1}, K)$ is then homeomorphic to the basic hyperbolic pair $(H, J(1,1))$.

By Lemma 2.4.2 replacing $J(p,q) \subset H$ in Section 8.4 with a similar circle constructed with parameters $m \geq 2$ and $n = 1$ then turns $(H, J(1,1))$ into a simple pair of index m.

Finally, to obtain a maximal simplicial collection of size $|\mathbb{T}| = 2$ it suffices to choose $K \subset \partial V$ with $|K \cap \partial D| = 4$ in which case the pair $R_{2,1}$ is simple of index p as in Figure 6, top left. □

Proof of Theorem 1(3) The claim follows from Propositions 8.2.1, 8.3.1 and 8.5.1. □

References

[1] **A J Casson**, **C M Gordon**, *Reducing Heegaard splittings*, Topology Appl. 27 (1987) 275–283 MR Zbl

[2] **M Cohen**, **W Metzler**, **A Zimmermann**, *What does a basis of $F(a,b)$ look like?*, Math. Ann. 257 (1981) 435–445 MR Zbl

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

[4] **J R Eisner**, *Knots with infinitely many minimal spanning surfaces*, Trans. Amer. Math. Soc. 229 (1977) 329–349 MR Zbl

[5] **M Eudave-Muñoz**, *On nonsimple* 3*-manifolds and* 2*-handle addition*, Topology Appl. 55 (1994) 131–152 MR Zbl

[6] **M Eudave-Muñoz**, *Non-hyperbolic manifolds obtained by Dehn surgery on hyperbolic knots*, from "Geometric topology", AMS/IP Stud. Adv. Math. 2.1, Amer. Math. Soc., Providence, RI (1997) 35–61 MR Zbl

[7] **D Gabai**, *Foliations and the topology of* 3*-manifolds, III*, J. Differential Geom. 26 (1987) 479–536 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] **C M Gordon**, **J Luecke**, *Non-integral toroidal Dehn surgeries*, Comm. Anal. Geom. 12 (2004) 417–485 MR Zbl

[10] **J Hempel**, 3*-manifolds*, Ann. of Math. Stud. 86, Princeton Univ. Press (1976) MR Zbl

[11] **W Jaco**, *Lectures on three-manifold topology*, CBMS Region. Conf. Ser. Math. 43, Amer. Math. Soc., Providence, RI (1980) MR Zbl

[12] **O Kakimizu**, *Finding disjoint incompressible spanning surfaces for a link*, Hiroshima Math. J. 22 (1992) 225–236 MR Zbl

[13] **Y Koda**, **M Ozawa**, *Essential surfaces of non-negative Euler characteristic in genus two handlebody exteriors*, Trans. Amer. Math. Soc. 367 (2015) 2875–2904 MR Zbl

[14] **R Myers**, *Excellent* 1*-manifolds in compact* 3*-manifolds*, Topology Appl. 49 (1993) 115–127 MR Zbl

[15] **P Przytycki**, **J Schultens**, *Contractibility of the Kakimizu complex and symmetric Seifert surfaces*, Trans. Amer. Math. Soc. 364 (2012) 1489–1508 MR Zbl

[16] **M Sakuma**, *Minimal genus Seifert surfaces for special arborescent links*, Osaka J. Math. 31 (1994) 861–905 MR Zbl

[17] **M Sakuma**, **K J Shackleton**, *On the distance between two Seifert surfaces of a knot*, Osaka J. Math. 46 (2009) 203–221 MR Zbl

[18] **M Scharlemann**, **A Thompson**, *Finding disjoint Seifert surfaces*, Bull. Lond. Math. Soc. 20 (1988) 61–64 MR Zbl

[19] **J Schultens**, *The Kakimizu complex is simply connected*, J. Topol. 3 (2010) 883–900 MR Zbl

[20] **W P Thurston**, *Three-dimensional manifolds, Kleinian groups and hyperbolic geometry*, Bull. Amer. Math. Soc. 6 (1982) 357–381 MR Zbl

[21] **Y Tsutsumi**, *Universal bounds for genus one Seifert surfaces for hyperbolic knots and surgeries with non-trivial JSJT-decompositions*, Interdiscip. Inform. Sci. 9 (2003) 53–60 MR Zbl

[22] **Y Tsutsumi**, *Hyperbolic knots with a large number of disjoint minimal genus Seifert surfaces*, Tokyo J. Math. 31 (2008) 253–258 MR Zbl

[23] **L G Valdez-Sánchez**, *Seifert Klein bottles for knots with common boundary slopes*, from "Proceedings of the Casson Fest", Geom. Topol. Monogr. 7, Geom. Topol. Publ., Coventry (2004) 27–68 MR Zbl

[24] **L G Valdez-Sánchez**, *Seifert surfaces for genus one hyperbolic knots in the* 3*-sphere*, Algebr. Geom. Topol. 19 (2019) 2151–2231 MR Zbl

[25] **F Waldhausen**, *On irreducible* 3*-manifolds which are sufficiently large*, Ann. of Math. 87 (1968) 56–88 MR Zbl

Department of Mathematical Sciences, University of Texas at El Paso
El Paso, TX, United States

lvsanchez@utep.edu

Received: 24 April 2023 Revised: 14 November 2023

Algebraic & Geometric Topology **25**:3 (2025) 1731–1791
DOI: 10.2140/agt.2025.25.1731
Published: 20 June 2025

Band diagrams of immersed surfaces in 4-manifolds

MARK HUGHES
SEUNGWON KIM
MAGGIE MILLER

We study immersed surfaces in smooth 4-manifolds via singular banded unlink diagrams. Such a diagram consists of a singular link with bands inside a Kirby diagram of the ambient 4-manifold, representing a level set of the surface with respect to an associated Morse function. We show that every self-transverse immersed surface in a smooth, orientable, closed 4-manifold can be represented by a singular banded unlink diagram, and that such representations are uniquely determined by the ambient isotopy or equivalence class of the surface up to a set of singular band moves which we define explicitly. By introducing additional finger, Whitney and cusp diagrammatic moves, we can use these singular band moves to describe homotopies or regular homotopies as well.

Using these techniques, we introduce bridge trisections of immersed surfaces in arbitrary trisected 4-manifolds and prove that such bridge trisections exist and are unique up to simple perturbation moves. We additionally give some examples of how singular banded unlink diagrams may be used to perform computations or produce explicit homotopies of surfaces.

57K45; 57K40

1 Introduction

Immersed surfaces are fundamental objects in low-dimensional topology, showing up frequently in the study of 4-manifolds. For example, immersed disks play a key role in Freedman's proof of the topological h-cobordism theorem and the homeomorphism classification of simply connected smooth 4-manifolds [6]. One reason for the prominent part they play lies in how abundant they are when compared to their embedded counterparts. In particular, maps of surfaces into smooth 4-manifolds can always be perturbed slightly to yield smooth immersions with transverse double points.

Despite their importance, immersed surfaces and their isotopies are difficult to describe explicitly outside of a few concrete examples. While diagrammatic techniques have been developed to describe both smooth 4-manifolds and embedded surfaces (see eg Carter and Saito [3; 4], Hughes, Kim and Miller [14], Kamada [17; 19], Meier and Zupan [26; 27] and Roseman [29]), methods of studying immersed surfaces diagrammatically have not been established as fully in the literature, aside from a few examples (see eg Kamada, Kawauchi, Kim and Lee [20] for a diagrammatic framework for representing immersed surfaces in $\mathbb{R}^4$ via marked graph diagrams).

In this paper, we introduce a new diagrammatic system for describing immersed surfaces in smooth, oriented, closed 4-manifolds, called *singular banded unlink diagrams*. Such a diagram consists of a Kirby diagram for the ambient 4-manifold along with a decorated singular (4-valent) link with bands attached away from vertices (see Section 2.2 for details). As a Kirby diagram of X is uniquely determined by a Morse function h and its gradient ∇h, given two singular banded unlink diagrams in the same Kirby diagram (induced by the same Morse function on X), it makes sense to ask whether they determine isotopic surfaces. Even with singular banded unlink diagrams in two different Kirby diagrams of X, we can still ask whether they describe equivalent surfaces. With this in mind, we define a set of moves, called *singular band moves*, in Figures 3 and 4, which allow us to relate the diagrams of any two immersed surfaces which are ambiently isotopic. When combined with Kirby moves to the ambient diagram, these moves are also sufficient to relate equivalent surfaces. That is, we show the equivalence

$$\frac{\{\text{singular banded unlink diagrams}\}}{\text{singular band moves}} \leftrightarrow \frac{\{\text{self-transversely immersed surfaces in 4-manifolds}\}}{\text{ambient diffeomorphism}}.$$

We make this equivalence precise in Corollary 2.40 (and for isotopy rather than diffeomorphism in Theorem 2.39). This work generalizes earlier results in [14], where the authors define *banded unlink diagrams* of smoothly *embedded* surfaces in smooth 4-manifolds, and present a family of moves (called *band moves*) to describe isotopies between such surfaces. More precisely, given a smoothly embedded surface Σ in a smooth oriented closed 4-manifold X and a self-indexing Morse function $h\colon X \to \mathbb{R}$, we obtain a diagram $\mathfrak{D}(\Sigma)$ which is well defined up to band moves and depends only on the ambient isotopy class of Σ inside X. Furthermore, given the diagram $\mathfrak{D}(\Sigma)$, we may recover the pair (X, Σ) up to diffeomorphism. If we also specify the Morse function $h\colon X \to \mathbb{R}$, then the surface $\Sigma \subset X$ is determined up to isotopy. In the special case that $X^4 = S^4$ and h is standard (ie h has no index 1, 2 or 3 critical points), these results are originally due to Swenton [32] and Kearton and Kurlin [22].

Unless otherwise stated, we will assume that X is a closed, smooth, oriented 4-manifold. Our main theorems are as follows:

Theorem 2.39 *Let Σ be a smoothly immersed, self-transverse surface in a 4-manifold X. Then any choice of a self-indexing Morse function $h\colon X \to \mathbb{R}$ (with one index 0 point) and a gradientlike vector field ∇h on X induces a singular banded unlink diagram $\mathfrak{D}(\Sigma)$ of (X, Σ) that is well defined up to singular band moves.*

Furthermore, let $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ be singular banded unlink diagrams of immersed surfaces Σ and Σ' in X.

(i) *The diagrams $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ are related by band moves and Kirby moves if and only if there is a diffeomorphism $(X,\Sigma) \cong (X,\Sigma')$.*

(ii) *If $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ are induced by the same self-indexing Morse function h and gradientlike vector field ∇h (which are suitably generic so as to ensure the underlying Kirby diagrams of $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ agree), then $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ are related by band moves if and only if Σ and Σ' are ambiently isotopic.*

In other words, if $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ are banded unlink diagrams whose underlying Kirby diagrams are identified, then Σ and Σ' are smoothly ambiently isotopic if and only if $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ are related by singular band moves.

In the opening paragraph of Theorem 2.39, we say that $\mathfrak{D}(\Sigma)$ is well defined only up to singular band moves, even though ∇h is specified. This is because, in order to obtain $\mathfrak{D}(\Sigma)$, we also need to choose a gradientlike vector field of $h|_\Sigma$, which is not canonically determined by $(h, \nabla h, \Sigma)$.

Note that part (ii) of Theorem 2.39 clearly implies part (i), so we will focus on proving part (ii). Furthermore, since Kirby diagrams of two 4-manifolds can be related by a sequence of Kirby moves if and only if they are diffeomorphic, we obtain the following corollary:

Corollary 2.40 *Let $\mathfrak{D}$ and $\mathfrak{D}'$ be singular banded unlink diagrams of surfaces Σ and Σ' self-transversely immersed in diffeomorphic 4-manifolds X and X'. There is a diffeomorphism taking (X,Σ) to (X',Σ') if and only if there is a sequence of singular band moves and Kirby moves taking $\mathfrak{D}$ to $\mathfrak{D}'$.*

Without much extra work, we may also extend Theorem 2.39 to consider homotopy instead of isotopy:

Corollary 2.41 *Let Σ and Σ' be self-transverse surfaces smoothly immersed in X, and let $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ be singular banded unlink diagrams in the same Kirby diagram of X.*

(i) *The surfaces Σ and Σ' are regularly homotopic if and only if $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ can be related by a sequence of singular band moves and the finger/Whitney moves illustrated in Figure 15.*

(ii) *The surfaces Σ and Σ' are homotopic (without specifying regularity) if and only if $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ are related by singular band moves, finger/Whitney moves and cusp moves as illustrated in Figure 15.*

One application of the authors' results in [14] was to prove the uniqueness of bridge trisections of surfaces in arbitrary trisected 4-manifolds up to perturbation. In Section 3.2, we define the notions of *bridge position* and *bridge trisections* for immersed surfaces in trisected 4-manifolds, and in Section 3.5 we prove an analogous uniqueness statement:

Theorem 3.36 *Let $(X^4,\mathcal{T})$ be a trisected 4-manifold. Let Σ be a self-transverse immersed surface in X^4. Then Σ can be isotoped into bridge position with respect to $\mathcal{T}$, yielding a bridge trisection of Σ*

with respect to $\mathcal{T}$. Moreover, any two bridge trisections of Σ with respect to $\mathcal{T}$ are related by $\mathcal{T}$-preserving isotopy, perturbations and vertex perturbations (and their inverses).

The moves referenced in Theorem 3.36 are defined in Section 3.1. For experts, we will say now that the perturbation move is the standard perturbation move that increases the number of disks of Σ in one section of the trisection, while vertex perturbation is supported in a neighborhood of the trisection surface and simply passes a self-intersection of Σ from one piece of the trisection to another.

Organization

In Section 2 we lay out the framework of singular banded unlink diagrams. We begin in Section 2.1 with a discussion on marked singular banded links. In Section 2.2, we describe how to use these decorated singular links to obtain immersed surfaces. In Section 2.3, we discuss two subclasses of immersed surfaces that will be needed to prove Theorem 2.39 and its corollaries in Section 2.4.

In Section 3 we turn our attention to bridge trisections. We review the theory of bridge trisections of embedded surfaces in Section 3.1. In Section 3.2, we adapt the notions of trivial tangles and bridge position to singular links, before defining bridge position for immersed surfaces in Section 3.3 and showing that every immersed surface in a smooth 4-manifold can be arranged in this position. It is here that we define the various moves on immersed bridge trisections referenced in Theorem 3.36. In Section 3.4, we then proceed to adapt the singular banded unlinks developed in Section 2 to bridge trisections, before using the uniqueness results for singular banded unlinks to prove Theorem 3.36 in Section 3.5.

In Section 4 we give some additional sample applications of the usefulness of singular banded unlink diagrams. In Section 4.1, we show how one may compute the Kirk invariant (see Schneiderman and Teichner [30]) of a spherical link using these diagrams. In Section 4.2, we prove that homologous immersed oriented surfaces with the same number of positive and negative self-intersections are stably isotopic (ie become isotopic after surgery along some collection of arcs). Finally, in Section 4.3, we show that certain 2-spheres embedded in S^4 can be trivialized by a single finger and Whitney move (recovering a fact originally proved by Joseph, Klug, Ruppik and Schwartz [16]).

Acknowledgements

This project began in Spring 2019 when Kim and Miller visited Hughes at Brigham Young University.

Miller thanks Peter Teichner and Mark Powell for useful discussions in Fall 2020, which influenced the viewpoint of this paper, particularly in Sections 2.3 and 2.4.

Hughes was supported by NSF grant DMS-2213295. Kim was supported by the Institute for Basic Science (IBS-R003-D1) and a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1C1C2004559). Miller was supported by NSF grant DGE-1656466 at Princeton University, NSF grant DMS-2001675 at Massachusetts Institute of Technology, a Clay Research Fellowship at Stanford University, and finally NSF grant DMS-2404810 and Simons Foundation travel gift MPS-TSM-00007679.

2 Singular banded unlink diagrams

2.1 Marked singular banded links

In this section we introduce marked singular banded links, which are the combinatorial objects we will use to describe self-transverse immersed surfaces in 4-manifolds. In what follows, all manifolds and maps between them should be assumed to be smooth. All isotopies of immersed (or embedded) submanifolds are assumed to be ambient isotopies unless otherwise specified. Note that we are isotoping the images of immersions rather than immersions themselves.

2.1.1 Marked singular links We begin by defining special singular links with additional data recorded at their double points.

Definition 2.1 Let M^3 be an orientable 3-manifold. A *singular link* L in M is the image of an immersion $\iota\colon S^1 \sqcup \cdots \sqcup S^1 \to M$ which is injective except at isolated double points that are not tangencies. At every double point p we include a small disk $v \cong D^2$ embedded in M such that $(v, v \cap L) \cong \bigl(D^2, \{(x,y) \in D^2 \mid xy = 0\}\bigr)$. We refer to these disks as the *vertices* of L.

(Equivalently, a singular link is a 4-valent fat-vertex graph smoothly embedded in M.) For now, our motivating idea is that M will correspond to some level set of a 4-manifold X, and the double points of a singular link L in M will correspond to the isolated double points an immersed surface in X. Because these double points are isolated, we expect the singularities of L to be resolved away from the level set M. We must make a choice of how to resolve each double point.

Definition 2.2 A *marked singular link* (L,σ) in M is a singular link L along with decorations σ on the vertices of L, as follows: say that v is a vertex of L, with $\partial v \cap \overline{(L \setminus v)}$ consisting of the four points p_1, p_2, p_3, p_4 in cyclic order. Choose a coorientation of the disk v. On the positive side of v, add an arc connecting p_1 and p_3. On the negative side of v, add an arc connecting p_2 and p_4. See Figure 1, left. A choice of σ involves making a fixed choice of decoration on v for all vertices v of L.

Note that, if L has n vertices, there are 2^n choices of decorations σ such that (L,σ) is a marked singular link.

Definition 2.3 Let (L,σ) be a marked singular link in a 3-manifold M. Let v be a vertex of L; say that on the positive side of v there is an arc with endpoints p_1 and p_3, and on the negative side of v there is an arc with endpoints p_2 and p_4.

Let L^+ denote the link in M obtained from (L,σ) by pushing the arc of L between p_1 and p_3 off v in the positive direction, and repeating for each vertex in L. We call L^+ the *positive resolution* of (L,σ) (see Figure 1).

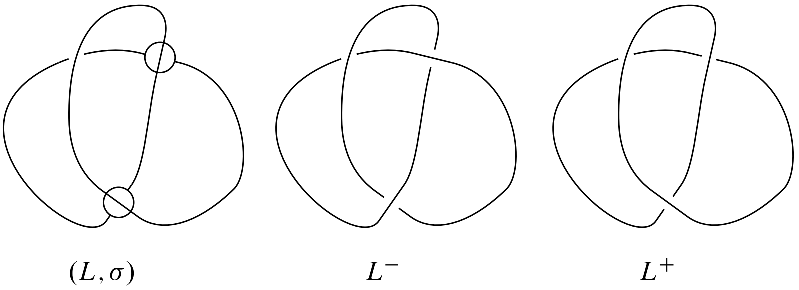


Figure 1: Left: a marked singular link (L,σ). Middle and right: the negative and positive resolutions of (L,σ), respectively.

Similarly, let L^- denote the link in M obtained from (L,σ) by pushing the arc of L between p_1 and p_3 off v in the negative direction, and repeating for each vertex in L. We call L^- the *negative resolution* of (L,σ) (see Figure 1).

Informally, L^+ is obtained from (L,σ) by turning the decorations of σ into new overstrands, while L^- is obtained by turning the decorations of σ into new understrands.

To ease notation, from now on we will always take singular links to be marked. We will generally not specify the decorations σ, and will instead write "L is a marked singular link", with σ implicitly fixed.

2.1.2 Banded singular links Let L be a singular link, and let Δ_L denote the union of the vertices of L. A *band* b attached to L is the image of an embedding $\phi\colon I\times I\hookrightarrow M\setminus\Delta_L$, where $I=[-1,1]$, and $b\cap L=\phi(\{-1,1\}\times I)$. We call $\phi\bigl(I\times\{\frac{1}{2}\}\bigr)$ the *core* of the band b. Let L_b be the singular link defined by

$$L_b=\bigl(L\setminus\phi(\{-1,1\}\times I)\bigr)\cup\phi(I\times\{-1,1\}).$$

Then we say that L_b is the result of performing *band surgery* to L along b. If B is a finite family of pairwise disjoint bands for L, then we will let L_B denote the link we obtain by performing band surgery along each of the bands in B. We say that L_B is the result of *resolving* the bands in B. Note that the self-intersections of L_B naturally correspond to those of L, so a choice of markings for L yields markings for L_B. A triple (L,σ,B), where (L,σ) is a marked singular link and B is a family of disjoint bands for L, is called a *marked singular banded link*. To ease notation, we may refer to the pair (L,B) as a *singular banded link* and implicitly remember that L is actually a *marked* singular link.

2.2 Singular banded links describing surfaces

In this section, we use marked singular banded links to describe surfaces in 4-manifolds. Thinking of M as a level set of the 4-manifold X, we'll begin by defining what the surface looks like in a product tubular neighborhood of M.

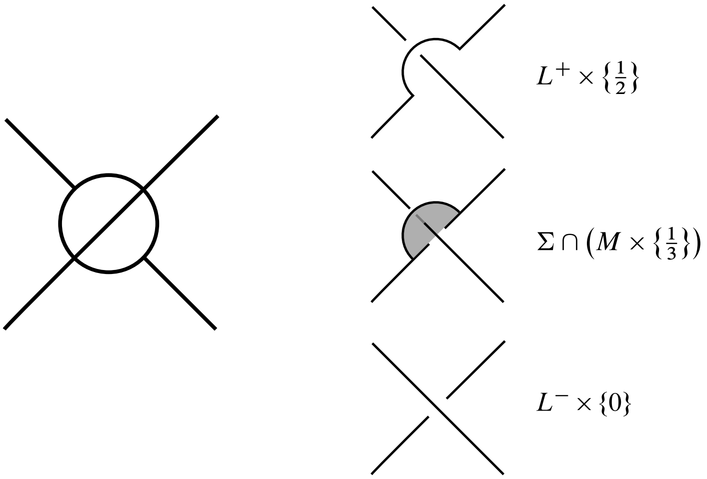

$L^+ \times \{\frac{1}{2}\}$
$\Sigma \cap (M \times \{\frac{1}{3}\})$
$L^- \times \{0\}$

Figure 2: Left: a vertex v of a marked singular link (L, B). Right: part of the surface Σ built from (L, B) near v.

2.2.1 Realizing surfaces in $M^3 \times [0, 1]$ Let (L, B) be a marked singular banded link in the oriented 3-manifold M. We will describe how to construct a surface Σ in $M \times [0, 1]$ using (L, B).

Recall first that L^- is the (nonsingular) link obtained by negatively resolving each vertex of L. Also notice that L^- differs from L^+ only in a neighborhood of the vertices of L, where at each vertex a single strand of L is pushed in the positive direction to give L^+, and the negative direction to give L^-. For each vertex v of L, these two opposite pushoffs form a bigon in a neighborhood of v, which bounds an embedded disk D_v. This disk D_v can be chosen so that its interior intersects L transversely in a single point near v. For each vertex v, select such a disk D_v (ensuring that all of these disks are pairwise disjoint), and let D_L denote the union of all of these embedded disks.

We can then define the surface $\Sigma \subset M \times [0, 1]$ by

$$\begin{aligned}
\Sigma \cap \big(M \times \big[0, \tfrac{1}{3}\big)\big) &= L^- \times \big[0, \tfrac{1}{3}\big), \\
\Sigma \cap \big(M \times \big\{\tfrac{1}{3}\big\}\big) &= (L^- \cup D_L) \times \big\{\tfrac{1}{3}\big\}, \\
\Sigma \cap \big(M \times \big(\tfrac{1}{3}, \tfrac{2}{3}\big)\big) &= L^+ \times \big(\tfrac{1}{3}, \tfrac{2}{3}\big), \\
\Sigma \cap \big(M \times \big\{\tfrac{2}{3}\big\}\big) &= \big(L^+ \cup B\big) \times \big\{\tfrac{2}{3}\big\}, \\
\Sigma \cap \big(M \times \big(\tfrac{2}{3}, 1\big]\big) &= L^+_B \times \big(\tfrac{2}{3}, 1\big].
\end{aligned}$$

In total, Σ is a surface properly immersed in $M \times [0, 1]$ with boundary $(L^- \times \{0\}) \sqcup (L^+{}_B \times \{1\})$, and with isolated transverse self-intersections all contained in $M \times \big\{\frac{1}{3}\big\}$ corresponding to the vertices of L.

Definition 2.4 Let $\overline{\Sigma}(L, B)$ be a surface properly immersed in $M \times [0, 1]$ obtained from Σ by smoothing corners. We refer to $\overline{\Sigma}(L, B)$ as a *surface segment realizing* (L, B).

Proposition 2.5 *Up to ambient isotopy of $M \times [0,1]$, the surface segment $\overline{\Sigma}(L,B)$ depends only on the singular banded link (L,B).*

Proof There is a unique way (up to isotopy) to smooth the corners of Σ in a neighborhood of $M \times \left\{\frac{1}{3}, \frac{2}{3}\right\}$. The disks D_v in $M \times \left\{\frac{1}{3}\right\}$ are determined up to isotopy by the links L^- and L^+, which are well defined up to isotopy in M. No other choices were made in constructing $\overline{\Sigma}(L,B)$. □

Note that, by rescaling the interval parameter, we can similarly define a surface segment realizing (L,B) in any product of the form $M \times [t_1,t_2]$. As above, the ambient isotopy class of $\overline{\Sigma}(L,B)$ will depend only on (L,B).

2.2.2 Morse functions and projections between level sets Before describing how to construct a closed realizing surface in a 4-manifold from a singular banded unlink, it will be convenient to take a brief detour to set up some useful notation. Let X be a closed, oriented, 4-manifold equipped with a self-indexing Morse function $h\colon X \to [0,4]$, where h has exactly one index 0 critical point. In what follows, it will be helpful to have a way of identifying subsets of distinct level sets $h^{-1}(t_1)$ and $h^{-1}(t_2)$.

Suppose then that $t_1 \le t_2$, and let $x_1,\dots,x_p$ denote the critical points of h which satisfy $t_1 \le h(x_j) \le t_2$. Let X_{t_1,t_2} denote the complement in X of the ascending and descending manifolds of the critical points $x_1,\dots,x_p$. Then the gradient flow of h defines a diffeomorphism $\rho_{t_1,t_2}\colon h^{-1}(t_1) \cap X_{t_1,t_2} \to h^{-1}(t_2) \cap X_{t_1,t_2}$.

Definition 2.6 We call ρ_{t_1,t_2} the *projection of $h^{-1}(t_1)$ to $h^{-1}(t_2)$*. Similarly, we call ρ_{t_1,t_2}^{-1} the *projection of $h^{-1}(t_2)$ to $h^{-1}(t_1)$*, which we likewise denote by ρ_{t_2,t_1}.

Note that, despite calling ρ_{t_1,t_2} the projection from $h^{-1}(t_1)$ to $h^{-1}(t_2)$, it is only defined on the complement of the ascending and descending manifolds of the critical points that sit between t_1 and t_2. These projection maps allow us to define local product structures away from the ascending and descending manifolds of the critical points of h.

2.2.3 Singular banded unlinks and closed realizing surfaces We are now able to define a closed realizing surface associated to a given singular banded *unlink*, which we define below. As above, let X be a closed, oriented 4-manifold equipped with a self-indexing Morse function $h\colon X \to [0,4]$, with exactly one index 0 critical point.

Definition 2.7 Let (L,B) be a marked singular banded link in the 3-manifold $M := h^{-1}\left(\frac{3}{2}\right)$ such that $L, B \subset X_{1/2,5/2}$. Suppose furthermore that $\rho_{3/2,1/2}(L^-)$ bounds a collection of disjoint embedded disks D_- in $h^{-1}\left(\frac{1}{2}\right)$, and that $\rho_{3/2,5/2}(L_B^+)$ bounds a collection of disjoint embedded disks D_+ in $h^{-1}\left(\frac{5}{2}\right)$. Then we say that (L,B) is a *singular banded unlink* in the manifold X.

In plain English, (L, B) is a singular banded unlink when both

(1) L^- is an unlink when viewed as a link in $h^{-1}\left(\frac{3}{2}\right)$ ("below the 2-handles"),

(2) L^+_B is an unlink when viewed as a link in $h^{-1}\left(\frac{5}{2}\right)$ ("above the 2-handles").

Fix $\varepsilon \in \left(0, \frac{1}{2}\right)$. Given a singular banded unlink (L, B) in $M = h^{-1}\left(\frac{3}{2}\right)$, and families of disks D_+ and D_- as in Definition 2.7, we can construct an immersed surface with corners $\Sigma \subset X$ as follows:

(i) $\Sigma \cap h^{-1}(t) = \varnothing$ for $t < \frac{1}{2}$ or $t > \frac{5}{2}$.

(ii) $\Sigma \cap h^{-1}\left(\frac{1}{2}\right) = D_-$.

(iii) $\Sigma \cap h^{-1}(t) = \rho_{1/2,t}(\partial D_-)$ for $t \in \left(\frac{1}{2}, \frac{3}{2} - \varepsilon\right)$.

(iv) $\Sigma \cap h^{-1}\left(\left(\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\right)\right)$ is a realizing surface segment in the product $h^{-1}\left(\left(\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\right)\right) \cong M \times \left(\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\right)$ for the singular banded link (L, B) in M.

(v) $\Sigma \cap h^{-1}(t) = \rho_{5/2,t}(\partial D_+)$ for $t \in \left(\frac{3}{2} + \varepsilon, \frac{5}{2}\right)$.

(vi) $\Sigma \cap h^{-1}\left(\frac{5}{2}\right) = D_+$.

That is, Σ consists from bottom to top of minimum disks, a realizing surface segment (which we recall has self-intersections and index 1 critical points) and maximum disks.

Note that the identification of $h^{-1}\left(\left(\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\right)\right)$ with $M \times \left(\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\right)$ in part (iv) above is made using the projection maps $\rho_{3/2,t} \colon h^{-1}\left(\frac{3}{2}\right) \to h^{-1}(t)$, which is a diffeomorphism for $t \in \left(\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\right)$ and small ε. Under this identification, the boundary of the realizing surface segment will be precisely $\rho_{5/2,3/2+\varepsilon}(\partial D_+) \sqcup \rho_{1/2,3/2-\varepsilon}(\partial D_-)$, and hence the surface Σ constructed above will be closed.

Definition 2.8 Let $\Sigma(L, B)$ be an immersed surface in X obtained from Σ by smoothing corners. We refer to $\Sigma(L, B)$ as a (*closed*) *realizing surface* for the singular banded unlink (L, B) in X.

The surface $\Sigma(L, B)$ is an immersed surface in X with isolated, transverse self-intersections. Note that $\Sigma(L, B)$ is obtained (up to isotopy) by smoothing the result of capping off the boundary components of $\overline{\Sigma}(L, B)$ by horizontal disks, which is possible exactly when (L, B) is a singular banded *unlink*.

Proposition 2.9 *Any two realizing surfaces for the singular banded unlink* (L, B) *are smoothly isotopic.*

Proof We first note that choosing a different value for ε changes Σ by an isotopy through realizing surfaces. Second, by Proposition 2.5 any two choices of surface segment $\overline{\Sigma}(L, B) \subset h^{-1}\left(\left[\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\right]\right)$ are isotopic, and this isotopy can be extended to the rest of $\Sigma \cap h^{-1}\left(\left(\frac{1}{2}, \frac{5}{2}\right)\right)$ using the projection maps ρ_{t_1,t_2}. Finally, any choice of embedded disks $\Sigma \cap h^{-1}\left(\frac{1}{2}\right)$ and $\Sigma \cap h^{-1}\left(\frac{5}{2}\right)$ are isotopic rel boundary in $h^{-1}\left(\left[0, \frac{1}{2}\right]\right)$ and $h^{-1}\left(\left[\frac{5}{2}, 4\right]\right)$, respectively, which follows from the fact that $h^{-1}\left(\left[0, \frac{1}{2}\right]\right) \cong B^4$ and $h^{-1}\left(\left[\frac{5}{2}, 4\right]\right) \cong \natural_k (S^1 \times B^3)$. □

As the realizing surface $\Sigma(L, B)$ is determined by the singular banded unlink (L, B) up to isotopy, we will often think of $\Sigma(L, B)$ as representing an isotopy class of immersed surfaces, rather than a particular representative.

2.2.4 Singular banded unlink diagrams and Kirby diagrams We now make sense of how to describe a realizing surface as in Section 2.2.3 via a Kirby diagram. If one is comfortable with these diagrams, then the contents of this subsection are clear from Definition 2.7: simply draw L and B inside a diagram for X in a natural way. We now review some basic notions about Kirby diagrams.

Let $h\colon X \to \mathbb{R}$ be a self-indexing Morse function with a unique index 0 critical point, and let n be the number of index 1 critical points of h. Fix a gradientlike vector field ∇h for h. Let $M = h^{-1}\bigl(\frac{3}{2}\bigr)$, and let L_2 be the intersection of M with the descending manifolds of the index 2 critical points of h. Perturb ∇h slightly if necessary so that this intersection is transverse, so that L_2 is a link in the 3-manifold $M \cong \#_n\, S^1 \times S^2$. To each component of L_2, assign the framing induced by the descending manifold of the associated index 2 critical point, so that L_2 is actually a framed link in M.

Fix an n-component unlink L_1 in S^3. Let V denote the complement of the unique (up to isotopy rel boundary) boundary-parallel disks bounded by L_1 in B^4. Then V is diffeomorphic to $\natural_n\, S^1 \times B^3$, and we can therefore find a diffeomorphism $\phi\colon V \to h^{-1}\bigl(\bigl[0, \frac{3}{2}\bigr]\bigr)$. By Laudenbach and Poénaru [24] and Laudenbach [23], the choice of ϕ is natural up to isotopy and moves that correspond to slides of L_1 (as a 0-framed link) in S^3. Moreover, ∂V can be naturally identified with the result of performing 0-surgery on S^3 along L_1, which we denote by $S^3_0(L_1)$. By perturbing ∇h, we may assume that $\phi^{-1}(L_2) \subset \partial V \cong S^3_0(L_1)$ is disjoint from the surgery solid tori, and hence we can think of $\phi^{-1}(L_2)$ as a link in S^3. By abuse of notation, we will also refer to this link as L_2.

Definition 2.10 Let $\mathcal{K} := (L_1, L_2)$ be a pair of disjoint links in S^3 with L_1 an unlink and L_2 framed. Suppose there is a 4-manifold X, a Morse function $h\colon X \to \mathbb{R}$ and a gradientlike vector field ∇h for h such that $h^{-1}\bigl(\frac{3}{2}\bigr)$ may be identified with $S^3_0(L_1)$ and the descending manifolds of the index 2 critical points of h meet $h^{-1}\bigl(\frac{3}{2}\bigr)$ in the framed link L_2. Then we call $\mathcal{K}$ a *Kirby diagram of X corresponding to* $(h, \nabla h)$.

Remark 2.11 In [28], the third author and Naylor showed that a smooth, closed, *nonorientable* 4-manifold X^4 is also determined up to diffeomorphism by (framed) attaching regions of 0-, 1- and 2-handles. If desired, one could thus make sense of diagrams of closed (immersed) surfaces in Kirby diagrams of nonorientable 4-manifolds. We choose not to pursue this explicitly in this paper for sake of simplicity.

Remark 2.12 Given h and ∇h, a Kirby diagram $\mathcal{K}$ corresponding to $(h, \nabla h)$ is well defined up to isotopy and slides over L_1 as long as there is no flowline of ∇h between two index 2 critical points of h. That is, generically we expect h and ∇h to determine a Kirby diagram.

Conversely, given $\mathcal{K}$, the triple $(X, h, \nabla h)$ is determined up to diffeomorphism.

Let $E(\mathcal{K})$ denote the complement $S^3 \setminus \nu(\mathcal{K})$ of a small tubular neighborhood of the links L_1 and L_2 that form a Kirby diagram $\mathcal{K}$. Then we may think of a link $L \subset E(\mathcal{K})$ as describing a link in $h^{-1}(t)$ for any $t \in (0, 3)$ via the projection map $\rho_{3/2,t}$.

Definition 2.13 A *singular banded unlink diagram* in the Kirby diagram $\mathcal{K} = (L_1, L_2)$ is a triple $(\mathcal{K}, L, B)$, where $L \subset E(\mathcal{K})$ is a marked singular link and $B \subset E(\mathcal{K})$ is a finite family of disjoint bands for L, such that L^- bounds a family of pairwise disjoint embedded disks in $h^{-1}\left(\frac{1}{2}\right)$, and L_B^+ bounds a family of pairwise disjoint embedded disks in $h^{-1}\left(\frac{5}{2}\right)$.

By comparing Definition 2.13 to Definition 2.7, we see that a singular banded unlink diagram describes an immersed realizing surface, as follows. We first note that we can identify $E(\mathcal{K})$ with a subset of $h^{-1}\left(\frac{3}{2}\right)$ in a natural way (ie via ∇h). Since the banded link $L \cup B$ is disjoint from L_1, it can be identified with a subset of $h^{-1}\left(\frac{3}{2}\right)$, which we denote by $L' \cup B'$. This subset avoids the descending manifolds of the index 2 critical points of h.

Since L'^- is disjoint from L_1, we can isotope it vertically downwards via the projection map $\rho_{3/2,t}$ from $h^{-1}\left(\frac{3}{2}\right)$ to $h^{-1}\left(\frac{1}{2}\right)$, where it can be capped off by a family of disjoint embedded disks in $h^{-1}\left(\frac{1}{2}\right)$. Similarly, we can extend the surgered link $L'^+_{B'}$ vertically upwards from $h^{-1}\left(\frac{3}{2}\right)$ to $h^{-1}\left(\frac{5}{2}\right)$, where it can be capped off by disks. As these families of disks are unique up to isotopy rel boundary, the surface we obtain in this way from the banded unlink diagram $(\mathcal{K}, L, B)$ is well defined up to isotopy. (See also Proposition 2.9.) We denote this surface by $\Sigma(\mathcal{K}, L, B)$.

Definition 2.14 We say that $\Sigma(\mathcal{K}, L, B)$ is a *realizing surface for* $(\mathcal{K}, L, B)$, or that $(\mathcal{K}, L, B)$ *describes the surface* $\Sigma(\mathcal{K}, L, B)$.

Definition 2.15 If Σ is a realizing surface of a singular banded unlink diagram $(\mathcal{K}, L, B)$, then we say that $(\mathcal{K}, L, B)$ is a *singular banded unlink diagram* for Σ, and we write $\mathcal{D}(\Sigma) := (\mathcal{K}, L, B)$. (In practice, we might drop the word "singular", since this will be clear when Σ is immersed.) Note that Σ determines $\mathcal{D}(\Sigma)$ uniquely up to isotopy, assuming that Σ is a realizing surface for some diagram.

Definition 2.16 Let Σ be a subset of X. Then we say that $h|_\Sigma$ is *Morse* if there is a surface F and an immersion $f : F \to X$ such that $\Sigma = f(F)$, and such that $h \circ f$ is a Morse function on F. An index k critical point of $h|_\Sigma$ is a point of the form $f(p)$, where p is an index k critical point of $h \circ f$.

Lemma 2.17 *Let X be a closed 4-manifold, and $\mathcal{K}$ a Kirby diagram for X. Then any immersed surface Σ in X is ambient isotopic to a realizing surface $\Sigma(\mathcal{K}, L, B)$ for some singular banded unlink diagram $(\mathcal{K}, L, B)$.*

Proof After a small ambient isotopy we may assume that $h|_\Sigma$ is Morse. Isotope all of the maxima of Σ vertically upward into $h^{-1}\left(\left(\frac{5}{2}, 4\right)\right)$ (generically, maxima of Σ do not lie in the descending manifolds of index 1 or 2 critical points of h). Similarly isotope the minima of Σ vertically downward into $h^{-1}\left(\left(0, \frac{3}{2}\right)\right)$. Isotope all of the index 1 critical points of $h|_\Sigma$ vertically into $h^{-1}\left(\left(\frac{3}{2}, \frac{5}{2}\right)\right)$ (again, index 1 critical points of $h|_\Sigma$ generically do not lie in the ascending manifolds of index 3 critical points or the descending manifolds of index 1 critical points). Finally, isotope the self-intersections of Σ to lie in $h^{-1}\left(\left(\frac{3}{2}, \frac{5}{2}\right)\right)$ in such a way that they do not coincide with index 1 critical points of $h|_\Sigma$.

Now flatten Σ as in [21]. In words, notice that h and $-\nabla h$, when restricted to Σ, generically induce a CW decomposition of Σ in which 0-cells are the index 0 critical points of $h|_\Sigma$, one point in the interior of each 1-cell is an index 1 critical point of $h|_\Sigma$, and one point in the interior of each 2-cell is an index 2 critical point of $h|_\Sigma$. Perturb, if necessary, so that self-intersections of Σ all lie outside the descending and ascending manifolds in Σ of index 1 critical points of $h|_\Sigma$.

The family of gradient flowlines of ∇h in X which originate on the ascending manifolds of an index 1 critical point of $h|_\Sigma$ is 2-dimensional, as is the family of gradient flowlines of $-\nabla h$ in X which originate on the descending manifolds of an index 1 critical point of $h|_\Sigma$. Thus, we may generically take them all to be disjoint and also disjoint from ascending and descending manifolds of index 2 points of h. (We discuss this more in Section 2.3. While this condition is generic, it is not natural — this lack of generality precisely corresponding to the singular band moves of Theorem 2.39.)

Fix $\varepsilon > 0$, and let $L^- = \Sigma \cap h^{-1}\bigl(\frac{3}{2} - \varepsilon\bigr)$. Isotope Σ near height $\frac{3}{2}$ so that $\Sigma \cap h^{-1}\bigl(\bigl[\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\bigr]\bigr)$ is of the form $L^- \times \bigl[\frac{3}{2} - \varepsilon, \frac{3}{2} + \varepsilon\bigr]$. A neighborhood of each 1-cell of Σ can be isotoped via $-\nabla h$ to a band in $h^{-1}\bigl(\frac{3}{2}\bigr)$ that is attached to a parallel copy of L^-. Let B be the collection of all such bands (one for each 1-cell in Σ).

Now isotope Σ near each self-intersection s of Σ as in Figure 2, right, ie make one of the sheets of Σ at s include a small region that is horizontal with respect to h and which contains s. Isotope this sheet via $-\nabla h$ to push this horizontal region to $h^{-1}\bigl(\frac{3}{2}\bigr)$, where it can be interpreted as a marked fat vertex as in Figure 2, left. Repeating for every self-intersection of Σ, we obtain a marked singular banded link L in $h^{-1}\bigl(\frac{3}{2}\bigr)$ whose negative resolution is L^-.

Now Σ intersects regions of X in the following ways:

$$\begin{aligned} h^{-1}\bigl(\bigl[0, \tfrac{3}{2} - \varepsilon\bigr]\bigr) \quad & \text{in boundary parallel disks with boundary } L^-, \\ h^{-1}\bigl(\bigl[\tfrac{3}{2} - \varepsilon, \tfrac{3}{2} + \varepsilon\bigr]\bigr) \quad & \text{in the realizing surface segment for } (L, B), \\ h^{-1}\bigl(\bigl[\tfrac{3}{2} + \varepsilon, \tfrac{5}{2}\bigr]\bigr) \quad & \text{in an embedded surface on which } h \text{ has no critical points}, \\ h^{-1}\bigl(\bigl[\tfrac{5}{2}, 4\bigr]\bigr) \quad & \text{in boundary parallel disks with boundary } L^+_B. \end{aligned}$$

We conclude that Σ is isotopic to $\Sigma(\mathcal{H}, L, B)$. □

Remark 2.18 In the proof of Lemma 2.17, we made several references to genericity. That is, we made several choices of how to perturb Σ in order to obtain $(\mathcal{H}, L, B)$. It may be helpful to imagine the lower-dimensional analogue of knots in S^3: every knot in S^3 is isotopic to one that projects to a knot diagram. However, not every knot in S^3 actually projects to a knot diagram. An arbitrary knot may, for example, have a projection that includes a cusp, self-tangency or triple point. These conditions are not generic and can be corrected by a slight perturbation, but therein involves a choice that can yield diagrams differing by a Reidemeister move (RI, RII or RIII, respectively). There are, of course, even "worse" conditions, such as a knot whose projection involves a quadruple intersection. However, this condition is even "less" generic, by which we mean:

- A generic knot in S^3 admits a projection with no triple points.
- A generic 1-parameter family of smoothly varying knots in S^3 admits projections with finitely many triple points but no quadruple points.
- A generic 2-parameter family of smoothly varying knots in S^3 admits projections with 1-dimensional families of triple points and finitely many quadruple points.

Thus, in a 1-parameter family of knots (ie a knot isotopy), we expect to obtain diagrams that differ by an RIII move (and similarly for RI and RII), but need never consider moves involving quadruple intersections.

Moving back to the 4-dimensional world, in order to understand to what extent a singular banded unlink diagram is well defined up to isotopy of an immersed surface, we must understand which nongeneric behaviors of projections we expect to see a finite number of times in a 1-dimensional family of immersed surfaces. We discuss this more formally in Sections 2.3 and 2.4.

2.2.5 Singular band moves The Kirby diagram $\mathcal{K}$ only determines the described 4-manifold X up to diffeomorphism. Therefore, $(\mathcal{K}, L, B)$ only determines the pair $(X, \Sigma(\mathcal{K}, L, B))$ up to diffeomorphism; it does not make sense to say that $(\mathcal{K}, L, B)$ determines $\Sigma(\mathcal{K}, L, B)$ up to isotopy. However, if we have already identified X with the manifold described by $\mathcal{K}$, then we can consider $\Sigma(\mathcal{K}, L, B)$ up to isotopy. In particular, given another singular banded unlink diagram $(\mathcal{K}, L', B')$ in the same Kirby diagram $\mathcal{K}$, there is a natural (up to isotopy) diffeomorphism between the 4-manifolds containing $\Sigma(\mathcal{K}, L', B')$ and $\Sigma(\mathcal{K}, L, B)$. Therefore, it *does* make sense to ask whether $\Sigma(\mathcal{K}, L, B')$ and $\Sigma(\mathcal{K}, L, B)$ are ambiently isotopic, regularly homotopic or homotopic in X. In this section, we define moves of singular banded unlink diagrams that describe ambient isotopies of immersed surfaces; in Sections 2.3 and 2.4, we show that indeed these moves are sufficient.

Definition 2.19 Let $\mathcal{D} := (\mathcal{K}, L, B)$ and $\mathcal{D}' := (\mathcal{K}, L', B')$ be singular banded unlink diagrams. We say that *$\mathcal{D}'$ is related to $\mathcal{D}$ by singular band moves* if $\mathcal{D}'$ is obtained from $\mathcal{D}$ by a finite sequence of the moves in Figures 3 and 4, which we call *singular band moves*. (This relationship is clearly symmetric.)

Specifically, the singular band moves (illustrated in Figures 3 and 4) are

(i) isotopy in $E(\mathcal{K})$,

(ii) cup/cap moves,

(iii) band slides,

(iv) band swims,

(v) slides of bands over components of L_2 (band/2-handle slide),

(vi) swims of bands about L_2 (band/2-handle swim),

(vii) slides of unlinks and bands over L_1,

(viii) sliding a vertex over a band (intersection/band slide),

(ix) passing a vertex past the edge of a band (intersection/band pass).

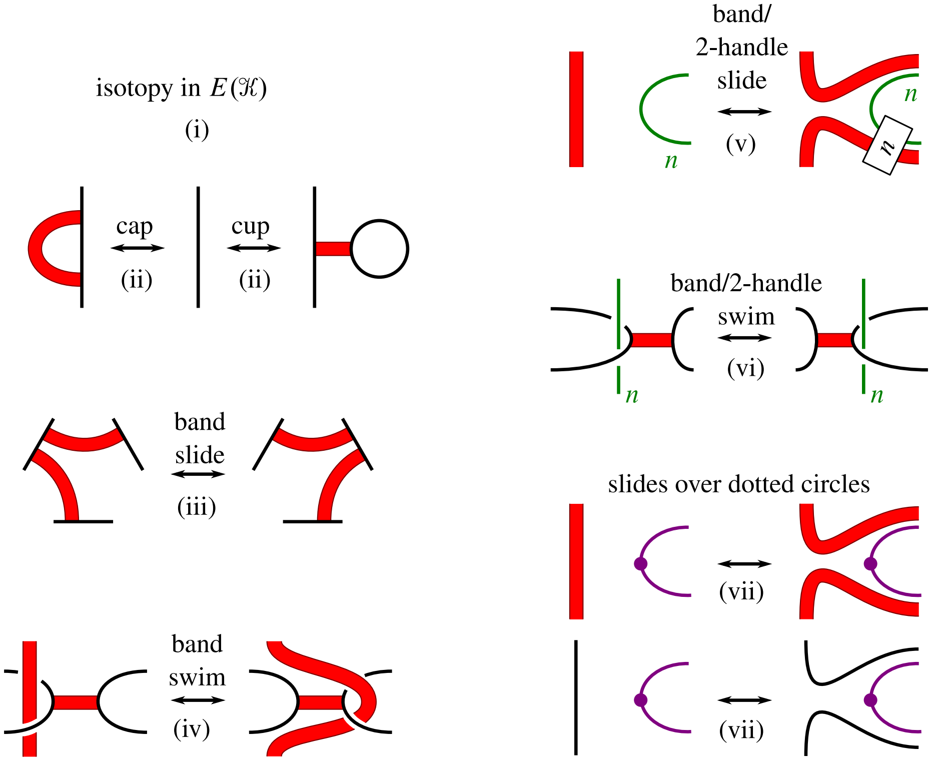


Figure 3: The band moves that do not involve the self-intersections of the described surface.

We may refer to moves (i)–(vii) (illustrated in Figure 3) as *band moves* (omitting the word "singular") since they do not involve the self-intersections of L. The remaining moves are illustrated in Figure 4.

Exercise 2.20 If $\mathcal{D}$ and $\mathcal{D}'$ are related by singular band moves, then $\Sigma(\mathcal{D})$ and $\Sigma(\mathcal{D}')$ are ambiently isotopic.

In the future, we will refer to moves by name rather than number to avoid confusion.

In Figures 5–10, we illustrate some other useful moves on singular banded unlink diagrams that are achievable by a combination of singular band moves. We call these moves $\star$ (Figure 5), the intersection/band swim (Figure 6), the upside-down intersection/band swim (Figure 7), the intersection pass (Figure 8), the intersection swim (Figures 9 and 10), the intersection/2-handle slide (Figure 11) and the intersection/2-handle swim (Figure 12).

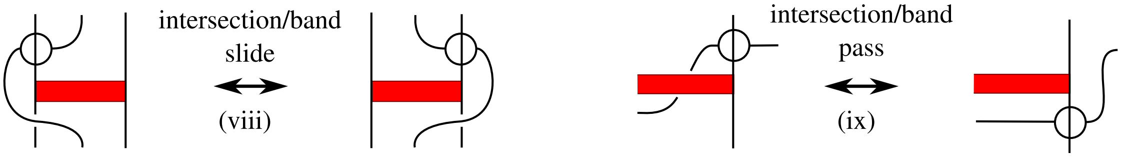


Figure 4: The singular band moves that involve self-intersections of the described surface.

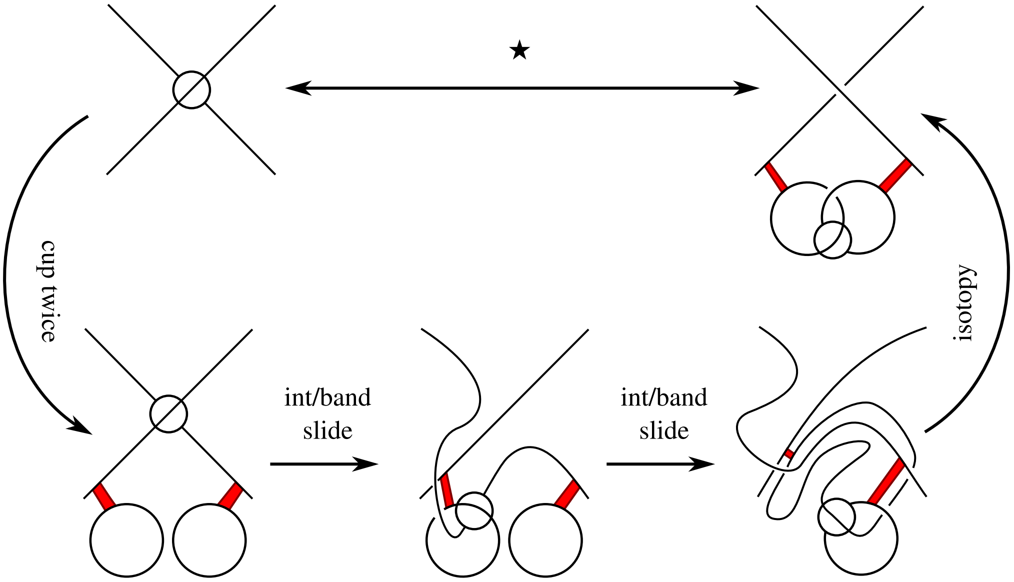


Figure 5: The ⋆ move moves a vertex onto two new unlink components (or the reverse). In Figures 7, 9 and 10, we see that the ⋆-move can be used (in conjunction with singular band moves) to achieve other seemingly natural moves.

In an earlier version of this paper, we included the intersection/band swim of Figure 6 as one of the singular band moves (as move (x)). Jablonowski [15] noticed that this move is redundant, so we have modified the list accordingly.

Remark 2.21 While the length of the list in Definition 2.19 may seem unwieldy, there is a general principle at play: singular band moves allow us to isotope a singular banded unlink (L, B) within $\mathscr{K}$; or to

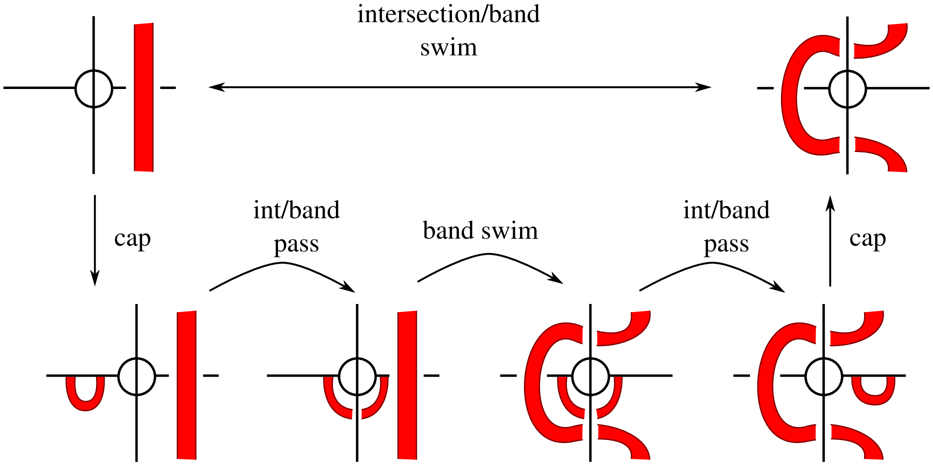


Figure 6: We can achieve an intersection/band swim by performing singular band moves. This sequence of moves was observed by Jablonowski [15].

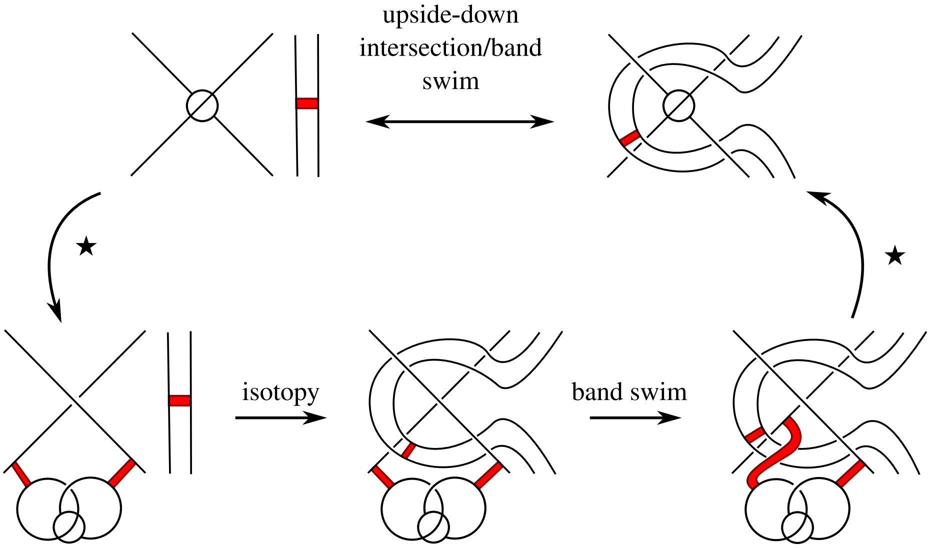


Figure 7: We can achieve the upside-down intersection/band swim by performing ⋆ and singular band moves.

push any vertex in L or band in B slightly into the past or future, do further isotopy there, and then push the vertex or band back into the present. In practice, when using these diagrams, we do not explicitly break a described isotopy into a sequence of the moves of Definition 2.19, just as how in practice one does not typically break an isotopy of a knot explicitly into a sequence of Reidemeister moves.

2.3 Ascending/descending manifolds and 0- and 1-standard surfaces

So far, we have only used singular banded unlink diagrams to describe realizing surfaces, which are incredibly nongeneric. One goal of this paper is to use singular banded unlink diagrams to describe any

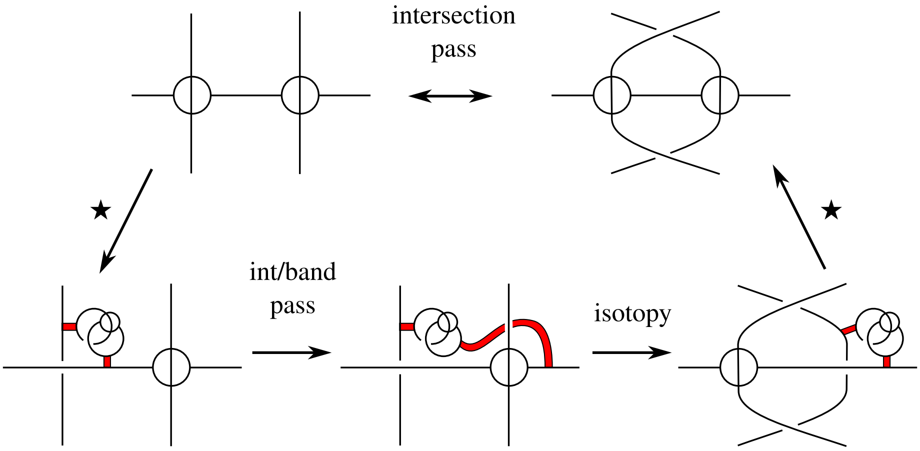


Figure 8: We can achieve an intersection pass by performing ⋆ and singular band moves.

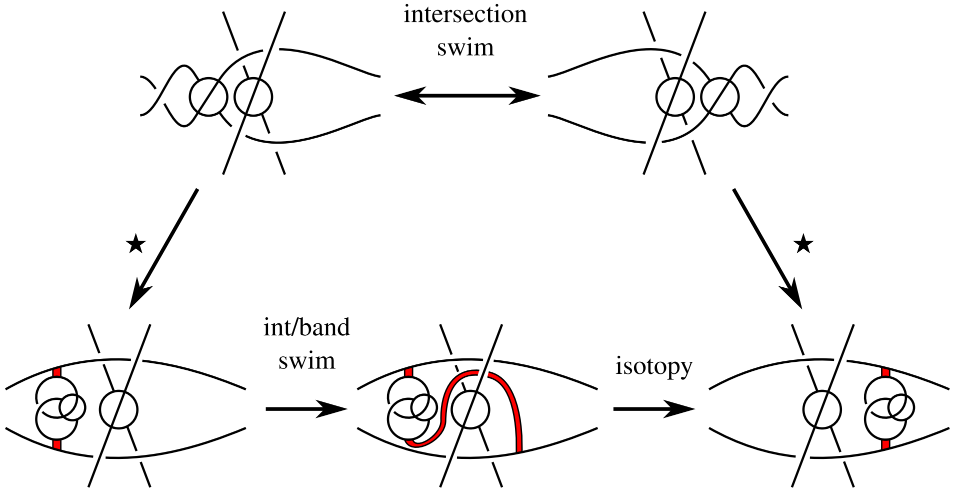


Figure 9: We can achieve an intersection swim by performing ⋆ and singular band moves.

self-transverse immersed surface Σ. In Lemma 2.17, we showed that any such Σ is isotopic to a realizing surface. However, it is not obvious that any two realizing surfaces isotopic to Σ have singular banded unlink diagrams that are related by singular band moves. In order to prove this, we must first restrict ourselves to understanding surfaces that intersect the ascending and descending manifolds of critical points of h in prescribed ways, but yet are still more generic than realizing surfaces.

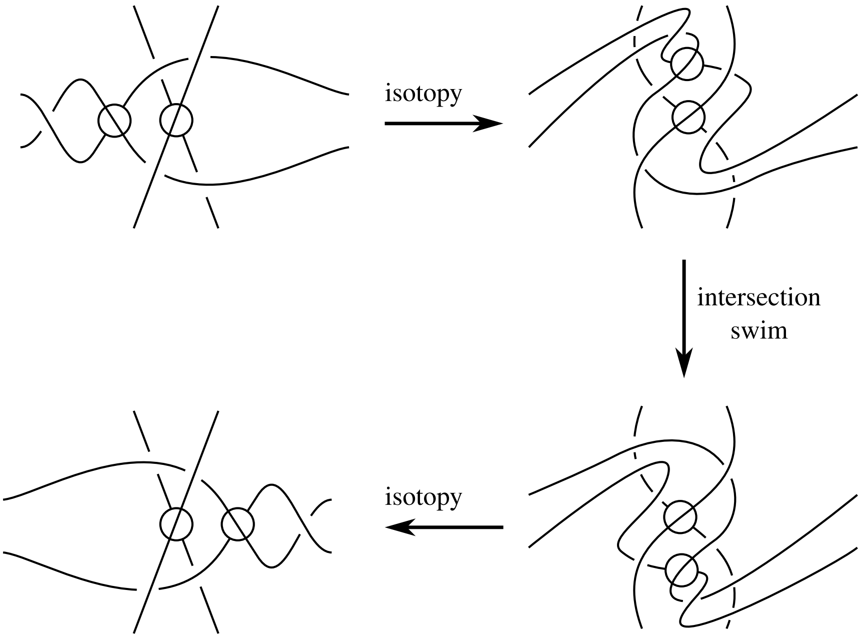


Figure 10: We achieve an alternative version of the intersection swim of Figure 9, in which one marking and one crossing are changed via isotopy and intersection swim.

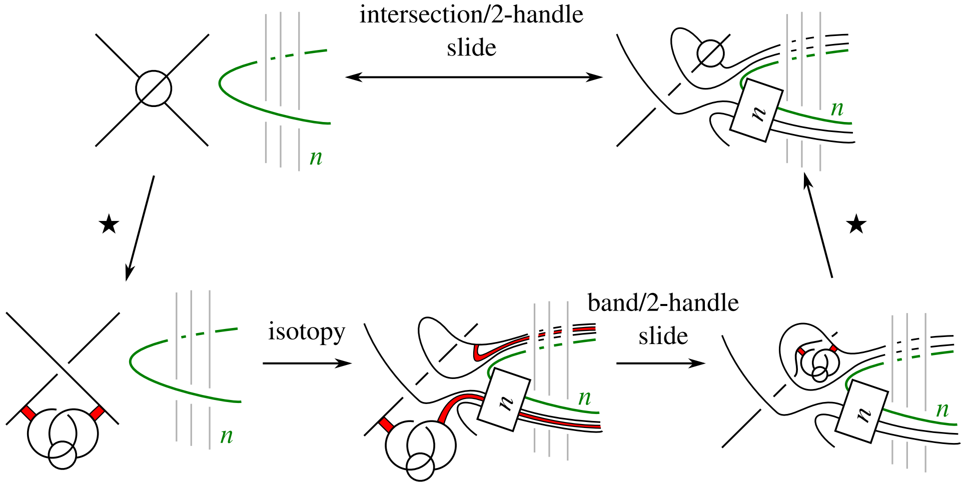


Figure 11: We achieve an intersection/2-handle slide by performing $\star$ and singular band moves.

We will now consider not only the ascending/descending manifolds of critical points of h, but also the ascending and descending manifolds of critical points of the restricted Morse function $h|_{\Sigma}$. From now on, fix a gradientlike vector field ∇h for the Morse function $h\colon X \to \mathbb{R}$, and let Z denote $X^4 \setminus \nu(\Sigma)$.

In order to obtain a gradientlike vector field on Σ itself, we choose a splitting $TX|_{\Sigma} = T\Sigma \oplus N$ and let $\mathrm{proj}_{T\Sigma}\colon TX|_{\Sigma} \to T\Sigma$ be the associated bundle projection. We can assume that the splitting is chosen so

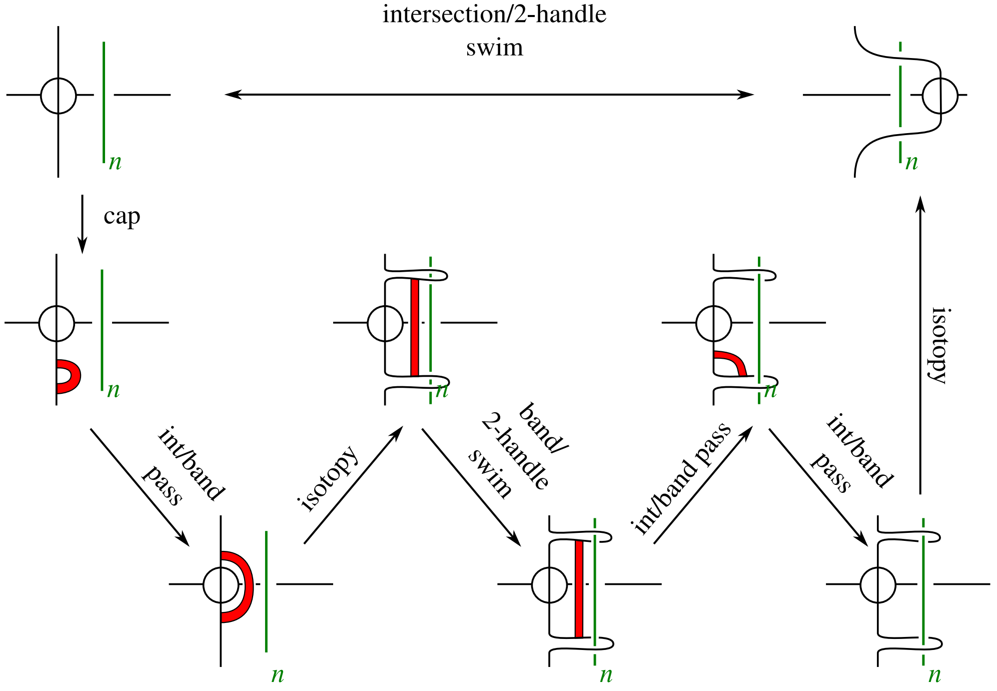


Figure 12: We achieve an intersection/2-handle swim by performing singular band moves.

that $\mathrm{proj}_{T\Sigma}(\nabla h)|_\Sigma$ is a gradientlike vector field for $h|_\Sigma$ on Σ, which we denote by $\nabla(h|_\Sigma)$. Note that this is actually *not* a vector field on the immersed surface Σ (although we could pull it back to a vector field on the abstract surface F), since there are two associated vectors at each point of self-intersection of Σ (the projections of ∇h onto the tangent planes of each local sheet); however, we think that the language "gradientlike vector field" is not confusing in this context. The vector field $\nabla(h|_\Sigma)$ is *not* canonically determined by h, ∇h and Σ, since to obtain it we have to choose a splitting of TX_Σ.

In what follows, we will often refer to the ascending or descending manifolds of critical points of $h|_\Sigma$ or of self-intersections of Σ. Unless we specify otherwise, assume that this always refers to the corresponding manifolds in X with respect to ∇h as defined above, rather than ascending or descending manifolds in Σ with respect to $\nabla(h|_\Sigma)$. These points are generally not critical points of h, but their ascending and descending manifolds can be studied as usual.

2.3.1 1-standard surfaces

Suppose that Σ is a self-transverse immersed surface in X. The following definition will be important as we consider 1-parameter families of immersed surfaces:

Definition 2.22 We say that Σ is 1-*standard* if the following are true:

(1) The surface Σ is disjoint from the critical points of h.

(2) The restriction $h|_\Sigma$ is Morse except for possibly at most one birth/death degeneracy, ie a point of Σ about which $h|_\Sigma$ can be represented as $h|_\Sigma(x,y)=x^2-y^3$ in some local coordinates on Σ.

(3) For $k\geq n+1$, the descending manifolds (with respect to ∇h) of index n critical points of h and index $n-1$ critical points of $h|_\Sigma$ are disjoint from the ascending manifolds of index k critical points of h and index $k-1$ critical points of $h|_\Sigma$. Moreover, self-intersections of Σ are disjoint from the ascending manifolds of index 3 critical points of h and descending manifolds of index 1 critical points of h. In other words, we ask for n-dimensional descending manifolds to be disjoint from $(4-n-1)$-dimensional ascending manifolds.

Remark 2.23 Definition 2.22 is essentially a list of all ascending/descending manifold pairs that we expect to be disjoint in a 1-parameter family of immersed surfaces by dimensional considerations, as explained in Proposition 2.24. This motivates the name "1-standard".

Proposition 2.24 *Let Σ_t be an isotopy between 1-standard surfaces Σ_0 and Σ_1. After an arbitrarily small perturbation of the isotopy Σ_t, we can assume that Σ_t is 1-standard for all t.*

Proof We prove that, after a small perturbation, Σ_t satisfies each property of Definition 2.22 for all t.

(1) The critical point set of h in $X\times I$ is 1-dimensional, while the isotopy Σ_t in $X\times I$ is 3-dimensional. Generically, we do not expect Σ_t to intersect a critical point of h for any t.

(2) This follows from Cerf's filtration on the space of surfaces (see eg [9, Chapter 1, Section 2]). This is a filtration on the space $C(F)$ of all smooth maps $F\to X^4$ for F a surface. The codimension-0

stratum consists of all maps $f\colon F \to X^4$ with $h|_{f(F)}$ Morse with critical points at distinct heights. The codimension-1 stratum includes f if either of the following is true:

- The restriction $h|_{f(F)}$ is Morse with exactly two critical points at the same height, but all other critical points sit at distinct heights.
- The restriction $h|_{f(F)}$ is Morse except for one birth or death degeneracy. This degeneracy and all critical points are at distinct heights.

Suppose Σ_0 has n points of self-intersection. Fix $2n$ points $x_1, y_2, \dots, x_n, y_n$ in F and choose $f_t\colon F \to X$ so that $f_t(F) = \Sigma_t$ and $f_t(x_i) = f_t(y_i)$ for all i and t. Now a small perturbation of the path f_t from f_0 to f_1 in $C(F)$ yields a path g_t that is completely contained in the codimension-0 and codimension-1 strata of Cerf's filtration with $g_0 = f_0, g_1 = f_1$. Since g_t lies in these strata, $g_t(F)$ has property (2) of Definition 2.22 for all t. Moreover, if the perturbation is sufficiently small, we may assume that $g_t(F)$ is an immersed surface with n transverse double points for all t, all of which are contained in a fixed small tubular neighborhood of Σ_t. (Recall that smooth or PL self-transversely immersed surfaces in 4-manifolds have tubular neighborhoods; use local coordinates to choose a tubular neighborhood near each of the finitely many self-intersections and then extend over the whole surface using the tubular neighborhood theorem.)

While g_t is a homotopy from g_0 to g_1, we may view its image as an isotopy between the singular submanifolds Σ_0 and Σ_1 in X. We must now check that this isotopy extends to an ambient isotopy of X. That is, while we have argued that we may perturb f_t to achieve property (2), we must explain why this perturbation may be achieved by perturbing the ambient isotopy from Σ_0 to Σ_1, since there is a distinction between the immersions f_t and their images $f_t(F) = \Sigma_t$. This is relatively standard (and indeed stated without proof in eg [7]): choose small disjoint closed disks D_{x_i}, D_{y_i} $(i = 1, \dots, n)$ in F, centered at x_i and y_i, respectively. We can fix a family of coordinates on a closed tubular neighborhood of $g_t(F)$ near the self-intersections such that, centered about $g_t(x_i) = g_t(y_i)$, we have a closed ball $B_i = g_t(D_{x_i}) \times g_t(D_{y_i})$ intersecting $g_t(F)$ in

$$\frac{(g_t(D_{x_i}) \times \{0\}) \cup (\{0\} \times g_t(D_{y_i}))}{(g_t(x_i) \times 0) \sim (0 \times g_t(y_i))}.$$

Now we may extend the isotopy $\Sigma_0 \to \Sigma_1$ that is the image of g_t to an isotopy ϕ_t of $\Sigma_0 \cup B_1 \cup \cdots \cup B_n$ by specifying that $\phi_t(g_0(a), g_0(b)) = (g_t(a), g_t(b))$ for all $a \in D_{x_i}, b \in D_{y_i}$, since $B_i = g_0(D_{x_i}) \times g_0(D_{y_i})$. Then $\phi_t(B_i) = g_t(D_{x_i}) \times g_t(D_{y_i})$. Since the B_i are balls, the isotopy $\phi_t|_{\bigcup_i B_i}$ extends to an ambient isotopy ψ_t of X. The composition $\psi_t^{-1}\phi_t$ then fixes B_i pointwise for each i.

Now, since $\Sigma_0 \cap (X \setminus \operatorname{int}(B_1 \sqcup \cdots \sqcup B_n))$ is an embedded submanifold (whose boundary is not tangent to the boundary of $X \setminus \operatorname{int}(B_1 \cup \cdots \cup B_n)$; ie $\Sigma_0 \cap (X \setminus \operatorname{int}(B_1 \sqcup \cdots \sqcup B_n))$ is neat in the sense of [12]) whose boundary is fixed by $\psi_t^{-1}\phi_t$, we may use usual isotopy extension to extend $\psi_t^{-1}\phi_t$ to an ambient isotopy. Then, since ψ_t^{-1} is an ambient isotopy (and hence a diffeotopy starting at the identity map), we conclude that ϕ_t extends to a diffeotopy starting at the identity map, ie an ambient isotopy.

We conclude that our original ambient isotopy from Σ_0 to Σ_1 may be perturbed to another ambient isotopy of Σ_0 to Σ_1 which satisfies property (2) of 1-standardness at all times.

(3) Note that both ascending and descending manifolds are parallel to ∇h, so, rather than counting transverse intersections, we count the dimension of the space of line intersections (parallel to ∇h) of these ascending and descending manifolds. (In other words, we count the dimension of the moduli space of unparametrized flowlines of $-\nabla h$ from one critical or intersection point to another.) An n-dimensional descending manifold and a $(4-k)$-dimensional ascending manifold thus have expected dimension

$$(n-1)+((4-k)-1)-(4-1)=n-k-1$$

as a space of lines. For $k \geq n+1$, this expected dimension is at most -2, so we conclude that we may perturb Σ_t (which by the previous item we see may be obtained by perturbing a path of immersions f_t in $C(F)$) to achieve property (3). □

2.3.2 0-standard surfaces In Remark 2.23, we explained that the definition of 1-standardness comes from studying generic 1-parameter families. That is, the conditions in Definition 2.22 are generically true for 1-parameter families of surfaces. We now define a slightly more restrictive condition on the surfaces we study, which we expect to be violated a finite number of times in a generic 1-parameter family.

Definition 2.25 We say that Σ is 0-*standard* if it is 1-standard and the following are true:

(1) The restriction $h|_\Sigma$ is Morse.

(2) Whenever p and q are either index 2 critical points of h, index 1 critical points of $h|_\Sigma$, or self-intersections of Σ (not necessarily of the same type), and $p \neq q$, the descending manifold of p is disjoint from the ascending manifold of q. In short: 2-dimensional descending manifolds are disjoint from 2-dimensional ascending manifolds.

Remark 2.26 Roughly speaking, a surface Σ is 0-standard if its index 1 critical points (viewed as bands) and self-intersections do not lie above each other, or above or below any index 2 critical points of h. This is all with respect to ∇h; we are *not* discussing $\nabla(h|_\Sigma)$. These forbidden conditions, allowed in a 1-standard surface, would cause a projection of Σ to a singular banded unlink diagram to not be well defined, motivating the cup/cap moves, band swims, band/2-handle slides and swims. Most of the other singular band moves are related to the choice of $\nabla(h|_\Sigma)$ (specifically the band slide, intersection/band slide and pass, and intersection pass). Isotopy in $E(\mathcal{K})$ and slides over L_1 correspond to horizontal isotopy.

Proposition 2.27 *Let Σ_t be an isotopy between* 0*-standard surfaces Σ_0 and Σ_1. After an arbitrarily small perturbation of the isotopy Σ_t, it is true that Σ_t is* 1*-standard for all t, and* 0*-standard for all but finitely many t.*

Proof It follows from Proposition 2.24 that 1-parameter families Σ_t of surfaces are generically 1-standard for all t. We now consider the conditions of Definition 2.25 separately.

(1) This is well known, due to Cerf (see eg [9, Chapter 1, Section 2]).

(2) A pair of complementary-dimension descending and ascending manifolds meet with expected dimension -1 (as a space of lines parallel to ∇h). Therefore, property (2) is generically true at all but finitely many times during a 1-parameter family of surfaces. □

Proposition 2.28 *Suppose Σ is 0-standard. Fix $\nabla(h|_\Sigma)$ with the property that, for p and q distinct index 1 points of $h|_\Sigma$ or self-intersections of Σ, the descending manifold of p with respect to $\nabla(h|_\Sigma)$ is disjoint from the ascending manifold of q with respect to $\nabla(h|_\Sigma)$. Then there is a singular banded unlink diagram $\mathcal{D}$ determined by Σ, ∇h and $\nabla(h|_\Sigma)$ up to isotopy and slides over the 1-handle circles L_1.*

Proof Since Σ is 0-standard (and hence 1-standard), we may vertically isotope Σ so that the minima of $h|_\Sigma$ lie below $h^{-1}\left(\frac{3}{2}\right)$, the maxima of $h|_\Sigma$ lie above $h^{-1}\left(\frac{5}{2}\right)$, and the self-intersections/bands of Σ lie in $h^{-1}\left(\left(\frac{3}{2},\frac{5}{2}\right)\right)$.

By assumption, the descending manifolds (using $\nabla(h|_\Sigma)$) of index 1 critical points of $h|_\Sigma$ end at index 0 points of $h|_\Sigma$ without meeting any index 1 points or self-intersections of Σ. Similarly, flowlines of $-\nabla(h|_\Sigma)$ originating at self-intersections of Σ also end at index 0 points of $h|_\Sigma$ without meeting any other index 1 critical points or self-intersections of Σ.

Now let S be the 1-skeleton of Σ determined by $\nabla(h|_\Sigma)$, ie the 1-complex with

(1) 0-cells at index 0 points of $h|_\Sigma$,

(2) 1-cells along the descending manifolds of index 1 critical point of $h|_\Sigma$,

(3) additional 1-cells consisting of pairs of flowlines of $-\nabla(h|_\Sigma)$ glued together at self-intersections of Σ.

Isotope Σ vertically so that the index 1 critical points of $h|_\Sigma$ and self-intersections of Σ lie disjointly in $h^{-1}\left(\frac{3}{2}\right)$. (Here we are implicitly using the fact that, since Σ is 0-standard, these points do not lie directly above one another nor above index 2 critical points of h.) Flatten Σ near $h^{-1}\left(\frac{3}{2}\right)$ to turn index 1 points of $h|_\Sigma$ into bands whose cores are contained in 1-cells of S.

Since Σ is 0-standard, the bands and self-intersections of $\Sigma \cap h^{-1}\left(\frac{3}{2}\right)$ are disjoint from the descending manifolds of index 2 critical points of h, ie they are disjoint from the attaching circles L_2 of the 2-handles in $\mathcal{H}$.

Then $\Sigma \cap h^{-1}\left(\frac{3}{2}\right)$ is a singular banded link (L, B), where L^- is isotopic to $\Sigma \cap h^{-1}\left(\frac{3}{2}-\varepsilon\right)$, and L_B^+ is isotopic to $\Sigma \cap h^{-1}\left(\frac{3}{2}+\varepsilon\right)$. We conclude that (L, B) is well defined up to isotopy in $h^{-1}\left(\frac{3}{2}\right) \setminus$(descending manifolds of index 2 critical points of h). Therefore, $(\mathcal{H}, L, B)$ is well defined up to slides of L and B over the dotted circles L_1 of $\mathcal{H}$. □

Corollary 2.29 *Let Σ_0 and Σ_1 be 0-standard surfaces. Suppose there is an isotopy Σ_t from Σ_0 to Σ_1 that is 0-standard for all t, with $\nabla(h|_{\Sigma_1})$ obtained from $\nabla(h|_{\Sigma_0})$ by the isotopy-induced map on $T\Sigma$. Then the singular banded unlink diagrams $\mathcal{D}_0$ and $\mathcal{D}_1$ for $\mathcal{H}_0$ and $\mathcal{H}_1$ produced by Proposition 2.28 are related by isotopy in $E(\mathcal{H})$ and slides over L_1.*

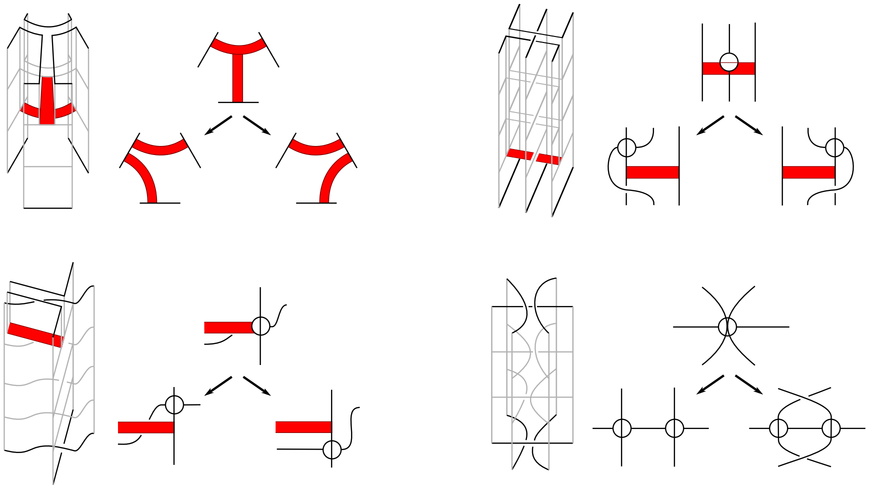


Figure 13: The cases of Proposition 2.30. At the left of each quadrant we draw a local model about the flowline of $-\nabla(h|_{\Sigma_{1/2}})$ that causes Proposition 2.28 to not apply. At the top right of each quadrant, we draw a schematic of the projection of $\Sigma_{1/2}$ to $E(\mathscr{K})$, where two bands, two self-intersections, or one of each coincide. We draw arrows to indicate the two diagrams that arise if we perturb $\Sigma_{1/2}$ to be 0-standard.

We can improve Proposition 2.28 by considering the difference between two choices for $\nabla(h|_\Sigma)$. First note that, if V_0 and V_1 are two such vector fields, then by considering the expected dimension of the space of flowlines between critical points of a Morse function on a surface, we find that V_0 and V_1 are isotopic through a sequence V_t of gradientlike vector fields for $\nabla(h|_\Sigma)$ with the property that, for all but finitely many t, V_t satisfies the conditions of Proposition 2.28. We can take the exceptional $V_{t_1},\dots,V_{t_n}$ to each satisfy the conditions of Proposition 2.28 except for one disallowed flowline from an index 1 point or self-intersection to another (not necessarily the same type).

Proposition 2.30 *Suppose V_t satisfies the conditions of Proposition 2.28 except for $t \neq \frac{1}{2}$. Let $\mathscr{D}_0$ and $\mathscr{D}_1$ be the singular banded unlink diagrams obtained from Σ as in Proposition 2.28 using V_0 and V_1, respectively. Then $\mathscr{D}_0$ and $\mathscr{D}_1$ are related by isotopy in $E(K)$, slides over L_1, and possibly a band slide, intersection/band slide, intersection/band pass or intersection pass.*

Proof Let p and q be the index 1 or self-intersection points in Σ with a flowline of $-V_{1/2}$ from p to q. The proof of Proposition 2.28 fails for $\Sigma_{1/2}$ precisely because p lying above q in Σ causes indeterminacy in the 1-skeleton S. There are then two choices (up to small isotopy through 0-standard surfaces) in how to perturb Σ near p to obtain a 0-standard surface. See Figure 13. The resulting two singular banded unlink diagrams differ by one of the following moves:

$$\left.\begin{array}{r}\text{band slide}\\ \text{intersection/band slide}\\ \text{intersection/band pass}\\ \text{intersection pass}\end{array}\right\}\ \text{if}\ \left\{\begin{array}{l} p \text{ and } q \text{ are index 1 points}\\ p \text{ is a self-intersection and } q \text{ is an index 1 point}\\ p \text{ is an index 1 point and } q \text{ is a self-intersection}\\ p \text{ and } q \text{ are self-intersections}\end{array}\right.$$

Letting $\mathcal{D}_t$ denote the diagram obtained using V_t for $t \neq \frac{1}{2}$, we conclude that $\mathcal{D}_{1/2-\varepsilon}$ and $\mathcal{D}_{1/2+\varepsilon}$ are either isotopic or isotopic after one of the above moves. The same is then true of $\mathcal{D}_0$ and $\mathcal{D}_1$ by Corollary 2.29. □

The following proposition and corollary now follow immediately from Propositions 2.28 and 2.30:

Proposition 2.31 *Suppose Σ is 0-standard. Then there is a singular banded unlink diagram $\mathcal{D}$ determined by $\Sigma, \nabla h$ up to isotopy in $E(\mathcal{K})$, slides over L_1, band slides, intersection/band slides, intersection/band passes and intersection passes.*

Corollary 2.32 *Let Σ_0 and Σ_1 be 0-standard surfaces. Suppose there is an isotopy Σ_t from Σ_0 to Σ_1 that is 0-standard for all t, with $\nabla(h|_{\Sigma_1})$ obtained from $\nabla(h|_{\Sigma_0})$ by the isotopy-induced map on $T\Sigma$. Then $\mathcal{D}_0$ and $\mathcal{D}_1$ are related by isotopy in $E(\mathcal{K})$, slides over L_1, band slides, intersection/band slides, intersection/band passes and intersection passes.*

2.4 Conclusion: uniqueness of singular banded unlink diagrams

2.4.1 Singular band moves and isotopy We are now in a position to prove our main results.

Theorem 2.33 *Let Σ_0 and Σ_1 be 0-standard self-transverse immersed surfaces. Suppose there exists an isotopy Σ_t such that Σ_t is 1-standard for all t, and 0-standard for all $t \neq \frac{1}{2}$.*

Set $\mathcal{D}_t := \mathcal{D}(\Sigma_t)$. Then $\mathcal{D}_0$ and $\mathcal{D}_1$ are related by singular band moves.

We break Theorem 2.33 into Propositions 2.34–2.37, in which we separately consider different ways in which $\Sigma_{1/2}$ may fail to be 0-standard.

Proposition 2.34 *Suppose that $\Sigma_{1/2}$ would be 0-standard if not for a single birth or death degeneracy. Then $\mathcal{D}_0$ and $\mathcal{D}_1$ are related by the singular band moves appearing in Proposition 2.30 and possibly a cup or cap move.*

Proof Combined with Corollary 2.32, this is a standard fact about the local model of a degenerate critical point appearing in a generic 1-parameter family of Morse functions. See eg [1]. □

Proposition 2.35 *Suppose that $\Sigma_{1/2}$ would be 0-standard if not for the descending manifold of p with respect to ∇h meeting the ascending manifold of q with respect to ∇h, where p and q are each index 1 critical points of $h|_\Sigma$ or self-intersections of Σ, and their ascending/descending manifolds intersect in their interiors (rather than in just Σ, as in Proposition 2.30). Then $\mathcal{D}_0$ and $\mathcal{D}_1$ are related by the singular band moves appearing in Proposition 2.30 and possibly a band swim, intersection/band swim, upside-down intersection/band swim or intersection swim.*

Proof The proof of Proposition 2.31 fails for $\Sigma_{1/2}$ because, when we attempt to project the 1-skeleton of Σ to $h^{-1}(\frac{3}{2})$, the edges corresponding to p and q will intersect. There are then two choices (up to small isotopy through 0-standard surfaces) in how to perturb Σ near p to obtain a 0-standard surface. See

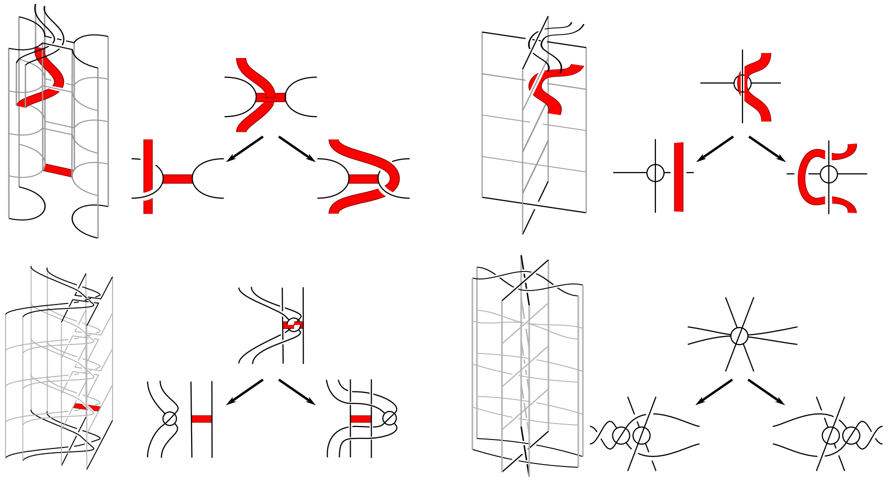


Figure 14: The cases of Proposition 2.35. At the left of each quadrant we draw a local model about the flowline that causes $\Sigma_{1/2}$ to not be 0-standard. At the top right of each quadrant, we draw a schematic of the projection of $\Sigma_{1/2}$ to $E(\mathcal{K})$, where two bands, two self-intersections, or one of each coincide. We draw arrows to indicate the two diagrams that arise if we perturb $\Sigma_{1/2}$ to be 0-standard.

Figure 14. The resulting two singular banded unlink diagrams differ by one of the following moves:

$$\left.\begin{array}{r}\text{band swim}\\ \text{intersection/band swim}\\ \text{upside-down intersection/band swim}\\ \text{intersection swim}\end{array}\right\}\quad\text{if}\quad\left\{\begin{array}{l}p \text{ and } q \text{ are index 1 points}\\ p \text{ is a self-intersection } q \text{ is an index 1 point}\\ p \text{ is an index 1 point } q \text{ is a self-intersection}\\ p \text{ and } q \text{ are self-intersections}\end{array}\right.$$

We conclude that $\mathcal{D}_{1/2-\varepsilon}$ and $\mathcal{D}_{1/2+\varepsilon}$ are either isotopic or isotopic after one of the above moves. The same is then true of $\mathcal{D}_0$ and $\mathcal{D}_1$ (up to the relevant moves) by Corollary 2.32. □

Proposition 2.36 *Suppose that $\Sigma_{1/2}$ would be 0-standard if not for the descending manifold of p intersecting the ascending manifold of q, where p is an index 1 critical point of $h|_\Sigma$ or a self-intersection of Σ and q is an index 2 critical point of h. Then $\mathcal{D}_0$ and $\mathcal{D}_1$ are related by the singular band moves appearing in Proposition 2.30 and possibly a band/2-handle slide or intersection/2-handle slide.*

Proof The proof of Proposition 2.31 fails because we cannot project the edge of the 1-skeleton of Σ corresponding to p to the level $h^{-1}\left(\frac{3}{2}\right)$. There are then two choices (up to small isotopy through 0-standard surfaces) in how to perturb Σ near p to obtain a 0-standard surface, with resulting singular banded unlink diagrams differing by a slide over a 2-handle. That is, the resulting two singular banded unlink diagrams differ by one of the following moves:

$$\left.\begin{array}{r}\text{band/2-handle slide}\\ \text{intersection/2-handle slide}\end{array}\right\}\quad\text{if}\quad\left\{\begin{array}{l}p \text{ is an index 1 point}\\ p \text{ is a self-intersection}\end{array}\right.$$

We conclude that $\mathcal{D}_{1/2-\varepsilon}$ and $\mathcal{D}_{1/2+\varepsilon}$ are either isotopic or isotopic after one of the above moves. The same is then true of $\mathcal{D}_0$ and $\mathcal{D}_1$ (up to the relevant moves) by Corollary 2.32. □

Proposition 2.37 *Suppose that $\Sigma_{1/2}$ would be 0-standard if not for the descending manifold of p intersecting the ascending manifold of q, where p is an index 2 critical point of h and q is either an index 1 critical point of $h|_\Sigma$ or a self-intersection of Σ. Then $\mathcal{D}_0$ and $\mathcal{D}_1$ are related by the singular band moves appearing in Proposition 2.30 and possibly a band/2-handle swim or intersection/2-handle swim.*

Proof The proof of Proposition 2.31 fails for $\Sigma_{1/2}$ because after we project the 1-skeleton of Σ to $h^{-1}\left(\frac{3}{2}\right)$, the edge corresponding to q will intersect the component of $L_2 \subset \mathcal{K}$ corresponding to p. There are then two choices (up to small isotopy through 0-standard surfaces) in how to perturb Σ near p to obtain a 0-standard surface, with resulting singular banded unlink diagrams differing by a swim through a 2-handle attaching circle. That is, the resulting two singular banded unlink diagrams differ by one of the following moves:

$$\left.\begin{array}{r}\text{band/2-handle swim}\\ \text{intersection/2-handle swim}\end{array}\right\} \quad \text{if} \quad \left\{\begin{array}{l}p \text{ is an index 1 point}\\ p \text{ is a self-intersection}\end{array}\right.$$

We conclude that $\mathcal{D}_{1/2-\varepsilon}$ and $\mathcal{D}_{1/2+\varepsilon}$ are either isotopic or isotopic after one of the above moves. The same is then true of $\mathcal{D}_0$ and $\mathcal{D}_1$ (up to the relevant moves) by Corollary 2.32. □

This completes the proof of Theorem 2.33, since Propositions 2.34–2.37 cover all of the cases in which $\Sigma_{1/2}$ is 1-standard and not 0-standard (of course, if $\Sigma_{1/2}$ is 0-standard then Theorem 2.33 follows from Corollary 2.32) except for the case that there are flowlines of $-\nabla h$ between index 2 critical points. However, h and ∇h are fixed during the isotopy, so this does not happen. □

Corollary 2.38 *Let Σ_0 and Σ_1 be 0-standard self-transverse immersed surfaces. Suppose there exists an isotopy Σ_t and values $t_1 < t_2 < \cdots < t_n \in (0,1)$ such that Σ_t is 0-standard for all $t \notin \{t_1, t_2, \dots, t_n\}$, and Σ_{t_i} is 1-standard for each $i = 1, 2, \dots, n$.*

Let $\mathcal{D}_t := \mathcal{D}(\Sigma_t)$. Then $\mathcal{D}_0$ and $\mathcal{D}_1$ are related by a sequence of singular band moves.

Proof For each $i = 1, \dots, n-1$, let s_i be a value in (t_i, t_{i+1}). By Corollary 2.32:

- $\mathcal{D}_0$ is related to $\mathcal{D}_{s_1}$ by singular band moves.
- $\mathcal{D}_{s_i}$ is related to $\mathcal{D}_{s_{i+1}}$ by singular band moves for $i = 1, \dots, n-1$.
- $\mathcal{D}_{s_{n-1}}$ is related to $\mathcal{D}_1$ by singular band moves.

We conclude that $\mathcal{D}_0$ and $\mathcal{D}_1$ are related by singular band moves. □

2.4.2 Proof of uniqueness theorems We finally prove that singular banded unlink diagrams of isotopic (resp. regularly homotopic, homotopic) surfaces exist for arbitrary immersed self-transverse surfaces and are well defined up to singular band moves. At this point, not much is left to do — the material in Section 2.4 is essentially the whole proof that diagrams exist and are unique up to singular band moves.

Theorem 2.39 *Let Σ be a self-transverse smoothly immersed surface in X. Then there is a singular banded unlink diagram $\mathfrak{D}(\Sigma)$, well defined up to singular band moves, such that Σ is isotopic to the closed realizing surface for $\mathfrak{D}(\Sigma)$. Moreover, if Σ is isotopic to Σ', then $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ are related by singular band moves.*

We say that $\mathfrak{D}(\Sigma)$ is *a singular banded unlink diagram for* Σ, or simply that $\mathfrak{D}(\Sigma)$ is a diagram for Σ.

Proof Via a small perturbation, Σ is isotopic to a 0-standard surface Σ_0. Set $\mathfrak{D}(\Sigma) := \mathfrak{D}(\Sigma_0)$. To show that $\mathfrak{D}(\Sigma)$ is well defined, suppose that Σ_1 is another 0-standard surface that is isotopic to Σ, and hence isotopic to Σ_0. By Proposition 2.27, there is an isotopy Σ_t from Σ_0 to Σ_1 such that Σ_t is 1-standard for all t and 0-standard for all but finitely many t. By Corollary 2.38, $\mathfrak{D}(\Sigma_0)$ and $\mathfrak{D}(\Sigma_1)$ are related by singular band moves.

Since this argument applies to any 0-standard surface Σ_1 isotopic to Σ, we conclude that, if Σ and Σ' are isotopic, then $\mathfrak{D}(\Sigma)$ and $\mathfrak{D}(\Sigma')$ are related by singular band moves. □

Corollary 2.40 *Let $\mathfrak{D}$ and $\mathfrak{D}'$ be singular banded unlink diagrams of surfaces Σ and Σ' immersed in diffeomorphic 4-manifolds X and X'. There is a diffeomorphism taking (X, Σ) to (X', Σ') if and only if there is a sequence of singular band moves and Kirby moves taking $\mathfrak{D}$ to $\mathfrak{D}'$.*

In addition, we can use these moves to describe homotopies of surfaces in terms of singular banded unlink diagrams.

Corollary 2.41 *Let $\mathfrak{D}$ and $\mathfrak{D}'$ be singular banded unlink diagrams for surfaces Σ and Σ' immersed in X. If Σ and Σ' are homotopic, then $\mathfrak{D}$ and $\mathfrak{D}'$ are related by a finite sequence of singular band moves and the following moves (illustrated in Figure 15):*

- *introducing or canceling two oppositely marked vertices (a "finger move" or "Whitney move"), as illustrated;*
- *replacing a nugatory crossing with a vertex or vice versa (a "cusp move"), as illustrated.*

In addition, if Σ and Σ' are **regularly** *homotopic, then $\mathfrak{D}$ and $\mathfrak{D}'$ are related by a finite sequence of singular band moves, finger moves and Whitney moves (ie a sequence of the given moves that does not include any cusp moves).*

Proof Say Σ and Σ' are homotopic and have self-intersection numbers s and s', respectively. By work of Hirsch [11] and Smale [31], Σ and Σ' are regularly homotopic if and only if $s = s'$.

After performing a cusp move on $\mathfrak{D}$, a realizing surface for the resulting diagram has self-intersection $s \pm 1$, with sign depending on the choice of cusp move. Perform $|s' - s|$ cusp moves of the appropriate sign to $\mathfrak{D}$ to obtain a diagram $\mathfrak{D}_2$ whose realizing surface Σ_2 has self-intersection number s'. Now Σ_2 and Σ' are regularly homotopic.

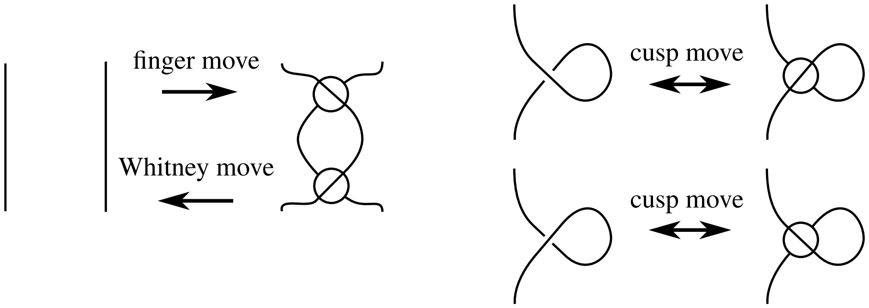


Figure 15: The new moves describing homotopy of a surface in a 4-manifold. There are two versions of the cusp move. One involves a positive self-intersection and one involves a negative self-intersection of the described immersed surface. To describe regular homotopy we only need finger and Whitney moves.

We recommend the reference [7] for exposition on regular homotopy of surfaces. In brief, there exists a sequence of finger moves on Σ_2 along framed arcs $\eta_1, \dots, \eta_n$ yielding a surface Σ_3, and a sequence of finger moves on Σ' along framed arcs $\eta'_1, \dots, \eta'_m$ yielding a surface Σ'', so that Σ_3 and Σ'' are ambiently isotopic.

We isotope η_1 to lie completely in $h^{-1}\left(\frac{3}{2}\right)$ (which may involve isotopy of Σ_2 inducing singular band moves on its singular banded unlink diagram according to Theorem 2.39) and then shrink η_1 to be short and contained in a neighborhood identical to the far left of Figure 15. Twist the diagram as necessary so

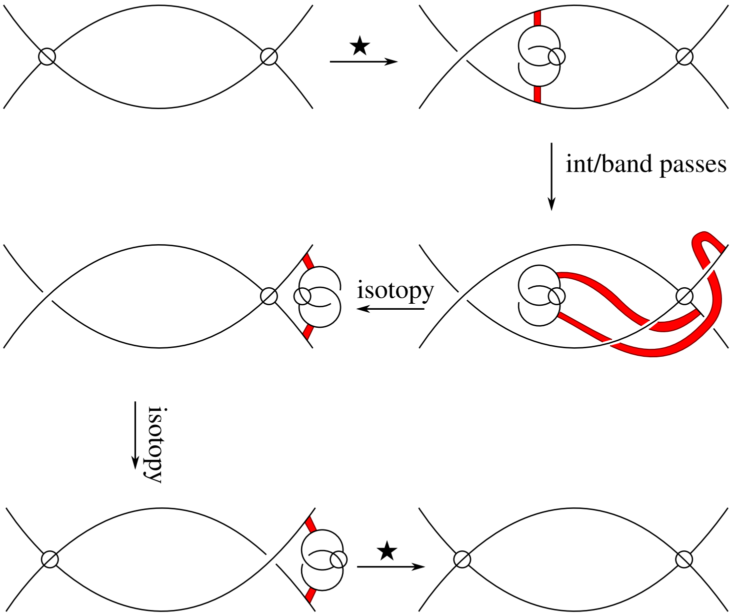


Figure 16: There are two seemingly different finger moves (differing in the decorations on the relevant vertices), but they yield singular banded unlink diagrams that differ by singular band moves.

that the framing of η_1 is untwisted. Then we perform a finger move to $\mathfrak{D}_2$ in that neighborhood. Repeat for each $i = 2, \dots, n$, and call the resulting diagram $\mathfrak{D}_3$. A realizing surface for $\mathfrak{D}_3$ is isotopic to Σ_3.

Now repeat for Σ' by performing singular band moves and finger moves to its diagram $\mathfrak{D}'$ until obtaining a diagram $\mathfrak{D}''$ whose realizing surface is isotopic to Σ''. Since Σ'' and Σ_3 are isotopic, by Theorem 2.39 it follows that $\mathfrak{D}_3$ and $\mathfrak{D}''$ are related by singular band moves.

We thus conclude that $\mathfrak{D}$ can be transformed into $\mathfrak{D}'$ by a sequence of singular band moves, cusp moves, finger moves and Whitney moves (which are the inverses to finger moves). □

Remark 2.42 When performing a finger move to a singular banded unlink diagram, there are seemingly two choices (related by a local symmetry) of how to mark the new vertices. However, the choices yield diagrams related by singular band moves, as shown in Figure 16.

3 Bridge trisections

3.1 Bridge trisections of embedded surfaces

In Section 3.2, we prove that self-transverse immersed surfaces in 4-manifolds can be put into *bridge position*, a notion introduced for embedded surfaces by Meier and Zupan [26; 27]. Meier and Zupan showed that a bridge trisection of a surface in S^4 (with respect to a standard trisection of S^4) is unique up to perturbation [26], using the work of Swenton [32] and Kearton and Kurlin [22] on banded unlink diagrams in S^4. The authors of this paper then used a general version of this theorem in arbitrary 4-manifolds to show that bridge trisections of surfaces in any trisected manifold are unique up to perturbation. In what follows, we will apply Theorem 2.39 to prove an analogous uniqueness result for bridge trisections of immersed surfaces. In this section, we will review the situation where the surface is embedded.

First, we recall the definition of a trisection of a closed 4-manifold. Similar exposition can be found in [14]. We do not require much knowledge of trisections; for more detailed exposition, the interested reader may refer to [8].

Definition 3.1 [8] Let X^4 be a connected, closed, oriented 4-manifold. A (g,k)*-trisection* of X^4 is a triple (X_1, X_2, X_3) where

(i) $X_1 \cup X_2 \cup X_3 = X^4$,

(ii) $X_i \cong \natural_{k_i} S^1 \times B^3$,

(iii) $X_i \cap X_j = \partial X_i \cap \partial X_j \cong \natural_g S^1 \times B^2$,

(iv) $X_1 \cap X_2 \cap X_3 \cong \Sigma_g$,

where Σ_g is a closed orientable surface of genus g. Here, g is an integer while $k = (k_1, k_2, k_3)$ is a triple of integers. If $k_1 = k_2 = k_3$, then the trisection is said to be *balanced*.

Briefly, a trisection is a decomposition of a 4-manifold into three elementary pieces, analogous to a Heegaard splitting of a 3-manifold into two elementary pieces.

Theorem 3.2 [8] *Any smooth, connected, closed, oriented* 4*-manifold* X^4 *admits a trisection. Moreover, any two trisections of* X^4 *are related by a stabilization operation.*

Note that, from the definition, Σ_g is a Heegaard surface for ∂X_i, inducing the Heegaard splitting $(X_i \cap X_j, X_i \cap X_k)$. By Laudenbach and Poénaru [24], X^4 is specified by its *spine*, $\Sigma_g \cup \bigcup_{i \neq j}(X_i \cap X_j)$. Therefore, we usually describe a trisection (X_1, X_2, X_3) by a *trisection diagram* $(\Sigma_g; \alpha, \beta, \gamma)$. Here each of α, β and γ consist of g independent curves on Σ_g (abusing notation to take Σ_g as both an abstract surface and the surface $X_1 \cap X_2 \cap X_3$ in X), which bound disks in the handlebodies $X_1 \cap X_2$, $X_2 \cap X_3$ and $X_1 \cap X_3$, respectively. Given (X_1, X_2, X_3), such a diagram is well defined up to slides of α, β and γ and automorphisms of Σ_g.

Definition 3.3 Let X^4 be a 4-manifold with trisection $\mathcal{T} = (X_1, X_2, X_3)$. We say that an isotopy f_t of X^4 is *$\mathcal{T}$-regular* if $f_t(X_i) = X_i$ for each $i = 1, 2, 3$ and for all t.

Definition 3.4 The *standard trisection of* S^4 is the unique $(0,0)$-trisection (X_1^0, X_2^0, X_3^0). View $S^4 = \mathbb{R}^4 \cup \infty$, with coordinates (x, y, r, θ) on $\mathbb{R}^4$, where (x, y) are Cartesian planar coordinates and (r, θ) are polar planar coordinates. Up to isotopy, $X_i^0 = \{\theta \in [\frac{2}{3}\pi i, \frac{2}{3}\pi(i+1)]\} \cup \infty$. Then $X_i^0 \cong B^4$, $X_i^0 \cap X_{i+1}^0 = \{\theta = \frac{2}{3}\pi(i+1)\} \cup \infty \cong B^3$, and $X_1^0 \cap X_j^0 \cap X_k^0 = \{r = 0\} \cup \infty \cong S^2$.

From a trisection (X_1, X_2, X_3) of X^4, we can obtain a handle decomposition of X^4 in which X_1 contains the 0- and 1-handles, X_2 is built from $(X_1 \cap X_2) \times I$ by attaching the 2-handles, and X_3 contains the 3- and 4-handles. The following definition encapsulates this construction:

Definition 3.5 Let $\mathcal{T} = (X_1, X_2, X_3)$ be a trisection of a 4-manifold X^4. Let $h\colon X^4 \to [0, 4]$ be a self-indexing Morse function. We say that h is *$\mathcal{T}$-compatible* if all of the following are true:

(i) $X_1 = h^{-1}([0, \frac{3}{2}])$.

(ii) $X_2 \subset h^{-1}([\frac{3}{2}, \frac{5}{2}])$ contains all of the index 2 critical points of h.

(iii) $X_1 \cup X_2$ contains the descending manifolds of all index 2 critical points of h.

Given any trisection $\mathcal{T}$, there always exists a Morse function compatible with $\mathcal{T}$ (see [8] or [25]).

Meier and Zupan used trisections to give a new way of describing a surface smoothly embedded in a 4-manifold.

Definition 3.6 [26; 27] Let $\mathcal{T} = (X_1, X_2, X_3)$ be a trisection of a closed 4-manifold X^4. Let S be a surface embedded in X^4. We say that S is in (b, c)-bridge position with respect to $\mathcal{T}$ if, for every $i \neq j \in \{1, 2, 3\}$:

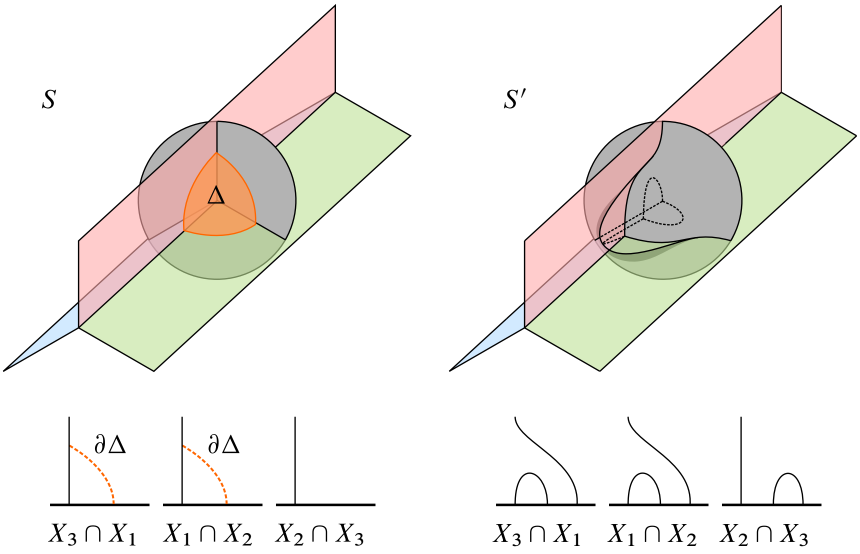


Figure 17: Left: a surface S in (b,c)-bridge position with respect to a trisection $\mathcal{T}$. We draw a neighborhood of an intersection of S with the central surface of $\mathcal{T}$. Right: we perturb S to obtain a surface S' in $(c', b+1)$-bridge position.

(i) $S \cap X_i$ is a disjoint union of c_i boundary parallel disks.

(ii) $S \cap X_i \cap X_j$ is a trivial tangle of b arcs.

Here b is an integer and $c = (c_1, c_2, c_3)$ is a triple of integers. Note that $\chi(S) = \sum c_i - b$.

Theorem 3.7 [26; 27] *Let S be a surface embedded in a 4-manifold X^4 with trisection $\mathcal{T} = (X_1, X_2, X_3)$. Then, for some c and b, S can be isotoped into (b,c)-bridge position with respect to $\mathcal{T}$. We may take $c_1 = c_2 = c_3$.*

Because a collection of boundary parallel disks in $\natural(S^1 \times B^3)$ is uniquely determined by its boundary (up to isotopy rel boundary), a surface S in bridge position is determined up to isotopy by $S \cap \big(\bigcup_{i \neq j} X_i \cap X_j\big)$.

There is a natural perturbation of a surface in bridge position, analogous to perturbation of a knot in bridge position within a 3-manifold. We define the simplest version of Meier and Zupan's original perturbation operation [26; 27].

Definition 3.8 Let $S \subset X^4$ be a surface in (b,c)-bridge position with respect to $\mathcal{T} = (X_1, X_2, X_3)$. Let S' be the surface obtained from S as in Figure 17. In words, we take a small disk D contained in $S \cap X_1$ whose boundary consists of an arc δ_1 in the interior of X_1, an arc δ_2 in $X_1 \cap X_2$, and an arc δ_3 in $X_3 \cap X_1$. We take a parallel copy Δ of D pushed off S away from δ_1, so Δ meets S in the arc $\delta_1 \subset \partial\Delta$ and the remaining boundary of Δ is an arc δ' in ∂X_1 that meets $X_1 \cap X_2 \cap X_3$ transversely in one point. Using the direction from which we obtained Δ from D, we frame Δ and isotope S along Δ to introduce two more intersection points between S and $X_1 \cap X_2 \cap X_3$. We call the resulting surface S' and say that S' is obtained from S by *elementary perturbation*. We likewise say that S is obtained from S' by *elementary deperturbation*.

We may exchange the roles of X_1, X_2 and X_3 cyclically when performing this operation, ie alternatively obtain S' from this compression operation in either X_2 or X_3. We still say S' is obtained from S by elementary perturbation and that S is obtained from S' by elementary deperturbation.

Proposition 3.9 [27, Lemma 5.2] *Let S be a surface in (b,c)-bridge position with respect to a trisection $\mathcal{T} = (X_1, X_2, X_3)$, with $c = (c_1, c_2, c_3)$. Let S' be obtained from S by elementary perturbation, using a disk in X_i. Then S' is in $(c', b{+}1)$-bridge position with respect to $\mathcal{T}$, with $c'_j = c_j$ for $j \neq i$ and $c'_i = c_i + 1$.*

In previous work, the authors of this paper showed that any two bridge trisections of a surface are related by elementary perturbations.

Theorem 3.10 [14] *Let S and S' be surfaces in bridge position with respect to a trisection $\mathcal{T}$ of a 4-manifold X^4. Suppose S is isotopic to S'. Then S can be taken to S' by a sequence of elementary perturbations and deperturbations, followed by a $\mathcal{T}$-regular isotopy.*

When $\mathcal{T}$ is the standard trisection of S^4, Theorem 3.10 is a result of Meier and Zupan [26].

3.2 Basic definitions for singular links and immersed surfaces

In Definition 3.6 of a bridge trisection of an embedded surface, we cut a 4-manifold into simple pieces so that an embedded surface is cut into systems of boundary-parallel disks. To describe immersed surfaces, we need to describe this notion with slightly different language.

Definition 3.11 Let $C_1, \dots, C_k$ be arcs properly immersed in a 3-manifold M^3. Assume that all intersections (including self-intersections) of $C_1, \dots, C_k$ are isolated points that are not tangencies. Let $V = (\partial M^3) \times I$ be a collar neighborhood of ∂M^3 and let $h\colon V \to I$ be projection onto the second factor.

We say that $(C_1, \dots, C_k)$ is a *trivial immersed tangle* if the following are satisfied:

(i) Each C_i is contained in V.

(ii) All self-intersections of C_i and intersections of C_i with C_j are contained in the interior of M.

(iii) There is an immersed tangle $(C'_1, \dots, C'_k)$ that is isotopic rel boundary to $(C_1, \dots, C_k)$ so that $h|C'_i$ is Morse with a single critical point for all i.

Definition 3.12 Let $D_1, \dots, D_k$ be 2-dimensional disks properly immersed in a 4-manifold X^4. Assume that all intersections (including self-intersections) of $D_1, \dots, D_k$ are isolated, transverse intersections contained in ∂X^4 (so $\partial\big(\bigcup D_i\big)$ is a singular link in ∂X). Let $V = \partial X \times I$ be a neighborhood of ∂X and let $h\colon V \to I$ be projection onto the second factor.

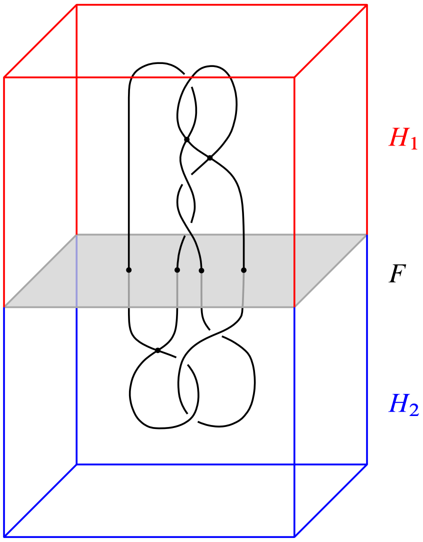


Figure 18: A singular link in bridge position.

We say that $(D_1, \dots, D_k)$ is a *trivial immersed disk system* if the following are satisfied (up to isotopy rel boundary):

(i) Each D_i is contained in V.

(ii) The restriction $h|D_i$ is Morse with a single critical point for all i.

Trivial immersed tangles and disk systems are the immersed analogue to systems of boundary parallel embedded tangles and disks. With immersed tangles we can easily define an analogue of bridge position for singular links.

Definition 3.13 Let L be a singular link in a 3-manifold M with a Heegaard splitting (H_1, H_2). Let $F := H_1 \cap H_2$.

We say that L is in *bridge position* with respect to F if $L \cap H_i$ is a trivial immersed tangle for $i = 1, 2$. See Figure 18. If (L, σ) is a marked singular link, then we say that (L, σ) is in *bridge position* if L is in bridge position.

We can perturb immersed tangles just as we perturb embedded tangles, but we must also account for vertices.

Definition 3.14 Let L be a marked singular link in a 3-manifold M with Heegaard splitting (H_1, H_2). Suppose L is in bridge position with respect to $\Sigma := H_1 \cup H_2$.

Let L' be a marked singular link obtained from L by perturbation near Σ, as in Figure 19. Note that we allow up to one vertex of L to be between the original intersection of L with Σ and the newly created pair of intersections. Then we say L' is obtained from L by *elementary perturbation*, and L is obtained from L' by *elementary deperturbation*.

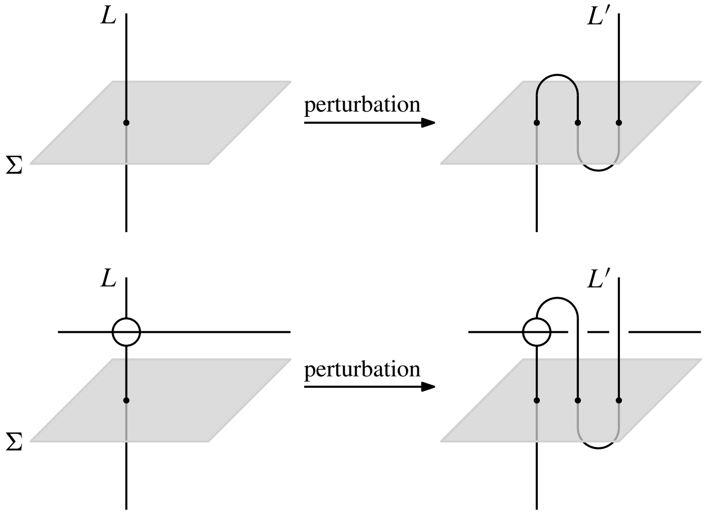


Figure 19: An elementary perturbation of a marked singular link in bridge position.

Let L'' be a marked singular link obtained from L by moving a vertex in L through Σ as in the local model shown in Figure 20. Then we say L'' is obtained from L (and vice versa) by *vertex perturbation*.

Theorem 3.15 *Let L and L' be isotopic marked singular links in a 3-manifold M with Heegaard splitting (H_1, H_2). Assume L and L' are in bridge position with respect to $\Sigma := H_1 \cap H_2$. Then there exists a marked singular link L'' that can be obtained from L and from L' by sequences of elementary perturbations, vertex perturbations and isotopies fixing Σ setwise.*

Proof When L and L' are nonsingular, this is a theorem of Hayashi and Shimokawa [10]. We will apply a version of this theorem for nonsingular banded links due to Meier and Zupan [26; 27] by using the following observation. First, recall from Section 2.1 that, if L is a marked singular link, then L^+ denotes the nonsingular link obtained by positively resolving the vertices of L.

Observation 3.16 *There exist disjoint framed arcs $a_1, \dots, a_n$ with endpoints on L^+ such that contracting L^+ along $a_1, \dots, a_n$ yields L.*

Similarly, let $a'_1, \dots, a'_n$ be framed arcs with endpoints on L'^+ such that contracting L'^+ along $a'_1, \dots, a'_n$ yields L'.

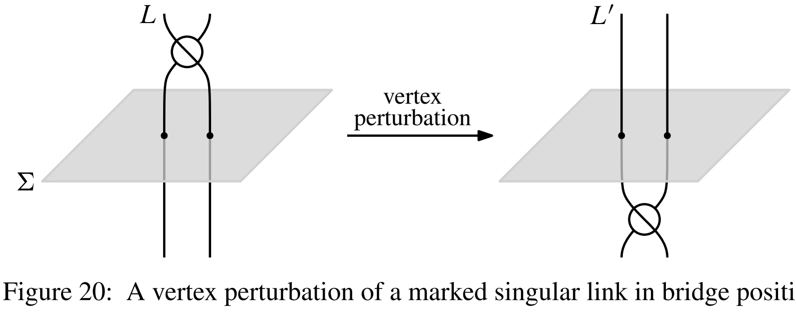


Figure 20: A vertex perturbation of a marked singular link in bridge position.

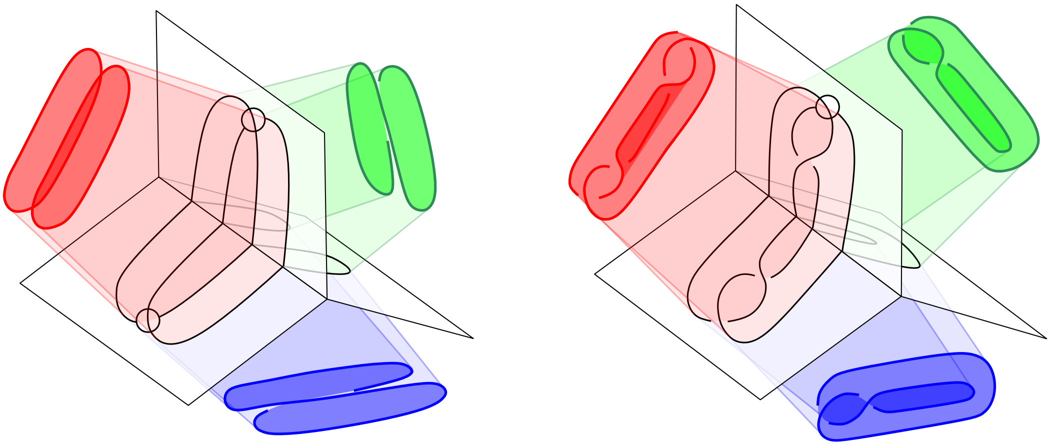


Figure 21: Two $((2,1,1),2)$-bridge trisections of immersed 2-spheres in S^4. Left: this 2-sphere has a pair of self-intersections of opposite sign. Right: this 2-sphere has a single self-intersection.

Now, by Meier and Zupan [26; 27], there exists a link J that can be obtained from L^+ and from L'^+ by elementary perturbations and isotopies fixing Σ setwise. Moreover, these isotopies and perturbations may be chosen to carry a_i and a_i' to framed arcs b_i and b_i', respectively, with endpoints on J, so that b_i and b_i' are parallel to Σ with surface framing, and are parallel to each other (though possibly on opposite sides of Σ). In Meier and Zupan's construction, during this sequence of perturbations and isotopies of L^+ (resp. L'^+), a_i (resp. a_i') never intersect Σ, so these perturbations and isotopies may be achieved by perturbations and isotopies of L (resp. L'). Let $\hat{J}$ and $\hat{J}'$ be the marked singular links obtained by contracting J along $\bigcup b_i$ and $\bigcup b_i'$, respectively, and with markings induced by those of L and L'. Then $\hat{J}'$ can be transformed into $\hat{J}$ by isotopy fixing Σ and a vertex perturbation for each pair a_i, a_i' separated in different components of $M \setminus \Sigma$. Therefore, the claim holds with $L'' = \hat{J}$. □

3.3 Bridge trisections of immersed surfaces

We now use the definitions from Section 3.2 to define bridge trisections of self-transverse immersed surfaces.

Definition 3.17 Let $\mathcal{T} = (X_1, X_2, X_3)$ be a trisection of a closed 4-manifold X^4. Let S be a self-transverse immersed surface in X^4. We say that S is in (b,c)-bridge position with respect to $\mathcal{T}$ if, for each $i \neq j \in \{1,2,3\}$:

(i) $S \cap X_i$ is a trivial immersed disk system of c_i disks.

(ii) $S \cap X_i \cap X_j$ is a trivial immersed tangle of b strands.

Here, b is a positive integer and $c = (c_1, c_2, c_3)$ is a triple of positive integers.

In Figure 21, we give some small examples of bridge trisections of 2-spheres immersed in S^4.

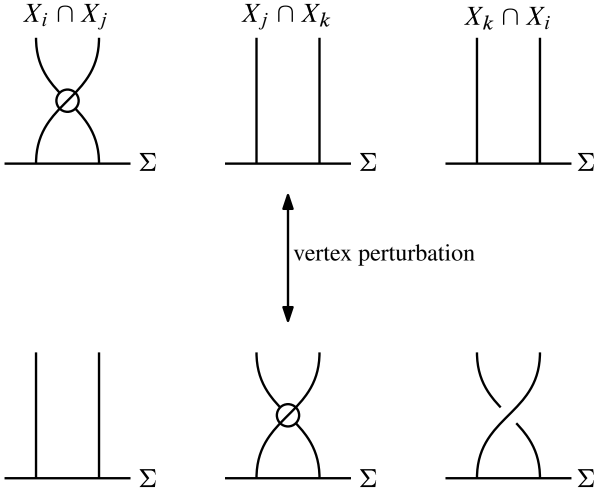


Figure 22: A vertex perturbation of a triplane diagram.

There is again a natural notion of perturbing an immersed surface in (b,c)-bridge position. More precisely, the notion of perturbing an embedded surface in bridge position works perfectly well for an immersed surface in bridge position. We write the definition below, believing that the value of transparency outweighs the cost of redundancy.

Definition 3.18 Let S be a self-transverse immersed surface in bridge position with respect to a trisection $\mathcal{T}=(X_1,X_2,X_3)$. In Figure 17, we depict a small neighborhood of a point in $S\cap\Sigma$ for $\Sigma:=X_1\cap X_2\cap X_3$. Let S' be the surface obtained from S as in Figure 17. We say that S' is obtained from S by *elementary perturbation*, and that S is obtained from S' by *elementary deperturbation*.

If S is in bridge position with respect to a trisection $\mathcal{T}=(X_1,X_2,X_3)$, then elementary perturbation and $\mathcal{T}$-regular isotopy cannot move a self-intersection of S from X_i to X_j for $i\neq j$. Thus, we introduce one new kind of perturbation for immersed surfaces in bridge position, based on the most elementary way one might move a self-intersection of S from X_i to X_j.

Definition 3.19 Let v be a vertex of the singular link $S\cap X_i\cap X_{i+1}$ for some i (where the indices are understood to be taken mod 3), so that v is a self-intersection of S. Suppose v has a neighborhood as in Figure 22, so that v is near $\Sigma:=X_1\cap X_2\cap X_3$. We may isotope S to move v into Σ and then into either $X_{i+1}\cap X_{i+2}$ or $X_{i-1}\cap X_i$, producing a new surface S' in (b,c)-bridge position. See Figures 22 and 23. We say that S' is obtained from S (and vice versa) by *vertex perturbation*.

Remark 3.20 Let S be an immersed surface in $(b;c_1,c_2,c_3)$-bridge position with respect to $\mathcal{T}=(X_1,X_2,X_3)$.

(1) If S' is obtained from S by elementary perturbation along a disk in X_i, then S' is in $(b+1;c_1',c_2',c_3')$-bridge position with $c_i'=c_i+1$ and $c_j'=c_j$ for $j\neq i$.

(2) If S' is obtained from S by vertex perturbation, then S' is in $(b;c_1,c_2,c_3)$-bridge position.

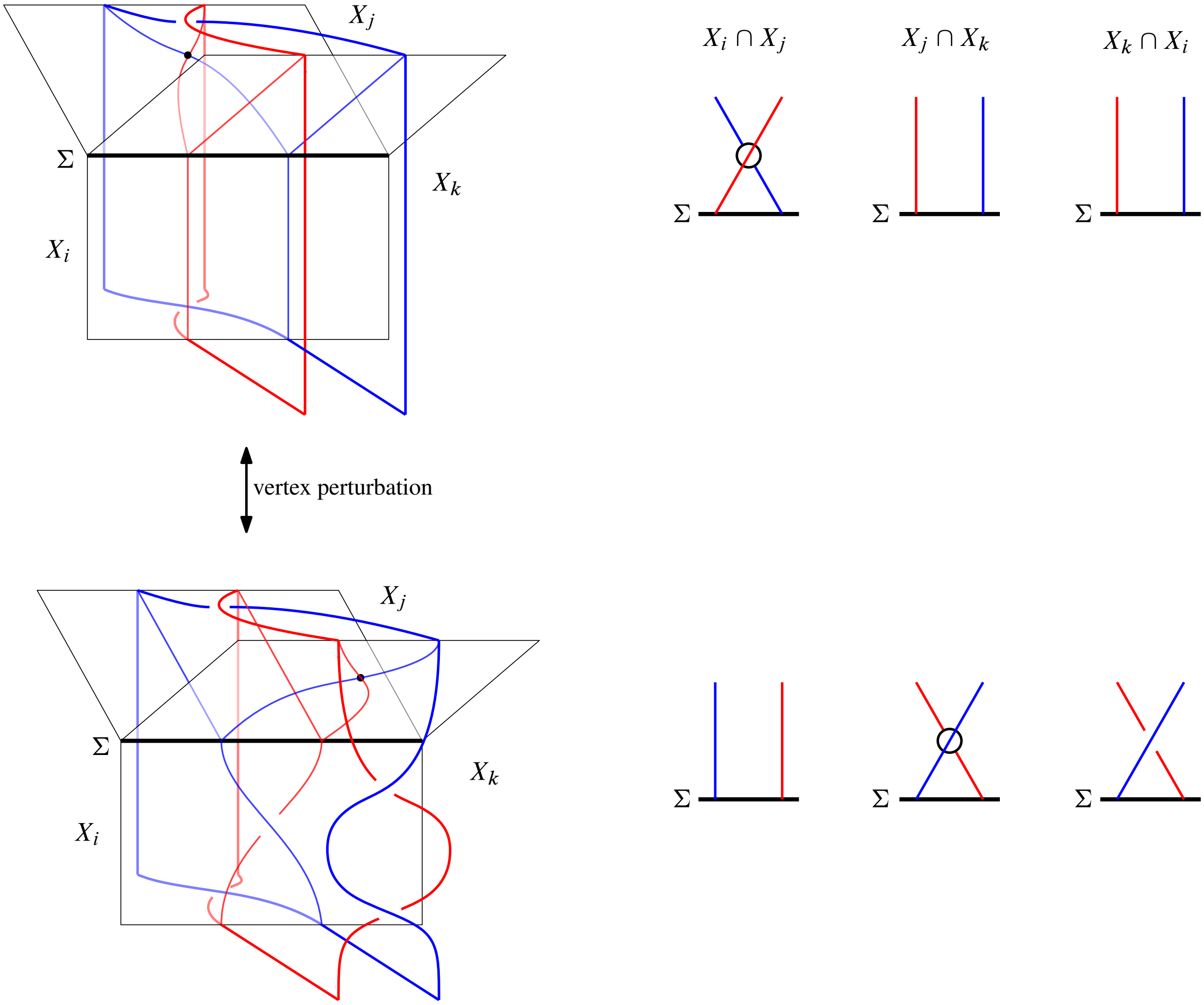


Figure 23: Pushing a self-intersection point from $X_i \cap X_j$ to $X_j \cap X_k$ during a vertex perturbation.

Definition 3.21 If a surface S' in bridge position with respect to a trisection $\mathcal{T}$ is obtained from a surface S in bridge position with respect to $\mathcal{T}$ by a sequence of elementary and vertex perturbations, then we simply say that S' is obtained from S by perturbation (with $\mathcal{T}$ implicit). If S' is obtained from S by a sequence of elementary perturbations and deperturbations and vertex perturbations, then we say that S' is obtained from S (or "related to S") by perturbation and deperturbation.

Theorem 3.22 *Let S be a self-transverse immersed surface in a 4-manifold X^4 with trisection $\mathcal{T} = (X_1, X_2, X_3)$. Then, for some c and b, S can be isotoped into (b, c)-bridge position with respect to $\mathcal{T}$.*

Proof Let h be a self-indexing Morse function of X^4 that is $\mathcal{T}$-compatible. Let (L, B) be a singular banded unlink diagram for S, so L is a singular link in $M := h^{-1}\left(\frac{3}{2}\right)$, and B is a set of bands for L in M. Let $H_1 := X_3 \cap X_1$ and $H_2 := X_1 \cap X_2$, so that $\Sigma := H_1 \cap H_2$ is a Heegaard surface for M.

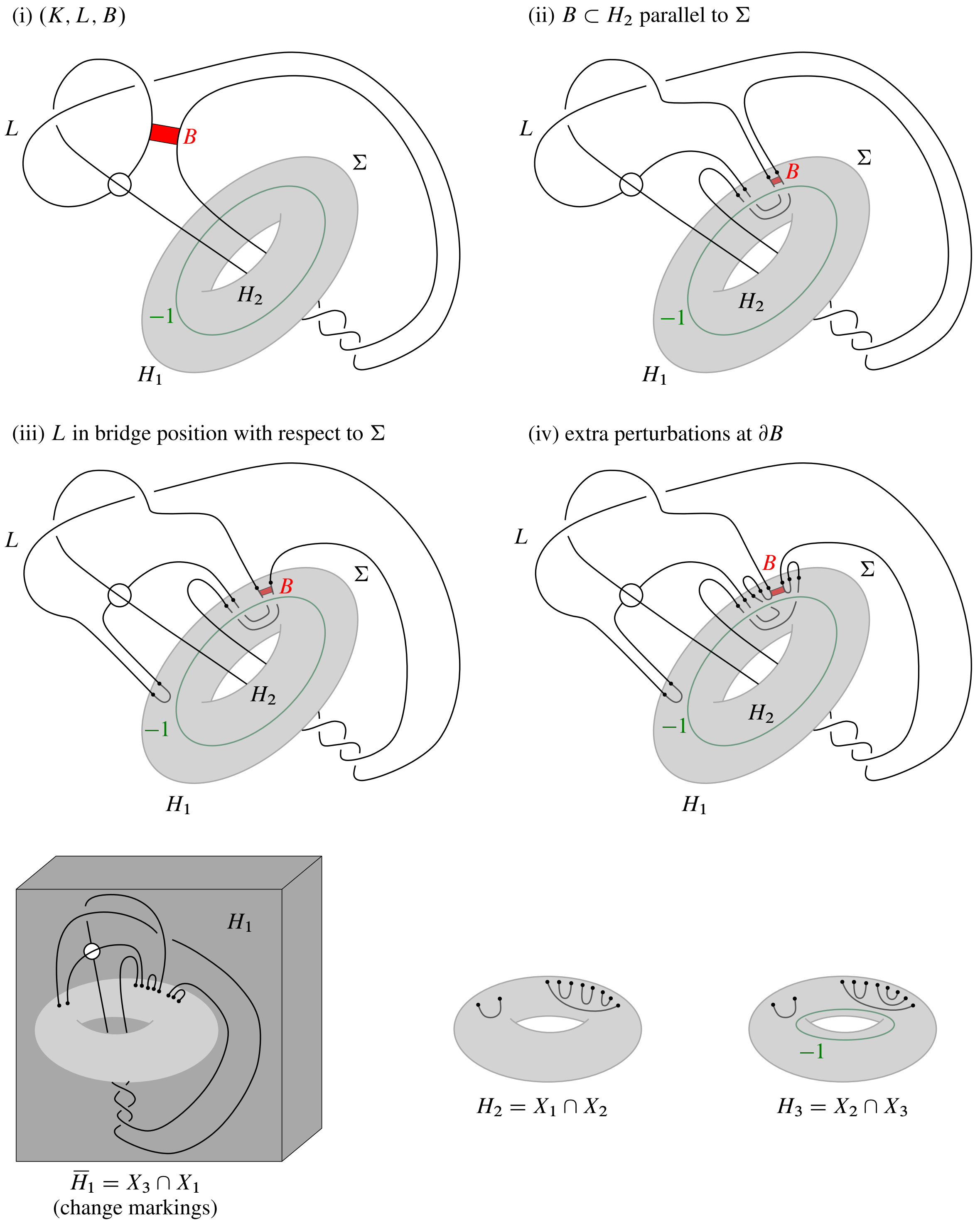


Figure 24: We illustrate how a surface that realizes a banded unlink diagram $(\mathcal{K}, L, B)$ may be isotoped to lie in bridge position. See the proof of Theorem 3.22.

By dimensionality, we may isotope L, B to be contained in $\Sigma \times [-1,1] \subset M$ (ie we isotope L and B to avoid a 1-skeleton of H_1 and H_2), with $\Sigma \times [-1,0] \subset H_1$, $\Sigma \times [0,1] \subset H_2$. Isotope L so that the vertices of L are disjoint from Σ, and so that B consists of short straight bands parallel to Σ in H_2 that are far from each other, as in Figure 24(ii). Let $\pi \colon \Sigma \times [0,1] \to [0,1]$ be the projection, and perform a small isotopy of L so that $\pi|_L$ is Morse. Isotope the index 0 critical points of $\pi|_L$ vertically with respect to π to be contained in H_1, and the index 1 critical points of $\pi|_L$ vertically with respect to π to be contained in H_2, isotoping horizontally first if necessary to avoid introducing self-intersections of L or intersections of L with B. Now L is in bridge position with respect to Σ. Perturb L again near ∂B as in Figure 24(iv), and isotope all bands in B to lie in H_2.

By Theorem 2.39, S is isotopic to $S' := \Sigma(L, B)$. We investigate the intersections of S' with the pieces of $\mathcal{T}$:

(i) $S' \cap X_1 = S' \cap h^{-1}\left(\frac{3}{2}\right)$ consists of the minimum disks of S'. All self-intersections of S' are contained in ∂X_1.

(ii) $S' \cap X_2$ contains the index 1 critical points of $h|_{S'}$. This surface is built from the singular tangle $L \cap H_2$ by extending vertically and then attaching bands according to B. By construction, these bandings are trivial and the components of $S' \cap X_2$ are boundary-parallel away from the intersections.

(iii) $S' \cap X_3$ contains the maximum disks of $h|_{S'}$. In particular, $(X_3, S' \cap X_3)$ can be strongly deformation retracted to $\left(h^{-1}\left(\left[\frac{5}{2}, 4\right]\right), S' \cap h^{-1}\left(\left[\frac{5}{2}, 4\right]\right)\right)$.

(iv) $S' \cap X_1 \cap X_2 = L \cap H_2$.

(v) $S' \cap X_2 \cap X_3$ is equivalent to the tangle obtained from $L^+ \cap H_2$ by surgery on B.

(vi) $S' \cap X_3 \cap X_1 = \overline{L \cap H_1}$. Note the reversed orientation; this is because H_1 is oriented as being in the boundary of X_1, but $X_3 \cap X_1$ is oriented as the boundary of X_3.

We conclude that S' is in (b, c)-bridge position with respect to $\mathcal{T}$ for some b and c. □

3.4 Bridge splittings of singular banded links

The proof of Theorem 3.22 motivates the following definition:

Definition 3.23 Let L be a singular link in a 3-manifold M, and let $B = b_1, \dots, b_n$ be a set of bands for L. Let F be a Heegaard surface for M. We say that the singular banded link (L, B) is in *bridge position* with respect to F if L is in bridge position with respect to F, and each band b_i is contained in a 3-ball U_i as in Figure 25, with $U_i \cap U_j = \varnothing$ for $i \neq j$.

The proof of Theorem 3.22 can be broken down into the following two lemmas, which are useful to state directly:

Lemma 3.24 *Let L be a singular link in a 3-manifold M, and let B be a set of bands for L. Fix a Heegaard surface F for M. Then (L, B) can be isotoped to lie in bridge position with respect to F.*

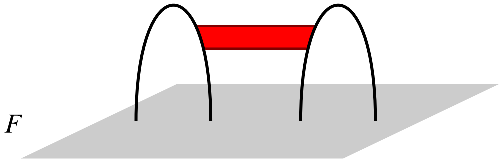


Figure 25: If a singular banded link (L, B) is in bridge position with respect to a Heegaard surface F, then every band in B has a neighborhood as pictured here. That is, every band in B has a neighborhood U containing two components C_1, C_2 of $L \setminus F$ (on which B has ends), meeting F in a disk, and not meeting any other bands in B or other components of $L \setminus F$. Moreover, $\overline{C}_1 \cup \overline{C}_2 \cup B$ may be isotoped rel $\partial(\overline{C}_1 \cup \overline{C}_2)$ in U to lie in F.

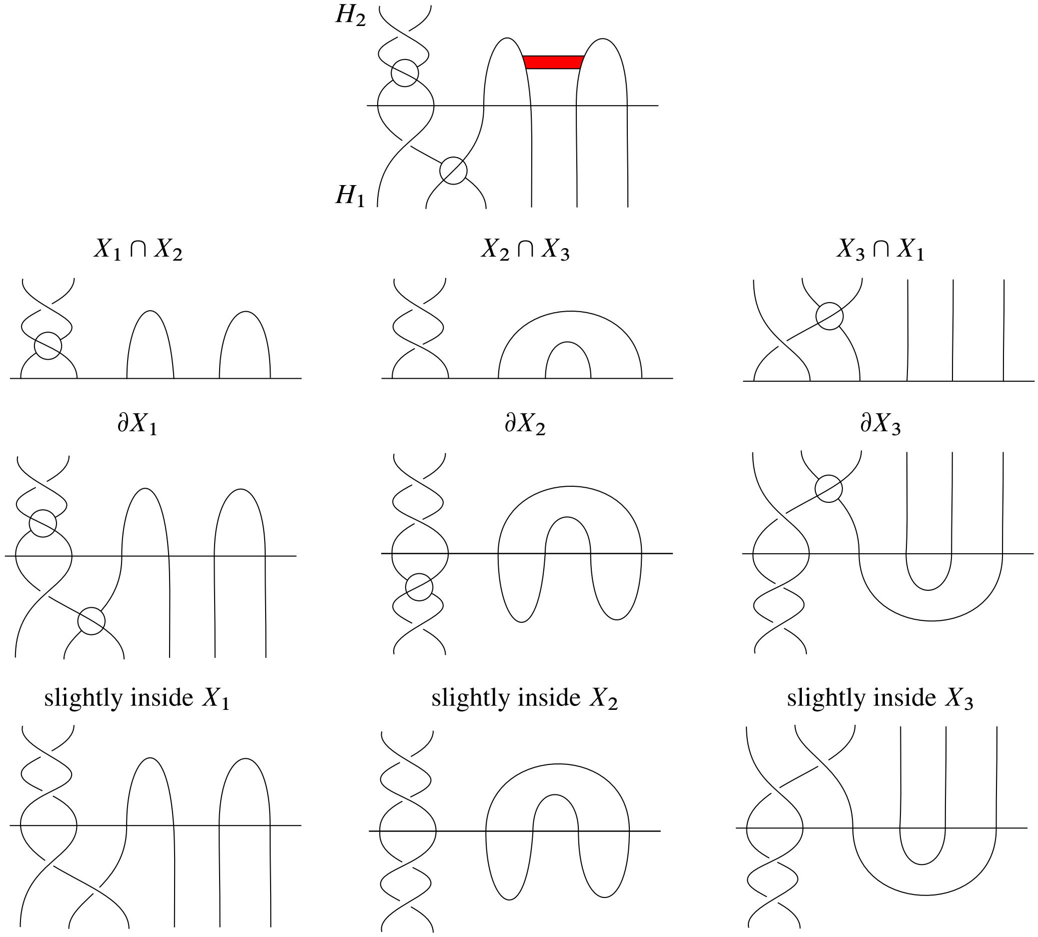


Figure 26: Top row: part of a singular banded unlink in bridge position. Second row: we obtain the singular tangles T_1, T_2 and T_3 as in Definition 3.26. Third row: the singular links that are the intersection of the associated bridge trisected surface with ∂X_1, ∂X_2 and ∂X_3. Bottom row: we draw the resolutions of these tangles in the interiors of X_1, X_2 and X_3. Note that vertices in $X_i \cap X_{i+1}$ are resolved negatively into X_i, while vertices in $X_{i-1} \cap X_i$ are resolved positively into X_i.

Lemma 3.25 *Let $\mathcal{T} = (X_1, X_2, X_3)$ be a trisection of a 4-manifold X^4. Let h be a $\mathcal{T}$-compatible Morse function on X^4, and $\mathcal{K}$ a Kirby diagram induced by h and a gradientlike vector field ∇h. Then $H_1 = X_3 \cap X_1$ and $H_2 = X_1 \cap X_2$ give a Heegaard splitting (H_1, H_2) for $h^{-1}\big(\frac{3}{2}\big)$, in which $\Sigma := H_1 \cap H_2 \subset E(\mathcal{K})$ is a Heegaard surface.*

Suppose a banded unlink $(\mathcal{K}, L, B)$ is in bridge position with respect to Σ. Then a realizing surface $\Sigma(\mathcal{K}, L, B)$ is in bridge position with respect to $\mathcal{T}$.

Definition 3.26 Let S be a self-transverse immersed surface in a 4-manifold X^4 with trisection $\mathcal{T} = (X_1, X_2, X_3)$. Assume S is in (b, c)-bridge position. We call the triple of singular marked tangles $(T_1, T_2, T_3) = (S \cap X_1 \cap X_2, S \cap X_2 \cap X_3, S \cap X_3 \cap X_1)$ a *bridge trisection diagram* of S. The markings of each tangle should be chosen so that:

- In X_i, cross-sections of S are the negative resolution of $S \cap X_i \cap X_{i+1}$.
- In X_i, cross-sections of S are the positive resolution of $S \cap X_{i-1} \cap X_i$.

Note that we choose the marking convention to be symmetric with respect to the trisection, even though in the construction of Theorem 3.22, we used a Morse function h in which the pieces X_1, X_2 and X_3 were not symmetric. If (L, B) is a singular banded unlink diagram for S and we follow the construction of Theorem 3.22, then we obtain a bridge trisection diagram (T_1, T_2, T_3) of S with:

(i) $T_1 = L \cap H_2$ with markings agreeing with those of L.

(ii) $T_2 = (L \cap H_2)^+_B$.

(iii) $T_3 = L \cap \overline{H}_1$ with markings *opposite* those of L.

We include a local example in Figure 26.

From a bridge trisection diagram of S, we can reconstruct a surface that is ambiently isotopic to S as usual. For convenience (to mirror the construction in Theorem 3.22), we assume all self-intersections lie in H_1 and H_2 (ie in ∂X_1 and not in $X_2 \cap X_3$).

Lemma 3.27 *Let S be a self-transverse immersed surface in a 4-manifold X^4 that is in bridge position with respect to a trisection $\mathcal{T} = (X_1, X_2, X_3)$. Assume that S has no self-intersections in $X_2 \cap X_3$.*

Let h be a $\mathcal{T}$-compatible Morse function on X^4, and fix a gradientlike vector field ∇h inducing a Kirby diagram $\mathcal{K}$. Then there is a singular banded unlink diagram $(\mathcal{K}, L, B)$ such that (L, B) is in bridge position with respect to the Heegaard surface $\Sigma = X_1 \cap X_2 \cap X_3 \subset E(\mathcal{K})$, and S is $\mathcal{T}$-regularly isotopic to the surface $\Sigma(\mathcal{K}, L, B)$.

Proof Isotope S to be 0-standard (with respect to $h, \nabla h$). Since S is in bridge position, we may take this isotopy to be $\mathcal{T}$-regular.

Let $L := S \cap h^{-1}\big(\frac{3}{2}\big)$. Recall $h^{-1}\big(\frac{3}{2}\big) = \partial X_1 = H_1 \cup H_2$, where $H_1 = X_3 \cap X_1$ and $H_2 = X_1 \cap X_2$. Then L is a singular link whose vertices are either in H_1 or H_2. Mark L so that the negative resolutions of the vertices in H_1 and the positive resolutions of the vertices in H_2 correspond to the resolutions of the immersed disk system $S \cap X_1$. Then L is a marked singular link and L^- is an unlink.

Now $S \cap X_2$ is a trivial immersed disk system with all intersections in $X_1 \cap X_2$. Let $\widetilde{X}_2$ be obtained from X_2 by deleting a small neighborhood of each intersection, so that $\widetilde{X}_2$ is still a 4-dimensional 1-handlebody, but $S \cap \widetilde{X}_2$ is a trivial embedded disk system D. Let $\widetilde{H}_2$ denote the closure of $(\partial \widetilde{X}_2) \setminus (X_2 \cap X_3)$.

Now D is a collection of boundary parallel disks in $\widetilde{X}_2$, and $\partial \widetilde{X}_2$ has a Heegaard splittings $(\widetilde{H}_2, X_2 \cap X_3)$, which in respect to ∂D is in bridge position. We proceed as in [26, Lemma 3.3]: For each component D_i of D, let a_i be one component of $\overline{\partial D \setminus (X_2 \cap X_3)}$. Then let y_i be an arc in $\partial \widetilde{X}_2$ parallel to $\partial D_i \setminus a_i$ with endpoints on ∂D, with framing induced by D_i. Isotope y_i in ∂X_2 into the Heegaard surface for $\partial \widetilde{X}_2$, twisting y_i around ∂D as necessary so that the framing of y_i agrees with the framing induced by the Heegaard surface. Finally, project y_i to ∂X_2, push slightly into H_2, and thicken (according to the framing of y_i) to obtain a band attached to $S \cap H_2$ (ie a band b_i in $h^{-1}\left(\frac{3}{2}\right)$ attached to L, with b_i in H_2 parallel to $H_1 \cap H_2$).

Repeat this for every component of D to obtain a collection B of bands for L. By construction, L_B^+ is an unlink when projected to $h^{-1}\left(\frac{5}{2}\right)$. More specifically, in $\mathcal{K}$ the link L_B^+ (projected to $h^{-1}\left(\frac{5}{2}\right)$) can be made to agree with the link $S \cap h^{-1}\left(\frac{5}{2}\right)$ via an isotopy rel boundary in H_2 and slides in H_2 over curves in $\mathcal{K}$. We conclude immediately that $(\mathcal{K}, L, B)$ is a singular banded unlink for some surface $S' := \Sigma(\mathcal{K}, L, B)$ in X. Moreover, S' is in bridge position with respect to $\mathcal{T}$, and by the above paragraph can be $\mathcal{T}$-regularly isotoped so that it agrees with S in $X_i \cap X_j$ for all $i \neq j$. Therefore, S and S' are $\mathcal{T}$-regularly isotopic. □

Remark 3.28 Fix a trisection $\mathcal{T} = (X_1, X_2, X_3)$ of X, a $\mathcal{T}$-compatible Morse function h and a gradientlike vector field ∇h, so that $(h, \nabla h)$ induce a Kirby diagram $\mathcal{K}$ of X in which $\Sigma := X_1 \cap X_2 \cap X_3$ is a Heegaard surface. Definition 3.26 and Lemma 3.27 can be combined to form the equivalence

$$\frac{\{\text{bridge trisections with respect to } \mathcal{T} \text{ with no self-intersections in } X_2 \cap X_3\}}{\mathcal{T}\text{-regular isotopy}} \leftrightarrow \frac{\{\text{SBUDs in } \mathcal{K} \text{ in bridge position with respect to } \Sigma\}}{\text{singular band moves preserving } \Sigma \text{ setwise}}.$$

The restriction of bridge position to not include self-intersections in $X_2 \cap X_3$ is merely a diagrammatic convenience from the viewpoint of singular banded unlinks diagrams (SBUDs).

Lemma 3.29 *Let S be in bridge position with respect to $\mathcal{T} = (X_1, X_2, X_3)$. There exists a sequence of perturbations of S yielding a surface S' in bridge position such that S' has no self-intersections in $X_2 \cap X_3$.*

To inductively prove Lemma 3.29, it is clearly sufficient to prove the following proposition:

Proposition 3.30 *Suppose there are $n > 0$ self-intersections of S in $X_2 \cap X_3$. Then, after $\mathcal{T}$-regular isotopy of S, there is a surface S' obtained from vertex perturbation on S such that S' has $n - 1$ self-intersections in $X_2 \cap X_3$.*

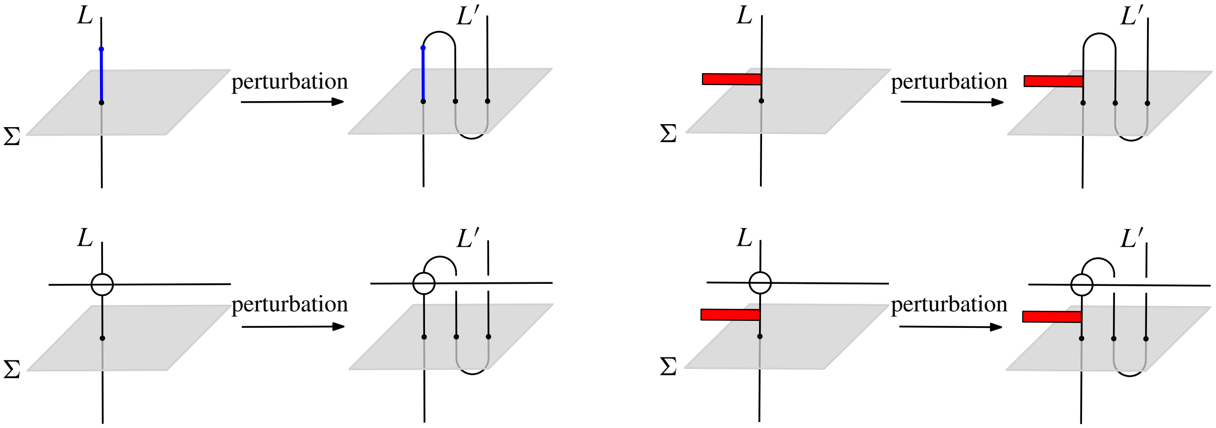


Figure 27: When performing a perturbation on the diagram in the top left, we allow the blue arc to intersect at most one band and one vertex, as shown in the other three diagrams.

Proof Following from Definition 3.11 of a trivial immersed tangle, some $\mathcal{T}$-regular isotopy of S can arrange for the tangle $T = S \cap X_2 \cap X_3$ to lie inside a collar neighborhood $\Sigma \times I \subset X_2 \cap X_3$ of $\partial(X_2 \cap X_3) = \Sigma$, so that projection to the I factor is Morse on T with one maximum on each arc component. Further isotope so that the self-intersections of S in $\Sigma \times I$ lie at different values of the I factor. In particular, one self-intersection c lies strictly closest to Σ. Then by $\mathcal{T}$-regular isotopy of S near Σ (sometimes called "mutual braid transposition" when performed diagrammatically), we can arrange for c to have a neighborhood as in Figure 22, and thus apply a vertex perturbation to S to obtain a surface S' with one less self-intersection in $X_2 \cap X_3$. □

3.5 Uniqueness of bridge trisections of immersed surfaces

Perturbation of bridge trisections is conveniently very similar to perturbation of a banded link in bridge position. When perturbing a banded link (L, B) with respect to a Heegaard surface Σ, we allow at most one band and one vertex to be between the intersection of L and Σ at which the perturbation is based and the two newly introduced intersections. See Figure 27.

Lemma 3.31 *Let $\mathcal{T} = (X_1, X_2, X_3)$ be a trisection of a 4-manifold X^4. Let h be a $\mathcal{T}$-compatible Morse function on X^4, and $\mathcal{K}$ a Kirby diagram induced by h. Let $H_1 := X_3 \cap X_1$ and $H_2 := X_1 \cap X_2$ give the usual Heegaard splitting (H_1, H_2) for $\mathcal{K}$, in which $\Sigma := H_1 \cap H_2$ is the Heegaard surface.*

Suppose a singular banded unlink diagram $(\mathcal{K}, L, B)$ is in bridge position with respect to Σ. Let $(\mathcal{K}, L', B')$ be obtained from $(\mathcal{K}, L, B)$ by perturbation near $L \cap \Sigma$. Then $\Sigma(\mathcal{K}, L', B')$ can be obtained from $\Sigma(\mathcal{K}, L, B)$ by perturbation followed by $\mathcal{T}$-regular isotopy.

Proof See Figure 28, top. □

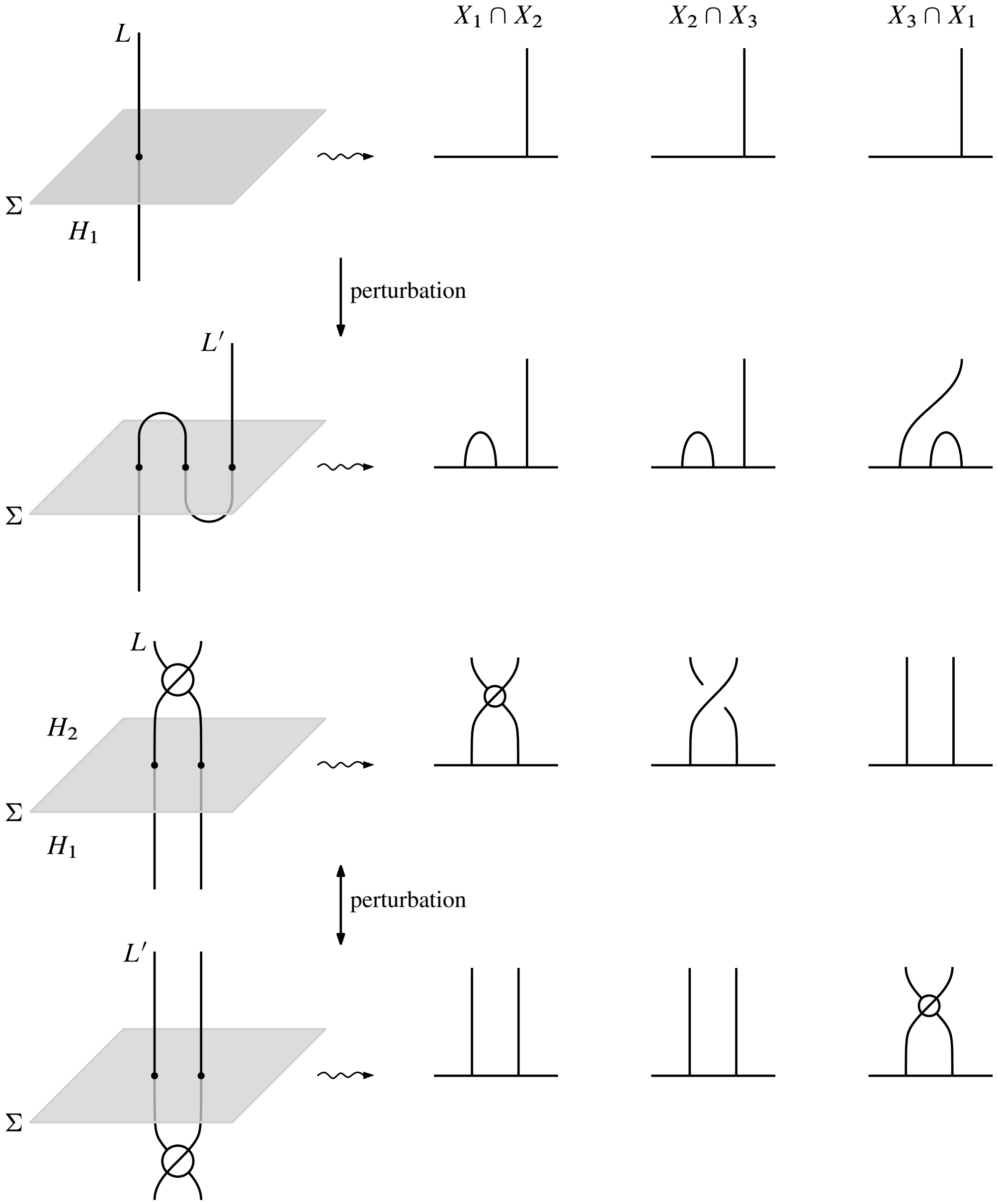


Figure 28: Perturbation of a singular banded unlink (L, B) in bridge position induces perturbation of $\Sigma(L, B)$. Top: elementary perturbation. Bottom: vertex perturbation.

Lemma 3.32 *Let $\mathcal{T} = (X_1, X_2, X_3)$ be a trisection of a 4-manifold X^4. Let h be a $\mathcal{T}$-compatible Morse function on X^4, and $\mathcal{K}$ a Kirby diagram induced by h. Let $H_1 := X_3 \cap X_1$ and $H_2 := X_1 \cap X_2$ give the usual Heegaard splitting (H_1, H_2) for $\mathcal{K}$, in which $\Sigma := H_1 \cap H_2$ is the Heegaard surface.*

Suppose a singular banded unlink diagram $(\mathcal{K}, L, B)$ is in bridge position with respect to Σ and that v is a vertex of L that is close to Σ as in Figure 20. Let $(\mathcal{K}, L', B')$ be obtained from $(\mathcal{K}, L, B)$ by isotoping

v through Σ. (We call this a *vertex perturbation of the banded link* (L, B).) *Then $\Sigma(\mathcal{H}, L', B')$ can be obtained from $\Sigma(\mathcal{H}, L, B)$ by one vertex perturbation followed by $\mathcal{T}$-regular isotopy.*

Proof See Figure 28, bottom. □

The following uniqueness of bridge splittings of banded links motivates the uniqueness of bridge trisections:

Theorem 3.33 *Let (L, B) and (L', B') be isotopic banded singular marked links in a 3-manifold M that has a Heegaard splitting (H_1, H_2). Assume that both (L, B) and (L', B') are in bridge position with respect to $\Sigma := H_1 \cap H_2$, and that B and B' are both contained in H_2. Then there exists a banded singular marked link (L'', B'') in bridge position with respect to Σ that can be obtained from both (L, B) and (L', B') by sequences of elementary perturbations, vertex perturbations and isotopies that fix Σ setwise.*

Theorem 3.33 is similar to a theorem for nonsingular banded links due to Meier and Zupan [26; 27].

Remark 3.34 Meier and Zupan study banded links by viewing each band as a framed arc with endpoints on a link. They give moves to perturb a link in order to make these framed arcs parallel to a bridge surface with correct framing. In the setting of singular banded links, we are able to use their proof by viewing both self-intersections and bands as framed arcs, applying the theorem and then contracting the self-intersection arcs to yield a singular link in bridge position.

Proof As in Theorem 3.15, there exist disjoint framed arcs $a_1, \dots, a_n$ with endpoints on L^+ such that contracting L^+ along $a_1, \dots, a_n$ yields L. Similarly, there exist framed arcs $a'_1, \dots, a'_n$ with endpoints on L'^+ such that contracting L'^+ along $a'_1, \dots, a'_n$ yields L'.

Now, by Meier and Zupan [26; 27], there exists a link J that can be obtained from L^+ and from L'^+ by elementary perturbations and isotopies fixing Σ setwise. Moreover, these isotopies and perturbations carry a_i and a'_i to framed arcs b_i and b'_i, respectively, with endpoints on J, so that b_i and b'_i are each parallel to Σ with surface framing, and either agree or could be isotoped to agree if the endpoints of b'_i were allowed to pass through Σ (ie b_i and b'_i are parallel and both lie close to Σ, but potentially on opposite sides). Moreover, during the perturbations and isotopies of L^+ (resp. L'^+), a_i (resp. a'_i) never intersect Σ, so these perturbations and isotopies may be achieved by perturbations and isotopies of L (resp. L').

Meier and Zupan's proof allows us to not only control the framed arcs a_i, a'_i, but also the framed arcs that are the cores of the bands B and B'. That is, by perhaps perturbing J even further, we may also assume that B and B' are taken to bands B_J, and B'_J whose i^{th} bands either agree or are parallel and close to Σ but on opposite sides, and that (J, B_J) and (J, B'_J) are both in bridge position. Let $\widehat{J}$ and $\widehat{J}'$ be the marked singular links obtained by contracting J along b_i and b'_i, respectively, and with markings induced by L and L'. Then $\widehat{J}'$ can be transformed into $\widehat{J}$ by isotopy fixing Σ and a vertex perturbation for each pair a_i and a'_i in different components of $M \setminus \Sigma$. Therefore, the claim holds with $L'' = \widehat{J}$ and $B'' = B_J$. □

Corollary 3.35 *If $\mathcal{D} = (L, B)$ and $\mathcal{D}' = (L', B')$ are isotopic banded unlink diagrams that are each in bridge position with respect to Σ, then $S := \Sigma(\mathcal{D})$ and $S' := \Sigma(\mathcal{D}')$ are related by elementary perturbation and deperturbation, vertex perturbation and $\mathcal{T}$-regular isotopy.*

Proof By Theorem 3.33, $\mathcal{D}$ and $\mathcal{D}'$ are related by a sequence of elementary perturbations and deperturbations, vertex perturbations and isotopies fixing Σ setwise. It is therefore sufficient to show that the claim is true if $\mathcal{D}'$ is obtained from $\mathcal{D}$ by a single one of these moves. We have already shown the claim to be true when $\mathcal{D}'$ is obtained from $\mathcal{D}$ by either a perturbation/deperturbation (Lemma 3.31), or a vertex perturbation (Lemma 3.32). So suppose that $\mathcal{D}'$ is obtained from $\mathcal{D}$ by an isotopy f_t of M that fixes Σ setwise.

The surface $\Sigma_{3/2} := \Sigma$ is a separating surface in $M = h^{-1}\bigl(\frac{3}{2}\bigr)$. For every $t \in [0, 4]$, there is a separating surface Σ_t in $h^{-1}(t)$ that is vertically above or below Σ. Then f_t can be extended to a horizontal isotopy of the whole 4-manifold X^4 that fixes every Σ_t horizontally. Since all index 2 critical points of h are contained in one component of $X^4 \setminus \bigcup_t \Sigma_t$, this isotopy can be chosen to take S to S'. Since this isotopy is horizontal, it fixes $X_1 = h^{-1}\bigl(\bigl[0, \frac{3}{2}\bigr]\bigr)$ and $X_2 \cup X_3 = h^{-1}\bigl(\bigl[\frac{3}{2}, 4\bigr]\bigr)$ setwise. Since this isotopy fixes $X_2 \cap X_3 = \bigcup_{[3/2,4]} \Sigma_t$ setwise, it also fixes X_2 and X_3 setwise. Therefore, this is a $\mathcal{T}$-regular isotopy. □

The main theorem of this section is that bridge position and hence bridge trisection diagrams are essentially unique. The proof uses Theorem 2.39.

Theorem 3.36 *Let S and S' be self-transverse immersed surfaces in bridge position with respect to a trisection $\mathcal{T} = (X_1, X_2, X_3)$ of a closed 4-manifold X^4. Suppose S is ambiently isotopic to S'. Then S can be taken to S' by a sequence of elementary perturbations and deperturbations, vertex perturbations and $\mathcal{T}$-regular isotopy.*

Proof Let $h\colon X \to [0, 4]$ be a $\mathcal{T}$-compatible Morse function on X^4. Let $\mathcal{K}$ be a Kirby diagram for X induced by h and a fixed choice of ∇h. As usual, we view $\Sigma := X_1 \cap X_2 \cap X_3$ as a Heegaard surface for the ambient space of $\mathcal{K}$, with the dotted circles of $\mathcal{K}$ contained in one handlebody H_1 of this splitting and the 2-handle circles of $\mathcal{K}$ contained in the other handlebody H_2.

By Lemma 3.29, we may $\mathcal{T}$-regularly isotope and perturb S and S' so that they do not include self-intersections in $X_2 \cap X_3$. Then, by Lemma 3.27, there are banded unlink diagrams $\mathcal{D} := (\mathcal{K}, L, B)$ and $\mathcal{D}' := (\mathcal{K}, L', B')$ such that (L, B) and (L', B') are in bridge position with respect to Σ and such that S and S' are $\mathcal{T}$-regular isotopic to $\Sigma(\mathcal{D})$ and $\Sigma(\mathcal{D}')$, respectively.

By Theorem 2.39, $\mathcal{D}$ and $\mathcal{D}'$ are related by a sequence of singular band moves. By Corollary 3.35, if $\mathcal{D}$ and $\mathcal{D}'$ are isotopic, then the theorem holds.

Assume that $\mathcal{D}'$ is obtained from $\mathcal{D}$ by one singular band move (other than isotopy). We will show that S' and S become $\mathcal{T}$-regular isotopic after some sequence of perturbations and deperturbations. The theorem will then hold via induction on the length of a sequence of band moves relating $\mathcal{D}$ and $\mathcal{D}'$.

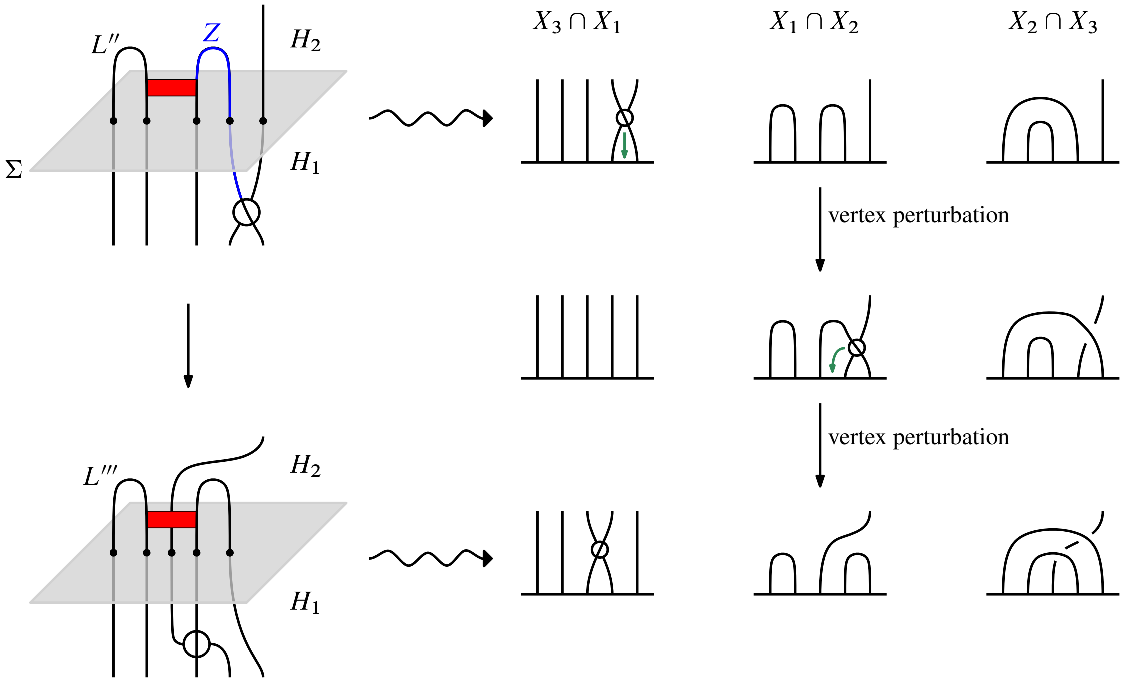


Figure 29: Left: the singular banded unlink (L''', B''') is obtained from (L'', B'') by an intersection/band pass. Right: we show that $\Sigma(L''', B''')$ (bottom) may be obtained from $\Sigma(L'', B'')$ (top) by two vertex perturbations and $\mathcal{T}$-regular isotopy.

Meier and Zupan [26] previously showed that the claim holds when the move turning $\mathfrak{D}$ into $\mathfrak{D}'$ is a cup, cap, band swim or band slide. The authors of this paper [14] showed the claim is true when the move is a 2-handle/band slide, 2-handle/band swim or dotted circle slides. These arguments were technically only made for nonsingular banded unlinks, so we repeat them in the singular setting for clarity, often repeating Meier and Zupan's arguments. In the following paragraphs, we consider every singular band move that might transform $\mathfrak{D}$ into $\mathfrak{D}'$.

Intersection/band pass Suppose $\mathfrak{D}'$ is obtained from $\mathfrak{D}$ by an intersection/band pass along a framed arc z in L between a vertex of L and a band in B. Isotope (L, B) so that z is as in Figure 29, top left. Then isotope the rest of L and B outside a neighborhood of z to obtain a banded link (L'', B'') in bridge position. This banded singular link is isotopic to (L, B), so, by Corollary 3.35, $S'' := \Sigma(L'', B'')$ is obtainable from S by (de)perturbations and $\mathcal{T}$-regular isotopy. Let (L''', B''') be obtained from (L'', B'') by performing the intersection/band pass along z, and let $S''' := \Sigma(L''', B''')$. Now the intersection of S''' with each $X_i \cap X_j$ is isotopic rel boundary to the intersection of S'' with $X_i \cap X_j$, so S''' is $\mathcal{T}$-regular isotopic to S''. Finally, by Corollary 3.35, we find that S''' can be transformed into S' by (de)perturbations and $\mathcal{T}$-regular isotopy.

Intersection/band slide Suppose $\mathfrak{D}'$ is obtained from $\mathfrak{D}$ by an intersection/band slide along a framed arc z in L between a vertex of L and a band in B. Isotope (L, B) so that z is short and contained

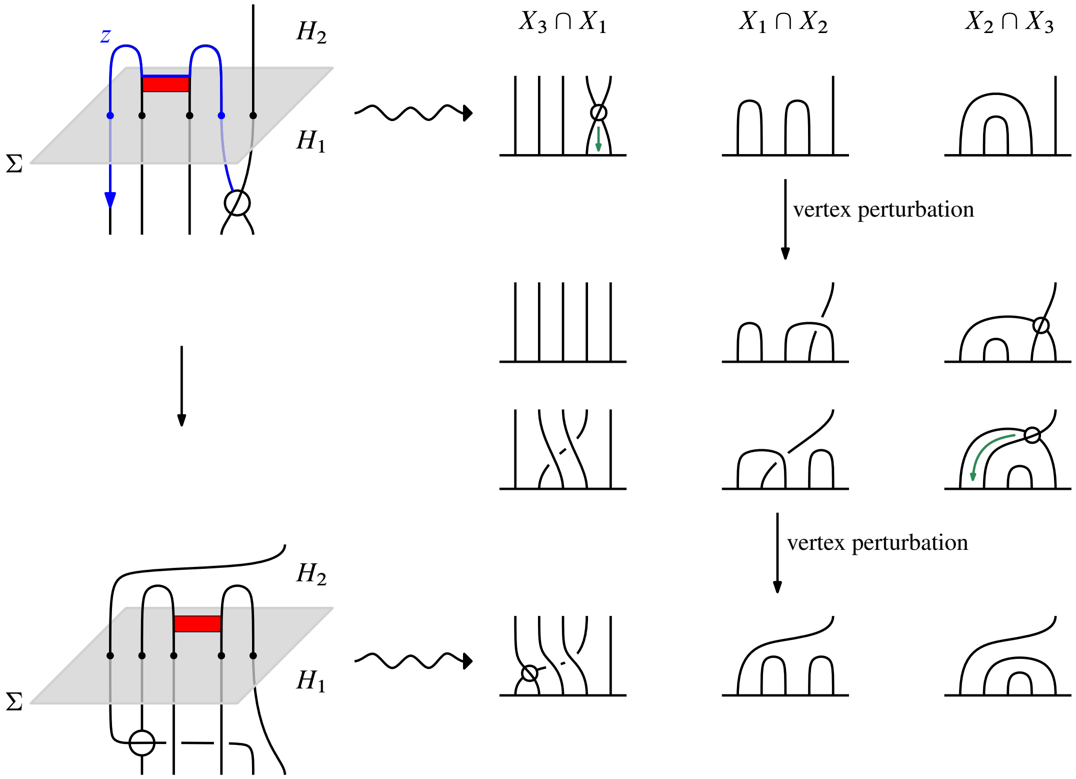


Figure 30: Left: the singular banded unlink (L''', B''') is obtained from (L'', B'') by an intersection/band slide. Right: we show that $\Sigma(L''', B''')$ (bottom) may be obtained from $\Sigma(L'', B'')$ (top) by two vertex perturbations and $\mathcal{T}$-regular isotopy.

in H_2 in a neighborhood as in Figure 30. Then isotope the rest of L and B outside this neighborhood to obtain a banded link (L'', B'') in bridge position. This banded singular link is isotopic to (L, B), so, by Corollary 3.35, $S'' := \Sigma(L'', B'')$ is obtainable from S by (de)perturbations and $\mathcal{T}$-regular isotopy. Let (L''', B''') be obtained from (L'', B'') by performing the intersection/band slide along z, and let $S''' := \Sigma(L''', B''')$. In Figure 30, we show that S''' can be obtained from S'' by perturbation and $\mathcal{T}$-regular isotopy. Finally, by Corollary 3.35, S''' can be transformed into S' by (de)perturbations and $\mathcal{T}$-regular isotopy.

Cup Suppose $\mathcal{D}'$ is obtained from $\mathcal{D}$ by a cup move. It does not matter in which direction we take the move, so assume that L' is obtained from L by adding a new unlink component O contained in a ball not meeting L or B, and B' is obtained from B by adding a trivial band b_O from L to O. By isotopy and intersection/band passes, we may take O to be in 1-bridge position with respect to Σ, and b_O to be in H_2, contained in a neighborhood as in Figure 31. Performing the cup move yields a diagram $\mathcal{D}''$ that is related to $\mathcal{D}'$ by isotopy and intersection/band passes; by Corollary 3.35 and the already-considered intersection/band pass case, $\Sigma(\mathcal{D}'')$ can be transformed into S' by perturbation and $\mathcal{T}$-regular isotopy. Finally, we observe that $\Sigma(\mathcal{D}'')$ is obtained from the (perturbed) surface S by perturbation (see Figure 31).

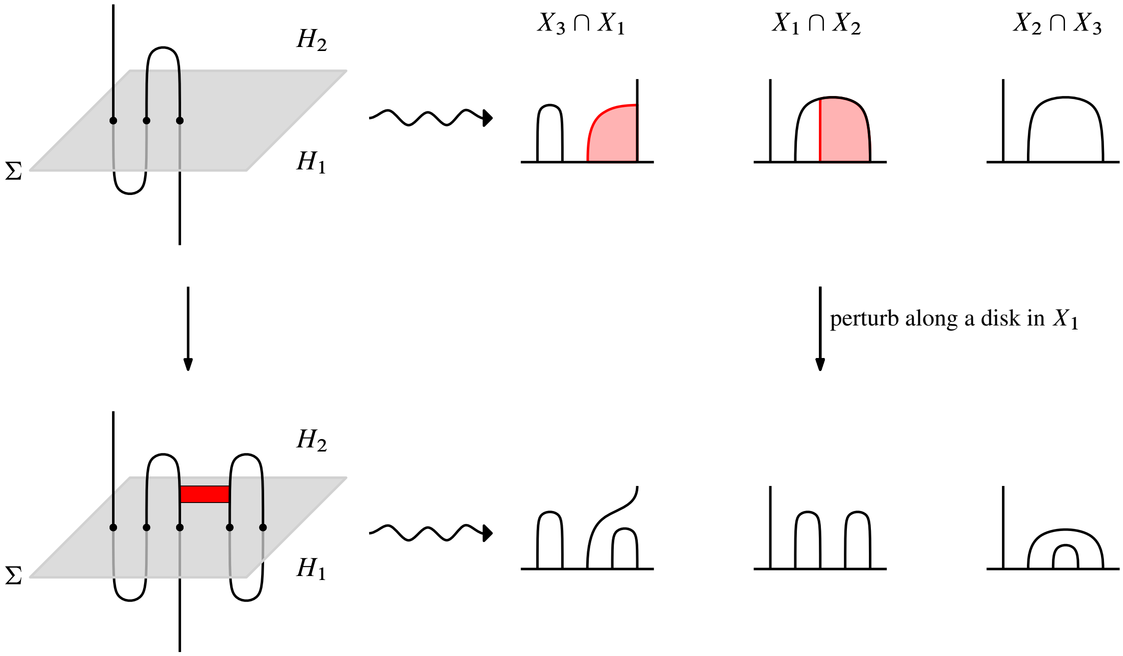


Figure 31: Left: the singular banded unlink (L''', B''') is obtained from (L'', B'') by a cup move. Right: we show that $\Sigma(L''', B''')$ (bottom) may be obtained from $\Sigma(L'', B'')$ (top) by an elementary perturbation and $\mathcal{T}$-regular isotopy.

Cap Suppose $\mathfrak{D}'$ is obtained from $\mathfrak{D}$ by a cap move. Again, it does not matter in which direction we take the move, so assume that $L' = L$ and B' is obtained from B by adding a trivial band b. By isotopy and intersection/band passes, we may take b to have a neighborhood as in Figure 32. Performing the cap move yields a diagram $\mathfrak{D}''$ that is related to $\mathfrak{D}'$ by isotopy and intersection/band passes; by Corollary 3.35 and the case for intersection/band pass, $\Sigma(\mathfrak{D}'')$ can be transformed into S' by perturbation and $\mathcal{T}$-regular isotopy. Finally, we observe that $\Sigma(\mathfrak{D}'')$ is obtained from the (perturbed) surface S by perturbation and deperturbation (see Figure 32).

Band swim Suppose $\mathfrak{D}'$ is obtained from $\mathfrak{D}$ by a band swim. Isotope $\mathfrak{D}$ to obtain a diagram in which the band swim looks as in Figure 33. Perform the band swim to obtain a diagram $\mathfrak{D}''$ that is related to $\mathfrak{D}'$ by isotopy; by Corollary 3.35 and the intersection/band swim case, $\Sigma(\mathfrak{D}'')$ can be transformed into S' by perturbation and $\mathcal{T}$-regular isotopy. Finally, we observe that $\Sigma(\mathfrak{D}'')$ is obtained from the (perturbed) surface S by $\mathcal{T}$-regular isotopy (see Figure 33).

Band slide Suppose $\mathfrak{D}'$ is obtained from $\mathfrak{D}$ by a band slide. Isotope $\mathfrak{D}$ to obtain a diagram in bridge position in which the desired band slide looks like Figure 34. By Corollary 3.35, the effect on S can be achieved by (de)perturbation and $\mathcal{T}$-regular isotopy. Call the result of the band slide $\mathfrak{D}''$; by Corollary 3.35, the surface $\Sigma(\mathfrak{D}'')$ can be transformed into S' by (de)perturbation and $\mathcal{T}$-regular isotopy. In Figure 34, we observe that $\Sigma(\mathfrak{D}'')$ is obtained from S by perturbation and deperturbation.

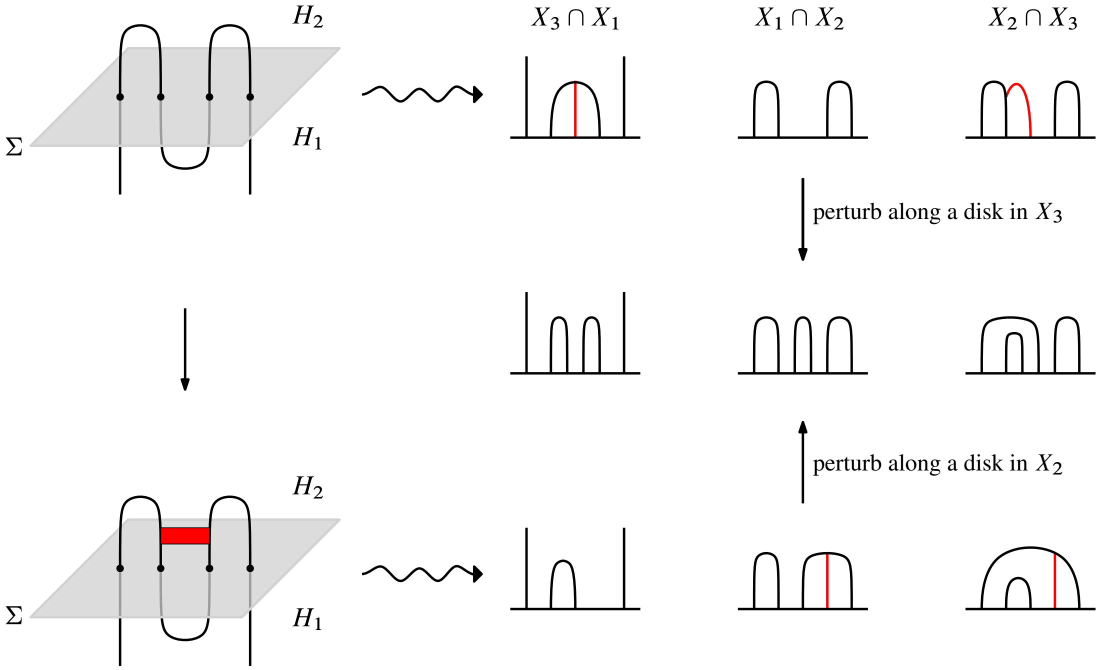


Figure 32: Left: the singular banded unlink $\mathscr{D}''$ is obtained from $\mathscr{D}$ by a cap move. Right: we show that $\Sigma(\mathscr{D}'')$ (bottom) may be obtained from $\Sigma(\mathscr{D})$ (top) by an elementary perturbation and deperturbation and $\mathscr{T}$-regular isotopy.

2-handle/band slide Suppose $\mathscr{D}'$ is obtained from $\mathscr{D}$ by sliding a band over a 2-handle via a framed arc z between a band in B and a 2-handle attaching circle in $\mathscr{K}$. As in the band slide case, we may perturb $\mathscr{D}$ so that z is contained in H_2 (see Figure 35). Now, performing the slide along z yields a diagram $\mathscr{D}''$ that is

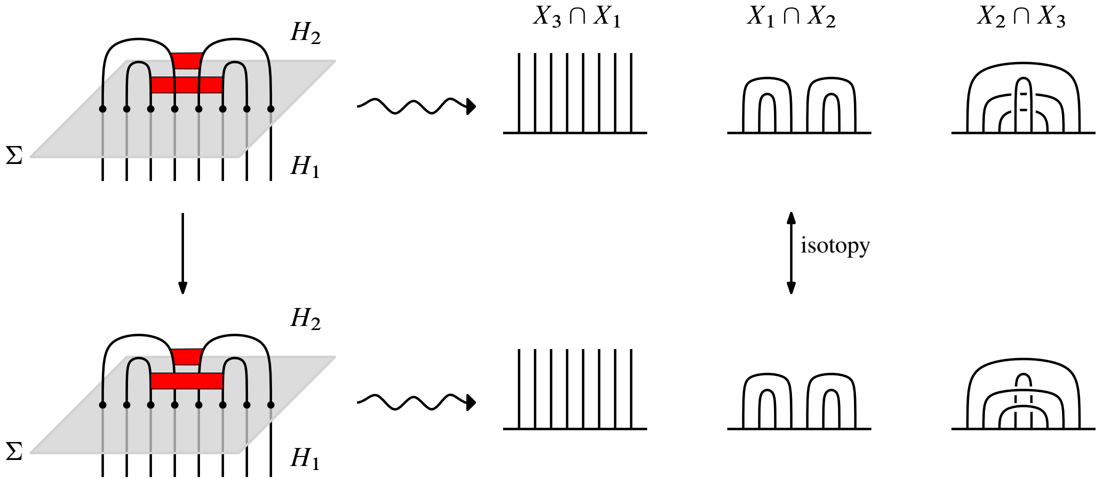


Figure 33: Left: the singular banded unlink $\mathscr{D}''$ is obtained from $\mathscr{D}$ by a band swim. Right: we show that $\Sigma(\mathscr{D}'')$ (bottom) may be obtained from $\Sigma(\mathscr{D})$ (top) by $\mathscr{T}$-regular isotopy.

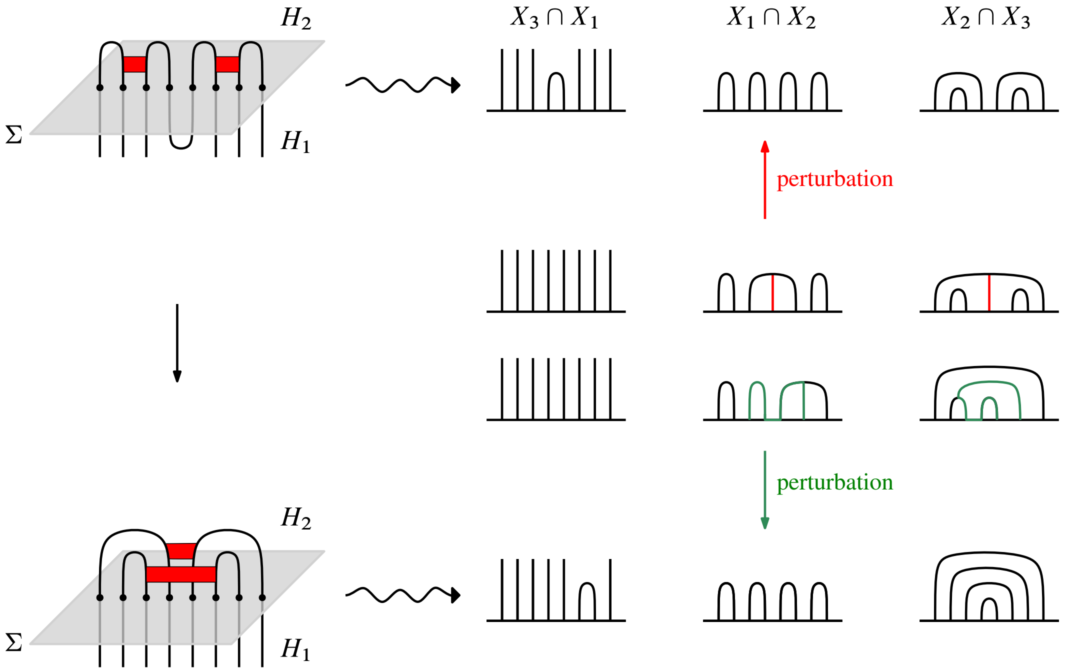


Figure 34: Left: the singular banded unlink $\mathscr{D}''$ is obtained from $\mathscr{D}$ by a band slide. Right: we show that $\Sigma(\mathscr{D}'')$ (bottom) may be obtained from $\Sigma(\mathscr{D})$ (top) by an elementary perturbation and deperturbation and $\mathscr{T}$-regular isotopy.

related to $\mathscr{D}'$ by isotopy; by Corollary 3.35, the surface $\Sigma(\mathscr{D}'')$ can be transformed into S' by perturbation and $\mathscr{T}$-regular isotopy. Finally, we observe that $\Sigma(\mathscr{D}'')$ is obtained from the (perturbed) surface S by $\mathscr{T}$-regular isotopy supported in X_2 and X_3.

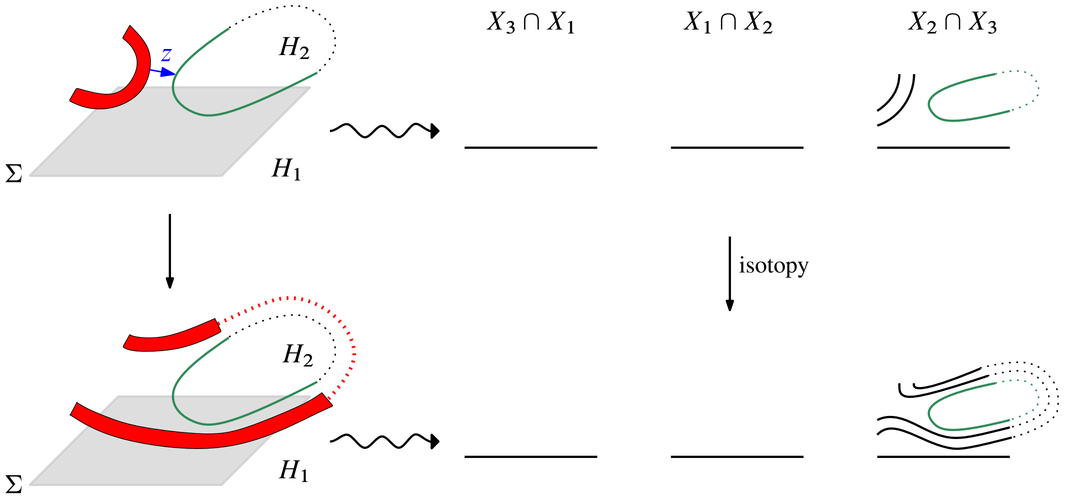


Figure 35: Left: the singular banded unlink $\mathscr{D}''$ is obtained from $\mathscr{D}$ by a 2-handle/band slide. Right: we show that $\Sigma(\mathscr{D}'')$ (bottom) may be obtained from $\Sigma(\mathscr{D})$ (top) by $\mathscr{T}$-regular isotopy.

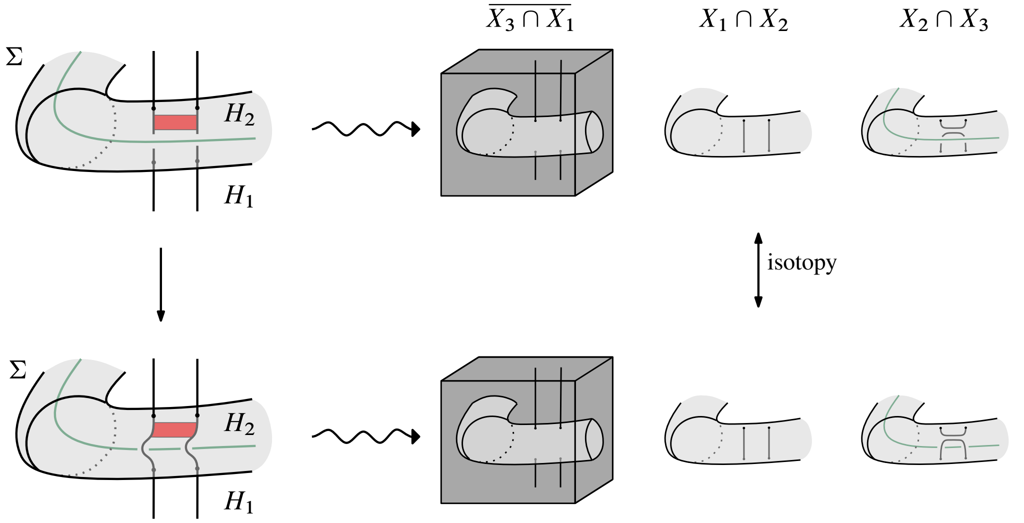


Figure 36: Left: the singular banded unlink $\mathscr{D}''$ is obtained from $\mathscr{D}$ by a 2-handle/band swim. Right: we show that $\Sigma(\mathscr{D}'')$ (bottom) may be obtained from $\Sigma(\mathscr{D})$ (top) by $\mathscr{T}$-regular isotopy.

2-handle/band swim Suppose $\mathscr{D}'$ is obtained from $\mathscr{D}$ by swimming a 2-handle through a band. Isotope $\mathscr{D}'$ so that the swim looks like the one in Figure 36. By Corollary 3.35, this can be achieved by (de)perturbations and $\mathscr{T}$-regular isotopy of S. Now, performing the swim along z yields a diagram $\mathscr{D}''$ that is related to $\mathscr{D}'$ by isotopy; by Corollary 3.35, the surface $\Sigma(\mathscr{D}'')$ can be transformed into S' by perturbation and $\mathscr{T}$-regular isotopy. Finally, we observe that $\Sigma(\mathscr{D}'')$ is obtained from the (perturbed) surface S by $\mathscr{T}$-regular isotopy supported in X_2 and X_3.

Slide of a band or L over a dotted circle This follows from Theorem 3.33, as slides over dotted circles are simply isotopies of the banded link (L, B) in $M_{3/2}$. □

4 Some example applications

In this (comparatively short) section, we give a few sample applications of the diagrammatic theory of singular banded unlink diagrams.

4.1 Calculating the Kirk invariant

Schneiderman and Teichner [30] classified all 2-component spherical links in S^4 up to link homotopy using the Kirk invariant $\sigma_i(F_1, F_2) := \lambda(F_i, F_i')$. Here $i \in \{1, 2\}$, F_i' is a parallel pushoff of F_i, and $\lambda(F_i, F_i')$ is Wall's intersection invariant. Furthermore, F_i denotes an oriented immersed 2-sphere in S^4, with F_1 and F_2 disjoint. The Kirk invariant takes values in $\mathbb{Z}[\mathbb{Z}] = \mathbb{Z}[x^{\pm}]$.

Schneiderman and Teichner showed that the set of all 2-component spherical links in S^4 up to link homotopy is a free R-module, where $R = \mathbb{Z}[z_1, z_2]/(z_1 z_2)$ is freely generated by the Fenn–Rolfsen link FR depicted in Figure 37.

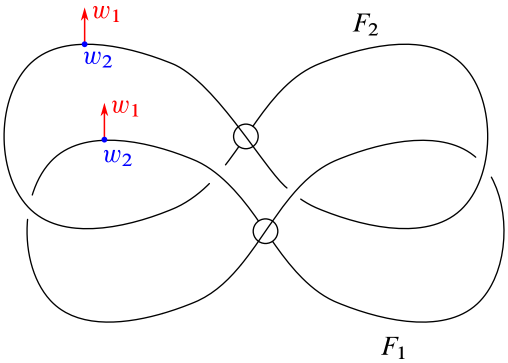


Figure 37: The Fenn–Rolfsen link. At the indicated points with arrows, a positive basis of the normal bundle is (w_1, w_2), where w_1 is the drawn arrow pointing upward and w_2 points out of the page toward the reader.

In this subsection, we show how to compute the Kirk invariant of FR. This computation can be adapted to compute Wall's self-intersection invariant for general 2-component spherical links in arbitrary closed orientable 4-manifolds. Since FR has a symmetry between its two components that reverses the orientation on one component, we have $\sigma_2 = -\sigma_1$ and thus only compute σ_1.

Consider the singular banded unlink diagram of $\mathrm{FR} = F_1 \sqcup F_2$ as in Figure 37. Choose a basepoint p far away from FR and an arc γ from p to a point q in F_1. Take a pushoff F_1' of F_1 that transversely intersects F_1; simultaneously push off γ to obtain an arc γ' from p to a point q' of F_1'.

We thus have two parallel arcs γ' and γ from p to F_1' and from p to F_1, respectively (as in Figure 38). Now delete a neighborhood of F_2 as in Figure 39.

Pick a vertex v between the diagrams of F_1 and F_1', and choose arcs η, η' contained in F_1 and F_1', respectively, from q and q', respectively, to v. Let C_v be the based loop obtained by concatenating γ, η,

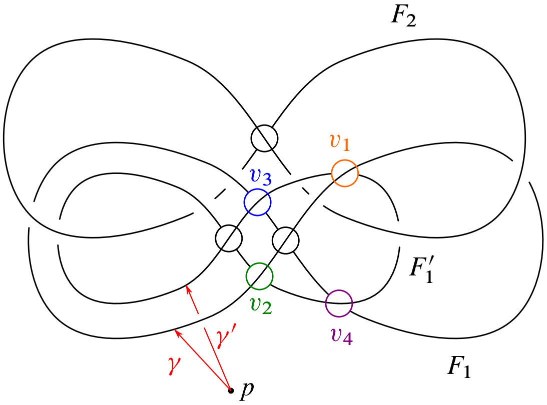


Figure 38: A parallel pushoff F_1' of F_1 that intersects F_1 transversely in four points yielding vertices v_1, v_2, v_3 and v_4 in the singular banded unlink diagram. The intersections have respective signs $s_{v_1} = 1$, $s_{v_2} = -1$, $s_{v_3} = -1$ and $s_{v_4} = 1$.

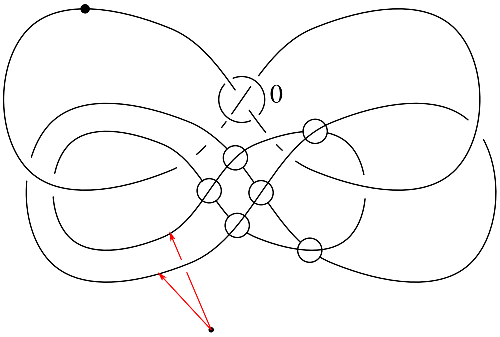


Figure 39: We delete a neighborhood of F_2. The resulting singular banded unlink diagram of $F_1 \cup F_1'$ is in a Kirby diagram with one 1-handle and one 2-handle.

$-\eta'$ and $-\gamma'$. There are four vertices v_1, v_2, v_3 and v_4 shared between the diagrams of F_1 and F_1'; see Figure 40 for potential loops C_{v_i} for all $i = 1, 2, 3, 4$. Note that each loop might pass through the other intersections in the singular banded unlink diagram, but we always can perturb each loop a little bit on the actual surface FR to miss the intersections.

Now each loop C_{v_i} represents some element of $H_1(S^4 - F_2) = \mathbb{Z}$. In addition, each vertex has a sign $s_{v_i} \in \{-1, +1\}$ given by the sign of the corresponding intersection of F_1 and F_1', which agrees with the sign of the crossing when the marking is resolved negatively. The values of $[C_{v_i}]$ and s_{v_i} are as follows:

i	s_{v_i}	$[C_{v_i}]$
1	1	0
2	-1	1
3	-1	-1
4	1	0

The Kirk invariant σ_1 is then given by

$$\sigma_1(\mathrm{FR}) = \sum_{i=1}^{4} s_{v_i} x^{[C_{v_i}]} = -x + 2 - x^{-1}.$$

The above computation generalizes for any singular banded unlink diagram of a 2-component spherical link (F_1, F_2) in S^4; use whiskers from a basepoint p to F_1 and a parallel pushoff F_1' intersecting F_1 in $v_1, \dots, v_n$ to form a loop C_{v_i} for each v_i representing $[C_{v_i}] \in H_1(S \setminus F_2) = \mathbb{Z}$. Then $\sigma_1(F_1, F_2) = \sum_{i=1}^{n} s_{v_i} x^{[C_{v_i}]}$.

4.2 Immersed surfaces and stabilization

Hosokawa and Kawauchi [13] showed that any pair of embedded oriented surfaces in S^4 become isotopic after some number of *stabilizations*.

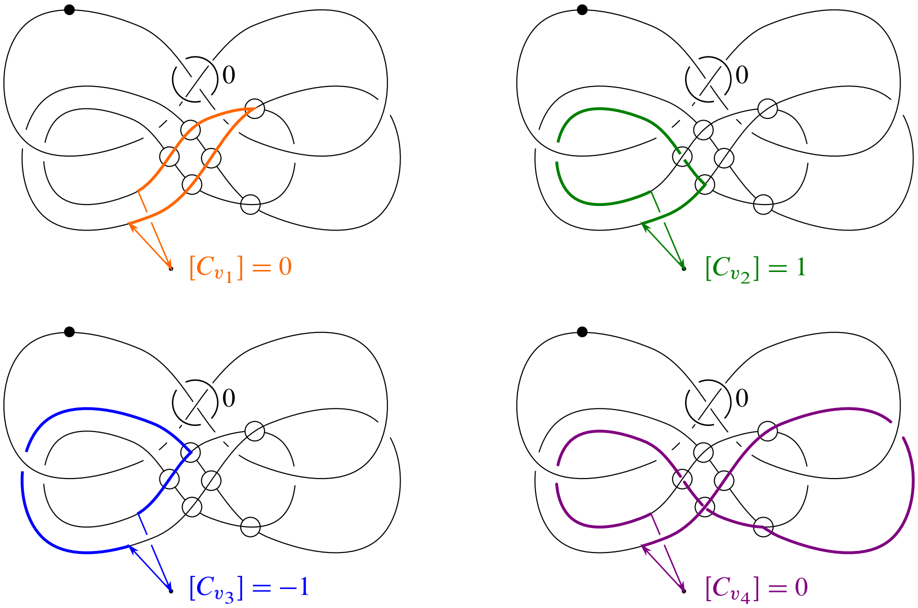


Figure 40: The loops C_{v_1}, C_{v_2}, C_{v_3} and C_{v_4}, respectively, represent the elements 0, 1, -1 and 0 in $H_1(S^4 \setminus F_2) = \mathbb{Z}$.

Definition 4.1 Let F be a connected, self-transversely immersed genus g oriented surface in S^4. Let γ be an arc with endpoints on F and which is normal to F near $\partial\gamma$, but with the interior of γ disjoint from F. Frame γ so that $\gamma \times D^2$ is a 3-dimensional 1-handle with ends on F, and so that surgering F along this 1-handle yields an oriented genus $g+1$ surface F'. Then we say F' is obtained from F by *stabilization.*

Remark 4.2 In Definition 4.1, there are two distinct ways to frame γ to obtain a 3-dimensional 1-handle with ends on F. However, one of these choices will yield a nonorientable surface after surgery, so in fact the framing of γ need not be specified.

More generally, Baykur and Sunukjian [2] extended this result for any pair of homologous embedded oriented surfaces in a closed orientable 4-manifold, and Kamada [18] extended it to immersed oriented surfaces in S^4 using singular braid charts. In this subsection, we extend these results in full generality, ie for any pair of homologous immersed surfaces in a closed orientable 4-manifold.

Theorem 4.3 *Let F and F' be oriented self-transversely immersed surfaces in a closed, orientable 4-manifold X which are homologous and have the same number of transverse double points of each sign. Then F and F' become isotopic after a sequence of stabilizations.*

To prove Theorem 4.3, we rely on the following diagrammatic lemma:

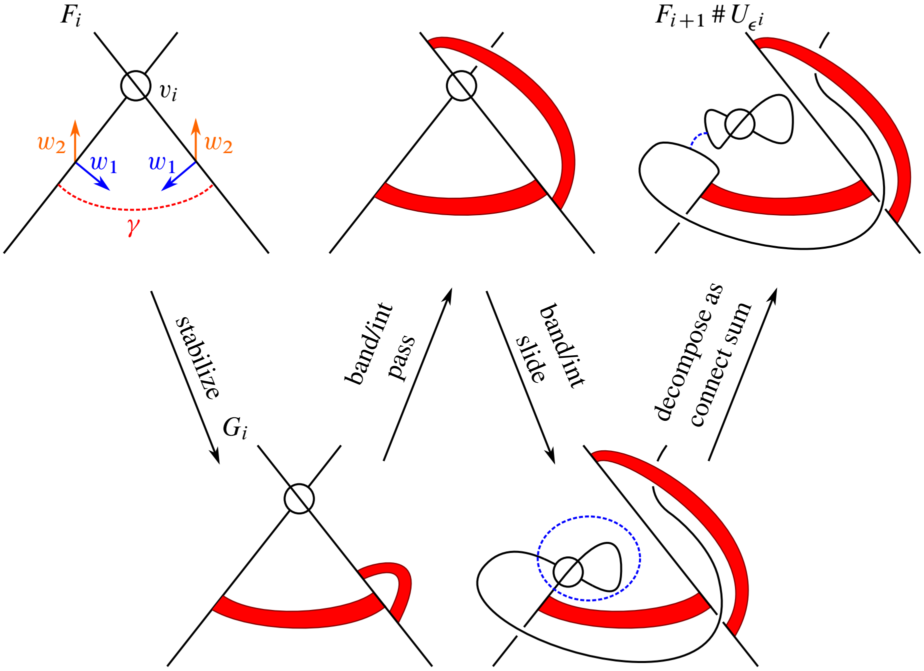


Figure 41: Top left: F_i is an oriented surface with $k-i>0$ transverse self-intersections. Here we draw part of a singular banded unlink diagram for F_i near a vertex v_i representing a self-intersection of F_i. (In this drawing, it is a negative self-intersection. Changing the marking at v_i yields a positive self-intersection.) We draw a positive normal basis (w_1, w_2) along each local sheet of F_i and indicate an arc γ along which we may stabilize F_i. From left to right following the arrows: we stabilize F_i to obtain a surface G_i, and then isotope G_i to realize a connect sum of a surface F_{i+1} with U_{ϵ^i}, where ϵ^i is the sign of the self-intersection represented by v_i.

Lemma 4.4 *Let F be an oriented self-transversely immersed surface in a closed, orientable 4-manifold. Suppose F has p positive and n negative self-intersections. After some number of stabilizations, F becomes isotopic to the connected-sum of an embedded surface with p copies of U_+ and n copies of U_-, where $U_\pm$ denotes the result of performing a cusp move to the embedded unknotted 2-sphere to create a $\pm$ self-intersection.*

Proof Let $(\mathcal{K}, L, B)$ be a singular banded unlink diagram of $F_0 := F$. Suppose that F has $k = p+n > 0$ self-intersections. Fix a vertex v_0 of L. Stabilize F_0 as in Figure 41, ie along an arc in $h^{-1}\left(\frac{3}{2}\right)$ that lies close to v_0. Call the resulting surface G_0. Now perform singular band moves as in Figure 41 to see that G_0 is isotopic to a connect sum $F_1 \# U_{\epsilon^0}$, where ϵ^0 is the sign of v_0 and F_1 is a self-transverse immersed surface with $k-1$ self-intersections.

If $k-1>0$, then repeat this argument on F_1 near another vertex v_1, stabilizing F_1 to obtain a surface G_1 that is isotopic to $F_2 \# U_{\epsilon^1}$, where F_2 has $k-2$ self-intersections. Note F is then stably isotopic to $F_2 \# U_{\epsilon^1} \# U_{\epsilon^0}$.

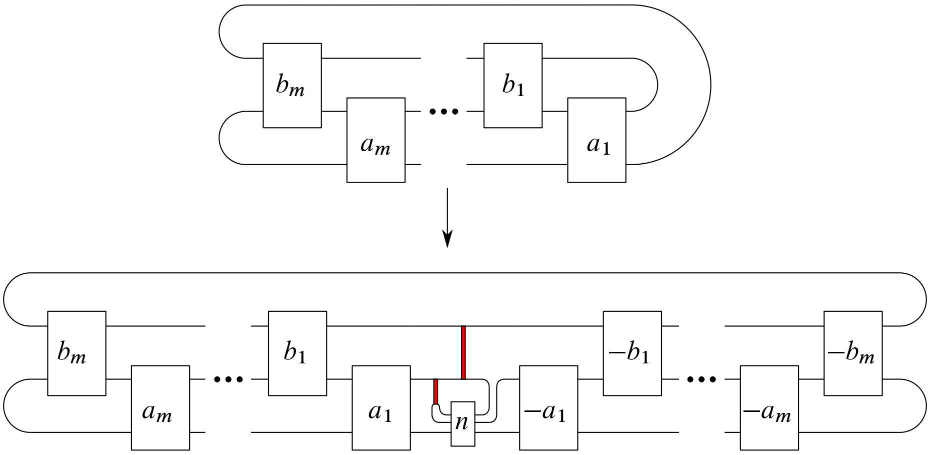


Figure 42: Top: a 2-bridge knot K in normal form. Here, a_i and b_i indicate signed numbers of whole twists (so each box has an even number of half-twists). Bottom: the n-twist spin $\tau^n K$ of K.

Repeat inductively to find that F is stably isotopic to $F_k \# (\#_p U_+) \# (\#_n U_-)$ for F_k an embedded surface, as desired. □

Proof of Theorem 4.3 By Lemma 4.4, F may be stabilized to a surface isotopic to $\hat{F}\#(\#_p U_+)\#(\#_n U_-)$, where $\hat{F}$ is an embedded surface and p and n are (respectively) the numbers of positive and negative

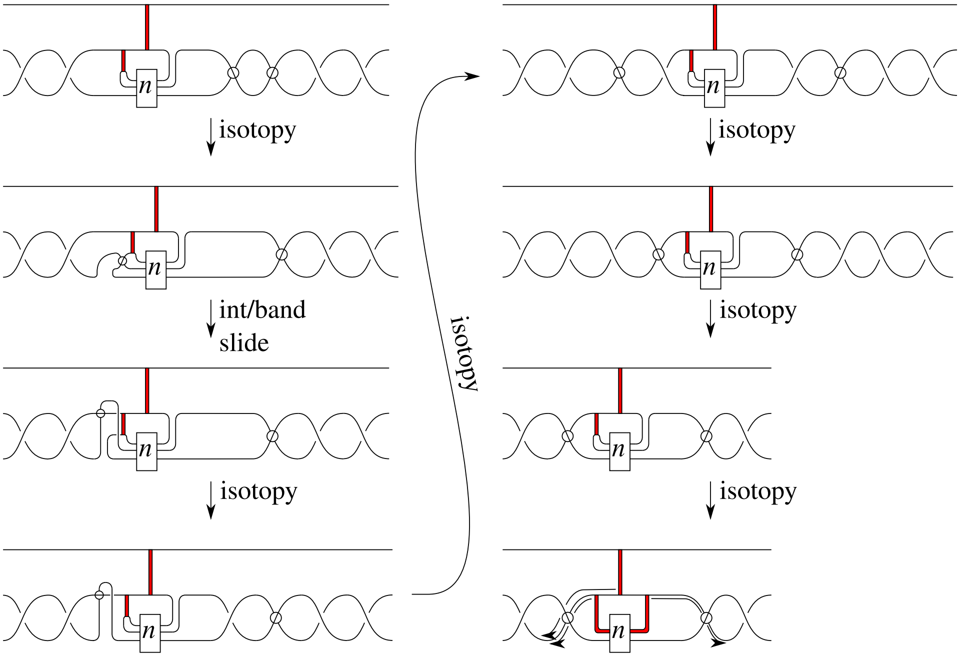


Figure 43: The first frame is (a portion of the diagram) obtained from Figure 42, bottom, by a finger move. We begin applying singular band moves with the goal of decreasing $|a_1|$ by 1. In the last frame we indicate three band/intersection passes that yield the first frame of Figure 44.

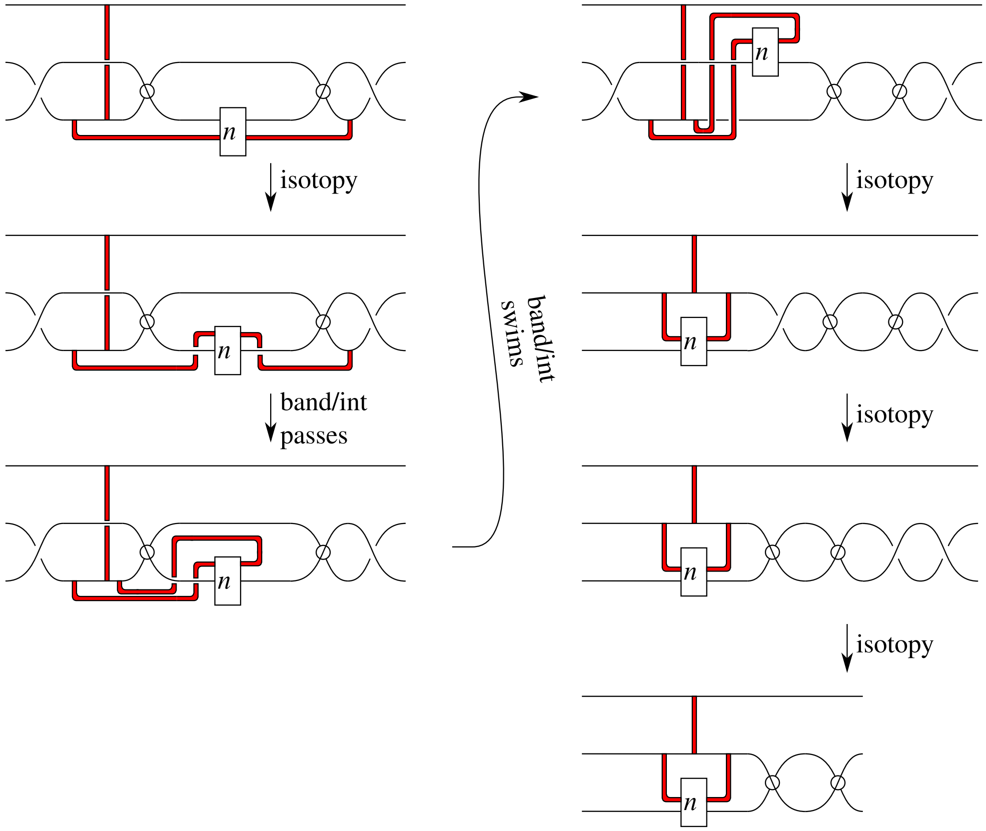


Figure 44: Continuing from Figure 43, we perform more singular band moves. In the last frame, the two vertices can be removed by a Whitney move, yielding the diagram from Figure 42, bottom, but with $|a_1|$ decreased by 1.

self-intersections of F. Applying the lemma also to F' (recalling that F' also has p positive and n negative self-intersections), we find that, after suitable stabilizations, F' becomes isotopic to

$$\widehat{F}' \# (\#_p U_+) \# (\#_n U_-)$$

for some embedded surface $\widehat{F}'$. Since $U_\pm$ is nullhomologous, $\widehat{F}$ and $\widehat{F}'$ are homologous to F and F' and hence to each other. Then, by [2], we know that $\widehat{F}$ and $\widehat{F}'$ (and hence F and F') are stably isotopic. □

4.3 Unknotting 2-knots with regular homotopies

Joseph, Klug, Ruppik and Schwartz [16] introduced the notion of the *Casson–Whitney number* of a 2-knot, which is half the minimal number of finger and Whitney moves needed to change a given 2-knot to an unknot. They showed that the Casson–Whitney number of any nontrivial twist spin of a 2-bridge knot is 1; ie that any nontrivial twist spin of a 2-bridge knot can be unknotted via one finger move followed by one Whitney move. In this subsection, we explicitly realize such a regular homotopy via singular banded unlink diagrams.

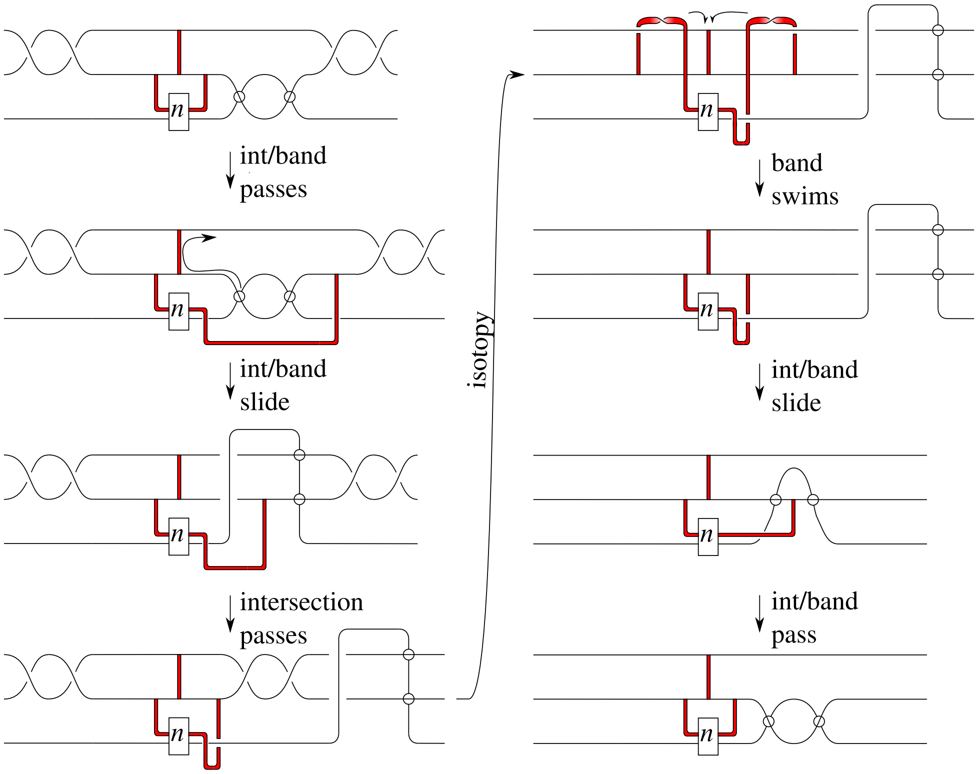


Figure 45: The first frame agrees with the last frame of Figure 44 after $|a_1|$ is decreased to zero. We can then perform singular band moves to the diagram to decrease $|b_1|$ by 1.

Theorem 4.5 [16] *The Casson–Whitney number of the n-twist spin ($|n| \neq 1$) $\tau^n K$ of a 2-bridge knot K is* 1.

Proof First, as in [16], we assume that the 2-bridge knot K is in normal form [5] with the number of half-twists in each twist region even, as in Figure 42. That is, using the standard correspondence between 2-bridge link diagrams and continued fraction expansion, we arrange for a diagram of K to correspond to a continued fraction $(a_1, b_1, \dots, a_m, b_m)$ of all even integers. We write $K = K(a_1, b_1, \dots, a_m, b_m)$.

Apply a finger move to the diagram of $\tau^n K$ in Figure 42 to obtain the first frame of Figure 43 (the visible twists are contained in the $\pm a_1$ twist boxes). In Figures 43 and 44, we show how to perform singular band moves with the result of decreasing $|a_1|$ by 1. Repeating this sequence, we eventually arrange for a_1 to become 0.

In Figure 45, we give another sequence of band moves (now assuming $a_1 = 0$) that decrease $|b_1|$ by 1. Repeating this sequence, we eventually arrange for $a_1 = b_1 = 0$.

We repeat these sequences of band moves to undo the twist boxes labeled $\pm a_2, \pm b_2, \dots, \pm a_m, \pm b_m$, and then finally apply a Whitney move to remove the two vertices and obtain a singular banded unlink diagram for the n-twist spin of the unknot. This is an unknotted sphere, so we conclude that the Casson–Whitney number of $\tau^n K$ is 1. □

References

[1] **V I Arnold**, **A N Varchenko**, **S M Guseĭn-Zade**, Особенности дифференцируемых отображений, Nauka, Moscow (1982) MR

[2] **R İ Baykur**, **N Sunukjian**, *Knotted surfaces in* 4*-manifolds and stabilizations*, J. Topol. 9 (2016) 215–231 MR Zbl

[3] **J S Carter**, **M Saito**, *Reidemeister moves for surface isotopies and their interpretation as moves to movies*, J. Knot Theory Ramifications 2 (1993) 251–284 MR Zbl

[4] **J S Carter**, **M Saito**, *Knotted surfaces and their diagrams*, Math. Surv. Monogr. 55, Amer. Math. Soc., Providence, RI (1998) MR Zbl

[5] **J H Conway**, *An enumeration of knots and links, and some of their algebraic properties*, from "Computational problems in abstract algebra", Pergamon, Oxford (1970) 329–358 MR Zbl

[6] **M H Freedman**, *The topology of four-dimensional manifolds*, J. Differential Geom. 17 (1982) 357–453 MR Zbl

[7] **M H Freedman**, **F Quinn**, *Topology of* 4*-manifolds*, Princeton Math. Ser. 39, Princeton Univ. Press (1990) MR Zbl

[8] **D Gay**, **R Kirby**, *Trisecting* 4*-manifolds*, Geom. Topol. 20 (2016) 3097–3132 MR Zbl

[9] **A Hatcher**, **J Wagoner**, *Pseudo-isotopies of compact manifolds*, Astérisque 6, Soc. Math. France, Paris (1973) MR Zbl

[10] **C Hayashi**, **K Shimokawa**, *Heegaard splittings of trivial arcs in compression bodies*, J. Knot Theory Ramifications 10 (2001) 71–87 MR Zbl

[11] **M W Hirsch**, *Immersions of manifolds*, Trans. Amer. Math. Soc. 93 (1959) 242–276 MR Zbl

[12] **M W Hirsch**, *Differential topology*, Graduate Texts in Math. 33, Springer (1976) MR Zbl

[13] **F Hosokawa**, **A Kawauchi**, *Proposals for unknotted surfaces in four-spaces*, Osaka Math. J. 16 (1979) 233–248 MR Zbl

[14] **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

[15] **M Jabłonowski**, *Minimal generating sets of moves for surfaces immersed in the four-space*, J. Knot Theory Ramifications 32 (2023) art. id. 2350071 MR Zbl

[16] **J M Joseph**, **M R Klug**, **B M Ruppik**, **H R Schwartz**, *Unknotting numbers of* 2*-spheres in the* 4*-sphere*, J. Topol. 14 (2021) 1321–1350 MR Zbl

[17] **S Kamada**, 2*-dimensional braids and chart descriptions*, from "Topics in knot theory", NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci. 399, Kluwer, Dordrecht (1993) 277–287 MR Zbl

[18] **S Kamada**, *Unknotting immersed surface-links and singular* 2*-dimensional braids by* 1*-handle surgeries*, Osaka J. Math. 36 (1999) 33–49 MR Zbl

[19] **S Kamada**, *Braid and knot theory in dimension four*, Math. Surv. Monogr. 95, Amer. Math. Soc., Providence, RI (2002) MR Zbl

[20] **S Kamada**, **A Kawauchi**, **J Kim**, **S Y Lee**, *Presentation of immersed surface-links by marked graph diagrams*, J. Knot Theory Ramifications 27 (2018) art. id. 1850052 MR Zbl

[21] **A Kawauchi**, **T Shibuya**, **S Suzuki**, *Descriptions on surfaces in four-space*, *I*: *Normal forms*, Math. Sem. Notes Kobe Univ. 10 (1982) 75–125 MR Zbl

[22] **C Kearton**, **V Kurlin**, *All* 2*-dimensional links in* 4*-space live inside a universal* 3*-dimensional polyhedron*, Algebr. Geom. Topol. 8 (2008) 1223–1247 MR Zbl

[23] **F Laudenbach**, *Sur les* 2*-sphères d'une variété de dimension* 3, Ann. of Math. 97 (1973) 57–81 MR

[24] **F Laudenbach**, **V Poénaru**, *A note on* 4*-dimensional handlebodies*, Bull. Soc. Math. France 100 (1972) 337–344 MR Zbl

[25] **J Meier**, **T Schirmer**, **A Zupan**, *Classification of trisections and the generalized property R conjecture*, Proc. Amer. Math. Soc. 144 (2016) 4983–4997 MR Zbl

[26] **J Meier**, **A Zupan**, *Bridge trisections of knotted surfaces in S^4*, Trans. Amer. Math. Soc. 369 (2017) 7343–7386 MR Zbl

[27] **J Meier**, **A Zupan**, *Bridge trisections of knotted surfaces in* 4*-manifolds*, Proc. Natl. Acad. Sci. USA 115 (2018) 10880–10886 MR Zbl

[28] **M Miller**, **P Naylor**, *Trisections of nonorientable* 4*-manifolds*, Michigan Math. J. 74 (2024) 403–447 MR Zbl

[29] **D Roseman**, *Reidemeister-type moves for surfaces in four-dimensional space*, from "Knot theory", Banach Center Publ. 42, Polish Acad. Sci. Inst. Math., Warsaw (1998) 347–380 MR Zbl

[30] **R Schneiderman**, **P Teichner**, *The group of disjoint* 2*-spheres in* 4*-space*, Ann. of Math. 190 (2019) 669–750 MR Zbl

[31] **S Smale**, *A classification of immersions of the two-sphere*, Trans. Amer. Math. Soc. 90 (1958) 281–290 MR Zbl

[32] **F J Swenton**, *On a calculus for* 2*-knots and surfaces in* 4*-space*, J. Knot Theory Ramifications 10 (2001) 1133–1141 MR Zbl

Department of Mathematics, Brigham Young University
Provo, UT, United States

Sungkyunkwan University
Suwon, South Korea

Department of Mathematics, The University of Texas at Austin
Austin, TX, United States

hughes@mathematics.byu.edu, seungwon.kim@skku.edu, maggie.miller.math@gmail.com

Received: 29 April 2023 Revised: 27 November 2023

Algebraic & Geometric Topology **25**:3 (2025) 1793–1838
DOI: 10.2140/agt.2025.25.1793
Published: 20 June 2025

Anosov flows and Liouville pairs in dimension three

THOMAS MASSONI

Building upon the work of Mitsumatsu and Hozoori, we establish a complete homotopy correspondence between three-dimensional Anosov flows and certain pairs of contact forms that we call *Anosov Liouville pairs*. We show a similar correspondence between projectively Anosov flows and bicontact structures, extending the work of Mitsumatsu and Eliashberg–Thurston. As a consequence, every Anosov flow on a closed oriented three-manifold M gives rise to a Liouville structure on $\mathbb{R} \times M$ which is well-defined up to homotopy, and which only depends on the homotopy class of the Anosov flow. Our results also provide a new perspective on the classification problem of Anosov flows in dimension three.

53E50, 57K33, 57K43; 37D20, 37D30

1 Introduction

Throughout this article, M denotes a closed, oriented, smooth manifold of dimension three. We will always assume that the Anosov and projectively Anosov flows on M under consideration are *oriented*, ie their stable and unstable foliations are oriented. This can always be achieved by passing to a suitable double cover of M. For simplicity, we will only consider smooth (ie $\mathcal{C}^\infty$) flows, as we are primarily interested in smooth contact and symplectic structures. Our main results hold for (projectively) Anosov flows generated by $\mathcal{C}^1$ vector fields with minor changes. Moreover, the structural stability of $\mathcal{C}^1$ Anosov vector fields (Robinson [29]) ensures that any Anosov flow generated by a $\mathcal{C}^1$ vector field is topologically equivalent to a smooth Anosov flow, and these two flows are dynamically identical. The definitions and basic properties of Anosov and projectively Anosov flows are recalled in Section 3.1.

The notion of Anosov flow, originally introduced by Anosov [1; 2] as a generalization of the geodesic flow on hyperbolic manifold, plays a central role in the theory of smooth dynamical systems. The interplay between the dynamical and topological properties of Anosov flows is particularly rich and striking in dimension three. We refer to the nice survey by Barthelmé [4] for many relevant results and references, and to the book by Fisher and Hasselblatt [14] for a more complete exposition. Eliashberg and Thurston [13], and independently Mitsumatsu [28], introduced the more general concept of a *conformally/projectively Anosov flow* on three-manifolds, and established a correspondence between such flows and *bicontact structures*, ie transverse pairs of contact structures with opposite orientations. Recently, Hozoori [21] extended this correspondence to Anosov flows, and showed that (oriented) Anosov flows can be completely characterized in terms of bicontact structures admitting a pair of contact forms satisfying a natural symplectic condition. More precisely, Hozoori showed:

Theorem [21, Theorem 1.1] *Let Φ be a nonsingular flow on a closed oriented 3-manifold M, generated by a vector field X. The flow Φ is oriented Anosov if and only if there exist transverse contact structures ξ_- and ξ_+, negative and positive, respectively, and contact forms α_- and α_+ for ξ_- and ξ_+, respectively, such that the 1-forms*

$$(1-t)\,\alpha_- + (1+t)\,\alpha_+ \quad \textit{and} \quad -(1-t)\,\alpha_- + (1+t)\,\alpha_+$$

are positively oriented Liouville forms on $[-1,1]_t \times M$.

Recall that a Liouville form on a smooth manifold with boundary V is a 1-form λ such that $\omega = d\lambda$ is symplectic, ie nondegenerate, and the Liouville vector field Z defined by $\omega(Z,\cdot\,) = \lambda$ is outward-pointing along the boundary of V. The pair (V,λ) is called a *Liouville domain*. The above theorem shows in particular that an Anosov flow on a 3-manifold M (under some suitable orientability assumptions recalled in Definition 3.1) gives rise to a Liouville structure on $[-1,1]\times M$ which is *not Weinstein*, since the latter manifold has a nontrivial third homology group and disconnected boundary. It is natural to ask:

Questions How do the Liouville structures constructed by Hozoori depend on the underlying Anosov flow? More precisely:

(1) For a given Anosov flow Φ, is the space of pairs of contact forms (α_-,α_+) as in the previous theorem path-connected?

(2) Does a path of Anosov flows induce a path of Liouville structures on $[-1,1]\times M$?

(3) Does every bicontact structure (ξ_-,ξ_+) supporting an Anosov flow admit a pair of contact forms (α_-,α_+) as in the previous theorem?

Here, we say that a bicontact structure (ξ_-,ξ_+) supports a nonsingular flow generated by a vector field X if $X \in \xi_- \cap \xi_+$ (in the more precise Definition 2.1, we also add a condition on the orientations of $\xi_\pm$).

In the present article, we give a complete answer to these questions and upgrade Hozoori's correspondence to a *homotopy equivalence* between the space of Anosov flows on M, and a space of suitable pairs of contact forms on M. To that extent, we will consider a *different* condition on the pair (α_-, α_+) than the one in Hozoori's theorem, and we first show:

Theorem 1 *Let Φ be a nonsingular flow on a closed oriented 3-manifold M, generated by a vector field X. The flow Φ is oriented Anosov if and only if there exists a pair of contact forms (α_-, α_+) on M such that $X \in \ker\alpha_- \cap \ker\alpha_+$, and the 1-forms*

$$e^{-s}\alpha_- + e^{s}\alpha_+ \quad \text{and} \quad -e^{-s}\alpha_- + e^{s}\alpha_+$$

are positively oriented Liouville forms on $\mathbb{R}_s \times M$.

In the terminology of Massot, Niederkrüger and Wendl [25, Definition 1], we say that a pair of contact forms (α_-, α_+) on a manifold M is a *Liouville pair* if the 1-form

$$\lambda := e^{-s}\alpha_- + e^{s}\alpha_+$$

is a positively oriented Liouville form on $\mathbb{R}_s \times M$. By positively oriented, we mean that the volume form $d\lambda \wedge d\lambda$ is compatible with the natural orientation on $\mathbb{R} \times M$ induced by the natural orientation on $\mathbb{R}$ and the orientation on M.

Warning At first glance, Theorem 1 seems almost identical to Hozoori's theorem. However, we warn the reader that the condition on (α_-, α_+) that we consider is *different* than Hozoori's one. Indeed, there exist pairs of contact forms (α_-, α_+) which are Liouville pairs as defined above, but such that

$$(1-t)\alpha_- + (1+t)\alpha_+$$

is *not* a Liouville form on $[-1,1]_t \times M$; see Lemma 5.6. It turns out that our condition enjoys some nice symmetries (see Lemma 2.4) which make it much easier to work with. For instance, our notion of Liouville pair is easier to characterize than Hozoori's one (compare Lemma 2.7 which involves a single equation between three quantities, and Lemma 5.2 which involves two independent equations between four quantities). More importantly, *we do not know if our main results* (*Theorems 3 and 10 below*) *are true for Hozoori's notion of Liouville pair.* The corresponding computations are much more complicated because of their lack of symmetry.

Theorem 1 motivates the following:

Definition 2 An *Anosov Liouville pair* (AL pair for short) on an oriented 3-manifold M is a pair of contact forms (α_-, α_+) such that both (α_-, α_+) and $(-\alpha_-, \alpha_+)$ are Liouville pairs. We denote by $\mathcal{AL} := \mathcal{AL}(M) \subset \Omega^1(M) \times \Omega^1(M)$ the space of Anosov Liouville pairs on M.

Notice that we do not assume that $\xi_\pm := \ker\alpha_\pm$ are transverse, since this is implied by the Liouville conditions; see Proposition 2.9. By Theorem 1, the intersection $\xi_- \cap \xi_+$ is spanned by an Anosov vector field. A positive time reparametrization of an Anosov flow remains Anosov, and we denote by

$\mathcal{AF} := \mathcal{AF}(M)$ the space of smooth oriented Anosov flows on M up to positive time reparametrization. Alternatively, $\mathcal{AF}$ can be viewed as the space of smooth unit Anosov vector fields on M for an arbitrary Riemannian metric on M, or the space of smooth 1-dimensional oriented foliations spanned by Anosov vector fields on M, together with some extra orientation data. Hence, there is a natural continuous *intersection map*,

$$\mathcal{I}\colon \mathcal{AL} \to \mathcal{AF}, \quad (\alpha_-, \alpha_+) \mapsto \ker\alpha_- \cap \ker\alpha_+,$$

which sends an AL pair to the 1-dimensional (oriented) distribution obtained by intersecting the underlying contact structures. Here, we endow the spaces $\mathcal{AL}$ and $\mathcal{AF}$ with the $\mathcal{C}^\infty$ topology. Denoting by $\mathcal{BC}$ the space of smooth bicontact structures on M and by $\mathbb{P}\mathcal{AF}$ the space of smooth oriented projectively Anosov flows on M up to positive time reparametrization, we have a similar intersection map,

$$\mathbb{P}\mathcal{I}\colon \mathcal{BC} \to \mathbb{P}\mathcal{AF}, \quad (\xi_-, \xi_+) \mapsto \xi_- \cap \xi_+,$$

as well as a *kernel map*,

$$\underline{\ker}\colon \mathcal{AL} \to \mathcal{BC}, \quad (\alpha_-, \alpha_+) \mapsto (\ker\alpha_-, \ker\alpha_+).$$

The main results of this paper, answering the Questions (1), (2) and (3) above, can be summarized as follows.

Theorem 3 *The maps in the commutative diagram*

$$\begin{tikzcd}
\mathcal{AL} \arrow[r, "\underline{\ker}"] \arrow[d, "\mathcal{I}"'] & \mathcal{BC} \arrow[d, "\mathbb{P}\mathcal{I}"] \\
\mathcal{AF} \arrow[r, hook] & \mathbb{P}\mathcal{AF}
\end{tikzcd}$$

satisfy the following properties.

- *$\mathcal{I}$ and $\mathbb{P}\mathcal{I}$ are acyclic Serre fibrations* (*Theorems 4.8 and 4.9*).
- *$\underline{\ker}$ is an acyclic Serre fibration onto its image* (*Theorem 4.13*).
- *The inclusion $\underline{\ker}(\mathcal{AL}) \subset \mathbb{P}\mathcal{I}^{-1}(\mathcal{AF})$ is strict in general* (*Theorem 3.15*), *but it is a homotopy equivalence* (*Theorem 4.15*).

Recall that an acyclic Serre fibration is a Serre fibration which is also a weak homotopy equivalence, or equivalently, whose fibers are weakly contractible. All the topological spaces under consideration have the homotopy type of a CW complex (see the beginning of Section 4), so these acyclic Serre fibrations are homotopy equivalences by the Whitehead theorem. Unpacking the notations,

- $\underline{\ker}(\mathcal{AL})$ is the space of bicontact structures (ξ_-, ξ_+) admitting contact forms α_-, α_+ such that (α_-, α_+) is an AL pair,
- $\mathbb{P}\mathcal{I}^{-1}(\mathcal{AF})$ is the space of bicontact structures supporting an Anosov flow.

We emphasize that the top row in the diagram of Theorem 3 *only involves concepts from contact and symplectic geometry.* This enables us to identify projectively Anosov flows with bicontact structures, and

Anosov flows with bicontact structures satisfying a quantitative constraint, coming from the existence of a suitable pair of contact forms. Moreover, the space of AL pairs for a fixed underlying bicontact structure is (weakly) contractible if nonempty. Hence, AL pairs can be thought of as *auxiliary data attached to bicontact structures*.

Our results can be summarized by the following slogan:

> The topological properties of the spaces $\mathcal{AF}$, $\mathbb{P}\mathcal{AF}$ and the inclusion $\mathcal{AF} \subset \mathbb{P}\mathcal{AF}$ can be translated into topological properties of the spaces $\mathcal{AL}$, $\mathcal{BC}$, and the map $\underline{\mathrm{ker}}\colon \mathcal{AL} \to \mathcal{BC}$, and vice versa.

One important missing piece in this correspondence between Anosov dynamics and contact topology is the mirror notion of *topological* or *orbit equivalence* of flows in the contact world.

Definition 4 Two Anosov flows $\Phi = \{\phi^t\}$ and $\Psi = \{\psi^t\}$ on M are *topologically equivalent*, or *orbit equivalent*, if there exist a homeomorphism $h\colon M \to M$ and a continuous map $\tau\colon \mathbb{R} \times M \to \mathbb{R}$ such that $\tau(t,x) \geq 0$ for $t \geq 0$, and

$$\psi^{\tau(t,x)} = h \circ \phi^t \circ h^{-1}(x)$$

for every $t \in \mathbb{R}$ and $x \in M$.

In other words, the topological equivalence h sends the oriented trajectories of ϕ onto the oriented trajectories of ψ, but does not necessarily preserves the parametrization. The structural stability of Anosov flows with smooth dependence on parameters (de la Llave, Marco and Moriyón [22, Theorem A.1]) implies that two smooth Anosov flows which are homotopic through smooth Anosov flows are topologically equivalent through a topological equivalence which is isotopic to the identity. We do not know if the converse is true.

Question 5 If two (smooth) Anosov flows on M are topologically equivalent (via a topological equivalence which is merely continuous), what can be said about the spaces of AL pairs supporting them? How to characterize topological equivalence in terms of AL pairs?

It is not clear to us how the (hyper)tight contact structure $\xi_\pm$ associated with a Anosov flow behave under topological equivalence. Solving these questions could have a significant impact in the understanding of Anosov flows from the perspective of contact geometry. For instance, a fundamental problem in 3-dimensional Anosov dynamics is the following:

Question 6 (Barthelmé [4]) On a closed 3-manifold, are there finitely many Anosov flows up to topological equivalence?

It is known by the work of Colin, Giroux and Honda [10] that an atoroidal 3-manifold carries finitely many isotopy classes of *tight* contact structures. Although toroidal (and irreducible) 3-manifolds can carry infinitely many isotopy classes of tight contact structures, all of them can be obtained from finitely many contact structures by performing *Lutz twists* along suitable tori; see [10]. The authors also show that there are finitely many tight contact structures for a prescribed *Giroux torsion*, up to isotopy and Dehn twists. Since the contact structures defined by (Anosov) Liouville pairs are by definition exactly semifillable, they are strongly fillable (Eliashberg [12, Corollary 1.4]), hence they have zero Giroux torsion (Gay [15, Corollary 3]). This observation plays an essential role in the recent solution of Question 6 for the class of $\mathbb{R}$-*covered* Anosov flows (Barthelmé and Mann [5]; Marty [23]).

We hope that this *coarse classification* of tight contact structures together with our homotopy correspondence could lead to important results in the classification of Anosov flows on 3-manifolds. To this end, it is crucial to understand the following.

Question 7 Let (α_-, α_+) be an AL pair on M. Fixing α_+, what can be said about the Anosov flow supported by an AL pair (α'_-, α_+), where α'_- is isotopic to α_-?

The main difficulty here is that a path $(\alpha^t_-)_{t\in[0,1]}$ of contact forms from $\alpha^0_- = \alpha_-$ to $\alpha^1_- = \alpha'_-$ might not induce a path of bicontact structures, as $\xi^t_- = \ker\alpha^t_-$ and $\xi_+ = \ker\alpha_+$ might fail to be transverse for some $t \in (0,1)$. Even if transversality holds, (α^t_-, α_+) might fail to be an AL pair. Nevertheless, one could try to analyze the failure of these properties for a generic path $(\alpha^t_-)_t$, and apply suitable modifications to it. We wish to explore this direction in future work.

A closely related question, already raised by Hozoori [21, Question 7.2] is the following.

Question 8 Let Φ_0 and Φ_1 be two Anosov flows on M and assume that they are homotopic through *projectively* Anosov flows. Equivalently, assume that there exist two AL pairs (α^0_-, α^0_+) and (α^1_-, α^1_+) supporting Φ_0 and Φ_1, respectively, such that their underlying bicontact structures are homotopic (through bicontact structures). Are Φ_0 and Φ_1 homotopic through Anosov flows, ie are (α^0_-, α^0_+) and (α^1_-, α^1_+) homotopic through AL pairs? Are Φ_0 and Φ_1 topologically equivalent?

From the point of view of Liouville geometry, it is natural to weaken the definition of AL pairs as follows.

Definition 9 A Liouville pair (α_-, α_+) on M is a *weak Anosov Liouville pair* (wAL pair for short) if it satisfies the following two conditions.

(1) The contact plane fields $\xi_\pm := \ker\alpha_\pm$ are everywhere transverse.

(2) The intersection $\xi_- \cap \xi_+$ is spanned by an Anosov vector field.

An *Anosov Liouville structure* (AL structure for short) on $V = \mathbb{R}_s \times M$ is a pair (ω, λ), where $\omega = d\lambda$ is a symplectic form and

$$\lambda = e^{-s}\alpha_- + e^s\alpha_+$$

for a weak Anosov Liouville pair (α_-, α_+). We call the triple (V, ω, λ) an *Anosov Liouville manifold.*

An Anosov flow Φ is *supported* by the AL structure (ω, λ) if the vector field X generating Φ satisfies $X \in \xi_- \cap \xi_+$.

Note that the definition of wAL pairs *does* make reference to the underlying Anosov flow, as opposed to AL pairs. By Theorem 1, there is an inclusion $\mathcal{AL} \subset \mathcal{AL}^w$, where $\mathcal{AL}^w$ denotes the space of wAL pairs on M. This inclusion is strict in general. The map $\mathcal{I}$ naturally extends to a map $\mathcal{I}^w : \mathcal{AL}^w \to \mathcal{AF}$, and similarly to the first bullet of Theorem 3, we have:

Theorem 10 *The map $\mathcal{I}^w : \mathcal{AL}^w \to \mathcal{AF}$ is an acyclic Serre fibration, hence a homotopy equivalence.*

Corollary 11 *Let Φ_0 and Φ_1 be two Anosov flows on M, supported by AL structures (ω_0, λ_0) and (ω_1, λ_1), respectively. If Φ_0 and Φ_1 are homotopic through Anosov flows, then (ω_0, λ_0) and (ω_1, λ_1) are homotopic through AL structures, and (V, ω_0, λ_0) and (V, ω_1, λ_1) are exact symplectomorphic.*

Here, an exact symplectomorphism $\psi : (V, \omega_0, \lambda_0) \to (V, \omega_1, \lambda_1)$ is a diffeomorphism such that $\psi^* \lambda_1 = \lambda_0 + df$ for some smooth function $f : V \to \mathbb{R}$. In Corollary 11, we can further assume that df has compact support.

Proof of Corollary 11 If Φ_0 and Φ_1 are homotopic through Anosov flows, Theorem 10 provides a *continuous* path of smooth AL structures from (ω_0, λ_0) to (ω_1, λ_1). This path can be smoothed while ensuring the existence of some number $A > 0$ such that the corresponding Liouville vector fields are all transverse to $\{\pm A\} \times M$. Then Cieliebak and Eliashberg [7, Proposition 11.8] provide an exact symplectomorphism ψ such that $\psi^* \lambda_1 - \lambda_0$ is compactly supported. □

Anosov Liouville manifolds have numerous interesting invariants coming from Floer theory, eg symplectic cohomology and wrapped Fukaya category. As an important consequence of Corollary 11, these are *invariants of the underlying Anosov flow*, and only depend on its homotopy class in the space of Anosov flows. Some of these invariants were studied in detail with Cieliebak, Lazarev and Moreno [8]. To our knowledge, this is the first thorough analysis of symplectic invariants of non-Weinstein Liouville manifolds.

One can also consider Liouville pairs (α_-, α_+) whose underlying contact planes are everywhere transverse. We call such pairs *transverse Liouville pairs*. They correspond to particular projectively Anosov flow that we call *semi-Anosov flows*, see Remark 3.12 below. General Liouville pairs (without the transversality assumption) are more complicated to understand, but their underlying contact planes can only intersect *positively*, see Remark 2.10 below. In the terminology of Colin and Firmo [9], they constitute *positive contact pairs*.

These geometric structures are summarized in the following diagram; the ones in blue are the main protagonists of this article. Liouville pairs and positive contact pairs will be investigated in forthcoming work [24]:

$$\begin{tikzcd}
\{\text{AL pairs}\} \arrow[r, phantom, "\subset"] \arrow[rrd, bend right, "\ker"'] & \{\text{weak AL pairs}\} \arrow[r, phantom, "\subset"] & \{\text{transverse Liouville pairs}\} \arrow[r, phantom, "\subset"] \arrow[d, "\ker"'] & \{\text{Liouville pairs}\} \arrow[d, "\ker"'] \\
 & & \{\text{bicontact structures}\} \arrow[r, phantom, "\subset"] & \{\text{positive contact pairs}\}
\end{tikzcd}$$

Acknowledgments

I am grateful to my PhD advisor John Pardon for his constant support and encouragement. I would like to thank Sergio Fenley, Jonathan Zung and Malo Jézéquel for insightful conversations about Anosov flows, and Surena Hozoori for multiple discussions about his work and fruitful exchanges. I am grateful to the referees for pointing out several typos and inaccuracies, as well as providing valuable suggestions.

2 Anosov Liouville pairs

2.1 Preliminary definitions

If X is a nonsingular vector field on M, we write

$$N_X := TM/\langle X\rangle.$$

An orientation on M naturally determines an orientation on the plane bundle $N_X \to M$. We denote by $\pi : TM \to N_X$ the quotient map. There is a correspondence between n-forms α on M satisfying $\iota_X\alpha = 0$ and n-forms $\overline{\alpha}$ on N_X. Moreover, a vector field Y on M induces a section $\overline{Y} := \pi(Y)$ on N_X. The operator $\mathcal{L}_X$, the Lie derivative along X, naturally induces an operator, still denoted by $\mathcal{L}_X$, on sections of N_X and on n-forms on N_X.

Definition 2.1 A *bicontact structure* on an oriented 3-manifold M is a pair of cooriented contact structures (ξ_-, ξ_+) such that ξ_- is negative, ξ_+ is positive and ξ_- and ξ_+ are transverse everywhere.

A nonsingular flow Φ on M generated by a vector field X is *supported* by a bicontact structure (ξ_-, ξ_+) if $X \in \xi_- \cap \xi_+$, and the following orientation compatibility condition holds. Let $\overline{\xi}_\pm \subseteq N_X$ be the image of $\xi_\pm$ under the quotient map $\pi : TM \to N_X$. The orientations on M, $\xi_\pm$ and X induce natural orientations on N_X and $\overline{\xi}_\pm$. We require that the orientation on N_X coincides with the one on $\overline{\xi}_- \oplus \overline{\xi}_+$ (see Figures 1 and 2).

Similarly, Φ is supported by a (weak) Anosov Liouville pair (α_-, α_+) if it is supported by the bicontact structure $(\xi_-, \xi_+) = (\ker\alpha_-, \ker\alpha_+)$.

Note that the definitions of bicontact structures and (weak) Anosov Liouville pairs still make sense if $\xi_\pm$, or $\alpha_\pm$, are merely $\mathcal{C}^1$. We will always assume that bicontact structures and (weak) Anosov Liouville pairs are smooth unless stated otherwise. Bicontact structures and (weak) Anosov Liouville pairs obviously constitute open subsets of the space of pairs of 2-plane fields on M and the space of pairs of 1-forms on M, respectively, since they are defined by open conditions.

If (ξ_-, ξ_+) is a bicontact structure supporting a nonsingular flow $\Phi = \{\phi^t\}$, then the bicontact structure obtained from (ξ_-, ξ_+) by reversing the coorientations of both ξ_- and ξ_+ supports Φ as well. Reversing the coorientation of ξ_- or ξ_+ only yields a bicontact structure supporting the reversed flow $\Phi^{-1} = \{\phi^{-t}\}$.

It is easy to deduce from Theorem 1 the very well-known corollary:

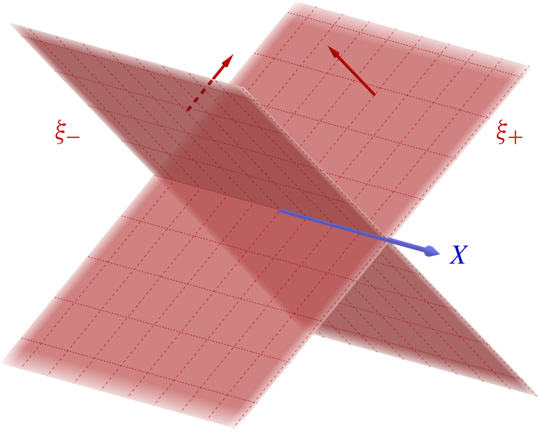


Figure 1: Coorientation convention for bicontact structures supporting a vector field or a flow.

Corollary 2.2 *The space of (smooth, $\mathcal{C}^1$) Anosov vector fields on M is open in the $\mathcal{C}^1$ topology.*

Proof Let X be an Anosov vector field on M and (α_-, α_+) be an AL pair supporting X. We choose a 1-form θ such that $\theta(X) \equiv 1$. If X' is another vector field which is sufficiently $\mathcal{C}^1$-close to X, the pair (α'_-, α'_+) defined by

$$\alpha'_\pm := \alpha_\pm - \frac{\alpha_\pm(X')}{\theta(X')}\theta$$

is an AL pair supporting X' and by Theorem 1, X' is Anosov. □

If (α_-, α_+) is an AL pair on M and $\sigma\colon M \to \mathbb{R}$ is a smooth function, it follows from the definition that

$$\sigma \cdot (\alpha_-, \alpha_+) := (e^{-\sigma}\alpha_-, e^{\sigma}\alpha_+)$$

is also an AL pair that defines the same bicontact structure as (α_-, α_+). These two AL pairs will be called *equivalent*. This defines an action of $\mathcal{C}^\infty(M, \mathbb{R})$ on the space of AL pairs.

Definition 2.3 A pair of contact forms (α_-, α_+) on M, negative and positive, respectively, is *balanced* if

$$\alpha_+ \wedge d\alpha_+ = -\alpha_- \wedge d\alpha_-.$$

In other words, (α_-, α_+) is balanced if $\alpha_\pm$ define opposite volume forms on M.

Lemma 2.4 *Two equivalent AL pairs on M define Liouville isomorphic Liouville structures on $\mathbb{R} \times M$. Any AL pair on M is equivalent to a (unique) balanced one.*

Proof Let (α_-, α_+) be an AL pair on M and $\lambda := e^{-s}\alpha_- + e^{s}\alpha_+$ be the corresponding Liouville form. If $\sigma \in \mathcal{C}^\infty(M, \mathbb{R})$ and $\lambda' := e^{-(s+\sigma)}\alpha_- + e^{s+\sigma}\alpha_+$, the diffeomorphism

$$\Psi\colon \mathbb{R} \times M \to \mathbb{R} \times M, \quad (s, x) \mapsto (s - \sigma(x), x),$$

satisfies $\Psi^*\lambda' = \lambda$. Moreover, if $f\colon M \to \mathbb{R}_{>0}$ is such that

$$\alpha_- \wedge d\alpha_- = -f\,\alpha_+ \wedge d\alpha_+,$$

then $\sigma \cdot (\alpha_-, \alpha_+)$ is balanced if and only if $\sigma = \frac{1}{4}\ln f$. □

As a straightforward application of Gray's stability theorem and the above lemma, we have:

Lemma 2.5 *Let (α_-, α_+) be an AL pair and let $\xi_+ := \ker\alpha_+$. If $\xi'_+ = \ker\alpha'_+$ is a contact structure homotopic to ξ_+, then there exists a path of AL pairs (α^t_-, α^t_+), $t \in [0,1]$, such that $(\alpha^0_-, \alpha^0_+) = (\alpha_-, \alpha_+)$ and $\alpha^1_+ = \alpha'_+$.*

Definition 2.6 A pair of contact forms (α_-, α_+) on M, negative and positive, respectively, is *closed* if $\alpha_- \wedge \alpha_+$ is a closed 2-form.

It is straightforward to check that a closed pair (α_-, α_+) is an AL pair (see also Lemma 2.7 below). As we will see in Proposition 3.13, closed AL pairs are in correspondence with volume preserving Anosov flows.

2.2 Elementary properties of Anosov Liouville pairs

The notion of Anosov Liouville pair can be conveniently characterized in the following way, which only involves the forms and their exterior differentials.

Lemma 2.7 *Let (α_-, α_+) be a pair of 1-forms on M. We write*

$$\alpha_+ \wedge d\alpha_+ = f_+\,\mathrm{dvol}, \quad \alpha_- \wedge d\alpha_- = -f_-\,\mathrm{dvol}, \quad d(\alpha_- \wedge \alpha_+) = f_0\,\mathrm{dvol},$$

where dvol *is any volume form on M and $f_\pm, f_0\colon M \to \mathbb{R}$ are smooth functions. Then (α_-, α_+) is an AL pair if and only if $f_\pm > 0$, and*

$$f_0^2 < 4f_- f_+. \tag{2-1}$$

Proof Following [25, Lemma 9.4], (α_-, α_+) is a Liouville pair if and only if for all constants $C_-, C_+ \geq 0$ with $(C_-, C_+) \neq (0,0)$,

$$(C_+\alpha_+ - C_-\alpha_-) \wedge (C_+ d\alpha_+ + C_- d\alpha_-) > 0,$$

which is equivalent to

$$C_+^2 f_+ + C_- C_+ f_0 + C_-^2 f_- > 0.$$

Applying this fact to (α_-, α_+) and $(-\alpha_-, \alpha_+)$, we obtain that (α_-, α_+) is an AL pair if and only if $f_\pm > 0$ and for every $x \in \mathbb{R}$,

$$x^2 f_+ + x f_0 + f_- > 0,$$

which is equivalent to (2-1) by the quadratic formula. □

Remark 2.8 The proof also shows that (α_-, α_+) is a Liouville pair if and only if

$$f_\pm > 0 \quad \text{and} \quad -f_0 < 2\sqrt{f_- f_+}.$$

We now use this criterion to show some natural geometric properties of Anosov Liouville pairs.

Proposition 2.9 *Let (α_-, α_+) be an Anosov Liouville pair. Then it defines a bicontact structure $(\xi_-, \xi_+) = (\ker\alpha_-, \ker\alpha_+)$. Moreover, if $X \in \xi_- \cap \xi_+$ is a nowhere vanishing vector field and $R_\pm$ is the Reeb vector field of $\alpha_\pm$, then $\{X, R_-, R_+\}$ is a basis at every point of M.*

Proof We first show that ξ_- and ξ_+ intersect transversally everywhere. Assume by contradiction that there exist a point $x \in M$ and two linearly independent vectors $X, Y \in T_x M$ such that $\alpha_\pm(X) = \alpha_\pm(Y) = 0$. In what follows, all the quantities will be implicitly evaluated at this point x. We can assume without loss of generality that $d\alpha_+(X, Y) > 0$ and $\mathrm{dvol}(X, Y, R_+) = 1$. We compute

$$\begin{aligned}
\alpha_+ \wedge d\alpha_+(X, Y, R_+) &= d\alpha_+(X, Y) = f_+,\\
\alpha_- \wedge d\alpha_-(X, Y, R_+) &= \alpha_-(R_+)\, d\alpha_-(X, Y) = -f_-,\\
\alpha_- \wedge d\alpha_+(X, Y, R_+) &= \alpha_-(R_+)\, d\alpha_+(X, Y),\\
\alpha_+ \wedge d\alpha_-(X, Y, R_+) &= d\alpha_-(X, Y),
\end{aligned}$$

hence

$$\begin{aligned}
f_0^2 - 4f_- f_+ &= (d\alpha_-(X, Y))^2 - 2\,\alpha_-(R_+)\, d\alpha_-(X, Y)\, d\alpha_+(X, Y)\\
&\qquad\qquad + \alpha_-(R_+)^2\, (d\alpha_+(X, Y))^2 + 4\,\alpha_-(R_+)\, d\alpha_-(X, Y)\, d\alpha_+(X, Y)\\
&= \big(d\alpha_-(X, Y) + \alpha_-(R_+)\, d\alpha_+(X, Y)\big)^2\\
&\geq 0,
\end{aligned}$$

contradicting (2-1).

For the second part, we write

$$\alpha_- \wedge d\alpha_+ = g_+\, \mathrm{dvol}, \quad \alpha_+ \wedge d\alpha_- = g_-\, \mathrm{dvol},$$

where $g_\pm \colon M \to \mathbb{R}$ are smooth functions (note that $f_0 = g_- - g_+$) and we compute

$$\begin{aligned}
\alpha_+ \wedge d\alpha_+(X, R_+, \cdot\,) &= -d\alpha_+(X, \cdot\,) = f_+\, \mathrm{dvol}(X, R_+, \cdot\,),\\
\alpha_- \wedge d\alpha_-(X, R_+, \cdot\,) &= -\alpha_-(R_+)\, d\alpha_-(X, \cdot\,) + d\alpha_-(X, R_+)\alpha_- = -f_-\, \mathrm{dvol}(X, R_+, \cdot\,),\\
\alpha_- \wedge d\alpha_+(X, R_+, \cdot\,) &= -\alpha_-(R_+)\, d\alpha_+(X, \cdot\,) = g_+\, \mathrm{dvol}(X, R_+, \cdot\,),\\
\alpha_+ \wedge d\alpha_-(X, R_+, \cdot\,) &= -d\alpha_-(X, \cdot\,) + d\alpha_-(X, R_+)\alpha_+ = g_-\, \mathrm{dvol}(X, R_+, \cdot\,).
\end{aligned}$$

Let us assume that $\mathrm{dvol}(X, R_-, R_+) = 0$ at a point $x \in M$. In what follows, all the quantities will be implicitly evaluated at this point x. Plugging in R_- in the first two of the four equations above yields

$$d\alpha_-(X, R_+) = d\alpha_+(X, R_-) = 0.$$

Note that X and R_+ are not colinear since $\alpha_+(X) = 0$ and $\alpha_+(R_+) = 1$. The last two of the four equations above imply $\alpha_-(R_+) \neq 0$ and

$$f_+ = \frac{1}{\alpha_-(R_+)} g_+, \quad f_- = -\alpha_-(R_+) g_-.$$

Finally,

$$f_0^2 - 4f_- f_+ = (g_- - g_+)^2 + 4g_- g_+ = (g_- + g_+)^2 \geq 0,$$

contradicting (2-1). □

Remark 2.10 A (non-Anosov) Liouville pair may not define a bicontact structure, namely $\xi_- = \ker \alpha_-$ and $\xi_+ = \ker \alpha_+$ may not be transverse everywhere. Nevertheless, the first part of the proof can easily be adapted to show that at a point where ξ_- and ξ_+ coincide, and their orientations coincide (and their coorientations are opposite). In the terminology of [9], (ξ_-, ξ_+) is a *positive pair* of contact structures. After a generic perturbation of α_- and/or α_+, the singular set $\Delta := \{x \in M : \xi_-(x) = \xi_+(x)\}$ is a smoothly embedded link in M. Moreover, it can be shown that $f_0 > 0$ along Δ, so the Liouville condition of Remark 2.8 is largely satisfied. We refer to our article [24] for detailed proofs of these facts and a thorough investigation of general Liouville pairs.

For any AL pair (α_-, α_+), if X (or dvol) is chosen so that $\mathrm{dvol}(X, R_-, R_+) = 1$, then

$$\begin{aligned} f_+ &= d\alpha_+(X, R_-) = \mathcal{L}_X \alpha_+(R_-), \quad g_+ = \alpha_-(R_+) f_+, \\ f_- &= d\alpha_-(X, R_+) = \mathcal{L}_X \alpha_-(R_+), \quad g_- = -\alpha_+(R_-) f_-. \end{aligned}$$

Moreover, if (α_-, α_+) is balanced, ie if $f_+ = f_-$, the condition (2-1) becomes

$$|\alpha_-(R_+) + \alpha_+(R_-)| < 2. \tag{2-2}$$

In fact, (balanced) AL pairs can be completely characterized by their Reeb vector fields.

Proposition 2.11 *Let (α_-, α_+) be a pair of contact forms on M, negative and positive, respectively, and assume that it is balanced. Then it is an AL pair if and only if (2-2) is satisfied.*

Proof We only have to show that under these hypothesis, the conclusions of Proposition 2.9 are satisfied, since these imply that $g_+ = \alpha_-(R_+) f_+$ and $g_- = -\alpha_+(R_-) f_-$ and Lemma 2.7 concludes the proof.

Assume first that ξ_- and ξ_+ are not transverse at a point $x \in M$. With the same notation as in the proof of Proposition 2.9, similar computations show that at this point,

$$\begin{aligned} \alpha_+ \wedge d\alpha_+(X, Y, R_+) &= d\alpha_+(X, Y) = f_+, \\ \alpha_- \wedge d\alpha_-(X, Y, R_+) &= \alpha_-(R_+)\, d\alpha_-(X, Y) = -f_-, \\ \alpha_+ \wedge d\alpha_+(X, Y, R_-) &= \alpha_+(R_-)\, d\alpha_+(X, Y) = f_+ \,\mathrm{dvol}(X, Y, R_-), \\ \alpha_- \wedge d\alpha_-(X, Y, R_-) &= d\alpha_-(X, Y) = -f_- \,\mathrm{dvol}(X, Y, R_-), \end{aligned}$$

hence by the first and third equalities,

$$\mathrm{dvol}(X, Y, R_-) = \alpha_+(R_-),$$

and by the second and fourth equalities,

$$\alpha_-(R_+)\alpha_+(R_-) = 1,$$

contradicting (2-2) by the inequality of arithmetic and geometric means.

Assuming now that $\mathrm{dvol}(X, R_-, R_+) = 0$ at a point $x \in M$, the proof of Proposition 2.9 showed that at this point,

$$d\alpha_-(X, R_+) = d\alpha_+(X, R_-) = 0,$$

and

$$g_+ = \alpha_-(R_+) f_+.$$

Similarly,

$$\alpha_+ \wedge d\alpha_+(X, R_-, \cdot\,) = -\alpha_+(R_-)\, d\alpha_+(X, \cdot\,) = f_+ \,\mathrm{dvol}(X, R_-, \cdot\,),$$

$$\alpha_- \wedge d\alpha_+(X, R_-, \cdot\,) = -d\alpha_+(X, \cdot\,) = g_+ \,\mathrm{dvol}(X, R_-, \cdot\,),$$

hence

$$f_+ = \alpha_+(R_-) g_+.$$

Once again, we obtain that

$$\alpha_-(R_+)\alpha_+(R_-) = 1,$$

contradicting (2-2). □

3 From Anosov flows to Anosov Liouville pairs and back

In this section, we adapt the proof of [21, Theorem 1.1] to the setting of Anosov Liouville pairs as defined in the introduction.

3.1 Anosov and projectively Anosov flows

We recall the definitions of Anosov and projectively Anosov flows with an emphasis on our orientation conventions, and recast them in terms of the existence of suitable 1-forms.

Definition 3.1 Let $\Phi = \{\phi^t\}_{t\in\mathbb{R}}$ be a flow on M generated by a nonsingular $\mathcal{C}^1$ vector field X.

- Φ is *Anosov* if there exists a continuous invariant *hyperbolic splitting*

$$TM = \langle X\rangle \oplus E^s \oplus E^u, \tag{3-1}$$

where E^s, E^u are 1-dimensional bundles such that for some (any) Riemannian metric g on M, there exist constants $C, a > 0$ such that for all $v \in E^s$ and $t \geq 0$,

$$\|d\phi^t(v)\| \leq Ce^{-at}\|v\|,$$

and for all $v \in E^u$ and $t \geq 0$,

$$\|d\phi^t(v)\| \geq Ce^{at}\|v\|.$$

E^s and E^u are called the (strong) stable and unstable bundles of Φ, respectively.

- Φ is *projectively Anosov* if there exists a continuous invariant splitting

$$TM/\langle X\rangle = N_X = \bar{E}^s \oplus \bar{E}^u, \tag{3-2}$$

where $\bar{E}^s$, $\bar{E}^u$ are 1-dimensional bundles such that for some (any) Riemannian metric $\bar{g}$ on N_X, there exist constants $C, a > 0$ such that for all unit vectors $v_s \in \bar{E}^s$, $v_u \in \bar{E}^u$, and $t \geq 0$,

$$\|d\phi^t(v_u)\| \geq Ce^{at}\,\|d\phi^t(v_s)\|.$$

Such a splitting is called a *dominated splitting*. We denote by $E^{ws} := \pi^{-1}(\bar{E}^s)$ and $E^{wu} := \pi^{-1}(\bar{E}^u)$ the weak-stable and weak-unstable bundles of Φ, respectively.

- In both cases, if the constant C can be chosen to be 1, the corresponding metrics g and $\bar{g}$ are called *adapted* to Φ.

- The Anosov (resp. projectively Anosov) flow Φ is *oriented* if E^s and E^u are oriented (resp. $\bar{E}^s$ and $\bar{E}^u$ are oriented) and their orientations are compatible with the splitting (3-1) (resp. the splitting (3-2)). See Figure 2.

Anosov flows are projectively Anosov, with dominated splitting $N_X = \pi(E^s) \oplus \pi(E^u)$. Every three dimensional (projectively) Anosov flow admits a smooth adapted metric; see [14, Proposition 5.1.5].[1] Anosov famously showed that the weak and strong stable/unstable bundles of an Anosov flow are *uniquely integrable*. Moreover, E^{ws} and E^{wu} integrate into *taut foliations* $\mathcal{F}^{ws}$ and $\mathcal{F}^{wu}$, respectively. This is not true for projectively Anosov flows; see [13, Example 2.2.9].

In the rest of the article, we implicitly assume that all of the Anosov and projectively Anosov flows under consideration are oriented.

The following definition appears in [21, Definition 3.11]; see also [21, Proposition 3.12].

Definition 3.2 Let Φ be a projectively Anosov flow on M generated by a vector field X and $\bar{g}$ be a Riemannian metric on N_X. The *expansion rates* in the stable and unstable directions for $\bar{g}$ are continuous functions $r_s, r_u \colon M \to \mathbb{R}$ defined by

$$r_s := \frac{\partial}{\partial t}\Big|_{t=0} \ln\|d\phi^t(\bar{e}_s)\|, \quad r_u := \frac{\partial}{\partial t}\Big|_{t=0} \ln\|d\phi^t(\bar{e}_u)\|,$$

where $\bar{e}_s$ and $\bar{e}_u$ are unit sections of $\bar{E}^s$ and $\bar{E}^u$, respectively, which are continuous and continuously differentiable along the flow Φ.[2] Moreover,

$$\mathcal{L}_X\bar{e}_s = -r_s\bar{e}_s, \quad \mathcal{L}_X\bar{e}_u = -r_u\bar{e}_u.$$

[1] The proof is given for Anosov flows but it easily generalizes to projectively Anosov flows in dimension three; note that it is not sufficient to integrate an arbitrary metric along the flow for a large time!

[2] In particular, the function $t \mapsto d\phi^t(\bar{e}_{s,u})$ is differentiable and has positive norm, so $r_{s,u}$ is well-defined.

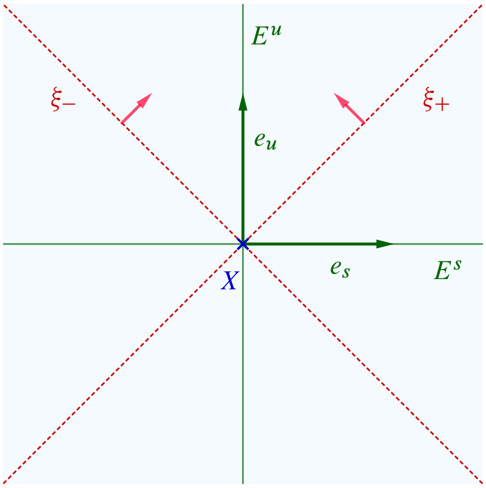


Figure 2: Orientation convention for the Anosov splitting. There is a similar picture for the dominated splitting. The vector field X points toward the reader; watch out.

The Lie derivative above means the following: if $e_{s,u}$ is a vector field on M which is a lift of $\bar{e}_{s,u}$ with the same regularity, the quantity $\mathcal{L}_X \bar{e}_{s,u} := \pi(\mathcal{L}_X e_{s,u})$ is a section of $\bar{E}^{s,u}$ which is independent of the choice of the lift. Here, $\mathcal{L}_X e_{s,u}$ denotes the usual Lie derivative along X, defined by

$$\mathcal{L}_X e_{s,u} := \frac{\partial}{\partial t}\Big|_{t=0} (\phi^t)^* e_{s,u}.$$

Notice that this involves the *pullback* along ϕ^t, and thus the differential of the *inverse* of the flow, which explains the presence of negative signs in the previous formulas.

Many natural quantities (eg functions, vector fields, 1-forms, Riemannian metrics) defined for (projectively) Anosov flows are continuous and can be upgraded to quantities which are continuous *and* continuously differentiable along the flow by considering the averaging

$$\frac{1}{T}\int_0^T (\phi^t)^* \_\_ \, dt$$

for some $T > 0$. Nevertheless, these quantities may not be $\mathcal{C}^1$. It is therefore natural to consider the following spaces (only the cases $k = 0, 1$ and $n = 0, 1$ will be relevant for us).

Definition 3.3 Let X be a smooth, nonsingular vector field on M, and $k \geq 0$ be a nonnegative integer.

- An n-form α on M is of class $\mathcal{C}^k_X$ if α is differentiable along X, and both α and $\mathcal{L}_X \alpha$ are of class $\mathcal{C}^k$. We denote by $\Omega^n_{X,k}$ the space of n-forms on M of class $\mathcal{C}^k_X$ satisfying $\iota_X \alpha = 0$ (which is vacuous for $n = 0$, ie for functions). We also denote by $\Omega^n_X = \Omega^n_{X,\infty} \subset \Omega^n$ the space of smooth n-forms satisfying $\iota_X \alpha = 0$.

- On $\Omega^n_{X,k}$, there is a natural norm defined by
$$|\alpha|_{\mathcal{C}^k_X} := |\alpha|_{\mathcal{C}^k} + |\mathcal{L}_X\alpha|_{\mathcal{C}^k},$$
making $(\Omega^n_{X,k}, |\cdot|_{\mathcal{C}^k_X})$ a Banach space. Ω^n_X is naturally a Fréchet space as a closed subspace of Ω^n.

These definitions naturally extend to sections of N_X and n-forms on N_X.

In Appendix A, we show some density results for these spaces which are particularly useful when dealing with (projectively) Anosov flows generated by $\mathcal{C}^1$ vector fields and can be used to bypass Hozoori's delicate approximation techniques in [21, Section 4]. The results in Appendix A are not needed (except in the proof of Theorem 4.4) if we restrict our attention to smooth Anosov flows in view of Lemma 3.5 below.

In dimension three, the definitions of Anosov and projectively Anosov flows can be rephrased in terms of the existence of certain 1-forms. The following lemma is essentially an adaptation of results of Mitsumatsu [28] and Hozoori [21; 20].

Lemma 3.4 *Let Φ be a smooth, nonsingular flow on M generated by a vector field X. Then*

(1) *Φ is oriented projectively Anosov if and only if there exist $(\alpha_s, \alpha_u) \in \Omega^1_{X,0} \times \Omega^1_{X,0}$ and continuous functions $r_u, r_s \colon M \to \mathbb{R}$ such that*
$$\bar{\alpha}_s \wedge \bar{\alpha}_u > 0, \quad \mathcal{L}_X\alpha_s = r_s\alpha_s, \quad \mathcal{L}_X\alpha_u = r_u\alpha_u,$$
and $r_s < r_u$. Here, $\bar{\alpha}_s$ and $\bar{\alpha}_u$ denote the 1-forms on N_X induced by α_s and α_u, respectively.

(2) *Φ is oriented Anosov if and only if there exist α_s, α_u, r_s, r_u as above such that $r_s < 0 < r_u$.*

(3) *Φ is oriented volume preserving Anosov if and only if there exist α_s, α_u, r_s, r_u as above such that $r_u + r_s = 0$.*

Moreover, $\ker\alpha_u = E^{ws}$ and $\ker\alpha_s = E^{wu}$.

Proof (1) This essentially follows from [21, Proposition 3.15]. We recall the main arguments. If Φ is projectively Anosov, we can choose an adapted metric and unit vector fields $\bar{e}_s$ and $\bar{e}_u \in N_X$ of class $\mathcal{C}^0_X$ such that $\bar{e}_s$ spans $\bar{E}^s$, $\bar{e}_u$ spans $\bar{E}^u$ and $(\bar{e}_s, \bar{e}_u)$ is positively oriented. The inequality $r_s < r_u$ follows from the definition of a dominated splitting and the fact that the metric is adapted. If $(\bar{\alpha}_s, \bar{\alpha}_u)$ denotes the dual basis of $(\bar{e}_s, \bar{e}_u)$, it induces a pair $(\alpha_s, \alpha_u) \in \Omega^1_{X,0} \times \Omega^1_{X,0}$, and the relations $\mathcal{L}_X\bar{e}_s = -r_s\bar{e}_s$ and $\mathcal{L}_X\bar{e}_u = -r_u\bar{e}_u$ imply the desired relations for α_s and α_u. Reciprocally, if (α_s, α_u) is such a pair, we define $(\bar{e}_s, \bar{e}_u)$ as the dual basis of $(\bar{\alpha}_s, \bar{\alpha}_u)$ and we easily check that it yields a projectively Anosov splitting of N_X. This is essentially because for $\star \in \{s, u\}$, $\phi^T_*\bar{e}_\star = \exp\bigl(\int_0^T r_\star \circ \phi^t\, dt\bigr)\bar{e}_\star$ and $r_s < r_u$, where $\Phi = \{\phi^t\}$. In particular, we have $\ker\alpha_s = E^{wu}$ and $\ker\alpha_u = E^{ws}$ since these bundles are uniquely determined by the flow.

(2) This follows from (1) and [21, Proposition 3.17].

(3) The forward direction follows from the proof of [28, Theorem 3]. Indeed, assuming that Φ is volume preserving Anosov, we can arrange that $r_u + r_s = 0$ in the following way. If dvol is a (smooth) volume

form preserved by Φ, then $\tau := \iota_X \mathrm{dvol}$ is a nondegenerate 2-form on N_X invariant under Φ. There exists an adapted metric g of class $\mathcal{C}^0_X$ for which the Anosov splitting is orthogonal and the volume form for the induced metric $\bar{g}$ on N_X is precisely τ.[3] Hence, if e_s and e_u are $\mathcal{C}^0_X$ unit vector fields spanning E^s and E^u, respectively, then $\tau(\bar{e}_s, \bar{e}_u) = 1$. Differentiating this equality along X yields $r_u + r_s = 0$ as desired, and we obtain (α_s, α_u) by dualization as before. For the reverse direction, $r_s < 0 < r_u$, since $r_s < r_u$ and $r_s = -r_u$, so Φ is Anosov by (2). Moreover, if θ is a smooth 1-form satisfying $\theta(X) \equiv 1$, then $\mathrm{dvol} := \alpha_s \wedge \alpha_u \wedge \theta$ is a $\mathcal{C}^0_X$ volume form preserved by X and by [22, Corollary 2.1], this volume form is smooth. □

It is well-known that in dimension three, the regularity of the weak-stable and weak-unstable bundles of a smooth (even $\mathcal{C}^2$) Anosov flow are $\mathcal{C}^1$. We have:

Lemma 3.5 *If Φ is Anosov and smooth, we can further assume that α_s, α_u, r_s and r_u as in Lemma 3.4 are $\mathcal{C}^1$, ie α_s and α_u are $\mathcal{C}^1_X$.*

Proof By [19, Corollary 1.8], E^{ws} and E^{wu} are $\mathcal{C}^1$ and an adapted metric can always be assumed to be smooth, so the construction in Lemma 3.4 yields $\mathcal{C}^1$ 1-forms α_s and α_u. The $\mathcal{C}^1$ regularity of r_s and r_u follows from a trick of Simić [30]. First, let us choose a $\mathcal{C}^1$ 1-form α_u such that $\ker \alpha_u = E^{ws}$ and $\mathcal{L}_X \alpha_u = r_u \, \alpha_u$, where r_u is continuous and positive. Fix a smooth vector field Z positively transverse to E^{ws}, so that $f := \alpha_u(Z) > 0$. Here, f is $\mathcal{C}^1$ and can be approximated by a smooth function $\tilde{f} > 0$ so that $h := \tilde{f}/f$ is $\mathcal{C}^1$-close to 1. Setting $\tilde{\alpha}_u := h\alpha_u$, $\tilde{\alpha}_u$ is $\mathcal{C}^1$ and satisfies $\mathcal{L}_X \tilde{\alpha}_u = \tilde{r}_u \, \tilde{\alpha}_u$, where $\tilde{r}_u$ is $\mathcal{C}^0$-close to r_u and can be assumed to be positive. We now show that $\tilde{r}_u$ is $\mathcal{C}^1$. Indeed, $\tilde{\alpha}_u(Z) = \tilde{f}$ and

$$\tilde{r}_u \tilde{f} = (\mathcal{L}_X \tilde{\alpha}_u)(Z) = \mathcal{L}_X(\tilde{\alpha}_u(Z)) - \tilde{\alpha}_u(\mathcal{L}_X Z) = X \cdot \tilde{f} - \tilde{\alpha}_u(\mathcal{L}_X Z),$$

and the last quantity is $\mathcal{C}^1$ since $X \cdot \tilde{f}$ and $\mathcal{L}_X Z$ are smooth and $\tilde{\alpha}_u$ is $\mathcal{C}^1$. The same argument applies to α_s. □

Remark 3.6 The proof actually shows more. Since $\tilde{r}_u = u + \tilde{\alpha}_u(V)$ for some smooth function u and some smooth vector field V, and $\mathcal{L}_X \tilde{\alpha}_u = \tilde{r}_u \, \tilde{\alpha}_u$, an immediate induction argument shows that for every integer $n \geq 0$, $\mathcal{L}^n_X \tilde{\alpha}_u$ and $\mathcal{L}^n_X \tilde{r}_u$ exist and are $\mathcal{C}^1$, where $\mathcal{L}^n_X := \mathcal{L}_X \circ \cdots \circ \mathcal{L}_X$ denotes the Lie derivative along X iterated n times. In fact, it is well known that in our setting, the individual leaves of the weak-(un)stable foliation are smooth (see [22, Lemma 2.1]).

Remark 3.7 The same argument works for smooth projectively Anosov flow whose weak-stable and weak-unstable distributions are $\mathcal{C}^1$. However, there are known examples of smooth projectively Anosov flows in dimension three whose weak distributions are not $\mathcal{C}^1$; see [13, Example 2.2.9].

[3] The induced metric $\bar{g}$ is the pushforward of the restriction of g to $E^s \oplus E^u$ along the projection $E^s \oplus E^u \to N_X$, which is an isomorphism. Concretely, if $\bar{v}_1, \bar{v}_2$ are vectors in N_X with lifts $v_1, v_2 \in E^s \oplus E^u$, then $\bar{g}(\bar{v}_1, \bar{v}_2) = g(v_1, v_2)$. Since X is orthogonal to $E^s \oplus E^u$ for g, the latter quantity does not depend on the choice of such lifts.

We call a pair of 1-forms (α_s, α_u) as in Lemma 3.4 (1) (resp. (2), (3)) a *defining pair* for the projectively Anosov (resp. Anosov, volume preserving Anosov) flow Φ. We further require defining pairs for (volume preserving) Anosov flows to be $\mathcal{C}^1$. We also impose the following conditions on orientations:

- The orientation on E^{ws}, induced by the orientation of X and the orientation on E^s or $\bar{E}^s$ which implicitly comes with Φ, agrees with the one induced by α_u,
- The orientation on E^{wu}, induced by the orientation of X and the orientation on E^u or $\bar{E}^u$ which implicitly comes with Φ, agrees with the one induced by $-\alpha_s$.[4]

Concretely, these properties mean that if $\star \in \{s, u\}$ and $\bar{e}_\star \in \bar{E}^\star$ forms an oriented basis, then $\alpha_\star(\bar{e}_\star) > 0$.

We denote by $\mathcal{D}_\Phi$ the space of defining pairs for Φ endowed with the $\mathcal{C}^0_X$ topology in the projectively Anosov case, and with the $\mathcal{C}^1$ topology in the (volume preserving) Anosov case.

Lemma 3.8 *The space $\mathcal{D}_\Phi$ of defining pairs for a projectively Anosov* (*resp. Anosov*, *volume preserving Anosov*) *flow Φ with its corresponding topology is contractible.*

Proof Let us fix a defining pair $(\alpha_s, \alpha_u) \in \mathcal{D}_\Phi$ for a projectively Anosov flow Φ generated by a vector field X. If $(\alpha'_s, \alpha'_u) \in \mathcal{D}_\Phi$ is any other defining pair, then $\ker \alpha_u = \ker \alpha'_u$ and $\ker \alpha_s = \ker \alpha'_s$, and the orientations on these spaces agree. Hence, there exist (unique) functions $\rho_s, \rho_u \colon M \to \mathbb{R}$ of class $\mathcal{C}^0_X$ such that $\alpha'_u = e^{\rho_u} \alpha_u$ and $\alpha'_s = e^{\rho_s} \alpha_s$, and they satisfy

$$r'_u - r'_s = X \cdot (\rho_u - \rho_s) + r_u - r_s > 0. \tag{3-3}$$

Here, r'_u and r'_s are such that $\mathcal{L}_X \alpha'_s = r'_s \alpha'_s$ and $\mathcal{L}_X \alpha'_u = r'_u \alpha'_u$.

Reciprocally, if $\rho_s, \rho_u \colon M \to \mathbb{R}$ are functions as above satisfying (3-3), then $(\alpha'_s, \alpha'_u) := (e^{\rho_s} \alpha_s, e^{\rho_u} \alpha_u)$ is also a defining pair for Φ.

It follows that $\mathcal{D}_\Phi$ is homeomorphic to

$$\mathcal{R} := \{(\rho_s, \rho_u) \in \mathcal{C}^0_X \times \mathcal{C}^0_X : X \cdot (\rho_u - \rho_s) + r_u - r_s > 0\},$$

and $\mathcal{R}$ is obviously convex, hence contractible. The proof for Anosov flows and volume preserving Anosov flows is similar. The condition on ρ_s and ρ_u becomes

$$X \cdot \rho_u + r_u > 0 \quad \text{and} \quad X \cdot \rho_s + r_s < 0 \tag{3-4}$$

if Φ is Anosov, and

$$X \cdot (\rho_u - \rho_s) + r_u - r_s > 0 \quad \text{and} \quad X \cdot (\rho_u + \rho_s) = 0 \tag{3-5}$$

if Φ is volume preserving Anosov. Both conditions (3-4) and (3-5) are convex in (ρ_s, ρ_u). □

[4] The somewhat strange minus sign is explained by the following remark. In the Euclidean plane $\mathbb{R}^2$ with its standard oriented basis (e_1, e_2), the dual basis (e_1^*, e_2^*) induces coorientations on the x and y-axis corresponding to the natural orientation on the x-axis and to the *opposite* of the natural orientation on the y-axis.

Remark 3.9 If Φ is a (projectively, volume preserving) Anosov flow generated by a vector field X, $f : M \to \mathbb{R}_{>0}$ is a positive function and Φ' is the (projectively, volume preserving) Anosov flow generated by fX, then $\mathcal{D}_\Phi = \mathcal{D}_{\Phi'}$. Indeed, if $(\alpha_s, \alpha_u) \in \mathcal{D}_\Phi$ then for $\star \in \{s, u\}$,

$$\mathcal{L}_{fX}\alpha_\star = f r_\star \alpha_\star.$$

Since the stable/unstable bundles of Φ' are the same as the ones of Φ, (α_s, α_u) is a defining pair for Φ' with expansion rates $r'_{s,u} = f r_{s,u}$.

Therefore, there is a well-defined notion of defining pairs for (projectively, volume preserving) oriented Anosov *line distributions*.

3.2 From Anosov flows to Anosov Liouville pairs

Throughout this section, we assume that Φ is a smooth Anosov flow on M and we construct an AL pair supporting Φ, proving the first part of Theorem 1. We choose a $\mathcal{C}^1$ defining pair $(\alpha_s, \alpha_u) \in \mathcal{D}_\Phi$ as in Lemma 3.4(2). Following [21, Section 4], we define

$$\alpha_- := \alpha_u + \alpha_s, \quad \alpha_+ := \alpha_u - \alpha_s. \tag{3-6}$$

Note that $\alpha_\pm$ is of class $\mathcal{C}^1_X$, and the orientation compatibility conditions of Definition 2.1 are satisfied. Let dvol be the $\mathcal{C}^1$ volume form on M defined by $\alpha_s \wedge \alpha_u = \iota_X \mathrm{dvol}$. We will make use of the elementary identities:

$$\alpha_s \wedge d\alpha_u = -r_u \,\mathrm{dvol}, \quad \alpha_u \wedge d\alpha_s = r_s \,\mathrm{dvol}, \quad \alpha_s \wedge d\alpha_s = 0, \quad \alpha_u \wedge d\alpha_u = 0.$$

The first one follows from

$$\iota_X(\alpha_s \wedge d\alpha_u) = -\alpha_s \wedge \iota_X d\alpha_u = -\alpha_s \wedge \mathcal{L}_X \alpha_u = -r_u\, \alpha_s \wedge \alpha_u = -r_u \,\mathrm{dvol},$$

and the three others can be obtained by similar computations. We easily deduce

$$\alpha_+ \wedge d\alpha_+ = (r_u - r_s)\,\mathrm{dvol}, \quad \alpha_- \wedge d\alpha_- = -(r_u - r_s)\,\mathrm{dvol}, \quad d(\alpha_- \wedge \alpha_+) = 2(r_u + r_s)\,\mathrm{dvol}.$$

Since $r_s < 0 < r_u$, α_- and α_+ are contact forms and the criterion of Lemma 2.7 is satisfied.[5] Therefore, (α_-, α_+) is a $\mathcal{C}^1$ AL pair supporting Φ.

Definition 3.10 A *standard AL pair* supporting Φ is a $\mathcal{C}^1$ AL pair obtained by the previous construction. We denote by $\mathcal{AL}^{\mathrm{std}}_\Phi$ the space of these AL pairs, endowed with the $\mathcal{C}^1$ topology.

There is an obvious (linear) homeomorphism between $\mathcal{D}_\Phi$ and $\mathcal{AL}^{\mathrm{std}}_\Phi$ induced by the map

$$(s, u) \mapsto (u + s, u - s).$$

Since $\mathcal{D}_\Phi$ is contractible by Lemma 3.8, $\mathcal{AL}^{\mathrm{std}}_\Phi$ is contractible as well.

[5] Note that $\lambda := e^{-s}\alpha_- + e^s\alpha_+$ is a Liouville form if and only if $r_u > 0$.

A standard AL pair (α_-, α_+) is not necessarily smooth by definition; it is smooth exactly when the weak-stable and weak-unstable foliations of Φ are smooth, which is a quite restrictive situation.[6] Nevertheless, any pair of smooth 1-forms (α'_-, α'_+) sufficiently $\mathcal{C}^1$-close to (α_-, α_+) and satisfying $\alpha'_\pm(X) = 0$ is a smooth AL pair supporting Φ. This shows the forward implication in Theorem 1.

Let $R_\pm$ denote the Reeb vector fields of $\alpha_\pm$, defined by

$$\alpha_\pm(R_\pm) = 1, \quad d\alpha_\pm(R_\pm, \cdot) = 0.$$

Rewriting these equations in terms of α_s and α_u, and using the equalities

$$\iota_X d\alpha_s = \mathcal{L}_X \alpha_s = r_s\,\alpha_s, \quad \iota_X d\alpha_u = \mathcal{L}_X \alpha_u = r_u\,\alpha_u,$$

one easily computes

$$\begin{aligned} \alpha_s(R_-) &= \frac{r_u}{r_u - r_s} > 0, & \alpha_u(R_-) &= \frac{-r_s}{r_u - r_s} > 0, \\ \alpha_s(R_+) &= \frac{-r_u}{r_u - r_s} < 0, & \alpha_u(R_+) &= \frac{-r_s}{r_u - r_s} > 0. \end{aligned}$$

Therefore,

- R_- is positively transverse to E^{ws} and E^{wu},
- R_+ is positively transverse to E^{ws} and negatively transverse to E^{wu},

and this remains true for a smoothing of (α_-, α_+) as above.[7] Since $\mathcal{F}^{ws}$ is a *taut* foliation, we obtain that $R_\pm$ has no contractible closed Reeb orbit, thus $\xi_\pm = \ker \alpha_\pm$ is *hypertight*. This was already observed by Hozoori [21, Theorem 1.11].

3.3 From Anosov Liouville pairs to Anosov flows

We now turn to the second part of the proof of Theorem 1. Let us assume that Φ is a smooth nonsingular flow on M generated by a vector field X and suppose that it is supported by an AL pair (α_-, α_+). By Proposition 2.9, (α_-, α_+) defines a bicontact structure (ξ_-, ξ_+) supporting X, so Φ is projectively Anosov and there exists a dominated splitting $N_X \cong \overline{E}^s \oplus \overline{E}^u$ as in Definition 3.1. We shall construct a defining pair (α_s, α_u) as in Lemma 3.4(2), implying that Φ is Anosov.

The proof of [13, Proposition 2.2.6] (see also [21, Remark 3.10]) shows that $\xi_\pm$ is everywhere transverse to E^{ws} and E^{wu}. By our orientation conventions, there exist two continuous functions $\sigma_s, \sigma_u \colon M \to \mathbb{R}$ such that

$$\ker\{e^{-\sigma_u}\alpha_- + e^{\sigma_u}\alpha_+\} = E^{ws}, \quad \ker\{e^{-\sigma_s}\alpha_- - e^{\sigma_s}\alpha_+\} = E^{wu}.$$

[6] By [17, Theorem 4.7], the smoothness of the weak-(un)stable implies that Φ is topologically equivalent to an *algebraic Anosov flow*, ie the suspension of an Anosov diffeomorphism of the 2-torus, or the geodesic flow on a closed hyperbolic surface, up to finite cover.

[7] These transversality properties for the Reeb vector fields are a key feature of Anosov flows and are not satisfied for projectively Anosov flows which are not Anosov, see [21, Theorem 6.3].

Note that σ_u and σ_s are also continuously differentiable along X.[8] Indeed, if $\bar{e}_s$ is any vector field of class $\mathcal{C}^0_X$ spanning $\bar{E}^s \subset N_X$, then

$$e^{\sigma_u}\alpha_+(\bar{e}_s) + e^{-\sigma_u}\alpha_-(\bar{e}_s) = 0,$$

hence

$$\sigma_u = \tfrac{1}{2}\ln\left(-\frac{\alpha_-(\bar{e}_s)}{\alpha_+(\bar{e}_s)}\right),$$

and this quantity is continuously differentiable along X; the same argument applies to σ_s.

We define[9]

$$\alpha_u := \frac{1}{2\sqrt{\cosh(\sigma_u-\sigma_s)}}(e^{-\sigma_u}\alpha_- + e^{\sigma_u}\alpha_+), \quad \alpha_s := \frac{1}{2\sqrt{\cosh(\sigma_u-\sigma_s)}}(e^{-\sigma_s}\alpha_- - e^{\sigma_s}\alpha_+),$$

so that

$$\ker\alpha_u = E^{ws}, \quad \ker\alpha_s = E^{wu},$$

and

$$\alpha_- = \frac{1}{\sqrt{\cosh(\sigma_u-\sigma_s)}}(e^{\sigma_s}\alpha_u + e^{\sigma_u}\alpha_s), \tag{3-7}$$

$$\alpha_+ = \frac{1}{\sqrt{\cosh(\sigma_u-\sigma_s)}}(e^{-\sigma_s}\alpha_u - e^{-\sigma_u}\alpha_s). \tag{3-8}$$

Note that α_u and α_s are continuously differentiable along X, and since E^{ws} and E^{wu} are invariant under Φ, there exist continuous functions $r_s, r_u\colon M \to \mathbb{R}$ such that

$$\mathcal{L}_X\alpha_s = r_s\,\alpha_s, \quad \mathcal{L}_X\alpha_u = r_u\,\alpha_u.$$

Moreover,

$$\alpha_- \wedge \alpha_+ = 2\,\alpha_s \wedge \alpha_u,$$

so $\bar{\alpha}_s \wedge \bar{\alpha}_u > 0$. We are left to show that $r_s < 0 < r_u$, which will follow from Lemma 2.7. Let dvol be the unique volume form on M such that $\alpha_s \wedge \alpha_u = \iota_X\,\text{dvol} =: \tau$.

Lemma 3.11 *With the same notation as in Lemma 2.7, we have*

$$f_+ = \frac{e^{-(\sigma_s+\sigma_u)}}{\cosh(\sigma_u-\sigma_s)}(X\cdot(\sigma_u-\sigma_s) + r_u - r_s), \tag{3-9}$$

$$f_- = \frac{e^{(\sigma_s+\sigma_u)}}{\cosh(\sigma_u-\sigma_s)}(-X\cdot(\sigma_u-\sigma_s) + r_u - r_s), \tag{3-10}$$

$$f_0 = 2(r_u + r_s). \tag{3-11}$$

[8] We cannot assume that they are $\mathcal{C}^1$ yet, since we do not know that Φ is Anosov!

[9] The seemingly strange conformal factors will greatly simplify some computations later, in particular the inequality (4-2) in the proof of Lemma 4.7.

Proof Although $\alpha_\pm$ are smooth, the quantities α_s, α_u, σ_s and σ_u are not $\mathcal{C}^1$ so we cannot compute $d\alpha_\pm$ directly by differentiating from (3-7) and (3-8). However, these quantities are differentiable along X and the functions f_0, f_- and f_+ can be computed from

$$\alpha_+ \wedge \mathcal{L}_X\alpha_+ = -f_+\tau, \quad \alpha_- \wedge \mathcal{L}_X\alpha_- = f_-\tau, \quad \mathcal{L}_X\alpha_- \wedge \alpha_+ + \alpha_- \wedge \mathcal{L}_X\alpha_+ = f_0\tau.$$

Moreover, the quantities $\mathcal{L}_X\alpha_\pm$ *can* be computed from (3-7) and (3-8) by differentiating along X. The calculations are left to the reader. □

Since $f_\pm > 0$, (3-9) and (3-10) imply

$$0 \le |X \cdot \sigma| < r_u - r_s,$$

where $\sigma := \sigma_u - \sigma_s$, and the inequality $f_0^2 < 4f_-f_+$ gives

$$(r_u + r_s)^2 \le \cosh^2(\sigma)(r_u + r_s)^2 < (r_u - r_s)^2 - (X \cdot \sigma)^2 \le (r_u - r_s)^2,$$

yielding $r_s < 0 < r_u$ as desired. This concludes the proof of Theorem 1.

Remark 3.12 Similar computations (and Lemma A.2) show that if Φ is a nondegenerate flow on M, the following are equivalent:

(1) Φ is supported by a transverse Liouville pair (α_-, α_+).

(2) Φ is projectively Anosov and admits a defining pair (α_s, α_u) with $r_u > 0$.

Note that in case, the Reeb vector fields for the standard construction of Section 3.2 are still transverse to the weak-unstable bundle of Φ, but are not necessarily transverse to the weak-stable bundle of Φ. We wish to call Φ a *semi-Anosov flow*. Our techniques would also show that the space of semi-Anosov flows on M is homotopy equivalent to the space of transverse Liouville pairs on M.

3.4 Volume preserving Anosov flows

Volume preserving Anosov flows, ie Anosov flows preserving a volume form, constitute a remarkable class of Anosov flows. They are topologically transitive, in the sense that they admit a dense orbit. A deep theorem of Asaoka [3] implies that on closed 3-manifold, every transitive Anosov flow is topologically equivalent to a volume preserving one. In this section, we show some striking connections between volume preserving Anosov flows and Anosov Liouville pairs.

Proposition 3.13 *Let Φ be a smooth nonsingular flow on M. Then Φ is a volume preserving Anosov flow if and only if it is supported by a closed AL pair.*

Proof Let us first assume that Φ preserves a (smooth) volume form dvol, and let $\tau := \iota_X \mathrm{dvol}$. Note that τ is *closed*. Let (ξ_-, ξ_+) be any bicontact structure supporting Φ, and $\alpha_\pm$ two contact forms such that $\ker \alpha_\pm = \xi_\pm$. There exists a smooth positive function $\kappa\colon M \to \mathbb{R}_{>0}$ such that

$$\alpha_- \wedge \alpha_+ = \kappa\, \tau.$$

The positivity of κ follows from our conventions on the coorientations of bicontact structures. Then, $\bigl(\alpha_-, \frac{1}{\kappa}\alpha_+\bigr)$ is a closed pair as in Definition 2.6, and it is automatically an AL pair in view of Lemma 2.7 since the corresponding function f_0 vanishes.

Let us now assume that Φ is supported by a closed AL pair (α_-, α_+). By Theorem 1, Φ is Anosov. Let θ be any smooth 1-form on M satisfying $\theta(X) \equiv 1$, where X is the vector field generating Φ, and define $\mathrm{dvol} := \alpha_- \wedge \alpha_+ \wedge \theta$. It is easy to check that it is a volume form, and

$$\mathcal{L}_X \mathrm{dvol} = \mathcal{L}_X(\alpha_- \wedge \alpha_+) \wedge \theta + \alpha_- \wedge \alpha_+ \wedge \mathcal{L}_X \theta = \alpha_- \wedge \alpha_+ \wedge d\theta(X, \cdot) = 0;$$

hence Φ preserves a smooth volume form. □

Remark 3.14 A special class of closed AL pairs is given by Geiges pairs, defined in [25, Section 8.5] as pairs of contact forms (α_-, α_+) satisfying

$$\alpha_+ \wedge d\alpha_+ = -\alpha_- \wedge d\alpha_- > 0, \quad \alpha_+ \wedge d\alpha_- = \alpha_- \wedge d\alpha_+ = 0.$$

Geiges pairs are called (-1)-Cartan structures in [20], and they are shown to be in correspondence with volume preserving Anosov flows. Here, we note that (α_-, α_+) is a $\mathcal{C}^1$ Geiges pair if and only if $(\alpha_- - \alpha_+, \alpha_- + \alpha_+)$ is a defining pair for the underlying volume preserving Anosov flow. As a result, the space of Geiges pairs supporting a given flow is contractible. Not every (smooth) volume preserving Anosov flow is supported by a smooth (or even $\mathcal{C}^2$) Geiges pair, as it would imply that the weak-stable and weak-unstable bundles are $\mathcal{C}^2$, so the flow would be smoothly equivalent to an algebraic Anosov flow; see [16, théorème A].

The previous proof shows more: for a volume preserving Anosov flow, *any* supporting bicontact structure can be realized as the kernel of an AL pair. Surprisingly, this is a characteristic feature of volume preserving Anosov flows.

Theorem 3.15 *Let Φ be a smooth Anosov flow on M. Then Φ preserves a volume form if and only if for every (smooth) supporting bicontact structure (ξ_-, ξ_+), there exists an AL pair (α_-, α_+) such that $\xi_\pm = \ker \alpha_\pm$.*

Proof The forward direction follows from the first part of the proof of Proposition 3.13. Let us assume that every (smooth) bicontact structure (ξ_-, ξ_+) supporting Φ is defined by a (smooth) AL pair, and let us fix a defining pair (α_s, α_u) for Φ with associated expansion rates r_s and r_u as in Lemma 3.4. Let $A > 0$

be a positive real number and $\{\alpha_u^n\}_{n\in\mathbb{N}}$ and $\{\alpha_s^n\}_{n\in\mathbb{N}}$ be sequences of smooth 1-forms converging to α_u and α_s, respectively, in the $\mathcal{C}^1$ topology. For every $n\in\mathbb{N}$, we define

$$\alpha_+^n := \alpha_u^n - e^{-A}\alpha_s^n, \quad \alpha_-^n := \alpha_u^n + e^{A}\alpha_s^n.$$

We also let

$$\alpha_+ := \alpha_u - e^{-A}\alpha_s, \quad \alpha_- := \alpha_u + e^{A}\alpha_s.$$

Then, for n sufficiently large (depending on A), (α_-^n, α_+^n) defines a (smooth) bicontact structure (ξ_-^n, ξ_+^n) supporting Φ. By assumption, there exists a smooth positive function $f_n\colon M\to\mathbb{R}_{>0}$ such that $(\alpha_-^n,\, f_n\alpha_+^n)$ is an AL pair defining (ξ_-^n, ξ_+^n). Lemma 2.7 will imply the following:

Claim *For every $\epsilon>0$, there exists a smooth function $h_\epsilon\colon M\to\mathbb{R}$ such that*

$$|X\cdot h_\epsilon + r_u + r_s| \le \epsilon. \tag{3-12}$$

Assuming Claim for now, it follows that if θ is a smooth 1-form such that $\theta(X)\equiv 1$, then for every closed orbit γ of X,

$$\int_\gamma (r_u + r_s)\,\theta = 0.$$

If the flow is transitive, a classical theorem of Livšic implies that there exists a continuous function $h\colon M\to\mathbb{R}$ which is differentiable along X and satisfies

$$X\cdot h + r_u + r_s = 0.$$

Writing $\mathrm{dvol}' := e^h\,\mathrm{dvol} = e^h\,\alpha_s\wedge\alpha_u\wedge\theta$, $\mathcal{L}_X\mathrm{dvol}' = 0$ so Φ preserves a positive continuous measure, and by [22, Corollary 2.1], this measure is a smooth volume form.

It turns out that the condition (3-12) *implies* that the flow is transitive. We have not been able to find a proof of this fact in the literature. We refer to Appendix B for a proof using the theory of Sinai–Ruelle–Bowen measures.

We now prove Claim. Let $\epsilon>0$ and choose $A>0$ such that

$$\frac{\sup_M(r_u - r_s)}{\cosh(A)} \le \epsilon/3.$$

Let e_s and e_u be $\mathcal{C}^1$ vector fields satisfying

$$\alpha_s(e_s) = 1, \quad \alpha_s(e_u) = 0,$$
$$\alpha_u(e_s) = 0, \quad \alpha_u(e_u) = 1,$$

so that $\alpha_s\wedge\alpha_u(e_s,e_u) = \mathrm{dvol}(X,e_s,e_u) = 1$. Since $(\alpha_-^n,\, f_n\alpha_+^n)$ is an AL pair for n large enough, Lemma 2.7 implies the inequality

$$\left| X\cdot \ln f_n + \frac{d(\alpha_-^n\wedge\alpha_+^n)(X,e_s,e_u)}{\alpha_-^n\wedge\alpha_+^n(e_s,e_u)} \right| < \frac{2\sqrt{-(\alpha_-^n\wedge d\alpha_-^n(X,e_s,e_u))\cdot(\alpha_+^n\wedge d\alpha_+^n(X,e_s,e_u))}}{|\alpha_-^n\wedge\alpha_+^n(e_s,e_u)|}.$$

One computes

$$\lim_{n\to\infty} \frac{d(\alpha_-^n \wedge \alpha_+^n)(X, e_s, e_u)}{\alpha_-^n \wedge \alpha_+^n(e_s, e_u)} = r_u + r_s,$$

$$\lim_{n\to\infty} \frac{2\sqrt{-(\alpha_-^n \wedge d\alpha_-^n(X, e_s, e_u)) \cdot (\alpha_+^n \wedge d\alpha_+^n(X, e_s, e_u))}}{|\alpha_-^n \wedge \alpha_+^n(e_s, e_u)|} = \frac{r_u - r_s}{\cosh(A)},$$

by first replacing $\alpha_\pm^n$ by $\alpha_\pm$ and then writing these expressions in terms of α_s and α_u. We obtain that for n large enough such that the above two sequences are $\epsilon/3$-close to their limits,

$$|X \cdot \ln f_n + r_u + r_s| \leq \frac{r_u - r_s}{\cosh(A)} + \epsilon/3 + \epsilon/3 \leq \epsilon,$$

hence $h_\epsilon := \ln f_n$ satisfies the required inequality and Claim is proved. □

Remark 3.16 The proof can be adapted to show that if every bicontact structure supporting Φ is realized as the kernel of a Liouville pair, then the determinant of the Poincaré return map for every closed orbit of Φ is bigger than or equal to 1. We expect that this should also imply that Φ is volume preserving.

4 Spaces of Anosov Liouville pairs and bicontact structures

This section is dedicated to the proof of Theorem 3 from the introduction. We first describe the main strategy in a more general setting. Let E and B be topological spaces and $f\colon E \to B$ be a continuous map. We can assume that E and B have the homotopy type of CW complexes. This is the case for the spaces we consider (eg $\mathcal{AL}$, $\mathcal{BC}$, $\mathcal{AF}$, $\mathbb{P}\mathcal{AF}$, etc.) as they are open subsets of Fréchet spaces.[10] To show that f is a homotopy equivalence, it is enough to show that it is a Serre fibration with (weakly) contractible fibers. However, it seems rather hard to show that the maps we care about (eg $\mathcal{I}$, $\mathbb{P}\mathcal{I}$) satisfy a homotopy lifting property, as this would require a careful understanding of how the stable and unstable bundles depend on the (projectively) Anosov flow. Instead, we choose a more indirect approach: we first show that these maps have contractible fibers, and we then show that they are *topological submersions*.

Definition 4.1 $f\colon E \to B$ is a *topological submersion* if it is surjective, and for every $x \in E$, there exists a neighborhood U of x in E such that if we write $y := f(x)$ and $V := U \cap f^{-1}(y)$, there exists a homeomorphism $U \xrightarrow{\sim} f(U) \times V$ making the following diagram commute:

$$\begin{tikzcd}
U \arrow[rr, "\sim"] \arrow[dr, "f"'] & & f(U) \times V \arrow[dl, "\mathrm{pr}_1"] \\
& f(U) &
\end{tikzcd}$$

Here, pr_1 denotes the projection onto the first factor.

[10] Indeed, Fréchet spaces are absolute neighborhood retracts (ANRs) by a theorem of Dugundji; an open subset of an ANR is an ANR; every ANR has the homotopy type of a CW complex by a theorem of Milnor and Whitehead. However, it is known that an open subset of an infinite dimensional Fréchet space is *not* a CW complex.

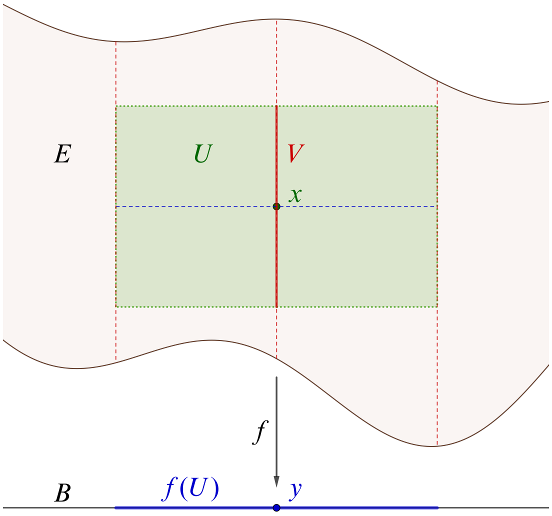


Figure 3: A topological submersion.

Fiber bundles are topological submersions, but the converse is not true since the product structure of topological submersions is only *local on the domain* and is not "uniform in the fibers"; see Figure 3. However, we have:

Lemma 4.2 *If $f : E \to B$ is a topological submersion with* (*weakly*) *contractible fibers*, *then f an acyclic Serre fibration.*

Proof Since projections are open and openness is a local property, f is open. By [26, Lemma 6], f is a *homotopic submersion* (see [26, Definition 1]), also known as a *Serre microfibration* [18]. The result then follows from [26, Corollary 13] (see also [31, Lemma 2.2]). □

4.1 Contractibility of fibers

In this section, we show:

Theorem 4.3 *Let Φ be a smooth Anosov flow on M. The spaces of AL pairs and weak AL pairs supporting Φ are both contractible.*

We also show a similar result for projectively Anosov flows:

Theorem 4.4 *Let Φ be a smooth projectively Anosov flow on M. The space of bicontact structures supporting Φ is contractible.*

We obtain a version for volume preserving Anosov flows as well:

Theorem 4.5 *Let Φ be a smooth volume preserving Anosov flow on M. The space of closed AL pairs supporting Φ is contractible.*

If Φ is a smooth Anosov flow on M,

- $\mathcal{AL}_\Phi$ denotes the space of smooth AL pairs on M supporting Φ, endowed with the $\mathcal{C}^\infty$ topology,
- $\mathcal{AL}^1_\Phi$ denotes the space of $\mathcal{C}^1$ AL pairs on M supporting Φ, endowed with the $\mathcal{C}^1$ topology.

Smooth AL pairs supporting Φ form a dense subset of $\mathcal{AL}^1_\Phi$. Recall that $\mathcal{D}_\Phi$ denotes the space of $\mathcal{C}^1$ defining pairs on M for Φ, and $\mathcal{AL}^{\mathrm{std}}_\Phi \subset \mathcal{AL}^1_\Phi$ denotes the space of $\mathcal{C}^1$ standard AL pairs supporting Φ, both endowed with the $\mathcal{C}^1$ topology. Theorem 4.3 will be a consequence of the contractibility of $\mathcal{AL}^{\mathrm{std}}_\Phi$ and the following two lemmas.

Lemma 4.6 *The natural map $\mathcal{AL}_\Phi \to \mathcal{AL}^1_\Phi$ is a homotopy equivalence.*

Proof This follows from some standard facts in algebraic topology. We will use that homotopy equivalences are *local* in the following sense:

Fact *A continuous map $f : X \to Y$ between topological spaces is a homotopy equivalence if there exists a numerable open cover $\mathcal{U}$ of Y satisfying*:

(1) *$\mathcal{U}$ is stable under finite intersections.*

(2) *For every $U \in \mathcal{U}$, $f : f^{-1}(U) \to U$ is a homotopy equivalence.*

See [11, Theorem 1] for a proof of this fact. Recall that an open cover is *numerable* if it admits a subordinate partition of unity. In our context, covers are automatically numerable since all the spaces under consideration are metrizable. We can simply cover $\mathcal{AL}^1_\Phi$ by sufficiently small open $\mathcal{C}^1$ balls and refine this cover by taking all possible finite intersections. These balls are convex as subsets of the space of pairs of $\mathcal{C}^1$ 1-forms on M, and so are finite intersections thereof, so all of the open subsets in our cover are contractible. Since smooth AL pairs supporting Φ are dense in $\mathcal{AL}^1_\Phi$, every open $\mathcal{C}^1$ ball in $\mathcal{AL}^1_\Phi$ intersects $\mathcal{AL}_\Phi$. The intersection of such a ball with $\mathcal{AL}_\Phi$ is also convex as a subset of $\Omega^1(M) \times \Omega^1(M)$, and so are finite intersections of such balls with $\mathcal{AL}_\Phi$. Thus, the conditions (1) and (2) of Fact are trivially satisfied. □

Lemma 4.7 *$\mathcal{AL}^{\mathrm{std}}_\Phi$ is a strong deformation retract of $\mathcal{AL}^1_\Phi$.*

Proof Let $(\alpha_-, \alpha_+) \in \mathcal{AL}^1_\Phi$. As in Section 3.3, there exist functions $\sigma_s, \sigma_u : M \to \mathbb{R}$ satisfying

$$\ker\{e^{-\sigma_u}\alpha_- + e^{\sigma_u}\alpha_+\} = E^{ws}, \quad \ker\{e^{-\sigma_s}\alpha_- - e^{\sigma_s}\alpha_+\} = E^{wu}. \tag{4-1}$$

If e_s and e_u are $\mathcal{C}^1$ vector fields such that $\bar{e}_s$ spans $\bar{E}^s$ and $\bar{e}_u$ spans $\bar{E}^u$, we can write

$$\sigma_s = \tfrac{1}{2}\ln\left(\frac{\alpha_+(e_u)}{\alpha_-(e_u)}\right), \quad \sigma_u = \tfrac{1}{2}\ln\left(-\frac{\alpha_-(e_s)}{\alpha_+(e_s)}\right),$$

so σ_s and σ_u are $\mathcal{C}^1$. Moreover, the map $\mathcal{S}\colon (\alpha_-,\alpha_+) \mapsto (\sigma_s,\sigma_u)$ is continuous in the $\mathcal{C}^1$ topology. As before, we define

$$\alpha_u := \frac{1}{2\sqrt{\cosh(\sigma_u-\sigma_s)}}(e^{-\sigma_u}\alpha_- + e^{\sigma_u}\alpha_+), \quad \alpha_s := \frac{1}{2\sqrt{\cosh(\sigma_u-\sigma_s)}}(e^{-\sigma_s}\alpha_- - e^{\sigma_s}\alpha_+).$$

The computations of Section 3.3 show that $(\alpha_s,\alpha_u) \in \mathcal{D}_\Phi$. Therefore, we obtain a continuous map $\mathcal{D}\colon (\alpha_-,\alpha_+) \mapsto (\alpha_s,\alpha_u)$. We now define a strong deformation retraction $r\colon \mathcal{AL}^1_\Phi \times [0,1] \to \mathcal{AL}^1_\Phi$. Let $(\alpha_-,\alpha_+) \in \mathcal{AL}^1_\Phi$, and $(\sigma_s,\sigma_u) = \mathcal{S}(\alpha_-,\alpha_+)$ and $(\alpha_s,\alpha_u) = \mathcal{D}(\alpha_-,\alpha_+)$ as before. For $t \in [0,1]$, we define

$$\begin{aligned}\alpha_-^t &:= \frac{1}{\sqrt{\cosh((1-t)\sigma)}}(e^{(1-t)\sigma_s}\alpha_u + e^{(1-t)\sigma_u}\alpha_s),\\ \alpha_+^t &:= \frac{1}{\sqrt{\cosh((1-t)\sigma)}}(e^{-(1-t)\sigma_s}\alpha_u - e^{-(1-t)\sigma_u}\alpha_s),\end{aligned}$$

where $\sigma = \sigma_u - \sigma_s$. We then have

- $(\alpha_-^0,\alpha_+^0) = (\alpha_-,\alpha_+)$,
- $(\alpha_-^1,\alpha_+^1) = (\alpha_u+\alpha_s,\alpha_u-\alpha_s) \in \mathcal{AL}^{\mathrm{std}}_\Phi$,
- if $(\alpha_-,\alpha_+) \in \mathcal{AL}^{\mathrm{std}}_\Phi$, then $(\alpha_-^t,\alpha_+^t) = (\alpha_-,\alpha_+)$ for every $t \in [0,1]$.

We claim that (α_-^t,α_+^t) is an AL pair for every $t \in [0,1]$. Indeed, by Lemmas 2.7 and 3.11, it is enough to show the inequality

$$\cosh^2((1-t)\sigma)(r_u+r_s)^2 + (1-t)^2(X\cdot\sigma)^2 < (r_u-r_s)^2. \tag{4-2}$$

It holds for $t = 0$ since (α_-,α_+) is an AL pair, and the left-hand side is obviously a nonincreasing function of t, so it holds for every $t \in [0,1]$.

We finally define $r((\alpha_-,\alpha_+),t) := (\alpha_-^t,\alpha_+^t)$ so that $r\colon \mathcal{AL}^1_\Phi \times [0,1] \to \mathcal{AL}^1_\Phi$ is continuous, and by the three bullets above, it is a strong deformation retraction of $\mathcal{AL}^1_\Phi$ onto $\mathcal{AL}^{\mathrm{std}}_\Phi$. □

Proof of Theorem 4.3 For AL pairs, combine Lemmas 3.8, 4.6 and 4.7. For weak AL pairs, the argument can be modified as follows. Lemma 4.6 can be easily adapted to the case of weak AL pairs. Lemma 4.7 can be adapted to show that the space of $\mathcal{C}^1$ weak AL pairs supporting a smooth Anosov flow Φ deformation retracts onto the space of pairs of the form $\alpha_\pm = \alpha_u \mp \alpha_s$, where (α_s,α_u) satisfies the conditions of a defining pair for Φ *without the condition* $r_s < 0$ (but still satisfies $r_s < r_u$ and $0 < r_u$; see Remark 3.12). The latter space is convex, hence contractible. □

Proof of Theorem 4.4 We only sketch how to modify the proof of Theorem 4.3 and we leave the details to the interested reader.

First of all, we shall introduce the space of $\mathcal{C}^0_X$ bicontact structures. Those are continuous pairs of codimension 1 distributions (ξ_-,ξ_+) which are continuously differentiable along X and which are defined by some 1-forms $\alpha_-, \alpha_+ \in \Omega^0_X$ satisfying

$$\bar{\alpha}_- \wedge \bar{\alpha}_+ > 0, \quad \bar{\alpha}_- \wedge \mathcal{L}_X \bar{\alpha}_- < 0, \quad \bar{\alpha}_+ \wedge \mathcal{L}_X \bar{\alpha}_+ > 0, \tag{4-3}$$

as forms on N_X.

For the purpose of the proof, we choose an arbitrary $\mathcal{C}^0_X$ vector field e_u such that $\bar{e}_u$ spans $\bar{E}^u \subset N_X$ and defines the prescribed orientation. This is equivalent to choosing a 1-form $\alpha_u \in \Omega^1_{X,0}$ such that $\ker \alpha_u = E^{ws}$ as oriented 2-plane fields, with the normalization $\alpha_u(e_u) \equiv 1$.

We denote by $\mathcal{BC}_\Phi$ (resp. $\mathcal{BC}^0_\Phi$) the space of smooth (resp. $\mathcal{C}^0_X$) bicontact structures supporting Φ. We write $\mathcal{BC}^{\text{std}}_\Phi \subset \mathcal{BC}^0_\Phi$ for the space of *standard* bicontact structures supporting Φ, of the form

$$\big(\ker(\alpha_u + \alpha_s),\, \ker(\alpha_u - \alpha_s)\big),$$

where α_u is fixed by the condition $\alpha_u(e_u) \equiv 1$.

Lemma 4.6 can be easily adapted to show that the natural map $\mathcal{BC}_\Phi \to \mathcal{BC}^0_\Phi$ is a homotopy equivalence, using Lemma A.2.

Lemma 4.7 can be modified as follows. For $(\xi_-, \xi_+) \in \mathcal{BC}^0_\Phi$, we denote by (α_-, α_+) the unique pair of 1-forms in Ω^0_X satisfying $\ker \alpha_\pm = \xi_\pm$ and $\alpha_\pm(e_u) = 1$. We define a $\mathcal{C}^0_X$ function $\sigma \colon M \to \mathbb{R}$ by

$$\ker\{e^{-\sigma}\alpha_- + e^{\sigma}\alpha_+\} = E^{ws},$$

so that

$$\alpha_u = \frac{1}{2\cosh(\sigma)}(e^{-\sigma}\alpha_- + e^{\sigma}\alpha_+),$$

and we define

$$\alpha_s := \frac{1}{2\cosh(\sigma)}(\alpha_- - \alpha_+)$$

so that $\alpha_s \in \Omega^1_{X,0}$, and

$$\ker \alpha_s = E^{wu}.$$

This readily implies that

$$\alpha_- = \alpha_u + e^{\sigma}\alpha_s, \quad \alpha_+ = \alpha_u - e^{-\sigma}\alpha_s.$$

Writing $\mathcal{L}_X \alpha_u = r_u \alpha_u$ and $\mathcal{L}_X \alpha_s = r_s \alpha_s$, where $r_s, r_u \colon M \to \mathbb{R}$ are continuous, the last two inequalities in (4-3) are equivalent to

$$|X \cdot \sigma| < r_u - r_s.$$

Moreover,

$$\alpha_- \wedge \alpha_+ = 2\cosh(\sigma)\alpha_s \wedge \alpha_u.$$

This shows that (α_s, α_u) is a defining pair for Φ that satisfies $\alpha_u(e_u) \equiv 1$. For $t \in [0, 1]$, we define

$$\alpha^t_- := \alpha_u + e^{(1-t)\sigma}\alpha_s, \qquad \xi^t_- := \ker \alpha^t_-,$$
$$\alpha^t_+ := \alpha_u - e^{-(1-t)\sigma}\alpha_s, \quad \xi^t_+ := \ker \alpha^t_+.$$

These formulas define a strong deformation retraction of $\mathcal{BC}^0_\Phi$ onto $\mathcal{BC}^{\text{std}}_\Phi$. Moreover, $\mathcal{BC}^{\text{std}}_\Phi$ is homeomorphic to the space of defining pairs (α_s, α_u) for Φ satisfying $\alpha_u(e_u) \equiv 1$, and one easily checks that this space is contractible. □

Proof of Theorem 4.5 The result essentially follows from Theorem 4.4. Let dvol be a smooth volume form preserved by Φ and $\tau := \iota_X \mathrm{dvol}$. If (α_-, α_+) is a closed AL pair supporting Φ, there exists a smooth positive function $\kappa := M \to \mathbb{R}_{>0}$ such that

$$\alpha_- \wedge \alpha_+ = \kappa\,\tau.$$

Moreover, $X \cdot \kappa = 0$ and since Φ is topologically transitive, κ is constant. One easily checks that the space of closed AL pairs supporting Φ is homotopy equivalent to the space of balanced pairs of contact forms (α_-, α_+) supporting Φ and satisfying $\alpha_- \wedge \alpha_+ = \tau$. We denote this space by $\mathcal{BC}_\Phi^\tau$. There is a natural continuous map $\mathcal{K}\colon \mathcal{BC}_\Phi^\tau \to \mathcal{BC}_\Phi$, obtained by taking kernels, which is surjective by Theorem 3.15. One easily checks that $\mathcal{K}$ is injective and that it is a homeomorphism. Theorem 4.4 finishes the proof. □

4.2 Homotopy equivalences

Let us recall some notation.

- $\mathcal{AL}$ denotes the space of smooth AL pairs on M.
- $\mathcal{AF}$ denotes the space of smooth Anosov flows on M, up to positive time reparametrization.
- $\mathbb{P}\mathcal{AF}$ denotes the space of smooth projectively Anosov flows on M, up to positive time reparametrization.

Recall that there is a continuous map,

$$\mathcal{I}\colon \mathcal{AL} \to \mathcal{AF}, \quad (\alpha_-, \alpha_+) \mapsto \ker \alpha_- \cap \ker \alpha_+,$$

where we identify an oriented 1-dimensional distribution with any flow spanned by it. Similarly, there is a continuous map

$$\mathbb{P}\mathcal{I}\colon \mathcal{BC} \to \mathbb{P}\mathcal{AF}, \quad (\xi_-, \xi_+) \mapsto \xi_- \cap \xi_+.$$

In this section, we show the main theorems of this article:

Theorem 4.8 *The map $\mathcal{I}$ is an acyclic Serre fibration.*

Our argument can easily be adapted to the case of projectively Anosov flows (this result might already be known to some experts):

Theorem 4.9 *The map $\mathbb{P}\mathcal{I}$ is an acyclic Serre fibration.*

Remark 4.10 With more work, it is possible to show that $\mathcal{I}$ is *shrinkable*: it is homotopy equivalent over $\mathcal{AF}$ to $\mathrm{id}\colon \mathcal{AF} \to \mathcal{AF}$. Concretely, this means that there exists a section s of $\mathcal{I}$ such that $s \circ \mathcal{I}$ is *fiberwise homotopic* to id. This implies that the space sections of $\mathcal{I}$ is nonempty and contractible. To prove this statement, one would need to upgrade the results of Section 4.1 to hold *in family over $\mathcal{AF}$*. A key ingredient is [22, Lemma 2.1], which shows that for smooth Anosov flows, the Anosov splitting depends continuously on the flow. We are not aware of a similar result for projectively Anosov flows.

We will need the following

Lemma 4.11 *$\mathcal{I}$ is a topological submersion.*

Proof By Theorem 1, $\mathcal{I}$ is surjective. We fix some auxiliary Riemannian metric g on M and identify $\mathcal{AF}$ with the space of unit Anosov vector fields on M. Let $\boldsymbol{\alpha}^0 = (\alpha_-^0, \alpha_+^0) \in \mathcal{AL}$ and let X be the unit vector field generating $\mathcal{I}(\boldsymbol{\alpha}^0)$, whose flow is denoted by Φ. We choose an arbitrary smooth 1-form θ such that $\theta(X) \equiv 1$. For a unit vector field X' sufficiently close to X (so that $\theta(X') > 0$) and $\boldsymbol{\alpha} = (\alpha_-, \alpha_+) \in \mathcal{AL}_\Phi$, we define

$$\alpha'_\pm := \alpha_\pm - \frac{\alpha_\pm(X')}{\theta(X')}\,\theta,$$

so that $\alpha'_\pm(X') = 0$. Since $\mathcal{AL} \subset \Omega^1 \times \Omega^1$ is open, we can find an open neighborhood $\mathcal{N}_{\boldsymbol{\alpha}^0}$ of $\boldsymbol{\alpha}^0$ in $\mathcal{AL}_\Phi$ and an open neighborhood $\mathcal{N}_\Phi$ of Φ in $\mathcal{AF}$ such that the map

$$\psi\colon \mathcal{N}_\Phi \times \mathcal{N}_{\boldsymbol{\alpha}^0} \to \mathcal{AL}, \quad (X', \boldsymbol{\alpha}) \mapsto \boldsymbol{\alpha}',$$

is well-defined. It is continuous and satisfies $\mathcal{I} \circ \psi = \mathrm{pr}_1$. Moreover, the restriction of ψ to $\{X\} \times \mathcal{N}_{\boldsymbol{\alpha}^0}$ is the inclusion $\mathcal{N}_{\boldsymbol{\alpha}^0} \subset \mathcal{AL}$. One easily checks that ψ is injective, has open image and has an inverse given by $\psi^{-1}(\boldsymbol{\alpha}') := (X', \boldsymbol{\alpha})$, where

$$\alpha_\pm := \alpha'_\pm - \alpha'_\pm(X)\theta,$$

and X' is the unit vector field spanning $\mathcal{I}(\boldsymbol{\alpha}')$. Therefore, ψ^{-1} is a local trivialization of $\mathcal{I}$ around $\boldsymbol{\alpha}^0$. □

Proof of Theorem 4.8 By Theorem 4.3 and Lemma 4.11, $\mathcal{I}$ is a topological submersion with contractible fibers, hence an acyclic Serre fibration by Lemma 4.2. □

Proof of Theorem 10 Lemma 4.11 and its proof hold verbatim for $\mathcal{I}^w$ so the previous proof applies to $\mathcal{I}^w$ as well. □

Proof of Theorem 4.9 By Theorem 4.4, we already know that the fibers of $\mathbb{P}\mathcal{I}$ are contractible so it is enough to adapt Lemma 4.11. It can be done by choosing an auxiliary smooth vector field Z, depending on an initial choice of $(\xi_-^0, \xi_+^0) \in \mathcal{BC}$, which is positively transverse to $\xi_\pm^0$ and satisfies $\theta(Z) \equiv 0$. We can uniquely choose contact forms $\alpha_\pm^0$ for $\xi_\pm^0$ by imposing $\alpha_\pm(Z) \equiv 1$. The proof of Lemma 4.11 can be reproduced with minor changes to provide a suitable trivialization near (ξ_-^0, ξ_+^0). □

4.3 The kernel map

Recall that we have a continuous map

$$\underline{\ker}\colon \mathcal{AL} \to \mathcal{BC}, \quad (\alpha_-, \alpha_+) \mapsto (\ker\alpha_-, \ker\alpha_+),$$

where the spaces $\mathcal{AL}$ and $\mathcal{BC}$ are endowed with the C^∞ topology.

Lemma 4.12 *The map $\underline{\ker}$ is open.*

Proof It easily follows from the openness of $\mathcal{AL}$ and $\mathcal{BC}$ in the space of smooth 1-forms on M and the space of smooth plane fields on M, respectively, and the following elementary fact. If $\mathring{\Omega}^1 \subset \Omega^1$ denotes the space of nowhere vanishing 1-forms on M and Π denotes the space of smooth cooriented plane fields on M, the natural map

$$\ker\colon \mathring{\Omega}^1 \to \Pi, \quad \alpha \mapsto \ker\alpha,$$

is open (for the $\mathcal{C}^\infty$ topology on the domain and target). Indeed, after trivializing the tangent bundle of M and fixing an auxiliary Riemannian metric, we can identify Π with the space of smooth maps $M \to S^3$ (via the unit normal vector) and $\mathring{\Omega}^1$ with the space of smooth maps $M \to \mathbb{R}^3 \setminus \{0\}$, so that ker becomes the composition with the standard projection $\mathbb{R}^3 \setminus \{0\} \cong \mathbb{R} \times S^3 \to S^3$. Ultimately, ker boils down to the projection $\mathcal{C}^\infty(M,\mathbb{R}) \times \mathcal{C}^\infty(M,S^3) \to \mathcal{C}^\infty(M,S^3)$ onto the second factor, which is clearly open. □

Theorem 4.13 *The map* $\underline{\ker}$ *is an acyclic Serre fibration onto its image.*

Proof As before, by Lemma 4.2, it is enough to show the following properties.

(1) The fibers of $\underline{\ker}$ over its image are contractible.

(2) $\underline{\ker}$ is a topological submersion onto its image.

We can simplify the situation by restricting to the space $\mathcal{AL}^b$ of *balanced* AL pairs, since there is a homeomorphism

$$\vartheta\colon \mathcal{C}^\infty(M,\mathbb{R}) \times \mathcal{AL}^b \xrightarrow{\sim} \mathcal{AL}, \quad (\sigma,(\alpha_-,\alpha_+)) \mapsto (e^{-\sigma}\alpha_-, e^{\sigma}\alpha_+),$$

and $\underline{\ker}$ is compatible with this homeomorphism in the obvious way.

Let us consider a bicontact structure (ξ_-,ξ_+) defined by a balanced AL pair (α_-,α_+), and let $\mathrm{dvol} := \alpha_+ \wedge d\alpha_+$. We also consider a vector field $X \in \xi_- \cap \xi_+$ normalized so that $\alpha_- \wedge \alpha_+ = \iota_X \mathrm{dvol}$.

To show (1), note that any other balanced AL pair defining (ξ_-,ξ_+) is of the form $(e^{\sigma}\alpha_-, e^{\sigma}\alpha_+)$ for a smooth function $\sigma\colon M \to \mathbb{R}$ satisfying

$$|2X\cdot\sigma + f_0| < 2,$$

where we use the notation of Lemma 2.7. By assumption, $|f_0| < 2$, so the space of σ such that $(e^{\sigma}\alpha_-, e^{\sigma}\alpha_+)$ is a balanced AL pair defining (ξ_-,ξ_+) is convex, hence contractible.

To show (2), we consider an open neighborhood $\mathcal{V}$ of (α_-,α_+) in $\mathcal{AL}^b$. We can find a smaller neighborhood $\mathcal{V}' \subset \mathcal{V}$ such that for every $(\alpha'_-,\alpha'_+) \in \mathcal{V}'$, the pair

$$(\widetilde{\alpha}'_-, \widetilde{\alpha}'_+) := \left(\frac{1}{\sqrt{f'}}\alpha'_-, \frac{1}{\sqrt{f'}}\alpha'_+\right)$$

is in $\mathcal{V}$, where

$$\alpha'_\pm \wedge d\alpha'_\pm = \pm f'\,\mathrm{dvol}.$$

Note that $(\widetilde{\alpha}'_-, \widetilde{\alpha}'_+)$ is a balanced AL pair satisfying $\widetilde{\alpha}'_+ \wedge d\widetilde{\alpha}'_+ = \mathrm{dvol}$. Since $\underline{\ker}$ is open by the previous lemma, $\mathcal{U}' := \underline{\ker}(\mathcal{V}') \subset \mathcal{BC}$ is an open neighborhood of (ξ_-, ξ_+). Let $\widetilde{\mathcal{V}}' \subset \mathcal{V}$ be the subspace of elements of the form $(\widetilde{\alpha}'_-, \widetilde{\alpha}'_+)$ for $(\alpha'_-, \alpha'_+) \in \mathcal{V}'$. One easily checks that $\underline{\ker}\colon \widetilde{\mathcal{V}}' \to \mathcal{U}'$ is injective and open. It is surjective by definition, hence it is a homeomorphism. By the previous paragraph, there is an open neighborhood of (α_-, α_+) in $\underline{\ker}^{-1}\{(\xi_-, \xi_+)\} \cap \mathcal{AL}^b$ homeomorphic to

$$\Sigma_\epsilon := \{\sigma\colon M \to \mathbb{R} : |X \cdot \sigma| < \epsilon\}$$

for some small $\epsilon > 0$. Therefore, after possibly shrinking ϵ, the map

$$\widetilde{\mathcal{V}}' \times \Sigma_\epsilon \to \mathcal{AL}^b, \quad ((\widetilde{\alpha}'_-, \widetilde{\alpha}'_+), \sigma) \mapsto (e^\sigma \widetilde{\alpha}'_-, e^\sigma \widetilde{\alpha}'_+),$$

induces a local trivialization of $\mathcal{AL}^b \to \underline{\ker}(\mathcal{AL})$ around (α_-, α_+).

This proves that $\underline{\ker}$ restricted to $\mathcal{AL}^b$ is a topological submersion with contractible fibers, and the same holds for $\underline{\ker}$ on $\mathcal{AL}$ via the homeomorphism ϑ. □

Remark 4.14 Since $\underline{\ker}$ is open, its image has the homotopy type of a CW complex.

There is an inclusion $\underline{\ker}(\mathcal{AL}) \subset \mathbb{P}\mathcal{I}^{-1}(\mathcal{AF})$ which is strict according to Theorem 3.15.[11] In more concrete terms, there exist bicontact structures supporting Anosov flows which cannot be represented as the kernel of an AL pair. Nevertheless, we have:

Theorem 4.15 *The inclusion* $\underline{\ker}(\mathcal{AL}) \subset \mathbb{P}\mathcal{I}^{-1}(\mathcal{AF})$ *is a homotopy equivalence.*

Proof It immediately follows from Theorems 4.8, 4.9, 4.13 and the commutative diagram

$$\begin{tikzcd}
\mathcal{AL} \arrow[r, "\sim"] \arrow[rrd, bend right, "\mathcal{I}"', "\sim"] & \underline{\ker}(\mathcal{AL}) \arrow[r, phantom, "\subset"] & \mathbb{P}\mathcal{I}^{-1}(\mathcal{AF}) \arrow[r, hook] \arrow[d, "\mathbb{P}\mathcal{I}"', "\wr"] \arrow[dr, phantom, "\lrcorner", very near start] & \mathcal{BC} \arrow[d, "\mathbb{P}\mathcal{I}"', "\wr"] \\
 & & \mathcal{AF} \arrow[r, hook] & \mathbb{P}\mathcal{AF}
\end{tikzcd}$$

Note that the corestriction of $\mathbb{P}\mathcal{I}$ over $\mathcal{AF}$ is also an acyclic Serre fibration, and all the spaces in this diagram have the homotopy type of CW complexes. □

5 Linear Liouville pairs

As explained in the introduction, there exist other possible definitions for Liouville pairs. The following one is used by some authors (eg [21; 28]).

[11] Indeed, any volume preserving Anosov flow can be perturbed near a closed orbit in such a way that the new Poincaré return map for this orbit has determinant different than 1, so the flow is not volume preserving anymore.

Definition 5.1 A pair of contact forms (α_-, α_+) on M is a *linear Liouville pair* if the 1-form

$$(1-t)\alpha_- + (1+t)\alpha_+$$

on $[-1,1]_t \times M$ is a positively oriented Liouville form.

The pair (α_-, α_+) is a *linear Anosov Liouville pair* (ℓAL pair for short) if both (α_-, α_+) and $(-\alpha_-, \alpha_+)$ are linear Liouville pairs.

Note that for this definition, $[-1,1] \times M$ is a Liouville *domain* instead of a Liouville *manifold*. In this section, we study some similarities and differences between Liouville pairs and linear Liouville pairs. In particular, we show that those are two different notions (Lemma 5.6). Moreover, a pair a contact forms which is both a Liouville pair and a linear Liouville pair defines Liouville structures in two different ways, and we show that they are homotopic (Proposition 5.8). We believe that this result is relevant since all of the natural constructions of Liouville pairs we are aware of satisfy both definitions. The linear formulation might be more convenient in some situations. The results in this section are independent from the main results of this article.

5.1 Elementary properties

The results in Section 2.2 can be adapted to ℓAL pairs. First of all, ℓAL pairs can be characterized in the following way (see Lemma 2.7):

Lemma 5.2 *Let (α_-, α_+) be a pair of contact forms on M. We write*

$$\alpha_+ \wedge d\alpha_+ = f_+ \,\mathrm{dvol}, \quad \alpha_- \wedge d\alpha_- = -f_- \,\mathrm{dvol}, \quad \alpha_- \wedge d\alpha_+ = g_+ \,\mathrm{dvol}, \quad \alpha_+ \wedge d\alpha_- = g_- \,\mathrm{dvol},$$

where dvol *is any volume form on M and $f_\pm, g_\pm \colon M \to \mathbb{R}$ are smooth functions. Then (α_-, α_+) is a* ℓAL *pair if and only if*

$$|g_-| < f_- \quad \text{and} \quad |g_+| < f_+. \tag{5-1}$$

Proof The pair (α_-, α_+) is a linear Liouville pair if and only if for every $t \in [-1,1]$,

$$(\alpha_+ - \alpha_-) \wedge \big(d\alpha_+ + d\alpha_- + t(d\alpha_+ - d\alpha_-)\big) > 0,$$

which is equivalent to

$$f_+ + f_- + g_- - g_+ + t(f_+ - f_- - g_- - g_+) > 0.$$

This inequality is satisfied for every $t \in [-1,1]$ if and only if it is satisfied for $t = -1$ and $t = 1$, which is equivalent to $g_+ < f_+$ and $-g_- < f_-$. □

Similarly to Proposition 2.9, we also have:

Proposition 5.3 *Let (α_-, α_+) be a ℓAL pair. Then it defines a bicontact structure*

$$(\xi_-, \xi_+) = (\ker \alpha_-, \ker \alpha_+).$$

Moreover, if $X \in \xi_- \cap \xi_+$ is a nowhere vanishing vector field, then (X, R_-, R_+) is a basis at every point of M.

Proof To see that ξ_- and ξ_+ are everywhere transverse, we argue by contradiction and assume that they coincide at a point $x \in M$. With the same notations as in the proof of 2.9, we readily get

$$\begin{aligned} f_+ &= d\alpha_+(X, Y), & g_+ &= \alpha_-(R_+)\, d\alpha_+(X, Y), \\ f_- &= -\alpha_-(R_+)\, d\alpha_-(X, Y), & g_- &= d\alpha_-(X, Y), \end{aligned}$$

hence

$$|g_- g_+| = |\alpha_-(R_+)\, d\alpha_-(X, Y)\, d\alpha_+(X, Y)| = f_- f_+,$$

contradicting (5-1).

Now, assuming that $\operatorname{dvol}(X, R_-, R_+) = 0$ at a point $x \in M$, the computations in the proof of Proposition 2.9 show

$$f_- = -\alpha_-(R_+) g_-, \quad f_+ = \frac{1}{\alpha_-(R_+)} g_+,$$

hence

$$|g_- g_+| = f_- f_+,$$

contradicting (5-1) once again. □

Remark 5.4 A main difference between Anosov Liouville pairs as in Definition 2 and linear Anosov Liouville pairs as in Definition 5.1 is that there does not seem to be a natural action of $\mathcal{C}^\infty(M, \mathbb{R})$ on ℓAL pairs. Moreover, we do not know if there is a natural modification making a ℓAL pair balanced.

The ℓAL pairs can also be characterized by their Reeb vector fields (see Proposition 2.11):

Proposition 5.5 *Let (α_-, α_+) be a pair of contact forms on M, negative and positive, respectively. Then it is a ℓAL pair if and only if*

$$|\alpha_-(R_+)| < 1 \quad \textit{and} \quad |\alpha_+(R_-)| < 1. \tag{5-2}$$

Proof If (α_-, α_+) is a ℓAL pair, then Proposition 5.3 and the computations in the proof of Proposition 2.9 imply

$$g_+ = \alpha_-(R_+)\, f_+, \quad g_- = -\alpha_+(R_-)\, f_-,$$

and (5-2) follows from Lemma 5.2.

Now, assuming that (α_-, α_+) satisfies (5-2), it is enough to prove that the conclusions of Proposition 5.3 are satisfied. This follows exactly from the proof of Proposition 2.11. □

Combining Proposition 2.11 and Proposition 5.5, we obtain that any *balanced* ℓAL pair is an AL pair. The converse is *not* true by the following lemma. It also implies that the two definitions of Liouville pairs (Definitions 2 and 5.1) are *different*:

Lemma 5.6 *Every smooth volume preserving Anosov flow on M admits a supporting balanced AL pair which is **not** a ℓAL pair, and whose underlying bicontact structure is **not** defined by a ℓAL pair.*

Proof Let (α_s, α_u) be a defining pair for a volume preserving Anosov flow $\Phi = \{\phi^t\}$. For $A \geq 1$, we define

$$\alpha_- := e^{-A}\alpha_u + e^{A}\alpha_s, \quad \alpha_+ := e^{A}\alpha_u - e^{-A}\alpha_s.$$

If dvol is such that $\iota_X \mathrm{dvol} = \alpha_s \wedge \alpha_u$, where X is the vector field generating the flow, then one easily computes

$$f_\pm = r_u - r_s = 2r_u, \quad g_\pm = -2\sinh(2A)r_u,$$

so (α_-, α_+) is a $\mathcal{C}^1$ closed balanced AL pair, but it is not a ℓAL pair since $|g_\pm| > f_\pm$. This remains true for a suitable smoothing of (α_-, α_+).

Let us assume for simplicity that the pair (α_-, α_+) as above is smooth. We show that there are no functions $h_\pm \colon M \to \mathbb{R}_{>0}$ such that $(h_-\alpha_-, h_+\alpha_+)$ is a ℓAL pair. Indeed, let us assume by contradiction that such functions exist. By Lemma 5.2 they would satisfy the following inequalities:

$$|\cosh(2A)h_+ X \cdot h_- - \sinh(2A)r_u h_- h_+| < r_u h_-^2,$$
$$|\cosh(2A)h_- X \cdot h_+ + \sinh(2A)r_u h_- h_+| < r_u h_+^2.$$

Writing $\rho_\pm := 1/h_\pm$, these are equivalent to

$$|X \cdot \rho_- + \tanh(2A)r_u\rho_-| < \frac{r_u}{\cosh(2A)}\rho_+, \quad |X \cdot \rho_+ - \tanh(2A)r_u\rho_+| < \frac{r_u}{\cosh(2A)}\rho_-.$$

Fixing a point $x \in M$, we define $y_\pm \colon \mathbb{R} \to \mathbb{R}_{>0}$ by

$$y_\pm(t) := \rho_\pm \circ \phi^t(x).$$

There exists $C > 0$ such that $0 < y_\pm < C$. Moreover, these functions satisfy

$$\left|\frac{d}{dt}y_- + ay_-\right| < \epsilon y_+, \quad \left|\frac{d}{dt}y_+ - ay_+\right| < \epsilon y_-,$$

where

$$a(t) := \tanh(2A)\, r_u \circ \phi^t(x) > 0, \quad \epsilon(t) := \frac{r_u \circ \phi^t(x)}{\cosh(2A)} > 0.$$

Since $A \geq 1$, we have $a > 2\epsilon$. It follows that for every $T > 0$,

$$\int_0^T ay_+\,dt \leq \int_0^T \epsilon y_-\,dt + y_+(T) + y_0(T) \leq \tfrac{1}{2}\int_0^T ay_-\,dt + 2C \leq \tfrac{1}{2}\int_0^T \epsilon y_+\,dt + 3C$$
$$\leq \tfrac{1}{4}\int_0^T ay_+\,dt + 3C,$$

and hence

$$\int_0^T ay_+ \, dt \leq 4C.$$

However, ay_+ is bounded from below by some positive constant, which contradicts the previous inequality for T large enough.

If (α_-, α_+) is only $\mathcal{C}^1$, this strategy still applies to a suitable smoothing of (α_-, α_+) which is sufficiently $\mathcal{C}^1$-close to (α_-, α_+). □

Remark 5.7 We also expect that there exist (unbalanced) ℓAL pairs which are not AL pairs, but the construction seems more delicate.

5.2 Induced Liouville structures

The "standard construction" of Section 3.2 yields a pair of contact forms which is both an AL pair and a ℓAL pair (after smoothing). If (α_-, α_+) is a pair of contact forms which is both a Liouville pair and a linear Liouville pair, we can consider two Liouville structures on $\mathbb{R}_s \times M$:

(1) The *completion* $\widehat{\lambda}_{\mathrm{lin}}$ of the Liouville domain $[-1,1]_t \times M$ with the "linear" Liouville form

$$\lambda_{\mathrm{lin}} := (1-t)\alpha_- + (1+t)\alpha_+.$$

(2) The Liouville structure induced by the "exponential" Liouville form

$$\lambda_{\exp} := e^{-s}\alpha_- + e^{s}\alpha_+.$$

Here, the completion of a Liouville domain V is obtained by attaching to ∂V the symplectization $[0,\infty) \times \partial V$ of the contact structure at the boundary. This procedure yields an open manifold $\widehat{V}$ with controlled geometry at infinity. See [7, Section 11.1] for a precise definition. The next proposition shows in particular that (1) and (2) above produce equivalent Liouville structures on $\mathbb{R} \times M$.

Proposition 5.8 *Let (α_-, α_+) be a pair of contact forms which is both a Liouville pair and a linear Liouville pair. The Liouville structures $\widehat{\lambda}_{\mathrm{lin}}$ and $\lambda_{\exp}$ on $\mathbb{R} \times M$ are Liouville homotopic.*

Proof We choose an arbitrary volume form dvol and we define $f_\pm$, $g_\pm$ as in Lemma 5.2. We also choose $A > 0$ such that $f_\pm < A$ and $|g_\pm| < A$. We proceed in three steps.

Step 1 (extending λ_{lin}) Let $\epsilon > 0$. We choose a smooth function $\phi = \phi_\epsilon \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ satisfying

- $\phi(s) = 0$ for $s \leq -1-\epsilon$,
- $\phi(s) > 0$ for $s > -1-\epsilon$,
- $\phi(s) = 1+s$ for $s \geq -1+\epsilon$,
- ϕ is nondecreasing and convex.

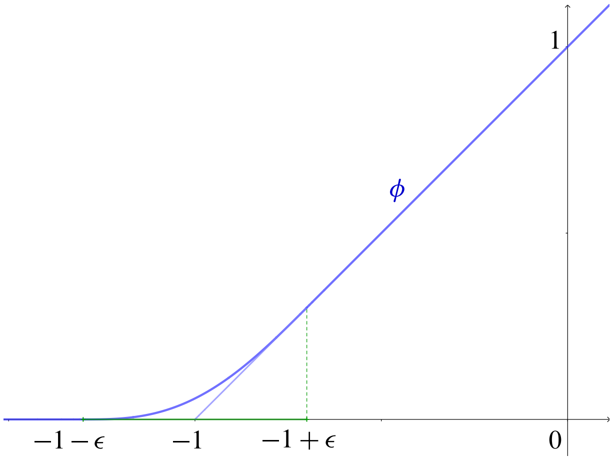


Figure 4: The function ϕ.

We claim that for ϵ sufficiently small, the 1-form

$$\lambda_0 := \phi(-s)\alpha_- + \phi(s)\alpha_+$$

on $\mathbb{R}_s \times M$ is Liouville. On $[-1+\epsilon, 1-\epsilon] \times M$, it coincides with λ_{lin} which is Liouville. On $(-\infty, -1-\epsilon] \times M$, it coincides with $(1-s)\alpha_-$ which is Liouville since α_- is a negative contact form. On $[1+\epsilon, \infty) \times M$, it coincides with $(1+s)\alpha_+$ which is Liouville since α_+ is positive contact form. If $s \in [1-\epsilon, 1+\epsilon]$, we compute

$$d\lambda_0 \wedge d\lambda_0 = \big\{(1+s)(f_+ - \phi'(-s)g_+) + \phi(-s)(\phi'(-s)f_- + g_-)\big\}\, ds \wedge \text{dvol} = F\, ds \wedge \text{dvol}.$$

Note that

$$0 \le \phi(-s) \le \epsilon, \quad 0 \le \phi'(-s) \le 1,$$

hence

$$F \ge (2-\epsilon)\min\{f_+, f_+ - g_+\} + 2\epsilon A,$$

where $f_+ > 0$ and $f_+ - g_+ > 0$ by Lemma 5.2; thus $F > 0$ for ϵ small enough. The case $s \in [-1-\epsilon, -1+\epsilon]$ is similar.

Step 2 (λ_0 is Liouville homotopic to $\widehat{\lambda}_{\text{lin}}$) We will use the following elementary fact: if two Liouville structures on a manifold with boundary V are Liouville homotopic, then their completions on $\widehat{V}$ are Liouville homotopic; see [7, Lemma 11.6]. By definition, $([-1-\epsilon, 1+\epsilon] \times M, \lambda_0)$ is a Liouville domain whose completion is exactly $(\mathbb{R} \times M, \lambda_0)$. Moreover, if ϵ is small enough, then $([-1-\epsilon, 1+\epsilon] \times M, \lambda_0)$ and $([-1+\epsilon, 1-\epsilon] \times M, \lambda_0) = ([-1+\epsilon, 1-\epsilon] \times M, \lambda_{\text{lin}})$ are Liouville domains which are Liouville homotopic (after identifying $[-1-\epsilon, 1+\epsilon]$ and $[-1+\epsilon, 1-\epsilon]$), and $([-1+\epsilon, 1-\epsilon] \times M, \lambda_{\text{lin}})$ and $([-1, 1] \times M, \lambda_{\text{lin}})$ are also Liouville homotopic (after identifying $[-1+\epsilon, 1-\epsilon]$ and $[-1, 1]$). This shows that $([-1-\epsilon, 1+\epsilon] \times M, \lambda_0)$ and $([-1, 1] \times M, \lambda_{\text{lin}})$ are Liouville homotopic (after identifying $[-1-\epsilon, 1+\epsilon]$ with $[-1, 1]$), and so are their completions.

Step 3 (λ_0 is Liouville homotopic to λ_{exp}) For $\tau \in [0,1]$, we set

$$\psi_\tau(s) := \tau e^s + (1-\tau)\phi(s), \quad \text{and} \quad \lambda_\tau := \psi_\tau(-s)\alpha_- + \psi_\tau(s)\alpha_+.$$

The family $\{\lambda_\tau\}_{\tau\in[0,1]}$ interpolates between λ_0 and $\lambda_1 = e^{-s}\alpha_- + e^s\alpha_+$. It is enough to show that for every $\tau \in (0,1)$, $d\lambda_\tau \wedge d\lambda_\tau$ is a positive volume form on $\mathbb{R} \times M$. By symmetry, it is enough to show it for $s \geq 0$. The computation of $d\lambda_\tau \wedge d\lambda_\tau$ reveals that the latter is equivalent to

$$f_+ + a_\tau(s)g_- - b_\tau(s)g_+ + a_\tau(s)b_\tau(s)f_- > 0, \tag{5-3}$$

where

$$a_\tau(s) := \frac{\psi_\tau(-s)}{\psi_\tau(s)}, \quad b_\tau(s) = \frac{\psi'_\tau(-s)}{\psi'_\tau(s)}.$$

It is easy to check that for $\tau \in (0,1)$ and $s \geq 0$, $0 \leq a_\tau(s) \leq 1$ and $0 \leq b_\tau(s) \leq 1$. Since (α_-, α_+) is both an exponential and a linear Liouville pair, we have that for every $a \in [0,1]$,

$$f_+ + ag_- - ag_+ + a^2 f_- > 0, \quad f_+ + ag_- - g_+ + af_- > 0,$$

so for every $a \in [0,1]$ and $b \in [a,1]$, we have

$$f_+ + ag_- - bg_+ + abf_- > 0. \tag{5-4}$$

By compactness, there exists $\delta > 0$, only depending on (α_-, α_+), such that for every $a \in [0,1]$ and $b \in [a-\delta, 1]$, the inequality (5-4) is satisfied. We claim that for every $\tau \in (0,1)$ and $s \geq 0$, we have

$$b_\tau(s) - a_\tau(s) \geq -\epsilon. \tag{5-5}$$

Indeed, fixing $\tau \in (0,1)$, we consider two cases.

Case 1 If $s \in [0, 1-\epsilon) \cup [1+\epsilon, \infty)$,

$$\psi_\tau(s) \geq \psi'_\tau(s), \quad \psi'_\tau(-s) \geq \psi_\tau(-s),$$

and (5-5) follows trivially since the left-hand side is nonnegative.

Case 2 If $s \in [1-\epsilon, 1+\epsilon)$,

$$\psi_\tau(s) = \tau e^s + (1-\tau)(1+s) \geq 1, \quad \psi'_\tau(s) = \tau e^s + (1-\tau) \geq 1,$$

and we compute

$$\begin{aligned} \psi'_\tau(-s)\psi_\tau(s) &= \tau^2 + \tau(1-\tau)\{(1+s)e^{-s} + e^s\phi'(-s)\} + (1-\tau)^2(1+s)\phi'(-s), \\ \psi'_\tau(s)\psi_\tau(-s) &= \tau^2 + \tau(1-\tau)\{e^s\phi(-s) + e^{-s}\} + (1-\tau)^2\phi(-s), \end{aligned}$$

hence

$$\begin{aligned} &\psi'_\tau(-s)\psi_\tau(s) - \psi'_\tau(s)\psi_\tau(-s) \\ &\qquad = (1-\tau)\Big\{\tau\big(\overbrace{e^s(\phi'(-s) - \phi(-s)) + se^{-s}}^{(1)}\big) + (1-\tau)\big(\underbrace{(1+s)\phi'(-s) - \phi(-s)}_{(2)}\big)\Big\}. \end{aligned}$$

Since $\phi'(-s) \geq 0$ and $0 \leq \phi(-s) \leq \epsilon$,

$$(1) \geq -e^{1+\epsilon}\epsilon + (1-\epsilon)e^{-(1+\epsilon)}, \quad (2) \geq -\epsilon.$$

For ϵ small enough, say $\epsilon \leq \frac{1}{100}$, (1) ≥ 0, and (5-5) follows.

This shows that for ϵ small enough, only depending on (α_-, α_+), the inequality (5-3) is satisfied for every $\tau \in (0, 1)$ and $s \geq 0$. The case $s \leq 0$ can be treated similarly. □

Remark 5.9 As mentioned in the introduction, we do not know if Theorems 4.3 and 4.8 are also true for linear Anosov Liouville pairs. The proof of Theorem 4.8 would immediately adapt to the linear case, provided that the space of ℓAL pairs supporting a given flow is (weakly) contractible. Our attempts at proving this fact for ℓAL pairs were fruitless because of the complexity and the lack of symmetry of the equations we obtained.

Appendix A Smoothing lemmas

This appendix concerns useful smoothing lemmas which are required to extend the results of this paper to Anosov flows generated by $\mathcal{C}^1$ vector fields, as their weak-stable and weak-unstable bundles are not necessarily $\mathcal{C}^1$. The approach can also be used to bypass Hozoori's delicate approximation techniques in [21, Section 4]. We state the results in greater generality than needed. M now denotes a closed n-dimensional manifold ($n \geq 1$) and X denotes a nonsingular vector field on M of class $\mathcal{C}^k$, $1 \leq k \leq \infty$ (without any Anosovity condition). We fix an arbitrary auxiliary metric on M.

The first smoothing lemma follows from [21, Lemma 4.3] and the regular approximation of differentiable functions by smooth ones.

Lemma A.1 *Let $f : M \to \mathbb{R}$ be a continuous function which is continuously differentiable along X. Then for every $\epsilon > 0$, there exists a smooth function $f^\epsilon : M \to \mathbb{R}$ satisfying*

$$|f^\epsilon - f|_{\mathcal{C}^0} \leq \epsilon \quad \textit{and} \quad |X \cdot f^\epsilon - X \cdot f|_{\mathcal{C}^0} \leq \epsilon.$$

In other words, with the notations of Definition 3.3, $\mathcal{C}^\infty$ is dense in $\mathcal{C}^0_X$. The same holds with $\mathcal{C}^\ell_X$ in place of $\mathcal{C}^0_X$, for $0 \leq \ell \leq k-1$. We will need a similar result for 1-forms on M.

Lemma A.2 *The space of $\mathcal{C}^k$ 1-forms on M vanishing along X is dense in $\Omega^1_{X,\ell}$ for $0 \leq \ell \leq k-1$.*

Proof This is a straightforward adaptation of the proof of [21, Lemma 4.3].

By compactness of M, we can find a positive real number $\tau > 0$ and a finite collection $\{(U_i, V_i, \phi_i)\}_{1 \leq i \leq N}$ where:

- $V_i \subset U_i \subset M$ are open subsets of M.
- $\{V_i\}_{1 \leq i \leq N}$ is a covering of M.
- $\phi_i : U_i \to (-2\tau, 2\tau)_t \times D$ is a $\mathcal{C}^k$ diffeomorphism such that $\phi_i(V_i) = (-\tau, \tau) \times D$ and $d\phi_i(X) = \partial_t$. Here, D denotes the open unit disk in $\mathbb{R}^{n-1}$.

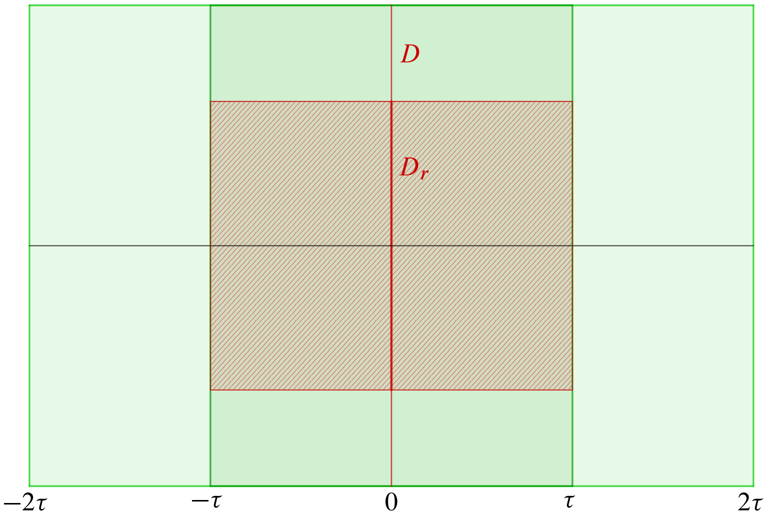


Figure 5: The nested open sets in the proof of Lemma A.2. The support of α'_i is contained in the hashed region.

Such a collection can be obtained by taking a finite collection of sufficiently small flow-boxes for X covering M.

Let $\{\psi_i\}_{1\leq i\leq N}$ be a partition of unity subordinate to the open covering $\{V_i\}_{1\leq i\leq N}$. For every i, the support of ψ_i is contained in V_i so we can find $0<r<1$ such that the support of $\psi_i\circ\phi_i^{-1}$ is contained in $(-\tau,\tau)\times D_r$, where D_r denotes the open disk of radius r.

Let $h\colon\mathbb{R}\to\mathbb{R}_{\geq 0}$ be a smooth bump function satisfying

- for $|t|\leq\tau$, $h(t)=1$,
- for $|t|\geq 2\tau$, $h(t)=0$,
- h is nondecreasing on $(-\infty,0)$ and nonincreasing on $(0,\infty)$.

Let $\alpha\in\Omega^1_{X,\ell}$. By definition, $\alpha=\sum_{i=1}^N\psi_i\alpha$. For every i, we write $\alpha_i:=\psi_i\alpha$ and $\alpha'_i:=(\phi_i)_*\alpha_i$. We have reduced the problem to a single flow-box $(-2\tau,2\tau)\times D$.

Let $\epsilon>0$. In what follows, the symbol "$\lesssim$" means "less than or equal to, up to a constant factor that does not depend on ϵ". For a fixed i, let β_i^0 be a smooth 1-form on D with support contained in D_r satisfying

$$|\beta_i^0-\alpha'_{i\,|\{0\}\times D}|_{\mathcal{C}^\ell}\leq\epsilon.$$

By definition, $\mathcal{L}_{\partial t}\alpha'_i$ is $\mathcal{C}^\ell$ and vanishes along ∂_t. Therefore, we can find a smooth 1-form η_i on $(-2\tau,2\tau)\times D$ with support contained in $(-\tau,\tau)\times D_r$ satisfying

$$\eta_i(\partial_t)=0,\quad |\eta_i-\mathcal{L}_{\partial t}\alpha'_i|_{\mathcal{C}^\ell}\leq\epsilon.$$

We can extend β_i^0 to a smooth 1-form β_i on $(-2\tau,2\tau)\times D$ by setting

$$\beta_i(\partial_t):=0,\quad \mathcal{L}_{\partial t}\beta_i:=\eta_i.$$

Finally, we define $\beta_i' := h\beta_i$. This is a smooth 1-form with support in $(-2\tau, 2\tau) \times D_r$. Note that at a point $(t,x) \in (-2\tau, 2\tau) \times D$, we have

$$(\beta_i - \alpha_i')_{(t,x)} = (\beta_i^0 - \alpha'_{i\,|\{0\}\times D})_x + \int_0^t (\eta_i - \mathcal{L}_{\partial t}\alpha_i')_{(s,x)}\, ds,$$

so

$$|\beta_i - \alpha_i'|_{\mathcal{C}^\ell} \le |\beta_i^0 - \alpha'_{i\,|\{0\}\times D}|_{\mathcal{C}^\ell} + 2\tau|\eta_i - \mathcal{L}_{\partial t}\alpha_i'|_{\mathcal{C}^\ell} \lesssim \epsilon, \quad |\mathcal{L}_{\partial t}\beta_i - \mathcal{L}_{\partial t}\alpha_i'|_{\mathcal{C}^\ell} = |\eta_i - \mathcal{L}_{\partial t}\alpha_i'|_{\mathcal{C}^\ell} \le \epsilon.$$

Since the support of α_i' is contained in $(-\tau, \tau) \times D_r$, we readily get

$$|\beta_{i\,|((-2\tau,-\tau)\cup(\tau,2\tau))\times D}|_{\mathcal{C}^\ell} \lesssim \epsilon.$$

Moreover,

- On $((-2\tau, -\tau) \cup (\tau, 2\tau)) \times D$,
$$\beta_i' - \alpha_i' = h\beta_i, \quad \mathcal{L}_{\partial t}\beta_i' - \mathcal{L}_{\partial t}\alpha_i' = (\partial_t h)\beta_i,$$
- On $(-\tau, \tau) \times D$,
$$\beta_i' - \alpha_i' = \beta_i - \alpha_i', \quad \mathcal{L}_{\partial t}\beta_i' - \mathcal{L}_{\partial t}\alpha_i' = \eta_i - \mathcal{L}_{\partial t}\alpha_i',$$

and we obtain

$$|\beta_i' - \alpha_i'|_{\mathcal{C}^\ell} \lesssim \epsilon, \quad |\mathcal{L}_{\partial t}\beta_i' - \mathcal{L}_{\partial t}\alpha_i'|_{\mathcal{C}^\ell} \lesssim \epsilon.$$

Finally, we define

$$\beta := \sum_i \phi_i^*\beta_i',$$

so that β is a $\mathcal{C}^k$ 1-form on M satisfying $\beta(X) = 0$, and

$$|\beta - \alpha|_{\mathcal{C}^\ell} \le \sum_{i=1}^N |\phi_i^*\beta_i' - \phi_i^*\alpha_i'|_{\mathcal{C}^\ell} \lesssim \sum_{i=1}^N |\beta_i' - \alpha_i'|_{\mathcal{C}^\ell} \lesssim \epsilon,$$

$$|\mathcal{L}_X\beta - \mathcal{L}_X\alpha|_{\mathcal{C}^\ell} \le \sum_{i=1}^N |\phi_i^*(\mathcal{L}_{\partial t}\beta_i') - \phi_i^*(\mathcal{L}_{\partial t}\alpha_i')|_{\mathcal{C}^\ell} \lesssim \sum_{i=1}^N |\mathcal{L}_{\partial t}\beta_i' - \mathcal{L}_{\partial t}\alpha_i'|_{\mathcal{C}^\ell} \lesssim \epsilon.$$

This finishes the proof. □

Appendix B Almost volume preserving Anosov flows

In this appendix, we prove a technical result used in the proof of Theorem 3.15. Let us recall the setup. Φ is a smooth Anosov flow on a closed oriented 3-manifold M, generated by a vector field X. For an adapted metric g, $r_u > 0$ and $r_s < 0$ denote the expansion rates in the unstable and stable directions, respectively. The divergence of X for this metric is simply $\operatorname{div}_g X = r_u + r_s$. We say that Φ is *almost volume preserving* if it satisfies one of the following equivalent conditions (compare with (3-12)):

(C1) For every $\epsilon > 0$, there exists a smooth function $f_\epsilon \colon M \to \mathbb{R}$ satisfying

$$|\operatorname{div}_g X + X \cdot f_\epsilon| \leq \epsilon.$$

(C2) For every $\epsilon > 0$, there exists a smooth volume form dvol_ϵ on M satisfying

$$|\operatorname{div}_\epsilon X| = \left|\frac{\mathcal{L}_X \mathrm{dvol}_\epsilon}{\mathrm{dvol}_\epsilon}\right| \leq \epsilon,$$

where $\operatorname{div}_\epsilon X$ denotes the divergence of X with respect to dvol_ϵ.

Proposition B.1 *If Φ is a smooth almost volume preserving Anosov flow on M, then Φ is volume preserving.*

Proof As noted in the proof of Theorem 3.15, it is enough to show that Φ is topologically transitive, ie Φ has a dense orbit. We closely follow the strategy from [27] that relies on some key properties of Sinai–Ruelle–Bowen (SRB) measures for Anosov diffeomorphisms. To adapt the proof to the case of Anosov *flows*, we rely on the results of [6].

Let $\Lambda \subset M$ be an attractor for Φ as defined in [6]. It exists thanks to Smale's spectral decomposition of the nonwandering set of Φ into basic sets. We will show that Λ is also an attractor for Φ^{-1}, implying that $\Lambda = M$. Since Φ is topologically transitive on Λ, it will be topologically transitive on M, as desired.

Let g be a Riemannian metric on M adapted to Φ (so that $r_s < 0 < r_u$), and let $\phi = \phi^1$ be the time-one map of Φ. By [6], there exists a unique Borel probability measure $\mu_u := \mu_{\varphi^{(u)}}$ on Λ satisfying the *Pesin entropy formula*

$$P(\Phi_{|\Lambda}, \varphi^{(u)}) = h_{\mu_u}(\phi) + \int_\Lambda \varphi^{(u)}\, d\mu_u = 0, \tag{B-1}$$

where $\varphi^{(u)} := -r_u$, $h_{\mu_u}(\phi)$ denotes the topological entropy of ϕ with respect to μ_u, and P denotes the topological pressure. Moreover, by [6, Theorem 5.5], this measure is ergodic on the basin W^s_Λ of Λ: for every continuous function $g \colon M \to \mathbb{R}$ and almost every point $x \in W^s_\Lambda$ with respect to the Lebesgue measure, one has

$$\int_\Lambda g\, d\mu_u = \lim_{T \to +\infty} \frac{1}{T} \int_0^T g(\phi^t(x))\, dt. \tag{B-2}$$

The hypothesis on Φ implies that for every $x \in M$,[12]

$$\lim_{T \to +\infty} \frac{1}{T} \int_0^T (r_u + r_s)(\phi^t(x))\, dt = 0. \tag{B-3}$$

[12] This exactly means that the sum of the nonzero *Lyapunov exponents* $\Lambda_u + \Lambda_s$ of Φ vanishes wherever they are both defined.

Indeed, let $\epsilon > 0$ and choose a smooth (hence bounded) function $f_\epsilon : M \to \mathbb{R}$ such that $|r_u + r_s + X \cdot f_\epsilon| < \epsilon$. Then we have

$$\begin{aligned}\limsup_{T\to+\infty}\left|\frac{1}{T}\int_0^T (r_u + r_s)\circ\phi^t\,dt\right| &\\ &\leq \limsup_{T\to+\infty}\frac{1}{T}\int_0^T |r_u + r_s + X\cdot f_\epsilon|\circ\phi^t\,dt + \limsup_{T\to+\infty}\frac{1}{T}\left|\int_0^T (X\cdot f_\epsilon)\circ\phi^t\,dt\right| \\ &\leq \epsilon + \limsup_{T\to+\infty}\frac{1}{T}|f_\epsilon\circ\phi^T - f_\epsilon| \\ &= \epsilon.\end{aligned}$$

Then, applying (B-2) to $g = r_u + r_s$ and using (B-3), we readily obtain

$$\int_\Lambda (r_u + r_s)\,d\mu_u = 0, \quad \text{and hence} \quad h_{\mu_u}(\phi) + \int_\Lambda r_s\,d\mu_u = 0.$$

However, since $h_{\mu_u}(\phi) = h_{\mu_u}(\phi^{-1})$, the left-hand side of the previous equation is exactly

$$h_{\mu_u}(\phi^{-1}) + \int_\Lambda \varphi^{(s)}\,d\mu_u,$$

where $\varphi^{(s)}$ plays the role of $\varphi^{(u)}$ for Φ^{-1}. This implies that

$$0 = h_{\mu_u}(\phi^{-1}) + \int_\Lambda \varphi^{(s)}\,d\mu_u \leq P(\Phi^{-1}_{|\Lambda}, \varphi^{(s)}) \leq 0,$$

where the first inequality follows from [6, Section 3] and the second from [6, Proposition 4.4]. Therefore, $P(\Phi^{-1}_{|\Lambda}, \varphi^{(s)}) = 0$ and Λ is an attractor for Φ^{-1} by [6, Theorem 5.6]. □

Remark B.2 The proof can be adapted to show that any almost volume preserving Anosov flow of class $\mathcal{C}^2$ on a closed manifold of any dimension is volume preserving.

Remark B.3 The main result of [27] asserts that an Anosov *diffeomorphism* of class $\mathcal{C}^2$ on a closed manifold M satisfying that at every periodic point the Poincaré return map has determinant one is volume preserving. The same result remains true for an Anosov flow whose Poincaré return map at every closed orbit has determinant one. As in the proof in [27], this condition and Anosov's *shadowing property* imply that the sum of the Lyapunov exponents $\Lambda_u + \Lambda_s$ vanishes almost everywhere, which is enough to conclude.

References

[1] **D V Anosov**, *Ergodic properties of geodesic flows on closed Riemannian manifolds of negative curvature*, Dokl. Akad. Nauk SSSR 151 (1963) 1250–1252 MR Zbl In Russian

[2] **D V Anosov**, *Geodesic flows on closed Riemannian manifolds of negative curvature*, Trudy Mat. Inst. Steklov. 90 (1967) 3–210 MR Zbl In Russian; translated in Proc. Steklov Inst. Math. 98 (1967) 1–235

[3] **M Asaoka**, *On invariant volumes of codimension-one Anosov flows and the Verjovsky conjecture*, Invent. Math. 174 (2008) 435–462 MR Zbl

[4] **T Barthelmé**, *Anosov flows in dimension* 3*: preliminary version*, lecture notes, Université de Montréal (2017) `http://www.crm.umontreal.ca/sms/2017/pdf/diapos/Anosov_flows_in_3_manifolds.pdf`

[5] **T Barthelmé**, **K Mann**, *Orbit equivalences of $\mathbb{R}$-covered Anosov flows and hyperbolic-like actions on the line*, Geom. Topol. 28 (2024) 867–899 MR Zbl

[6] **R Bowen**, **D Ruelle**, *The ergodic theory of Axiom A flows*, Invent. Math. 29 (1975) 181–202 MR Zbl

[7] **K Cieliebak**, **Y Eliashberg**, *From Stein to Weinstein and back: symplectic geometry of affine complex manifolds*, AMS Colloquium Publications 59, Amer. Math. Soc., Providence, RI (2012) MR Zbl

[8] **K Cieliebak**, **O Lazarev**, **T Massoni**, **A Moreno**, *Floer theory of Anosov flows in dimension three*, preprint (2022) arXiv 2211.07453

[9] **V Colin**, **S Firmo**, *Paires de structures de contact sur les variétés de dimension trois*, Algebr. Geom. Topol. 11 (2011) 2627–2653 MR Zbl

[10] **V Colin**, **E Giroux**, **K Honda**, *Finitude homotopique et isotopique des structures de contact tendues*, Publ. Math. Inst. Hautes Études Sci. 109 (2009) 245–293 MR Zbl

[11] **T tom Dieck**, *Partitions of unity in homotopy theory*, Compos. Math. 23 (1971) 159–167 MR Zbl

[12] **Y Eliashberg**, *A few remarks about symplectic filling*, Geom. Topol. 8 (2004) 277–293 MR Zbl

[13] **Y M Eliashberg**, **W P Thurston**, *Confoliations*, University Lecture Series 13, Amer. Math. Soc., Providence, RI (1998) MR Zbl

[14] **T Fisher**, **B Hasselblatt**, *Hyperbolic flows*, Eur. Math. Soc., Berlin (2019) MR Zbl

[15] **D T Gay**, *Four-dimensional symplectic cobordisms containing three-handles*, Geom. Topol. 10 (2006) 1749–1759 MR Zbl

[16] **E Ghys**, *Déformations de flots d'Anosov et de groupes fuchsiens*, Ann. Inst. Fourier (Grenoble) 42 (1992) 209–247 MR Zbl

[17] **E Ghys**, *Rigidité différentiable des groupes fuchsiens*, Inst. Hautes Études Sci. Publ. Math. 78 (1993) 163–185 MR Zbl

[18] **M Gromov**, *Partial differential relations*, Ergebnisse der Math. 9, Springer (1986) MR Zbl

[19] **B Hasselblatt**, *Regularity of the Anosov splitting and of horospheric foliations*, Ergodic Theory Dynam. Systems 14 (1994) 645–666 MR Zbl

[20] **S Hozoori**, *On Anosovity, divergence and bi-contact surgery*, Ergodic Theory Dynam. Systems 43 (2023) 3288–3310 MR Zbl

[21] **S Hozoori**, *Symplectic geometry of Anosov flows in dimension* 3 *and bi-contact topology*, Adv. Math. 450 (2024) art. id. 109764 MR Zbl

[22] **R de la Llave**, **J M Marco**, **R Moriyón**, *Canonical perturbation theory of Anosov systems and regularity results for the Livšic cohomology equation*, Ann. of Math. 123 (1986) 537–611 MR Zbl

[23] **T A Marty**, *Skewed Anosov flows in dimension* 3 *are Reeb-like*, J. Eur. Math. Soc. (online publication March 2025)

[24] **T Massoni**, *Taut foliations and contact pairs in dimension three*, preprint (2024) arXiv 2405.15635

[25] **P Massot**, **K Niederkrüger**, **C Wendl**, *Weak and strong fillability of higher dimensional contact manifolds*, Invent. Math. 192 (2013) 287–373 MR Zbl

[26] **G Meigniez**, *Submersions, fibrations and bundles*, Trans. Amer. Math. Soc. 354 (2002) 3771–3787 MR Zbl

[27] **F Micena**, *Some sufficient conditions for transitivity of Anosov diffeomorphisms*, J. Math. Anal. Appl. 515 (2022) art. id. 126433 MR Zbl

[28] **Y Mitsumatsu**, *Anosov flows and non-Stein symplectic manifolds*, Ann. Inst. Fourier (Grenoble) 45 (1995) 1407–1421 MR Zbl

[29] **R C Robinson**, *Structural stability of C^1 flows*, from "Dynamical systems" (A Manning, editor), Lecture Notes in Math. 468, Springer (1975) 262–275 MR Zbl

[30] **S Simić**, *Codimension one Anosov flows and a conjecture of Verjovsky*, Ergodic Theory Dynam. Systems 17 (1997) 1211–1231 MR Zbl

[31] **M Weiss**, *What does the classifying space of a category classify?*, Homology Homotopy Appl. 7 (2005) 185–195 MR Zbl

Department of Mathematics, Princeton University
Princeton, NJ, United States

Current address: *Department of Mathematics, Stanford University*
Stanford, CA, United States

tmassoni.math@gmail.com

Received: 8 June 2023 Revised: 27 August 2023

Algebraic & Geometric Topology **25**:3 (2025) 1839–1876
DOI: 10.2140/agt.2025.25.1839
Published: 20 June 2025

Hamiltonian classification of toric fibres and symmetric probes

JOÉ BRENDEL

In a toric symplectic manifold, regular fibres of the moment map are Lagrangian tori which are called *toric fibres*. We discuss the question of which two toric fibres are equivalent up to a Hamiltonian diffeomorphism of the ambient space. On the construction side of this question, we introduce a new method of constructing equivalences of toric fibres by using a symmetric version of McDuff's probes. On the other hand, we derive some obstructions to such equivalence by using Chekanov's classification of product tori together with a lifting trick from toric geometry. Furthermore, we conjecture that (iterated) symmetric probes yield all possible equivalences and prove this conjecture for $\mathbb{C}^n$, $\mathbb{C}P^2$, $\mathbb{C} \times S^2$, $\mathbb{C}^2 \times T^*S^1$, $T^*S^1 \times S^2$ and monotone $S^2 \times S^2$.

This problem is intimately related to determining the *Hamiltonian monodromy group* of toric fibres, ie determining which automorphisms of the homology of the toric fibre can be realized by a Hamiltonian diffeomorphism mapping the toric fibre in question to itself. For the above list of examples, we determine the Hamiltonian monodromy group for all toric fibres.

53D12; 53D20

1 Introduction

1.1 Symmetric probes

Probes were introduced by McDuff [23] to prove that some toric fibres are displaceable. Probes are rational segments in the base of a toric base polytope which hit the boundary integrally transversely in one point. The latter condition implies that one can perform symplectic reduction on a probe and obtain a two-disk as reduced space; see also the exposition of Abreu and Macarini [2]. Toric fibres map to circles in the reduced space, where displaceability questions are easy to settle since they boil down to area arguments.

Symmetric probes are rational segments in which *both* endpoints hit the boundary of the moment polytope integrally transversely. They were introduced in a follow-up paper to [23] by Abreu, Borman and McDuff [1] to settle some more subtle displaceability questions. Here, we use them to a different end. The reduced space associated to a symmetric probe is a two-sphere and the quotient map takes toric fibres to orbits of the standard S^1-action on the two-sphere. Observe that — except for the equator — orbits of this circle action in S^2 appear in pairs which are Hamiltonian isotopic. Our main observation is that, since Hamiltonian isotopies in reduced spaces can be lifted, this proves that toric fibres corresponding to such pairs of circles are Hamiltonian isotopic, as well. This is illustrated in Figure 1.

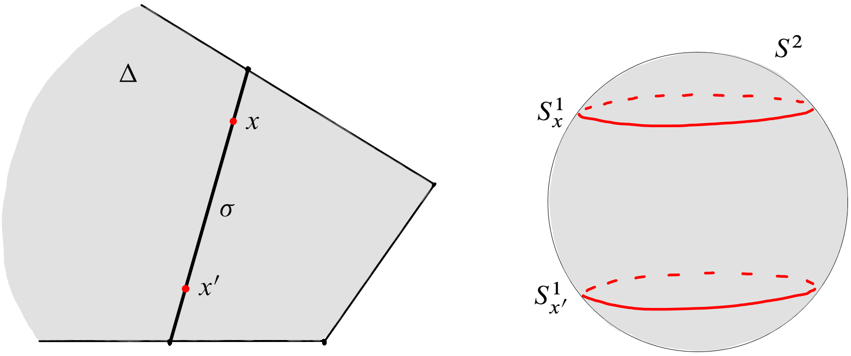


Figure 1: A symmetric probe σ in a moment polytope Δ with points x and x' at equal distance to the boundary. The toric fibres $T(x)$ and $T(x')$ map to the circles $S^1_x, S^1_{x'} \subset S^2$ under symplectic reduction.

To state this formally, let us introduce some notation. Let (X^{2n}, ω) be a (not necessarily compact) toric symplectic manifold with moment map $\mu\colon X \to \mathbb{R}^n$ and moment polytope $\mu(X) = \Delta$. For $x \in \operatorname{int} \Delta$ the set $T(x) = \mu^{-1}(x)$ is a Lagrangian torus, called a *toric fibre*. A *symmetric probe* $\sigma \subset \Delta$ is a rational segment intersecting $\partial\Delta$ integrally transversely in the in interior of two facets; see also [1, Definition 2.2.3]. An intersection of a rational line and a rational hyperplane is called *integrally transverse* if their union contains a $\mathbb{Z}$-basis of $\mathbb{Z}^n$. See also Definition 3.1 and the surrounding discussion or [23, Section 2.1] for more details.

Theorem A *Let (X, ω) be a toric symplectic manifold and let $\sigma \subset \Delta$ be a symmetric probe in its moment polytope. Furthermore, let $x, x' \in \sigma$ be two points at equal distance to the boundary $\partial\Delta$. Then $T(x)$ and $T(x')$ are Hamiltonian isotopic.*

1.2 Classification of toric fibres

Deciding which two given Lagrangians L and L' in (X, ω) can be mapped to one another by a symplectomorphism or by a Hamiltonian diffeomorphism is a central question in symplectic geometry. In many situations, it is quite hopeless to give a full classification — even constructing examples of Lagrangians that are not equivalent to known ones (so-called *exotic* Lagrangians) is an active area of research where many questions are open; see for example Auroux [5], Chekanov and Schlenk [14], and Vianna [30; 31]. In this paper we care about the following classification question of Lagrangian submanifolds.

Question 1.1 *In a toric symplectic manifold (X, ω), give a classification of toric fibres up to Hamiltonian diffeomorphisms of the ambient space.*

Remark 1.2 One can ask the same questions for symplectomorphisms of the ambient space. In this paper we focus on the case of Hamiltonian diffeomorphisms. See also Remark 4.2.

Although Question 1.1 is a much less ambitious question than a full classification of all Lagrangian tori (since we exclude exotic tori a priori) of X, it is open except for a few special cases and surprisingly absent from the literature. To our knowledge, it has only been answered for $\mathbb{C}^n$ (where toric fibres are simply product tori) by Chekanov [12, Theorem A] and for $\mathbb{C}P^2$ by Shelukhin, Tonkonog and Vianna [27, Proposition 7.1].

Let us make some conventions. From now on, we call $T(x)$ and $T(x')$ *equivalent* and write $T(x) \cong T(x')$ if they can be mapped to one another by a Hamiltonian diffeomorphism of the ambient space. Furthermore, let

$$\mathfrak{H}_x = \{x' \in \operatorname{int}\Delta \mid T(x) \cong T(x')\}, \tag{1}$$

the set of toric fibres equivalent to $T(x)$. A first guess may be that $\mathfrak{H}_x = \{x\}$, since the zero-section in T^*T^n is nondisplaceable (see McDuff and Salamon [24, Section 11.3]) and thus this is true if we restrict our attention to Hamiltonian diffeomorphisms supported in a Weinstein neighbourhood of $T(x)$. However, a glance at S^2 shows that this guess is wrong, since one can use the topology of the ambient space to obtain nontrivial equivalences of toric fibres. More generally, by Theorem A, symmetric probes (and their concatenations) can be used to construct equivalences of toric fibres up to Hamiltonian diffeomorphisms. Let us also point out that symmetric probes are abundant in arbitrary toric manifolds — at least close to the boundary of the moment polytope; see Section 5.7. We conjecture that the method of constructing equivalent toric fibres by symmetric probes gives a complete answer to the classification question.

Conjecture 1.3 *Two toric fibres $T(x), T(x') \subset X$ are equivalent if and only if they are equivalent by a sequence of symmetric probes.*

In Section 5, we verify this conjecture for $\mathbb{C}^n$ and $\mathbb{C}P^2$ (where the classification was previously known), for $\mathbb{C} \times S^2$, $\mathbb{C}^2 \times T^*S^1$, $T^*S^1 \times S^2$ and for monotone $S^2 \times S^2$ (where we classify toric fibres). The classification of toric fibres in nonmonotone $S^2 \times S^2$ is more intricate and is given in [7].

On the side of obstructions to Hamiltonian equivalence, we prove the following.

Theorem B *If toric fibres $T(x), T(x') \subset X$ of a compact toric manifold X are Hamiltonian isotopic, then the following three invariants agree*:

$$d(x) = d(x'), \quad \#_d(x) = \#_d(x'), \quad \Gamma(x) = \Gamma(x'). \tag{2}$$

The invariant $d(x) \in \mathbb{R}$ is the *integral affine distance* of x to the boundary of the moment polytope. The invariant $\#_d(x) \in \mathbb{N}_{\geqslant 1}$ is the number of facets of Δ realizing the minimal distance $d(x)$. Both of these invariants are *hard* in the symplectic sense. The last invariant is the subgroup

$$\Gamma(x) = \mathbb{Z}\langle \ell_1(x) - d(x), \dots, \ell_N(x) - d(x)\rangle \subset \mathbb{R} \tag{3}$$

and it is soft. Here $\ell_i(x)$ denotes the integral affine distance of x to the i^{th} facet of Δ. Since these invariants are derived from Chekanov's invariants [12, Theorem A] of product tori in $\mathbb{R}^{2n} = \mathbb{C}^n$, we call them *Chekanov invariants*.

Let us outline the proof of Theorem B. Suppose $T(x), T(x') \subset X$ are Hamiltonian isotopic fibres. By a construction going back to Delzant [18], we can view X as a symplectic quotient of $\mathbb{C}^N$, where N is the number of facets of Δ. The preimages of the tori $T(x)$ and $T(x')$ under the symplectic quotient map are the product tori $T(\ell(x)), T(\ell(x')) \subset \mathbb{C}^N$, where $\ell = (\ell_1, \dots, \ell_N)$. The Hamiltonian isotopy mapping $T(x)$ to $T(x')$ lifts to a Hamiltonian isotopy of $\mathbb{C}^N$ mapping $T(\ell(x))$ to $T(\ell(x'))$. This means that Chekanov's invariants for product tori have to agree on $T(\ell(x))$ and $T(\ell(x'))$, which yields the statement. To our knowledge, this *lifting trick* first appeared in [2] to prove nondisplaceability of certain fibres and it was also heavily used in [6]. It is not obvious to us how to prove Theorem B directly, ie without using the lifting trick. The first two invariants are clearly related to the area and the number of nontrivial Maslov two J-holomorphic disks of minimal area with boundary on the corresponding tori, respectively. It is not obvious how to pursue this due to the lack of monotonicity, although an approach in the spirit of [27] may be promising, especially in dimension four; see the remark in [27, Section 5.6].

The invariants in Theorem B are not complete, even in very simple examples such as $\mathbb{C}P^2$; see Example 4.4. We suspect that the first two invariants are all there is in terms of hard obstructions, but that the soft invariant $\Gamma(\,\cdot\,)$ is far from optimal — this is the case in all examples where we know the classification.

1.3 Examples

Let us give some examples of symmetric probes. In dimension two, there are not many toric spaces. The main examples are $T^*S^1 = S^1 \times \mathbb{R}$ equipped with the standard exact symplectic form and moment map given by projection to the $\mathbb{R}$-coordinate, $\mathbb{C} = \mathbb{R}^2$ equipped with the standard symplectic form and moment map $z \mapsto \pi|z|^2$, and S^2 equipped with the height function. We normalize the symplectic form ω_{S^2} such that $\int_{S^2} \omega_{S^2} = 2$ meaning that the corresponding moment polytope is $[-1, 1]$. In the two-dimensional setting, symmetric probes are not interesting and the classification of toric fibres boils down to simple area arguments. However, some four-dimensional products of the above examples (equipped with the product symplectic and toric structures) already contain nontrivial probes.

In $\mathbb{C}^2$, there is one nontrivial probe in direction $(1, -1)$, which can be used to show that $T(x, y) \cong T(y, x)$, which also follows from the fact that all elements in $\mathrm{U}(2)$ can be realized by Hamiltonian diffeomorphisms. These are all possible equivalences in $\mathbb{C}^2$, as was shown by Chekanov [12]. In $T^*S^1 \times S^2$, all directions $(k, 1)$ for $k \in \mathbb{Z}$ give symmetric probes; see Figure 2. This proves that $T(x, y)$ is Hamiltonian isotopic to all $T(x + 2ky, \pm y)$. Note that this also follows from a suspension argument due to Polterovich [25, Example 6.3.C] and the discussion of Mak and Smith in [22, Section 1.3]. The example $\mathbb{C} \times S^2$ is obtained from the previous one by a vertical symplectic cut and we will see in Section 5.3 that there are slightly more equivalences between toric fibres. In $\mathbb{C}P^2$ and monotone $S^2 \times S^2$, it is easy to see that symmetric probes realize all equivalences of toric fibres coming from symmetries of the moment polytope. In all of these examples, the method by probes is sharp and the classification of toric fibres is discussed in detail in Section 5.

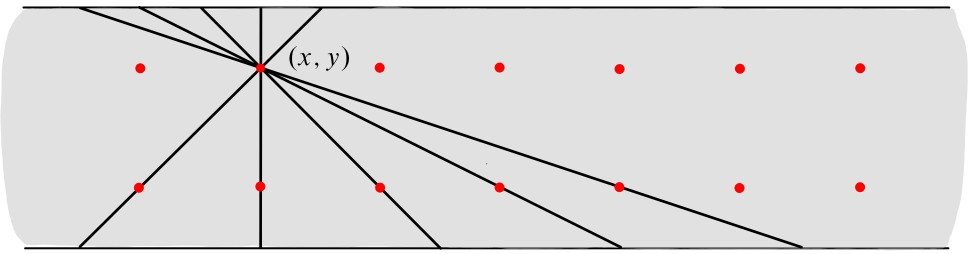


Figure 2: The set $\mathfrak{H}_{(x,y)}$ for $T(x,y) \subset T^*S^1 \times S^2$ and some symmetric probes.

In dimensions $\geqslant 6$, the situation is quantitatively different from the above examples. Indeed, the set $\mathfrak{H}_x$ has accumulation points in Δ for many $x \in \operatorname{int} \Delta$, see Corollary 5.16. This already occurs in the case of $\mathbb{C}^3$, treated by Chekanov [12]; see also Theorem 4.3. In essence, this is due to the existence of a symmetric probe in direction $(1, 1, -1)$ (or coordinate permutations thereof); see Figure 7. In Section 5.6, we show that one can recover Chekanov's classification using symmetric probes. The property that $\mathfrak{H}_x$ has accumulation points is not exclusive to dimension six and above. In fact, in the forthcoming [7], we show that this occurs in $S^2 \times S^2$ equipped with any nonmonotone symplectic form.

In light of this, it would be very interesting to characterize the toric manifolds having the property that there exists $x \in \operatorname{int} \Delta$ with $\mathfrak{H}_x$ not discrete.

1.4 Hamiltonian monodromy of toric fibres

Let $x \in \sigma \subset \Delta$ be the midpoint of a symmetric probe. The corresponding toric fibre $T(x)$ projects to the equator of the sphere obtained as a reduced space and thus we do not get any equivalence with another toric fibre by the above method. However, we still get information about $T(x)$. Indeed, we can lift a Hamiltonian isotopy mapping the equator in the reduced sphere to itself but changing the orientation of the equator. By lifting such a Hamiltonian isotopy, we obtain a Hamiltonian isotopy mapping $T(x)$ to itself with nontrivial homological monodromy, meaning that it induces a nontrivial map in $\operatorname{Aut} H_1(T(x); \mathbb{Z})$. An explicit formula for this monodromy map in terms of data related to the symmetric probe σ is given in (22).

Definition 1.4 Let $L \subset (X, \omega)$ be a compact Lagrangian submanifold. The *Hamiltonian monodromy group* is given by

$$\mathcal{H}_L = \{(\phi|_L)_* \in \operatorname{Aut} H_1(L; \mathbb{Z}) \mid \phi \in \operatorname{Ham}(X, \omega),\ \phi(L) = L\}. \tag{4}$$

The analogous monodromy group for symplectomorphisms was computed by Chekanov for product tori and Chekanov tori in [12, Theorem 4.5] and, in that case, the Hamiltonian monodromy group actually agrees with it. To our knowledge this is the first occurrence of this kind of question in the literature. See also Yau [32] for related results and Hu, Lalonde and Leclercq [21] which establishes that weakly exact Lagrangian manifolds have trivial Hamiltonian monodromy group. See Porcelli [26] for recent progress in the same direction. Another recent work is Augustynowicz, Smith and Wornbard [4] which makes significant progress in case L is a monotone Lagrangian torus and provides an excellent overview of the topic in its introduction.

Let $\xi_1, \dots, \xi_N \in \mathbb{Z}^n$ be the set of inward pointing primitive normal vectors to the facets of Δ, as in (6). The vectors ξ_i naturally determine homology classes $\xi_i \in H_1(T(x))$ for every toric fibre $T(x)$. See for example the discussion surrounding (16). Let $\mathfrak{D}(x)$ be the subset of those normal vectors realizing the minimal integral affine distance of x to facets,

$$\mathfrak{D}(x) = \{\xi_i \mid \ell_i(x) = d(x)\}. \tag{5}$$

We call elements of this subset *distinguished classes*. Note that $\#\mathfrak{D}(x) = \#_d(x)$. The following is an obstruction result for Hamiltonian monodromy of toric fibres.

Theorem C *Let $T(x) \subset X$ be a toric fibre in a compact toric manifold. Every element in the Hamiltonian monodromy group $\mathcal{H}_{T(x)}$ acts by a permutation on the set $\mathfrak{D}(x)$ of distinguished classes.*

This theorem again follows from Chekanov's work [12, Theorem 4.5] together with the lifting trick discussed in Section 1.2. In fact, we get a stronger statement; see Theorem 4.7. The number $\#_d(x)$ of distinguished classes is maximal if $T(x)$ is monotone, since all integral affine distances are equal in that case. In fact, in the monotone case, we recover [4, Theorem 2] for Hamiltonian diffeomorphisms; see Corollary 4.9. Note that Theorem C does not require monotonicity.

In terms of examples, we give a complete description of $\mathcal{H}_L$ for all toric fibres in $S^2 \times S^2$, $\mathbb{C}P^2$, $\mathbb{C} \times S^2$, $\mathbb{C}^2 \times T^*S^1$ and $T^*S^1 \times S^2$, and show that all Hamiltonian monodromy elements can be realized by symmetric probes as outlined above.

1.5 Outline

In Section 2, we review the relevant toric geometry and in particular we discuss *toric reduction*, a version of symplectic reduction which is compatible with the toric structure and on which we rely to prove Theorems A, B and C. Section 3 is the heart of this paper, where we discuss symmetric probes and prove Theorem A. In Section 4, we discuss obstructions to the equivalence of toric fibres and prove Theorem B. Furthermore, we discuss obstructions to which Hamiltonian monodromy can be realized for toric fibres. Section 5 is dedicated to examples and serves to illustrate the results of the previous sections.

Acknowledgements

We cordially thank Jonny Evans, Joontae Kim and Felix Schlenk for many useful discussions. We are grateful to Jack Smith for generously sharing his insights on Hamiltonian monodromy and offering explanations about [4]. This work was started at Université de Neuchâtel, partially supported by SNF grant 200020-144432/1, and continued at Tel Aviv University, partially supported by the Israel Science Foundation grant 1102/20 and by the ERC Starting Grant 757585.

2 Some toric symplectic geometry

In this section, we review toric geometry with special emphasis on a certain type of symplectic reduction, which we call *toric reduction*. Toric reduction generalizes probes as well as Delzant's construction of toric symplectic manifolds, both of which heavily feature in this paper.

2.1 Toric manifolds

A symplectic manifold (X^{2n}, ω) together with a moment map $\mu\colon X \to \mathfrak{t}^*$ is called *toric* if μ generates an effective Hamiltonian action of the n-torus T^n. By $\mathfrak{t}^*$ we denote the dual of the Lie algebra $\mathfrak{t}$ of T^n. Choosing an identification $T^n \cong \mathbb{R}^n/\mathbb{Z}^n$ induces an identification $\mathfrak{t}^* \cong \mathbb{R}^n$ and, depending on context, we will use both the invariant way and the coordinate-dependent way of seeing things. Note that some symplectic manifolds admit distinct toric structures and hence we are really concerned with the triple (X, ω, μ) when we say *toric manifold* although we may just write X or (X, ω) for simplicity.

A classical result by Delzant [18] states that if X is *compact* toric,[1] then the image $\Delta = \mu(X)$ is a so-called *Delzant polytope*, and that Delzant polytopes (up to integral affine transformations) classify toric manifolds up to equivariant symplectomorphism. There are many classical references for toric manifolds, eg [3; 11; 20], and we refer to these for details. We revisit part of Delzant's result in Section 2.3.

Due to Delzant's theorem, the moment polytope associated to a toric manifold X is a crucial object of study. We view it as

$$\Delta = \{x \in \mathfrak{t}^* \mid \ell_i(x) \geqslant 0\}, \quad \ell_i(x) = \langle x, \xi_i \rangle + \lambda_i. \tag{6}$$

Here, we view the vectors ξ_i in $\mathfrak{t}$ and $\langle \cdot, \cdot \rangle$ denotes the natural pairing of $\mathfrak{t}$ and its dual. Note that $\mathfrak{t}$ contains a natural lattice Λ obtained as the kernel of the exponential map $\exp\colon \mathfrak{t} \to T^n$. Similarly, the dual $\mathfrak{t}^*$ contains the dual lattice Λ^*. If we choose a basis, we can identify $\Lambda \cong \mathbb{Z}^n$ and dually $\Lambda^* \cong \mathbb{Z}^n$. Again, depending on context, we use both the invariant viewpoint and the coordinate-dependent one. Furthermore, since Δ is rational (with respect to Λ^*), we can choose the vectors ξ_i to be primitive in Λ.

Definition 2.1 A vector $v \in \Lambda$ in a lattice Λ is called *primitive* if $\alpha v \notin \Lambda$ for all $0 < \alpha < 1$.

Together with (6), this condition uniquely determines ξ_i and λ_i in terms of Δ and vice versa. As we have mentioned above, Δ is a *Delzant polytope*, meaning that at every vertex the vectors ξ_i determining the facets meeting at that vertex form a basis of the lattice Λ over the integers. There is a natural symmetry group acting on $\Delta \subset \mathfrak{t}^*$ without changing the toric manifold determined by Δ.

Definition 2.2 The *integral affine transformations* of $(\mathfrak{t}^*, \Lambda^*) \cong (\mathbb{R}^n, \mathbb{Z}^n)$ are the elements in the group

$$\operatorname{Aut} \Lambda^* \ltimes \mathfrak{t}^* \cong \mathrm{GL}(n; \mathbb{Z}) \ltimes \mathbb{R}^n. \tag{7}$$

[1] Many authors include compactness in the definition of *toric*, but we do not.

The elements in $\operatorname{Aut}\Lambda^* \cong \mathrm{GL}(n;\mathbb{Z})$ correspond to base changes in the torus T^n, whereas the translation part $\mathfrak{t}^* \cong \mathbb{R}^n$ corresponds to adding constant elements to the moment map. Neither of these transformations changes the Hamiltonian T^n-action.

2.2 Toric reduction

In this paragraph we are interested in symplectic reduction with respect to subtori of a toric T^n-action. We call symplectic reduction of this type *toric reduction*. The symplectic quotient of this operation inherits a toric structure with moment polytope obtained by intersecting Δ with an affine rational subspace in $\mathfrak{t}^*$. Roughly speaking, toric reductions are in bijection with inclusions (which are compatible in the sense of Definition 2.3) of the moment polytope of the reduced space into the moment polytope of the initial space. Although we could not find a precise statement of sufficient generality in the literature, this idea is hardly new — see for example [2]. In fact, as we will discuss in Section 2.3, the Delzant construction and McDuff's probes are special cases of Theorem 2.4. What may be new is the precise formulation we give in Definition 2.3 of the conditions for this reduction to yield a smooth symplectic quotient in terms of the geometry of Δ.

Let X be a toric manifold and Δ its moment polytope. Note that symplectic reduction with respect to the full T^n-action is pointless. Indeed, the reduced spaces are zero-dimensional. However, it is quite fruitful to perform symplectic reduction with respect to a subtorus $K \subset T^n$. Dually, we may look at affine rational subspaces $V \subset \mathbb{R}^n \cong \mathfrak{t}^*$. Indeed, to any affine rational subspace V we can associate its complementary torus

$$K_V = \exp(V^0), \quad V^0 = \{\xi \in \mathfrak{t} \mid \langle x - x', \xi\rangle = 0,\ x, x' \in V\} \subset \mathfrak{t}, \tag{8}$$

and vice versa. Rationality of V is equivalent to the compactness of K_V. The subspace V is a level set of the natural projection $\mathfrak{t}^* \to \mathrm{Lie}(K_V)^*$, meaning that the moment map $\mu_{K_V} : X \to \mathrm{Lie}(K_V)^*$ generating the induced K_V-action on X has level set $\mu^{-1}(\Delta \cap V)$ for some suitable level. Thinking in terms of V and instead of K_V or μ_{K_V} has the advantage that both the subtorus and the level at which we wish to carry out reduction are fixed by a choice of V. Furthermore, one can easily read off the integral affine geometry of the pair (Δ, V) whether the action of K_V on $\mu^{-1}(\Delta \cap V)$ is free (and hence the reduction admissible). Obviously, this is not always the case, since V may contain a vertex of Δ for example.

Definition 2.3 Let Δ be a Delzant polytope and let V be an affine rational subspace. We call the pair (Δ, V) *reduction-admissible* if, for every face $F \subset \Delta$ intersecting V, the union of (the linear part of) F and (the linear part of) V contains a basis of the lattice Λ^*.

Analogously we call a polytope $\Delta' \subset \Delta$ *reduction-admissible* if it is obtained as the intersection $\Delta' = \Delta \cap V$ of Δ with a reduction-admissible V. Note that one only needs to check reduction-admissibility at the faces F of the smallest dimension for which $V \cap F$ is nonempty, ie at the vertices of the polytope Δ'.

Theorem 2.4 (toric reduction) *Let $\Delta \subset \mathbb{R}^n$ be a Delzant polytope and $V \subset \mathbb{R}^n$ an affine rational subspace such that the pair (Δ, V) is reduction-admissible. Then the action of $K_V = \exp(V^0)$ on $Z = \mu^{-1}(\Delta \cap V)$ is free and the reduced space $X' = Z/K_V$ is itself toric with moment polytope $\Delta' = \Delta \cap V$.*

Proof Let $e_1^*, \dots, e_n^* \in \mathbb{R}^n = \mathfrak{t}^*$ be the standard basis. Reduction-admissibility implies that, up to applying an integral affine transformation, we may assume that

(9) $$V = \operatorname{span}_{\mathbb{R}}\{e_1^*, \dots, e_i^*\}, \quad F = \operatorname{span}_{\mathbb{R}}\{e_j^*, \dots, e_n^*\}, \quad j \leqslant i+1.$$

In this normal form, we have $V^0 = \operatorname{span}_{\mathbb{R}}\{e_{i+1}, \dots, e_n\}$ and hence $K_V = \{1\} \times T^{n-i}$. This subtorus acts freely on $\mu^{-1}(F)$. Since this holds for any facet F intersecting V, the action of K_V is free and thus symplectic reduction is admissible.

The quotient manifold carries a residual T^n/K_V-action. It is effective, since the T^n-action on X is. Since μ is invariant under the T^n-action, it is in particular invariant under the induced K_V-action and thus its restriction to $Z = \mu^{-1}(\Delta \cap V)$ factors through the quotient by K_V and has image $\Delta' = \Delta \cap V$. It is not hard to check that the map obtained in this way is a moment map generating the T^n/K_V-action on the quotient. For dimensional reasons, the resulting action is toric. □

Let $M = T^n/K_V$ be the torus acting by the residual action. Note that, by definition, Δ' is contained in $\mathfrak{t}^*$ instead of $\operatorname{Lie}(M)^* = \mathfrak{m}^*$. However, one can pick an identification of $(\mathfrak{m}^*, \Lambda_M^*)$ with $(V, \Lambda \cap V)$ and, up to an element in the integral affine transformations of $(\mathfrak{m}^*, \Lambda_M^*)$, this yields a well-defined polytope $\Delta' \subset \mathfrak{m}^*$. Conversely, given an integral affine embedding

(10) $$\iota\colon (\mathfrak{m}^*, \Lambda_M) \hookrightarrow (\mathfrak{t}^*, \Lambda), \quad \iota(\Delta') = \iota(\mathfrak{t}^*) \cap \Delta$$

such that $(\Delta, \iota(\mathfrak{m}^*))$ is reduction-admissible, there is a symplectic reduction from X to X'. To summarize, there is a short exact sequence of tori,

(11) $$0 \to K_V \hookrightarrow T^n \xrightarrow{\Xi} M \to 0,$$

where T^n acts on X and M acts on the reduced space X' such that the reduction map $p\colon Z \to X'$ is equivariant with respect to the T^n- and M-actions, meaning that

(12) $$p(t.x) = \Xi(t).p(x), \quad t \in T^N,\ x \in X.$$

In particular, orbits are mapped to orbits under toric reduction. This will be used in Section 2.4.

2.3 Delzant construction

The Delzant construction gives a recipe for constructing a toric manifold (X, ω, μ) from a compact Delzant polytope Δ. We review it here, since it will be used in Section 4, and refer to [20] for details. Actually, the Delzant construction is a special case of toric reduction as discussed in Section 2.2 where X is obtained as a symplectic quotient of some $\mathbb{C}^N$ equipped with its standard toric structure.

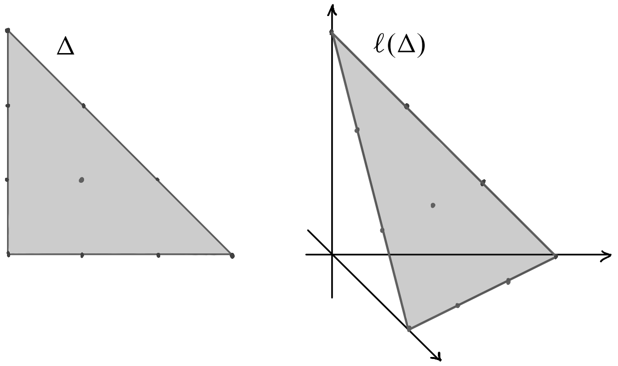


Figure 3: The idea of the Delzant construction in the case of $X = \mathbb{C}P^2$. The complement of $\operatorname{im}\ell$ generates the circle action by which the symplectic reduction is performed.

Let $\Delta \subset \mathfrak{t}^*$ be a Delzant polytope with N facets. Since Δ is compact, we have $N > n$. Let $(\mathbb{C}^N, \omega_0)$ be the standard symplectic vector space equipped with the moment map

$$\mu_0 \colon \mathbb{C}^N \to (\mathfrak{t}^N)^* \cong \mathbb{R}^N, \quad (z_1, \dots, z_N) \mapsto (\pi|z_1|^2, \dots, \pi|z_N|^2), \tag{13}$$

which generates the standard T^N-action on $\mathbb{C}^N$ by rotation in the factors. Its image is the positive orthant $\mathbb{R}^N_{\geqslant 0}$. Instead of starting with the subtorus $K \subset T^N$ by which to reduce, we start by defining an inclusion

$$\ell \colon \mathfrak{t}^* \hookrightarrow \mathbb{R}^N, \quad x \mapsto (\ell_1(x), \dots, \ell_N(x)), \tag{14}$$

which maps Δ to $\mathbb{R}^N_{\geqslant 0}$. The components ℓ_i defined in (6) are the functions measuring the integral affine distance of a given point to the facets of Δ. The map ℓ is an integral affine embedding as in (10) and the subtorus K by which we reduce is given $K = \exp(\operatorname{im}\ell)^0 \subset T^N$. Using the Delzant condition on Δ, it is easy to check that the inclusion $\ell(\Delta) \subset \mathbb{R}^N_{\geqslant 0}$ is admissible in the sense of Definition 2.3. Thus the toric symplectic manifold (X, ω, μ) is obtained as symplectic quotient $X = \mu_0^{-1}(\ell(\Delta))/K$.

Let us illustrate this by a simple example.

Example 2.5 (complex projective plane) Let $\Delta \subset \mathfrak{t}^* = \mathbb{R}^2$ be the simplex defined by

$$\ell_1(x) = x_1 + 1, \quad \ell_2(x) = x_2 + 1, \quad \ell_3(x) = -x_1 - x_2 + 1. \tag{15}$$

This simplex is Delzant and since $N = 3$, we will obtain X as a symplectic reduced space of $\mathbb{C}^3$. The map ℓ is depicted in Figure 3. The orthogonal complement $(\operatorname{im}\ell)^\perp$ is spanned by $(1,1,1)$ and thus $K = \{(t,t,t) \mid t \in S^1\} \subset T^3$ and $\mu_K(z) = \pi(|z_1|^2 + |z_2|^2 + |z_3|^2)$. We conclude that the symplectic reduction $\mu_K^{-1}(3) = S^5(3) \to \mathbb{C}P^2$ corresponds to the Hopf fibration map. The symplectic form one obtains by this procedure is the Fubini–Study form $\omega_{\mathbb{C}P^2}$ with normalization $\int_{\mathbb{C}P^1} \omega_{\mathbb{C}P^2} = 3$.

2.4 Toric fibres

Every toric manifold X^{2n} contains an n-parametric family of Lagrangian tori called *toric fibres*.

Definition 2.6 Let $x \in \operatorname{int} \Delta$ be a point in the interior of a toric moment polytope. The corresponding preimage $T(x) = \mu^{-1}(x)$ is called a *toric fibre*.

Example 2.7 (product tori) The toric fibres of the standard toric structure (13) are *product tori*

$$\mu_0^{-1}(a_1, \dots, a_N) = S^1(a_1) \times \cdots \times S^1(a_N) \subset \mathbb{C}^N,$$

where $a_i > 0$. Here, $S^1(a) \subset \mathbb{C}$ denotes the circle bounding a disk of area a.

Toric fibres are orbits with trivial stabilizer of the T^n-action. This means that the torus action gives a canonical identification $T^n \cong T(x)$ and

$$\Lambda = \ker(\exp\colon \mathfrak{t} \to T^n) = \pi_1(T(x)) = H_1(T(x); \mathbb{Z}). \tag{16}$$

Let us now discuss what happens to toric fibres under toric reduction. In general, let $p\colon Z \to X$ be the quotient map of a symplectic reduction. If $L \subset X$ is Lagrangian, then $p^{-1}(L)$ is Lagrangian as well and we call it the *lift* of L. Conversely, any Lagrangian contained in Z is automatically invariant under the group action and projects to a Lagrangian in the reduced space. Adopting our notation from Section 2.2, let X' be a quotient obtained from X by toric reduction and let $\iota(\Delta') \subset \Delta$ be the inclusion of the corresponding moment polytopes. Furthermore, we denote the reduction map by $p\colon Z \to X'$ and the toric fibres in X by $T(\cdot)$ and those in X' by $T'(\cdot)$.

Proposition 2.8 *In the above notation, we have the following correspondence of toric fibres in X and X',*

$$p^{-1}(T'(x)) = T(\iota(x)) \subset X, \quad x \in \operatorname{int} \Delta'. \tag{17}$$

Proof This follows directly from the definition of the moment map μ' on the quotient X'. □

In later sections, we will heavily use the second relative homotopy/homology groups of toric fibres, which is why we will discuss them here. Recall from (6) that the vectors $\xi_i \in \Lambda$ are defined as orthogonal vectors to the facets of Δ. We prove the following well-known fact using the Delzant construction together with Proposition 2.8.

Proposition 2.9 *Let $(X, T(x))$ be a pair of a toric symplectic manifold and a toric fibre. Then $\pi_2(X, T(x)) \cong \mathbb{Z}^N$, where N is the number of facets of Δ. Furthermore, there is a canonical basis $D_1, \dots, D_N \in \pi_2(X, T(x))$ bounding the classes $\partial D_i = \xi_i \in \Lambda = \pi_1(T(x))$.*

Proof By the Delzant construction and Proposition 2.8, the toric fibre $T(x)$ lifts to a product torus $T(\ell(x)) \subset \mathbb{C}^N$ under the reduction map $p\colon Z \to X$, where N is the number of facets of Δ. Let $\widetilde{D}_1, \dots, \widetilde{D}_N$ be the obvious basis of $\pi_2(\mathbb{C}^N, T(\ell(x)))$. Note that these can be chosen to lie in $Z \subset \mathbb{C}^N$ since the image of Z under the moment map μ_0 is equal to the image of the embedding ℓ from (14). Furthermore, reduction maps induce isomorphisms of relative homotopy groups; see for example the proof of [28, Proposition 3.2]. This shows that $\pi_2(\mathbb{C}^N, T(\ell(x)))$ and $\pi_2(X, T(x))$ are isomorphic, and we denote the image of $\widetilde{D}_i$ under the isomorphism by D_i. In order to compute the boundary operator ∂, consider the commutative diagram

$$\begin{tikzcd}
\pi_2(\mathbb{C}^N, T(\ell(x))) \arrow[r, "\partial'"] \arrow[d, "p_*"] & \pi_1(T(\ell(x))) \arrow[d, "p_*"] \\
\pi_2(X, T(x)) \arrow[r, "\partial"] & \pi_1(T(x))
\end{tikzcd} \tag{18}$$

The boundary operator ∂' is an isomorphism mapping $\widetilde{D}_i$ to the i^{th} standard basis vector e_i and therefore it suffices to understand p_* on the fundamental group. Recall from the discussion surrounding (11) that p is equivariant in the sense that $p(t.z) = \Xi(t).p(z)$ for all $t \in T^N$ and $z \in \mathbb{C}^N$. In the special case of the Delzant construction, one can easily check that $\Xi_*(e_i) = \xi_i$ and thus this proves the last claim. □

The homotopy long exact sequence for the pair $(X, T(x))$ gives a short exact sequence,

$$0 \to \pi_2(X) \to \pi_2(X, T(x)) \to \pi_1(T(x)) \to 0. \tag{19}$$

Indeed, the higher homotopy groups of the torus vanish and toric manifolds are simply connected. In homology (with integer coefficients) we obtain the same short exact sequence,

$$0 \to H_2(X) \to H_2(X, T(x)) \to H_1(T(x)) \to 0. \tag{20}$$

Indeed, the maps $H_*(T(x)) \to H_*(X)$ are zero, since there is a contractible subset $\Omega \subset X$ such that $T(x) \subset \Omega \subset X$. Take for example $\Omega = \mu^{-1}(\operatorname{int}\Delta \cup U)$, where U is a small neighbourhood of a vertex of Δ. There are obvious identifications of the respective groups in (19) and (20) which commute with the maps of these short exact sequences and thus we use homology and homotopy groups interchangeably.

Note that this discussion yields a very effective way to read off $\pi_2(X) = H_2(X)$ from the moment polytope of a toric manifold. It is the kernel of ∂, ie the lattice of integral relations among the vectors $\xi_1, \dots, \xi_N$ orthogonal to the facets of Δ. This in turn has a nice geometric interpretation in terms of the singular fibration structure of the moment map $\mu\colon X \to \Delta$. Indeed, when moving from the interior of the moment polytope to the interior of a facet F_i, the circle $S^1(\xi_i) \subset T^n$ collapses, where by $S^1(\xi_i)$ we have denoted the circle generated by the orthogonal vector $\xi \in \Lambda \subset \mathfrak{t}$ to the facet F_i. The canonical basis $D_1, \dots, D_N \in \pi_2(X, T(x))$ corresponds to the disks coming from these circles collapsing, which explains $\partial D_i = \xi_i$. Furthermore, let $\sum_i a_i D_i$ be an integral combination of such disks with $\sum_i a_i \xi_i = 0$. The latter condition means that the corresponding concatenation of curves representing the ξ_i bound in the fibre $T(x)$. Thus they define a homotopy class in $\pi_2(X)$, which illustrates $\ker\partial = \pi_2(X)$.

3 Symmetric probes

Symmetric probes were first defined in [1], where they were used to a different end. Let (X, ω, μ) be a toric symplectic manifold with moment polytope Δ.

Definition 3.1 A *symmetric probe* $\sigma \subset \Delta$ is a reduction-admissible line segment; see Definition 2.3.

Let us unpack this definition and introduce some notation. By $l \subset \mathfrak{t}^*$ we denote the line containing the symmetric probe σ, by $v \in \Lambda^*$ a primitive directional vector of l, and by F and F' the facets of Δ which σ intersects. We choose F and F' so that v points away from F and towards F'. See Figure 4 for an illustration of the set-up. Note that symmetric probes indeed do intersect facets, and not lower-dimensional faces. Definition 3.1 implies that there is a basis of Λ^* contained in the unions $l \cup F$ and $l \cup F'$, respectively. This means that, locally, all intersections of symmetric probes with a facet are equivalent under integral affine transformations. After choosing a basis, we can work in $(\mathbb{R}^n, \mathbb{Z}^n)$ and assume that

$$v = e_n^*, \quad F = \operatorname{span}_{\mathbb{R}}\{e_1^*, \dots, e_{n-1}^*\}. \tag{21}$$

This follows from the fact that $\mathrm{GL}(n;\mathbb{Z})$ acts transitively on the set of bases of $\mathbb{Z}^n$. We take (21) to be the normal form of an intersection of a symmetric probe with a facet. McDuff [23] calls these intersections *integrally transverse* and we refer to her paper for a detailed discussion of this notion. In the above notation we have $\langle v, \xi\rangle = -\langle v, \xi'\rangle = 1$, where $\xi, \xi' \in \Lambda$ are the normal vectors to F and F', respectively. By the normal form (21), it follows that we can assume $\xi = e_n$, which implies that $\xi' = \sum_{i=1}^{n-1} k_i e_i - e_n$. The numbers $k_1, \dots, k_{n-1} \in \mathbb{Z}$ completely determine the toric structure of a neighbourhood of the symmetric probe σ and they are topological invariants of the torus bundle coming from the reduction map $\mu^{-1}(\sigma) \to S^2$ appearing in the proof of Theorem 3.2.

Theorem 3.2 *Let $\sigma \subset \Delta$ be a symmetric probe and $x, y \in \sigma$ be a pair of points lying at equal distance to the boundary of σ. Then the toric fibres $T(x)$ and $T(y)$ are Hamiltonian isotopic by a Hamiltonian isotopy inducing the map*

$$\Phi_\sigma \colon H_1(T(x)) \to H_1(T(y)), \quad a \mapsto a + \langle v, a\rangle(\xi' - \xi) \tag{22}$$

on the first homology of the toric fibres.

In particular, this proves Theorem A. In (22), we have used the identification $\Lambda = H_1(T(x)) = H_1(T(y))$ induced by the torus action. The map Φ_σ is an involution and its $(+1)$-eigenspace is $(n-1)$-dimensional and given by the complement $\sigma^0 = v^0 \subset \mathfrak{t} = \Lambda \otimes \mathbb{R}$. Its (-1)-eigenspace is spanned by $\xi' - \xi$. Note also that Φ_σ is uniquely determined by $\sigma \subset \Delta$ and more precisely by an arbitrarily small neighbourhood of σ in Δ. Indeed, if we exchange ξ and ξ', then v changes its sign by our convention.

In case $x = y$, we obtain an interesting corollary about the Hamiltonian monodromy group (see Definition 1.4) of the corresponding toric fibre.

Corollary 3.3 *Let x be the midpoint of a symmetric probe σ. Then the Hamiltonian monodromy group $\mathcal{H}_{T(x)}$ contains the element Φ_σ.*

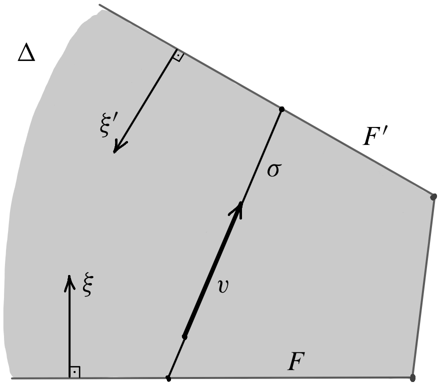


Figure 4: A symmetric probe $\sigma \subset \Delta$ and the surrounding notation.

Proof of Theorem 3.2 Since $\sigma \subset \Delta$ is reduction-admissible, we can perform toric reduction by Theorem 2.4. The reduced space is a copy of S^2 with a standard symplectic form of total area equal to the integral affine length of σ. Under the reduction, the fibres $T(x)$ and $T(y)$ are mapped to a pair of circles $S^1_x, S^1_y \subset S^2$ which are orbits of the residual Hamiltonian circle action on S^2; see Proposition 2.8. Since x and y are at equal distance to the boundary of σ, the circles S^1_x and S^1_y bound disks of the same area and thus can be exchanged by a Hamiltonian isotopy φ on S^2. Lift this Hamiltonian isotopy from S^2 to X by lifting its Hamiltonian function by the reduction map to $\mu^{-1}(\sigma)$ and extending it (for example by cut-off) to the total space. See for example [2, Lemma 3.1] or [6, Lemma 3.1] for details on lifting Hamiltonian isotopies.

Let us now compute the map induced by φ on Λ. We work with homotopy groups here, but the problem is exactly the same in homology by the discussion in Section 2.4. Let $d_x, d'_x \in \pi_2(S^2, S^1_x)$ and $d_y, d'_y \in \pi_2(S^2, S^1_y)$ be the generators of relative homotopy groups such that d_x and d_y contain the south pole, d'_x and d'_y contain the north pole, and $d_x + d'_x = d_y + d'_y = [S^2]$ for a chosen orientation on S^2. The map φ_* induced by the Hamiltonian isotopy φ on relative homotopy groups satisfies $\varphi_* d_x = d'_y$ and $\varphi_* d_y = d'_x$. Furthermore, the map Φ_σ is uniquely determined by the properties

$$\Phi_\sigma(\xi) = \xi', \quad \Phi_\sigma|_{v^0} = \mathrm{id}_{v^0}, \tag{23}$$

where $v^0 \subset \Lambda$ denotes the elements on which $v \in \Lambda^*$ vanishes. Indeed, ξ is transverse to v^0 since $\langle v, \xi \rangle = 1$. We show that the lift of φ satisfies (23), which proves the claim. The second property in (23) follows from the K_σ-equivariance of the lift of φ where $K_\sigma = \exp \sigma^0$ is the complementary torus of the probe σ. Indeed, $K_\sigma \subset T^n$ is the subtorus with respect to which the symplectic reduction $\mu^{-1}(\sigma) \to S^2$ is carried out — see also (8) and the proof of Theorem 2.4 — and thus any Hamiltonian isotopy lifted from the reduced space is equivariant with respect to this group action. For the first property in (23), note that the map $p_* \colon \pi_2(\mu^{-1}(\sigma), T(x)) \to \pi_2(S^2, S^1_x)$ induced by symplectic reduction is an isomorphism. See for example the proof of [28, Proposition 3.2]. Therefore $\pi_2(\mu^{-1}(\sigma), T(x_\sigma))$ is generated by D_x and D'_x with $\pi_*(D_x) = d_x$ and $\pi_*(D'_x) = d'_x$, and similarly for $T(y)$. This allows us to conclude that the lift of φ maps D_x to D'_y and D_y to D'_x. Since $\partial_x D_x = \partial_y D_y = \xi$ and $\partial_y D_y = \partial_y D'_y = \xi'$, this finishes the proof. □

Note that we have actually computed the map induced on relative second homology,

$$H_2(X, T(x)) \to H_2(X, T(y)), \quad b \mapsto b + \langle v, \partial b\rangle (D' - D), \tag{24}$$

where D and D' denote the homology classes of the canonical basis in Proposition 2.9 corresponding to F and F', respectively. Note also that the lift of φ in the proof of Theorem 3.2 depends on the extension of the Hamiltonian function to X and is thus not uniquely defined by φ.

Remark 3.4 By choosing a suitable cut-off of the lifted Hamiltonian function in the proof of Theorem 3.2, one can choose the Hamiltonian isotopy to be supported in an arbitrarily small neighbourhood of $\sigma \subset \Delta$.

4 Chekanov invariants

The main idea of this section is to use the Delzant construction to lift toric fibres of certain toric manifolds to product tori in some $\mathbb{C}^N$ via Proposition 2.8 and to make use of the various results on product tori in [12]. In particular, this yields strong obstructions to the equivalence of toric fibres (Theorem B) and their Hamiltonian monodromy (Theorem C). As we shall discuss, similar results can be obtained *by hand* (ie avoiding the lifting trick) via displacement energy and versal deformations, which comes in handy in case X cannot be seen as a toric reduction of $\mathbb{C}^N$. However, we note that the approach by hand runs into the question of determining the displacement energy of toric fibres, which turns out to be very subtle in general; see for example the papers [1; 23] for detailed discussions of the (qualitative) question of displaceability and [6, Section 3] for the quantitative question about displacement energy. In case X can be seen as a toric reduction of $\mathbb{C}^N$, this question can be completely avoided by the lifting trick.

Definition 4.1 A toric symplectic manifold X is called *of reduction type* if it can be obtained as a toric reduction of some $\mathbb{C}^N$.

By the Delzant construction in Section 2.3, all compact toric manifolds are of reduction type. The space $X = \mathbb{C} \times S^2$, which will be discussed in Section 5.3, is an example of a noncompact space which is of reduction type.

Before moving to the Chekanov invariants, let us point out the following.

Remark 4.2 (classification up to symplectomorphisms) We focus on equivalence of toric fibres up to Hamiltonian diffeomorphisms. One may ask an analogue of Question 1.1 for the group of symplectomorphisms. Note that a toric symplectic manifold X is simply connected whenever its moment polytope has at least one vertex, meaning that the distinction between the two classification questions is, at best, a question about connected components of $\mathrm{Symp}(X, \omega)$. In fact, both classifications agree in all simply connected examples we consider in Section 5 of this paper. This is not always true, as the following example illustrates. Let X be the space obtained from $\mathbb{C}P^2$ by three small toric blow-ups of the same size $\varepsilon > 0$ at the vertices of the original moment triangle. The resulting symplectic manifold is toric and its moment polytope is a hexagon with three long and three short edges. Near each of the short edges, there is a nondisplaceable toric fibre, as was proved in [16, Section 5.5]. In particular, these three nondisplaceable fibres are not equivalent under Hamiltonian diffeomorphisms. However they are symplectomorphic. Indeed, their base points can be permuted by integral affine symmetries of the moment polytope and such symmetries lift to symplectomorphisms of the corresponding toric manifold; see for example [8, Lemma 4.3] for a proof of this well-known fact.

In particular, Conjecture 1.3 is false for equivalence up to symplectomorphisms.

4.1 Equivalence of toric fibres

As we have seen in Example 2.7, the product tori

$$T(a) = T(a_1, \dots, a_N) = S^1(a_1) \times \dots \times S^1(a_N) \subset \mathbb{C}^N \tag{25}$$

are a special case of toric fibres. Chekanov has given a classification of product tori[2] up to symplectomorphism in [12, Theorem A]. A complete set of invariants is given by

$$d(a) = \min\{a_1, \dots, a_N\}, \tag{26}$$

$$\#_d(a) = \#\{i \in \{1, \dots, N\} \mid a_i = d(a)\}, \tag{27}$$

$$\Gamma(a) = \mathbb{Z}\langle a_1 - d(a), \dots, a_N - d(a)\rangle, \tag{28}$$

where we write $a = (a_1, \dots, a_N) \in \mathbb{R}^N_{>0}$. The first invariant is a positive real number and corresponds to the displacement energy[3] $d(a) = e(\mathbb{C}^N, T(a))$. The second invariant is a positive integer less than or equal to N (with equality if $T(a)$ is monotone) which comes from versal deformations and displacement energy. As it turns out, versal deformations of product tori are given as the minimum of $\#_d(a)$ linear functionals, and they contain no other information beyond this number. The third invariant $\Gamma(a) \subset \mathbb{R}$ is a subgroup of $\mathbb{R}$ generated by $N - \#_d(a)$ elements and is a purely *soft* invariant. In fact, it is the set of symplectic areas of disks with vanishing Maslov class $m(\cdot)$. Note that in the case of $\mathbb{C}^N$, the symplectic form has a primitive λ and thus we can express $\Gamma(a)$ as

$$\Gamma(a) = \left\{ \int_\gamma \lambda \in \mathbb{R} \;\middle|\; \gamma \in H_1(T(a)),\, m(\gamma) = 0 \right\}. \tag{29}$$

This invariant can be more explicitly expressed as $\Gamma(a) = \mathbb{Z}\langle a_1 - d(a), \dots, a_n - d(a)\rangle$.

Theorem 4.3 (Chekanov) *The product tori $T(a)$ and $T(a')$ are symplectomorphic in $\mathbb{C}^N$ if and only if*

$$d(a) = d(a'), \quad \#_d(a) = \#_d(a'), \quad \Gamma(a) = \Gamma(a'). \tag{30}$$

[2] Chekanov calls these tori *elementary tori*.

[3] In the original paper, Chekanov uses the first Ekeland–Hofer capacity instead.

Let us now get back to the case of toric fibres and prove Theorem B. Recall from Section 1.2 that the *Chekanov invariants* of a toric fibre $T(x) \subset X$ are defined in terms of the integral affine distances $\ell(x) = (\ell_1(x), \dots, \ell_N(x))$ of the point x to the facets of Δ,

$$d(x) = \min\{\ell_1(x), \dots, \ell_N(x)\}, \tag{31}$$

$$\#_d(x) = \#\{i \in \{1, \dots, N\} \mid \ell_i(x) = d(x)\}, \tag{32}$$

$$\Gamma(x) = \mathbb{Z}\langle \ell_1(x) - d(x), \dots, \ell_N(x) - d(x)\rangle. \tag{33}$$

Proof of Theorem B We prove the result for all toric manifolds of reduction type; see Definition 4.1. Let X be a toric manifold of reduction type and $T(x), T(x') \subset X$ be toric fibres which are equivalent under Hamiltonian isotopies. Recall from Section 2.3 that we may view X as a toric reduction of $\mathbb{C}^N$, where the inclusion map of the moment polytope Δ of X into $\mathbb{R}^N$ is given by the map $\ell(x) = (\ell_1(x), \dots, \ell_N(x))$. Furthermore, the toric fibre $T(x)$ lifts to the product torus $T(\ell(x))$ in $\mathbb{C}^N$ by Proposition 2.8 and similarly for $T(x')$. Since $T(x) \cong T(x')$, we obtain that $T(\ell(x)) \cong T(\ell(x'))$. Indeed, Hamiltonian isotopies can be lifted through symplectic reductions by lifting the corresponding Hamiltonian function and extending it to $\mathbb{C}^N$ by cut-off. It is easy to see that any such lift will map the lift of $T(x)$ to the lift of $T(x')$. Theorem B now follows from Theorem 4.3. □

The Chekanov invariants are not complete, as the following example illustrates.

Example 4.4 Let $\mathbb{C}P^2$ the complex projective plane equipped with the toric structure described in Example 2.5 and with moment polytope Δ, and set

$$x = \left(-\frac{5}{10}, -\frac{2}{10}\right), \quad x' = \left(-\frac{5}{10}, \frac{1}{10}\right) \in \Delta. \tag{34}$$

Since $\ell(x) = (1 + x_1, 1 + x_2, 1 - x_1 - x_2)$, we obtain

$$\ell(x) = \left(\frac{5}{10}, \frac{8}{10}, \frac{17}{10}\right), \quad \ell(x') = \left(\frac{5}{10}, \frac{11}{10}, \frac{14}{10}\right) \in \mathbb{R}^3_{\geqslant 0}. \tag{35}$$

By the classification of toric fibres in $\mathbb{C}P^2$ from [27, Proposition 7.1] — see also Section 5.2 — the fibres $T(x)$ and $T(x')$ are not Hamiltonian isotopic. However, their Chekanov invariants agree. Indeed, we find

$$d(x) = d(x') = \frac{1}{2}, \quad \#_d(x) = \#_d(x') = 1, \quad \Gamma(x) = \Gamma(x') = \mathbb{Z}\left\langle \frac{3}{10} \right\rangle. \tag{36}$$

4.2 Hamiltonian monodromy

Let $\phi \in \mathrm{Ham}(X, \omega)$ be a Hamiltonian diffeomorphism of a toric manifold (X, ω) mapping a toric fibre $T(x)$ to a toric fibre $T(x')$. Then one can consider the map induced on relative second homology,

$$\phi_* \colon H_2(X, T(x)) \to H_2(X, T(x')). \tag{37}$$

We call this map *ambient monodromy*. In the same vein as in Section 4.1, we derive obstructions to which maps ϕ_* can be obtained in this way by using the Delzant construction to lift Hamiltonian isotopies. Note that by setting $x = x'$ and by projecting to the first homology (see (20)), we can extract information about the Hamiltonian monodromy question as a special case.

The key result by Chekanov is [12, Theorem 4.5].

Theorem 4.5 (Chekanov) *Let $T(a), T(a') \subset \mathbb{C}^N$ be product tori. An isomorphism*

$$\Phi\colon H_1(T(a)) \to H_1(T(a')) \tag{38}$$

can be realized as $(\phi|_{T(a)})_* = \Phi$ by a symplectomorphism $\phi \in \mathrm{Symp}(\mathbb{C}^N, \omega_0)$ mapping $T(a)$ to $T(a')$ if and only if the following conditions hold:

$$\Phi(\mathscr{D}(a)) = \mathscr{D}(a'), \quad \Phi^* m_{T(a')} = m_{T(a)}, \quad \Phi^* \sigma_{T(a')} = \sigma_{T(a)}. \tag{39}$$

Here $m_{T(a)} \in H^1(T(a);\mathbb{Z})$ and $\sigma_{T(a)} \in H^1(T(a);\mathbb{R})$ are the Maslov class and the symplectic area class, respectively. By $\mathscr{D}(a) \subset H_1(T(a))$ we denote the set of *distinguished classes*. In the standard basis $e_1, \dots, e_N \in H_1(T(a))$ the basis vector e_i is called a *distinguished class* if the corresponding component in $a = (a_1, \dots, a_N)$ is minimal, ie if $a_i = d(a)$.

Let us now move to toric fibres. Recall from Proposition 2.8 that for any toric fibre $T(x) \subset X$, the relative second homology $H_2(X, T(x))$ has a canonical basis $D_1, \dots, D_N$, where D_i corresponds to the i^{th} facet of the moment polytope Δ of X.

Definition 4.6 Let $T(x) \subset X$ be a toric fibre. The *distinguished classes* of $T(x)$ are the elements of the set

$$\mathscr{D}(x) = \{D_i \mid \ell_i(x) = d(x)\} \subset H_2(X, T(x)), \tag{40}$$

ie elements of the canonical basis for which the distance of $x \in \operatorname{int}\Delta$ to the corresponding facet of Δ is minimal.

Recall that there is a canonical inclusion $H_2(X) \subset H_2(X, T(x))$, meaning that there is a distinguished subspace which is independent of the choice of x. We prove the following.

Theorem 4.7 *Let $T(x), T(x') \subset X$ be toric fibres in a compact toric manifold X such that there exists a Hamiltonian diffeomorphism $\phi \in \mathrm{Ham}(X, \omega)$ mapping $T(x)$ to $T(x')$. Then the induced map*

$$\phi_*\colon H_2(X, T(x)) \to H_2(X, T(x')) \tag{41}$$

on relative homology groups satisfies

$$\phi_*(\mathscr{D}(x)) = \mathscr{D}(x'), \quad \phi^* m_{T(x')} = m_{T(x)}, \quad \phi^* \sigma_{T(x')} = \sigma_{T(x)}, \quad \phi_*|_{H_2(X)} = \mathrm{id}. \tag{42}$$

Proof The second and third identity in (42) are general facts about the Maslov and the symplectic area class. The last identity is straightforward since Hamiltonian diffeomorphisms are isotopic to the identity on X and hence the map $\phi_*\colon H_*(X) \to H_*(X)$ is the identity. For the first identity in (42), we again use the Delzant construction together with lifting the Hamiltonian isotopy. The following groups are canonically isomorphic:

$$H_2(X, T(x)) \cong H_2(\mathbb{C}^N, T(\ell(x))) \cong H_1(T(\ell(x))), \tag{43}$$

See the proof of Proposition 2.9, where this is proved for the corresponding (relative) homotopy groups. Thus the map $H_1(T(\ell(x))) \to H_1(T(\ell(x')))$ induced by the lifted Hamiltonian diffeomorphism is conjugate to ϕ_* by the canonical isomorphism (43). It is easy to see that the distinguished classes $\mathcal{D}(x)$ of $T(x)$ are by definition mapped under (43) to the distinguished classes $\mathcal{D}(\ell(x))$ of the product torus $T(\ell(x))$ and thus the first identity in (42) follows from Theorem 4.5. □

It seems reasonable to guess that these constraints are sufficient. More precisely, note that there is a canonical identification $H_2(X, T(x)) = H_2(X, T(x'))$ for any two points $x, x' \in \operatorname{int}\Delta$. Then we conjecture the following.

Conjecture 4.8 *An isomorphism $\Phi \in \operatorname{Aut} H_2(X, T(x))$ can be realized as ambient monodromy of a Hamiltonian diffeomorphism mapping $T(x)$ to $T(x')$ if and only if the identities in* (42) *hold.*

We show that this conjecture holds in all examples discussed in Section 5. In fact, we use the ambient monodromy and Theorem 4.7 to classify toric fibres and determine the Hamiltonian monodromy groups in these examples. The area class $\sigma_{T(x)}$ determines x and hence proving this conjecture gives, in particular, an answer to Question 1.1.

Let us now move to the ordinary Hamiltonian monodromy group of toric fibres, see Definition 1.4. To derive information about $\mathcal{H}_{T(x)}$ from Theorem 4.7, fix $x = x'$ and let $\phi \in \operatorname{Ham}(X, \omega)$ be a Hamiltonian isotopy such that $\phi(T(x)) = T(x)$. Note that the ambient monodromy ϕ_* determines the map $(\phi|_{T(x)})_* \in \operatorname{Aut} H_1(T(x))$ by the short exact sequence (20).

Proof of Theorem C Any element in the Hamiltonian monodromy group $\mathcal{H}_{T(x)}$ comes from an ambient monodromy element ϕ_* by (20) and hence the theorem follows directly from Theorem 4.7 where the set of distinguished classes in $H_1(T(x))$ is given by

$$\partial\mathcal{D}(x) = \{\xi_i \mid \ell_i(x) = d(x)\} \subset H_1(T(x)), \tag{44}$$

where $\xi_i \in \mathfrak{t} \cong H_1(T(x))$ is a primitive defining vector of the i^{th} facet of Δ. Indeed, recall from Proposition 2.9 that the boundary of a canonical basis element D_i is ξ_i. □

It follows from Theorem C that if the distinguished classes span the lattice $H_1(T(x))$, then $\mathcal{H}_{T(x)}$ is a subgroup of the group of permutations on $\#_d(x)$ elements. In particular, the Hamiltonian monodromy group is finite in this case. See also [4, Theorem 1]. In contrast, we shall see that the Hamiltonian monodromy group is infinite in some examples; see Sections 5.3, 5.4 and 5.5. The number $\#_d(x)$ is maximal if $T(x)$ is the monotone toric fibre of a (monotone) toric manifold X. In that case, we obtain the obstructive statement of [4, Theorem 2] for the group of Hamiltonian diffeomorphisms as a special case of Theorem 4.7.

Corollary 4.9 *Let $T(x) \subset X$ be a monotone toric fibre. Then any element in $\mathcal{H}_{T(x)}$ acts as a permutation on the set $\{\xi_1, \dots, \xi_N\}$ of defining vectors of the polytope and the corresponding ambient monodromy acts as the identity on $H_2(X)$.*

4.3 Displacement energy and versal deformations of toric fibres

In this subsection, we discuss obstructions for the equivalence of toric fibres and their Hamiltonian monodromy relying on versal deformations instead of the lifting trick employed in the proofs of Theorems B and 4.7. This comes in handy in cases where X cannot be seen as a toric reduction of some $\mathbb{C}^N$, and we will use them in Sections 5.4 and 5.5. Note that the direct approach by versal deformations has the drawback that it requires a computation of the displacement energy of toric fibres, at least on an open dense subset. See Assumption 4.12.

Let us briefly discuss displacement energy and versal deformations. We refer to [12] and especially [14; 15] for more details. The displacement energy of a compact subset $A \subset (X, \omega)$ is defined as the infimum of the Hofer norm taken over all Hamiltonian isotopies displacing A from itself,

$$e(X, A) = \inf\{\|H\| \mid \phi_1^H(A) \cap A = \varnothing\}, \tag{45}$$

and by convention $e(X, A) = \infty$ if the infimum is taken over the empty set. The displacement energy is a symplectic invariant and we will use it only if A is a Lagrangian.

For a compact Lagrangian $L \subset X$, Chekanov introduced a way to strengthen a given symplectic invariant by looking at the invariant on Lagrangian neighbours of L. This is called *versal deformation* of L. Perturbing L in a Weinstein neighbourhood, we find that nearby Lagrangians correspond to graphs of closed one-forms on L. Furthermore, we can associate to every such perturbation an element in $H^1(L; \mathbb{R})$, by taking its (Lagrangian) flux. Two such perturbations are Hamiltonian isotopic (with support in the Weinstein neighbourhood of L) if and only if they map to the same element in $H^1(L; \mathbb{R})$. Thus we obtain a continuous bijection between locally supported Hamiltonian isotopy classes of Lagrangian neighbours of L and a neighbourhood of the origin of $H^1(L; \mathbb{R})$. As the flux description suggests, this correspondence is independent of the chosen Weinstein neighbourhood.

We may postcompose any symplectic invariant with the map from $U \subset H^1(L; \mathbb{R})$ to classes of nearby Lagrangians. Here, we use displacement energy to obtain a function $U \to \mathbb{R} \cup \{\infty\}$. By taking its germ, we obtain

$$\mathscr{E}_L \colon H^1(L; \mathbb{R}) \to \mathbb{R} \cup \{\infty\}. \tag{46}$$

Definition 4.10 We call the function (46) the *displacement energy germ* of $L \subset X$.

The displacement energy germ is a symplectic invariant in the sense that if $\phi \in \mathrm{Symp}(X, \omega)$, then

$$\mathscr{E}_L \circ \phi|_L^* = \mathscr{E}_{\phi(L)}, \tag{47}$$

where $\phi|_L^*$ is the transpose of the isomorphism $(\phi|_L)_*\colon H_1(L) \to H_1(\phi(L))$. In particular, this can be used to derive obstructions to Hamiltonian monodromy.

Proposition 4.11 *Let $L \subset X$ be a compact Lagrangian submanifold. If $\Phi \in \mathscr{H}_L$ is an element in the Hamiltonian monodromy group, then $\mathscr{E}_L \circ \Phi^* = \mathscr{E}_L$.*

Let us discuss this in more detail in the special case where $L = T(x) \subset X$ is a toric fibre of a toric manifold (X, ω). Coming up with a versal deformation of toric fibres is straightforward. Indeed, a versal deformation of $T(x)$ is obtained by varying the base point, $a \mapsto T(x + a)$ for small enough a, where we identify $H^1(T(x); \mathbb{R}) \cong \mathfrak{t}^*$ as usual via the T^n-action. Thus the crucial point in computing $\mathscr{E}_{T(x)}$ is finding the displacement energy of toric fibres $e(X, T(x))$ as a function of $x \in \mathrm{int}(T(x))$. Let us make the following assumption.

Assumption 4.12 On an open and dense subset of the moment polytope Δ, we assume that

$$e(X, T(x)) = d(x) = \min\{\ell_1(x), \dots, \ell_N(x)\}. \tag{48}$$

Here, $d(\cdot)$ denotes the integral affine distance to the boundary of Δ as in (14). Recall that the functionals $\ell_i(\cdot) = \langle \cdot, \xi_i \rangle + \lambda_i$ measure the integral affine distance of x to the i^{th} facet of Δ. Let f and g be two functions defined on a vector space V. Since equalities on open and dense subsets will come up quite often and are in fact sufficient for our purposes, we write $f \simeq g$ if f and g agree on an open and dense subset of V.

Let us briefly discuss why Assumption 4.12 is reasonable. First, we note that $e(X, T(x)) \geqslant d(x)$ whenever X is compact toric, and more generally, whenever X can be seen as the toric reduction of some $\mathbb{C}^N$. This follows again from toric reduction and the lifting trick; see also [6, Section 3.2]. Indeed, if $T(x) \subset X$ can be displaced with energy e, then so can the corresponding product torus $T(\ell(x)) \subset \mathbb{C}^N$ obtained by Proposition 2.8. The displacement energy of the latter is precisely given by $d(x) = \min\{\ell_1(x), \dots, \ell_N(x)\}$. Although this inequality may fail to be sharp (for example for nondisplaceable tori), in all the examples we know of, it fails only on the complement of an open dense subset, meaning that Assumption 4.12 still holds. Furthermore, the assumption holds for all compact *monotone* toric symplectic manifolds of dimension $\leqslant 18$ as was checked computationally. The monotone case in arbitrary dimension is related to the so-called Ewald conjecture. See [23] or [6, Section 3.4] for a detailed discussion. The following proposition is [6, Proposition 4.3].

Proposition 4.13 *Under Assumption 4.12, the displacement energy germ of $T(x)$ is given by*

$$\mathscr{E}_{T(x)}(a) \simeq \min_{i \in I(x)} \{\ell_i(x + a)\}, \tag{49}$$

where $I(x) \subset \{1, \dots, N\}$ is the subset of indices for which $\ell_i(x)$ is minimal.

Under Assumption 4.12, we can prove the symplectically hard part of Theorem B and a weaker form of the hard part of Theorem 4.7, where ambient monodromy is replaced by the map induced on first homology.

Theorem 4.14 *Let X be a toric manifold for which Assumption 4.12 holds. Let ϕ be a Hamiltonian diffeomorphism mapping a toric fibre $T(x)$ to a toric fibre $T(x')$. Then we have*

$$d(x) = d(x'), \quad \#_d(x) = \#_d(x'). \tag{50}$$

Furthermore, the map $(\phi|_{T(x)})_* \colon H_1(T(x)) \to H_1(T(x'))$ acts by a permutation on distinguished classes, $(\phi|_{T(x)})_* \mathscr{D}(x) = \mathscr{D}(x')$.

Proof Let $U \subset \operatorname{int} \Delta$ be an open dense subset such that (48) holds for all $x \in U$. For $x, x' \in U$, we have $d(x) = d(x')$. If $x \notin U$ or $x' \notin U$, use Proposition 4.13 to see that

$$\min_{i \in I(x)} \{\ell_i(x+a)\} \simeq \min_{i \in I(x')} \{\ell_i(x'+a)\}.$$

Thus these two continuous functions of a are actually equal near $a = 0$, and they yield $d(x)$ and $d(x')$, respectively, when evaluated at $a = 0$. The second invariance property in (50) similarly follows from (47) and Proposition 4.13 by noting that $\#_d(x) = \# I(x)$. The claim about $(\phi|_{T(x)})_*$ follows from (47) and Proposition 4.13. Indeed, recall that the distinguished classes of $T(x)$ are the vectors ξ_i for which the corresponding ℓ_i is minimal; see (44). □

To illustrate that the methods of this paragraph can be applied to a broader set of examples than toric fibres, we include the following example.

Example 4.15 (Vianna tori in $\mathbb{C}P^2$) Using Proposition 4.11, one can show that all Vianna tori in $\mathbb{C}P^2$, except for the first and the second one, have trivial Hamiltonian monodromy groups. The Vianna tori in $\mathbb{C}P^2$ form a countable family of monotone Lagrangian tori which are not pairwise symplectomorphic. They are in bijection with so-called *Markov triples*, ie triples of natural numbers solving the *Markov equation*. We refer to [30] for a detailed description. We denote the Vianna torus corresponding to a Markov triple (a,b,c) by $T(a,b,c) \subset \mathbb{C}P^2$. This torus appears as a monotone fibre of an *almost toric fibration* of $\mathbb{C}P^2$ with base diagram given by a certain triangle $\Delta_{a,b,c}$. On an open and dense subset of a neighbourhood of the origin in $H^1(T(a,b,c);\mathbb{R})$, the displacement energy germ $\mathscr{E}_{T(a,b,c)}$ has level sets given by scalings of $\partial \Delta_{a,b,c}$. This means that the versal deformation *sees* the corresponding almost toric base diagram, and thus the integral affine equivalence class of $\Delta_{a,b,c}$ is an invariant of $T(a,b,c)$. In particular, this can be used to distinguish the Vianna tori as was noted by Chekanov and Schlenk in private communications. For a proof of this claim, see the forthcoming paper [9].

Using Proposition 4.11, we note that a necessary condition for $T(a,b,c)$ to admit nontrivial monodromy is that the corresponding almost toric base diagram admits some integral affine symmetry. Such a symmetry can only exist if at least two vertices are of the same integral affine type, ie if the same Markov number appears at least twice in the same triple. This is only the case for $(1,1,1)$ and $(1,1,2)$. The former is the Clifford torus which has Hamiltonian monodromy group isomorphic to the dihedral group D_6 and the latter is the first nontrivial Vianna torus $T(1,1,2)$ having monodromy group isomorphic to $\mathbb{Z}_2$. For all other Vianna tori, we obtain $\mathscr{H}_{T(a,b,c)} = \{1\}$. In particular, the Hamiltonian monodromy group does not contain enough information to distinguish Vianna tori.

5 Examples

In Sections 5.1–5.5, we classify toric fibres and determine their Hamiltonian monodromy in some examples. With the exception of $\mathbb{C}^2 \times T^*S^1$, our examples are four-dimensional. This comes from the fact that the classification question in dimensions $\geqslant 6$ is qualitatively very different — provided the moment polytope has at least one vertex. Indeed, in that case, there are toric fibres $T(x)$ for which $\mathfrak{H}_x$ has accumulation points; see Corollary 5.16.

The proofs of the results of this section all follow the same pattern. Equivalences and monodromy elements are constructed by symmetric probes. The main ingredients for the obstructive side are Theorems B and 4.7 applied to the ambient monodromy map $\phi_*\colon H_2(X, T(x)) \to H_2(X, T(x'))$ induced by a Hamiltonian diffeomorphism ϕ. The conceptual reason why constraints on ambient monodromy give constraints on equivalences of toric fibres is the observation that the symplectic area class of a toric fibre determines $x \in \Delta$. These methods probably apply to most four-dimensional toric manifolds, with the computational complexity increasing with the number of edges of the moment polytope. The examples we chose are diverse in the sense that $S^2 \times S^2$ and $\mathbb{C}P^2$ are compact toric and thus the Delzant construction can be used directly; $\mathbb{C} \times S^2$ is noncompact, but still a toric reduction of $\mathbb{C}^3$; the spaces $\mathbb{C}^2 \times T^*S^1$ and $T^*S^1 \times S^2$ are noncompact and cannot be seen as toric reductions of any $\mathbb{C}^N$. However, the latter is a toric reduction of the former. In the case of the former, we apply the direct methods from Section 4.3. Note also that the spaces $\mathbb{C}^2 \times T^*S^1$ and $T^*S^1 \times S^2$ are not simply connected, whence the classification up to symplectomorphisms is drastically different from the classification up to Hamiltonian diffeomorphisms.

In Section 5.6, we revisit Chekanov's classification result and prove Conjecture 1.3 for $\mathbb{C}^n$. In Section 5.7, we collect some remarks on how to construct symmetric probes in arbitrary toric manifolds.

Let us point out that all monodromy results for *monotone* toric fibres in this section also follow from the methods developed in [4].

5.1 The case of monotone $X = S^2 \times S^2$

Let $S^2 \times S^2$ be equipped with the monotone product symplectic structure $\omega = \omega_{S^2} \oplus \omega_{S^2}$, where ω_{S^2} is the area form with normalization $\int_{S^2} \omega_{S^2} = 2$. Then the corresponding moment polytope is given by the square $\Delta = [-1,1] \times [-1,1]$. There are probes with four different directional vectors. The probes with $v = e_1^*, e_2^*$ are admissible everywhere in the interior of the polytope. The probes with $v = e_1^* + e_2^*$

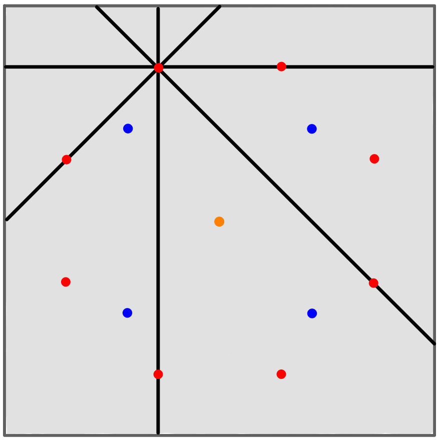


Figure 5: Some symmetric probes in the monotone $S^2 \times S^2$. Points of the same colour denote equivalent fibres.

and $v = e_1^* - e_2^*$ are admissible everywhere except for the two main diagonals of the square. Note that the equivalences of toric fibres generated by these probes can also be read off from the symmetries of $\Delta = [-1,1] \times [-1,1]$. Let us turn to the classification of toric fibres.

Proposition 5.1 *The classification of toric fibres of monotone $S^2 \times S^2$ is given by*

$$\mathfrak{H}_x = \{(\pm x_1, \pm x_2), (\pm x_2, \pm x_1)\}, \quad x = (x_1, x_2) \in \operatorname{int} \Delta. \tag{51}$$

Note that the sets $\mathfrak{H}_x$ contain eight elements if $x_1 \neq x_2$ and both are nonzero, four elements if $x_1 = x_2$ or if one of the x_i is zero, and one element (the monotone fibre) if $x_1 = x_2 = 0$. See Figure 5.

Proof The constructive side follows either from the symmetric probes listed above or from the symmetries of Δ. For the obstructions, we will use Chekanov's invariants as expressed in Theorems B and 4.7. Let $T(x) = T(x_1, x_2)$. By Chekanov's first invariant from Theorem B, we can restrict our attention to the set $\mathfrak{D}_x$ lying at distance $d(x)$ to the boundary of Δ. This set is the boundary of a square of size $2(1 - d(x))$ and it is stratified by the second Chekanov invariant. Indeed, we have $\#_d(x) = 2$ whenever x is a vertex of $\mathfrak{D}_x$ and $\#_d(x) = 1$ elsewhere on $\mathfrak{D}_x$. This means that we are left with proving the result on the interior of the four edges of $\mathfrak{D}_x$. Using the symmetries of $\mathfrak{D}_x \subset \Delta$, we can restrict our attention to the segment $[0, x_2) \times \{x_2\}$ since this is a fundamental domain for $\mathfrak{D}_x$ under the symmetries.

Claim *If $T(x_1, x_2) \cong T(x_1', x_2)$ for some $x_1, x_1' \in [0, x_2)$, then $x_1 = x_1'$.*

We use Theorem 4.7 to prove the claim. Suppose there is $\phi \in \operatorname{Ham}(X, \omega)$ mapping $T(x_1, x_2)$ to $T(x_1', x_2)$. This induces

$$\phi_*\colon H_2(X, T(x_1, x_2)) \to H_2(X, T(x_1', x_2)). \tag{52}$$

Let D_1, D_2, D_3 and D_4 be the canonical basis of $H_2(X, T(x_1, x_2))$ as in Proposition 2.9 where D_1 is the disk corresponding to the facet $\{1\} \times [-1, 1]$ and the remaining ones are ordered in the anticlockwise direction. Let D_1', D_2', D_3' and D_4' be the corresponding basis elements for $H_2(X, T(x_1', x_2))$. The distinguished classes are $\mathfrak{D}(x_1, x_2) = \{D_2\}$ and $\mathfrak{D}(x_1', x_2) = \{D_2'\}$, meaning that Theorem 4.7 yields $\phi_* D_2 = D_2'$. Set

$$\phi_* D_1 = a_1 D_1' + a_2 D_2' + a_3 D_3' + a_4 D_4', \quad a_i \in \mathbb{Z}. \tag{53}$$

Since the Maslov class is preserved, we obtain $a_1 + a_2 + a_3 + a_4 = 1$, meaning that $a_4 = 1 - a_1 - a_2 - a_3$. Since the induced map $(\phi|_{T(x_1,x_2)})_*$ is invertible, we deduce that $\det(\partial\phi_* D_1, \partial\phi_* D_2) = \pm 1$ which yields $a_1 - a_3 = \pm 1$. Preservation of symplectic area, $\int_{D_1} \omega = \int_{\phi_* D_1} \omega$, yields

$$1 - x_1 = a_1(1 - x_1') + a_2(1 - x_2) + a_3(1 + x_1') + a_4(1 + x_2) \tag{54}$$

since the areas of the disks D_i' are just given by the distances of (x_1', x_2) to the respective facets of Δ. In case $a_1 - a_3 = +1$, we use the above relations on the a_i to find $a_3 = a_1 - 1$ and $a_4 = 2 - 2a_1 - a_2$, and thus

$$x_1 - x_1' = 2x_2(a_1 + a_2 - 1) \in 2x_2\mathbb{Z}. \tag{55}$$

Since $|x_1 - x_1'| < x_2$, we conclude $x_1 = x_1'$. In case $a_1 - a_3 = -1$, we find by the same reasoning,

$$x_1 + x_1' = 2x_2(a_1 + a_2) \in 2x_2\mathbb{Z}. \tag{56}$$

Since $0 \leqslant x_1 + x_1' < 2x_2$, we deduce $x_1' = -x_1$ and hence $x_1 = x_1' = 0$ □

Proposition 5.2 *Let $0 \leqslant x_1 \leqslant x_2$. Then the Hamiltonian monodromy group of the toric fibre*

$$T(x_1, x_2) \subset S^2 \times S^2$$

in the monotone $S^2 \times S^2$ is given by

$$\mathcal{H}_{T(x_1,x_2)} = \begin{cases} \left\langle \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \right\rangle \cong \mathbb{Z}_2 & \text{if } x_1 = 0 \text{ and } x_2 \neq 0, \\ \left\langle \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\rangle \cong \mathbb{Z}_2 & \text{if } x_1 = x_2 \neq 0, \\ \left\langle \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \right\rangle \cong \mathbb{Z}_2 \times \mathbb{Z}_2 & \text{if } x_1 = x_2 = 0, \end{cases} \tag{57}$$

and by $\mathcal{H}_{T(x_1,x_2)} = \{1\}$ in all other cases.

Note that any other toric fibre is Hamiltonian isotopic to a fibre with $0 \leqslant x_1 \leqslant x_2$, meaning that its Hamiltonian monodromy group is conjugate to one of the above. Thus the only isomorphism types of groups which appear are $\mathbb{Z}_2$, $\mathbb{Z}_2 \times \mathbb{Z}_2$ and the trivial group. The astute reader may have wondered why $\mathcal{H}_{(0,0)}$ is not the full symmetry group of $\Delta = [-1, 1] \times [-1, 1]$. This comes from the fact that some of these symmetries act nontrivially on $H_2(S^2 \times S^2)$ (by exchanging the obvious generators) and thus they can be realized by symplectomorphisms, but not by Hamiltonian diffeomorphisms. We refer to [4], where the monodromy group generated by symplectomorphisms is determined for monotone toric fibres.

Proof Again, the construction side can be obtained by symmetric probes and Theorem 3.2. For the obstruction side, let $T(x_1,x_2)$ be a toric fibre of $S^2 \times S^2$ and $\phi \in \mathrm{Ham}(S^2 \times S^2)$ a Hamiltonian diffeomorphism mapping this fibre to itself. We again analyze the map

$$\phi_* \in \mathrm{Aut}\, H_2(X, T(x_1,x_2))$$

and use the fact that ϕ_* determines $(\phi|_{T(x_1,x_2)})_*$. Let us start with the case of the monotone fibre $T(0,0) \subset S^2 \times S^2$, which is also a special case of [4, Theorem 2]. In this case, the distinguished classes are $\mathfrak{D}(0,0) = \{D_1, D_2, D_3, D_4\}$. Therefore Theorem 4.7 implies that the ambient monodromy is a permutation of these classes. Since $D_1 + D_3, D_2 + D_4 \in H_2(X)$, these two classes must be preserved under ϕ_* which implies the claim. In the case $x_1 = x_2 \neq 0$, the distinguished classes are $\mathfrak{D}(x_1,x_1) = \{D_1, D_2\}$, and hence only permutations of D_1 and D_2 are permitted by Theorem 4.7. Now let $0 \leqslant x_1 < x_2$. Then the set of distinguished classes is $\mathfrak{D}(x_1,x_2) = \{D_2\}$, and hence the ambient monodromy map takes D_2 to D_2. We set $x_1 = x_1'$ in (55) and (56). In the first case, we find that $a_1 + a_2 = 1$ and a computation using the expressions for a_3 and a_4 from the proof of Proposition 5.1, this yields that the monodromy is trivial. In the second case, we find that $x_1 = 0$ and $a_1 + a_2 = 0$ and a similar computation shows that the monodromy maps $e_1 \mapsto -e_1$ in that case. We conclude that the monodromy group is trivial whenever $x_1 \neq 0$ and that it is generated by the map $e_1 \mapsto -e_1$ and $e_2 \mapsto e_2$ if $x_1 = 0$. □

Remark 5.3 If $S^2 \times S^2$ is equipped with a nonmonotone symplectic form, the classification as well as the Hamiltonian monodromy is drastically different. Indeed, some equivalence classes $\mathfrak{H}_x$ of fibres have accumulation points in Δ and some fibres have infinite Hamiltonian monodromy groups. We refer to [7] for details.

5.2 The case of $X = \mathbb{C}P^2$

Let $\mathbb{C}P^2$ be equipped with the symplectic form and moment polytope as in Example 2.5. We give the classification of toric fibres and the Hamiltonian monodromy groups without proof since the proofs are the same as for $S^2 \times S^2$. The classification of toric fibres was first given in [27, Proposition 7.1]. Note that all equivalences and Hamiltonian monodromies in the case of the monotone $S^2 \times S^2$ are induced by symmetries of the moment polytope. The same holds in the case of $\mathbb{C}P^2$.

Proposition 5.4 *Toric fibres $T(x), T(y) \subset \mathbb{C}P^2$ are equivalent if and only if x can be mapped to y by an integral symmetry of Δ. Similarly, the Hamiltonian monodromy group $\mathcal{H}_{T(x)}$ consists of transformations induced by integral symmetries of Δ fixing the point x.*

5.3 The case of $X = \mathbb{C} \times S^2$

Let $\mathbb{C} \times S^2$ be equipped with the symplectic form $\omega = \omega_{\mathbb{C}} \oplus \omega_{S^2}$. We normalize the moment map such that its moment polytope is given by $\Delta = \mathbb{R}_{\geqslant -1} \times [-1,1]$. There are symmetric probes with directional vector $e_2^* + k e_1^*$ for every $k \in \mathbb{Z}$. The probe with $k = 0$ is admissible everywhere. For $k = \pm 1$, the probes

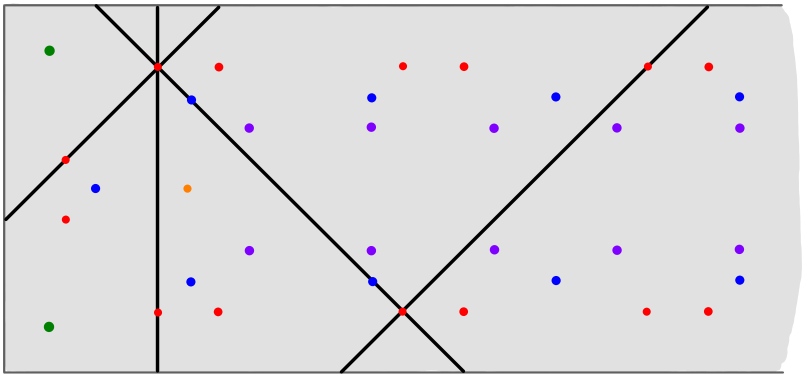


Figure 6: Some symmetric probes in $\mathbb{C} \times S^2$. Points of the same colour denote equivalent fibres.

are admissible everywhere except when they hit a vertex of Δ. The symmetric probes with $k \notin \{-1,0,1\}$ are admissible whenever they hit both facets $\mathbb{R}_{\geqslant -1} \times \{1\}$ and $\mathbb{R}_{\geqslant -1} \times \{-1\}$. As we shall see the latter types of symmetric probes are obsolete as all results can be proven using only those with $k \in \{-1,0,1\}$.

Proposition 5.5 *The classification of fibres in (X,ω) is as follows. For $x = (x_1,0) \in \operatorname{int} \Delta$ with $x_1 \geqslant 0$, we have $\mathfrak{H}_x = \{x\}$, ie the corresponding toric fibre is not equivalent to any other fibres. For $x = (x_1, \pm x_1)$ with $x_1 < 0$, we have $\mathfrak{H}_x = \{(x_1,x_1),(x_1,-x_1)\}$. For $x = (0,x_2)$ with $x_2 > 0$, we have*

$$\mathfrak{H}_x = \{(2nx_2, \pm x_2) \mid n \in \mathbb{N}\} \cup \{(-x_2,0)\}. \tag{58}$$

For $x = (x_1, \pm x_1)$ with $x_1 > 0$, we have

$$\mathfrak{H}_x = \{((2n+1)x_1, \pm x_1) \mid n \in \mathbb{N}\}. \tag{59}$$

For $x = (x_1,x_2)$ with $0 < x_1 < x_2$, we have

$$\mathfrak{H}_x = \{(\pm x_1 + 2nx_2, \pm x_2) \mid n \in \mathbb{N}\} \cup \{(-x_2, \pm x_1)\}. \tag{60}$$

All $y \in \operatorname{int} \Delta$ are in one of the above $\mathfrak{H}_x$.

Proof For the construction of the equivalences, we use concatenations of the probes with directional vectors $e_2^* + ke_1^*$ for $k \in \{-1,0,1\}$, as we discuss below on a case by case basis. For the obstruction side, note that we cannot directly apply Theorems B and 4.7, since X is noncompact. However, X is of reduction type; see Definition 4.1. Indeed, the toric reduction of $\mathbb{C}^2$ to S^2 coming from the Delzant construction yields a toric reduction of $\mathbb{C} \times \mathbb{C}^2$ to $\mathbb{C} \times S^2$. Therefore, the results of Theorems B and 4.7 still apply.

By the first Chekanov invariant from Theorem B, the polytope Δ decomposes into subsets $\mathfrak{D}_x$ of constant distance $0 < d(x) \leqslant 1$ to the boundary $\partial\Delta$. First, let $0 < d(x) < 1$. The toric fibres of the type $(x_1, \pm x_1)$ with $x_1 < 0$ are the only ones having $\#_d = 2$, which distinguishes them from all others. Note that for any other $x \in \Delta$ with $d(x) < 1$, the torus $T(x)$ is equivalent by symmetric probes to exactly one torus on the segment $[0,x_2] \times \{x_2\}$ with $x_2 = 1 - d(x)$. It is easy to see that by probes with directional vectors $e_2 + e_1$

and $e_2 - e_1$, any fibre is equivalent to one on the segment $[-x_2, x_2] \times \{x_2\}$. Now note that fibres on this segment come in equivalent pairs as can be seen by noting that (x_1, x_2) is equivalent to $(-x_2, -x_1)$ which is equivalent (by a vertical probe) to $(-x_2, x_1)$ which in turn is equivalent to $(-x_1, x_2)$. Thus the problem of classifying fibres with $d(x) < 1$ boils down to classifying fibres on the segment $[0, x_2] \times \{x_2\}$.

Claim *If $T(x_1, x_2) \cong T(x_1', x_2)$ for $x_1, x_1' \in [0, x_2]$ and $x_2 = 1 - d(x)$, then $x_1 = x_1'$.*

To prove the claim, we follow the same strategy as in the proof of Proposition 5.1. Suppose there is $\phi \in \mathrm{Ham}(X, \omega)$ mapping $T(x_1, x_2)$ to $T(x_1', x_2)$ and let ϕ_* be the ambient monodromy induced by this map. Let D_1, D_2 and D_3 be the canonical basis of $H_2(X, T(x_1, x_2))$ where D_1 is the disk corresponding to the facet $\{-1\} \times [-1, 1]$ and the remaining ones ordered in the anticlockwise direction. Let $D_1', D_2', D_3' \in H_2(X, T(x_1', x_2))$ be the disks obtained by the same convention. The distinguished classes are $\mathfrak{D}(x_1, x_2) = \{D_3\}$ and $\mathfrak{D}(x_1', x_2) = \{D_3'\}$ meaning that $\phi_* D_3 = D_3'$. Set

$$\phi_* D_1 = a_1 D_1' + a_2 D_2' + a_3 D_3', \quad a_i \in \mathbb{Z}. \tag{61}$$

By the invariance of the Maslov class, we obtain $a_1 + a_2 + a_3 = 1$. Since the induced map $(\phi|_{T(x_1,x_2)})_*$ is invertible, we deduce that $\det(\partial \phi_* D_1, \partial \phi_* D_3) = \pm 1$ which yields $a_1 = \pm 1$. Preservation of symplectic area, $\int_{D_1} \omega = \int_{\phi_* D_1} \omega$, yields

$$1 + x_1 = a_1(1 + x_1') + a_2(1 + x_2) + a_3(1 - x_2). \tag{62}$$

In the case $a_1 = -1$, we find

$$x_1 + x_1' = 2x_2(a_2 - 1), \tag{63}$$

from which we deduce that $x_1 = x_1'$. Similarly, for $a_1 = 1$, we find

$$x_1 - x_1' = 2x_2 a_2, \tag{64}$$

which implies the same conclusion and thus proves the claim.

Let us now turn to the case $d(x) = 1$, ie tori of the form $T(x_1, 0)$ with $x_1 \geqslant 0$. Note that $T(0, 0)$ is the only monotone fibre and thus not equivalent to any other fibre. We will show that the same remains true for $x_1 > 0$. Indeed, if $T(x_1, 0)$ and $T(x_1', 0)$ were equivalent, then the same arguments as above apply to the ambient monodromy ϕ_* except that now $\mathfrak{D}(x_1, 0) = \{D_2, D_3\}$. Equations (63) and (64) for $x_2 = 0$ imply the claim. Equivalently, one can use [19, Theorem 1.1] to find that $e(X, T(x_1, 0)) = 1 + x_1$, meaning that these fibres are also distinguished by their displacement energy. □

Proposition 5.6 *Let $(x_1, x_2) \in \operatorname{int} \Delta$. The Hamiltonian monodromy group of the toric fibre*

$$T(x_1, x_2) \subset \mathbb{C} \times S^2$$

is given by

$$\mathcal{H}_{T(x_1,x_2)} = \begin{cases} \left\langle \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \right\rangle \cong \mathbb{Z}_2 & \text{if } x_1 = -x_2 \text{ and } x_2 > 0, \\ \left\langle \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right\rangle \cong \mathbb{Z}_2 & \text{if } x_1 \leqslant 0 \text{ and } x_2 = 0, \\ \left\langle \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} \right\rangle \cong \mathbb{Z}_2 \ltimes \mathbb{Z} & \text{if } x_1 > 0 \text{ and } x_2 = 0, \end{cases} \tag{65}$$

and is given by a group conjugated to the above if $T(x_1,x_2)$ is equivalent to one of the above cases according to Proposition 5.5, and by $\mathcal{H}_{T(x_1,x_2)}=\{1\}$ in all other cases.

Proof If $x_1=-x_2$ and $x_2>0$, we have $\mathfrak{D}(x_1,x_2)=\{D_1,D_3\}$ and thus any monodromy matrix has to permute $\partial D_1=e_1$ and $\partial D_3=-e_2$. This permutation is realized by a symmetric probe with directional vector $e_1^*+e_2^*$. For the case $x_1\leqslant 0$ and $x_2=0$, note that the vertical symmetric probe realizes the claimed monodromy. In terms of obstructions, note that $\mathfrak{D}(0,0)=\{D_1,D_2,D_3\}$, which yields the claim for $T(0,0)$. In case $x_1<0$, note that $T(x_1,0)$ is Hamiltonian isotopic to $T(0,x_1)$ to which we can apply (63) and (64) to find that the only possible monodromy for $T(0,x_1)$ is $e_1\mapsto -e_1$ and $e_2\mapsto e_2$. Under the conjugation induced by the equivalence of $T(x_1,0)$ and $T(0,x_1)$, this yields the answer. Now let $x_1\in(0,x_2]$ and $x_2>0$. Then the Hamiltonian monodromy group is trivial by (63) and (64). Now let us turn to the case of the infinite monodromy groups for the fibres $T(x_1,0)$ with $x_1>0$. The two generators given in (65) correspond to the vertical probe and the probe with directional vector $-e_1^*+e_2^*$. Since $\mathfrak{D}(x_1,0)=\{D_2,D_3\}$, we distinguish the cases $D_2\mapsto D_2, D_3\mapsto D_3$ and $D_2\mapsto D_3, D_3\mapsto D_2$. Let us first restrict our attention to the former case. We use (61) with $D_i=D_i'$. Recall that $a_1=\pm 1$. Since (63) cannot be satisfied for $x_1=x_1'$ and $x_2=0$, we deduce that $a_1=1$. A computation shows that

$$(\phi|_{T(x_1,0)})_*e_1=\partial\phi_*D_1=e_1+2a_2e_2, \tag{66}$$

which proves the claim in case $\det(\phi|_{T(x_1,0)})_*=1$ (this corresponds to the powers of the product of the two generators given in (65)). The case $D_2\mapsto D_3$ and $D_3\mapsto D_2$ is completely analogous. □

5.4 The case of $X=\mathbb{C}^2\times T^*S^1$

Let $X=\mathbb{C}^2\times T^*S^1$ be equipped with the exact symplectic form

$$\omega=d\lambda=\omega_{\mathbb{C}^2}\oplus\omega_{T^*S^1}=d\lambda_{\mathbb{C}^2}\oplus d\lambda_{T^*S^1}$$

and the product toric structure with moment polytope $\Delta=\mathbb{R}^2_{\geqslant 0}\times\mathbb{R}$. Note that X is not a toric reduction of $\mathbb{C}^N$ and hence we cannot apply the techniques from Sections 4.1 and 4.2 as in the previous examples. Instead, we rely on Section 4.3, meaning that we need to compute the displacement energy of toric fibres.

Lemma 5.7 *The displacement energy of toric fibres is given by*

$$e(X,T(x_1,x_2,x_3))=\min\{x_1,x_2\}. \tag{67}$$

In particular, equality (48) (*and thus also Assumption 4.12*) *holds for all toric fibres in X.*

Proof The upper bound is obvious, either by using probes or the fact that $e(\mathbb{C}^2,T(x_1,x_2))=\min\{x_1,x_2\}$. For the lower bound, we use Chekanov's theorem [13], which we briefly recall here. Let $L\subset X$ be a compact Lagrangian submanifold of a tame symplectic manifold and let $J\in\mathcal{J}(X,\omega)$ be a tame almost complex structure. Furthermore, denote by $\sigma(X,L;J)$ the infimum of symplectic areas over all nonconstant J-holomorphic disks with boundary on L. If this set is empty, set $\sigma(X,L;J)=\infty$. If the set is not empty, we obtain a strictly positive value which is attained by Gromov compactness. The quantity $\sigma(X;J)$ is defined similarly for J-holomorphic spheres in X. Then Chekanov's theorem gives the lower bound

$$e(X,L) \geqslant \min\{\sigma(X;J),\sigma(X,L;J)\}. \tag{68}$$

Now let $X=\mathbb{C}^2\times T^*S^1$ and $L=T(x_1,x_2,x_3)$ a toric fibre. Note that X is aspherical, and thus $\sigma(X;J)=\infty$. Let $J_0\in\mathcal{J}(X,\omega)$ be the complex structure obtained from the identification $X=\mathbb{C}^2\times\mathbb{C}^\times$. There are two obvious families of J_0-holomorphic disks,

$$u_{\alpha_1,\alpha_2}(z)=\left(z,\sqrt{\frac{x_2}{\pi}}e^{i\alpha_1},e^{x_3+i\alpha_2}\right), \tag{69}$$

$$v_{\alpha_1,\alpha_2}(z)=\left(\sqrt{\frac{x_1}{\pi}}e^{i\alpha_1},z,e^{x_3+i\alpha_2}\right),\quad \alpha_1,\alpha_2\in S^1. \tag{70}$$

These disks have area $\int u^*_{\alpha_1,\alpha_2}\omega=x_2$ and $\int v^*_{\alpha_1,\alpha_2}\omega=x_1$ for all $\alpha\in S^1$. We show that the minimal one among these two disks realizes the minimum $\sigma(X,L;J_0)$. For a similar argument, see [5, Lemma 4]. Now let $u\colon(D,\partial D)\to(X,L)$ be a nontrivial J_0-holomorphic disk. First note that the map $p_2\circ u$, where $p_2\colon X\to T^*S^1\cong\mathbb{C}^*$ is the projection, is constant. This follows from the maximum principle. Indeed, by the maximum principle this map takes values in the unit disk. Since 0 is not contained in its image, we can precompose it with $z\mapsto\frac{1}{z}$, to see that it actually takes values in the unit circle. Since it is holomorphic, it is actually constant. By considering $p_1\circ u$, where $p_1\colon(X,J_0)\to(\mathbb{C}^2,i\oplus i)$ is the natural projection, it is sufficient to understand holomorphic disks in $\mathbb{C}^2$ with boundary on the product torus $T(x_1,x_2)$. The group $\pi_2(\mathbb{C}^2,T(x_1,x_2))$ is generated by the two coordinate disks, D_1 and D_2. We have $[p_1\circ u]=k_1D_1+k_2D_2$, where k_1 and k_2 are the algebraic intersection numbers with coordinate axes. By positivity of intersections, we deduce $k_1,k_2\geqslant 0$. Since $\int_{D_1}\omega_{\mathbb{C}^2}=x_1$ and $\int_{D_2}\omega_{\mathbb{C}^2}=x_2$, we obtain

$$\sigma(X,L;J_0)\geqslant\min\{k_1x_1+k_2x_2\mid k_1,k_2\geqslant 0,k_1k_2\neq 0\}=\min\{x_1,x_2\}. \tag{71}$$

This minimum is realized by the disks u_{α_1,α_2} or v_{α_1,α_2}. □

Using Theorem 4.14, we can now classify toric fibres and determine their Hamiltonian monodromy groups.

Proposition 5.8 *The classification of toric fibres $T(x)=T(x_1,x_2,x_3)$ in $X=\mathbb{C}^2\times T^*S^1$ is given by*

$$\mathfrak{H}_x=\{(x_1,x_2,x_3+k(x_2-x_1)),(x_2,x_1,x_3+k(x_2-x_1))\mid k\in\mathbb{Z}\}. \tag{72}$$

Furthermore, all Hamiltonian monodromy groups are trivial except when $x_1=x_2$, in which case

$$\mathcal{H}_{T(x_1,x_1,x_3)}=\left\langle\begin{pmatrix}1&0&1\\0&1&-1\\0&0&1\end{pmatrix},\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}\right\rangle. \tag{73}$$

Proof The equivalences are easy to construct using probes with direction $e_1^* - e_2^* + ke_3^*$ with $k \in \mathbb{Z}$. Let ϕ be a Hamiltonian diffeomorphism mapping $T(x)$ to $T(x')$. First note that the long exact sequence for relative homology looks quite different than in the compact toric case. Indeed, we obtain

$$0 \to H_2(X, T(x)) \to H_1(T(x)) \to H_1(X) \to 0, \tag{74}$$

and $H_1(T(x)) \cong H_2(X, T(x)) \oplus H_1(X) \cong \mathbb{Z}^2 \oplus \mathbb{Z}$. Then the induced map $(\phi|_{T(x)})_*$ on the first homology is of the form

$$(\phi|_{T(x)})_* = \begin{pmatrix} a_1 & a_2 & b_1 \\ a_3 & a_4 & b_2 \\ 0 & 0 & 1 \end{pmatrix}, \quad a_i, b_j \in \mathbb{Z}. \tag{75}$$

This follows from the fact that ϕ induces the identity on $H_1(X)$. First suppose that $x_1 = x_2$. Then, by Theorem 4.14 and Lemma 5.7, we obtain $x_1' = x_2' = x_1$ and $(\phi|_{T(x)})_*$ either swaps the first two coordinates or acts by the identity on them. By preservation of the Maslov index, we obtain $b_2 = -b_1$. Note that all monodromies of these tori can be realized by symmetric probes of direction $e_1^* - e_2^* + ke_3^*$, proving (73). To prove that $x_3 = x_3'$, compute

$$x_3 = \int_{e_3} \lambda|_{T(x)} = \int_{(\phi|_{T(x)})_* e_3} \lambda|_{T(x')} = b_1(x_1' - x_2') + x_3, \tag{76}$$

proving the claim. In the case $x_1 \neq x_2$, suppose without loss of generality that $x_1 < x_2$ and $x_1' < x_2'$. By Theorem 4.14 and Lemma 5.7, we obtain $x_1 = x_1'$ and $a_1 = 1$, $a_3 = 0$ and $a_4 = 1$, $a_2 = 0$ or $a_4 = -1$, $a_2 = -1$. The latter case is actually impossible. Indeed, if $a_4 = -1$, $a_2 = -1$, then

$$x_2 = \int_{e_2} \lambda|_{T(x)} = \int_{(\phi|_{T(x)})_* e_2} \lambda|_{T(x')} = 2x_1' - x_2', \tag{77}$$

contradicting $x_1 < x_2, x_1' < x_2'$. The rest of the proof is as in the case $x_1 = x_2$ and the claim $b_1 = 0$ follows from (76). □

5.5 The case of $X = T^*S^1 \times S^2$

Let $X = T^*S^1 \times S^2$ be equipped with the product symplectic form $\omega = \omega_{T^*S^1} \oplus \omega_{S^2}$. The moment polytope is $\Delta = \mathbb{R} \times [-1, 1]$. Note that X is not a toric reduction of any $\mathbb{C}^N$, but it is a toric reduction of $\mathbb{C}^2 \times T^*S^1$, meaning that we can use Proposition 5.8 together with the lifting trick.

Proposition 5.9 *The classification of toric fibres $T(x) = T(x_1, x_2)$ in $X = T^*S^1 \times S^2$ is given by*

$$\mathfrak{H}_x = \{(x_1 + 2kx_2, \pm x_2) \mid k \in \mathbb{Z}\}. \tag{78}$$

Furthermore, all Hamiltonian monodromy groups are trivial, except for $x_2 = 0$, in which case,

$$\mathcal{H}_{T(x_1,0)} = \left\{ \begin{pmatrix} 1 & 0 \\ 2k & \pm 1 \end{pmatrix} \middle| k \in \mathbb{Z} \right\}. \tag{79}$$

Proof Again, all constructions immediately follow from symmetric probes. For the obstructions, we view X as a toric reduction of $\mathbb{C}^2 \times T^*S^1$. Perform toric reduction on $\mathbb{C}^2 \times T^*S^1$ with respect to the plane $V = \{x_1 + x_2 = 1\} \subset \mathbb{R}^3$ to obtain X. This corresponds to the Hamiltonian $H = \pi|z_1|^2 + \pi|z_2|^2$. The classification of toric fibres follows immediately from Propositions 2.8 and 5.8 and the lifting trick, as in the proof of Theorem B. The obstructions to monodromy similarly follow from Proposition 5.8 and the lifting trick, as in the proof of Theorem 4.7. □

5.6 Chekanov's classification revisited

In this subsection, we prove Conjecture 1.3 in the case of $\mathbb{C}^n$. The classification of product tori goes back to Chekanov — see Theorem 4.3 — meaning that we only need to prove that all equivalences of toric fibres can be realized by iterated symmetric probes.

Theorem 5.10 *Product tori $T(x), T(y) \subset \mathbb{C}^n$ are Hamiltonian isotopic if and only if they are equivalent by a sequence of symmetric probes.*

Before proving this result, let us revisit Chekanov's classification. In $\mathbb{C}^2$, it states that

$$\mathfrak{H}_x = \{(x_1, x_2), (x_2, x_1)\}, \quad (x_1, x_2) \in \mathbb{R}^2_{>0}. \tag{80}$$

In $\mathbb{C}^n$ with $n \geqslant 3$, however, the situation is much richer. Note for example that all tori $T(1, 2, k)$ with $k \in \mathbb{N}_{\geqslant 2}$ are Hamiltonian isotopic, since their Chekanov invariants agree. The set $\mathfrak{H}_x$ even has accumulation points in many cases; see Corollary 5.16. To discuss this further, we slightly reformulate Chekanov's invariants. Since coordinate permutations can be realized by Hamiltonian isotopies, we may assume that $T(x)$ is given under the form

$$T(x) = T(\underbrace{\underline{x}, \dots, \underline{x}}_{\#_d(x)}, \underline{x} + \hat{x}_1, \dots, \underline{x} + \hat{x}_s), \tag{81}$$

for $\hat{x}_i > 0$ and $s = n - d(x)$. We call $\hat{x} = (\hat{x}_1, \dots, \hat{x}_s) \in \mathbb{R}^s_{>0}$ the *reduced vector* associated to x.

Theorem 4.3 can be reformulated as follows.

Corollary 5.11 *Product tori $T(x), T(y) \subset \mathbb{C}^n$ are Hamiltonian isotopic if and only if $d(x) = d(y)$, $\#_d(x) = \#_d(y)$ and there is $M \in \mathrm{GL}(s; \mathbb{Z})$ such that $M\hat{x} = \hat{y}$.*

Proof Note that the $\hat{x}_i$ are exactly the nontrivial generators of the lattice $\Gamma(x)$,

$$\Gamma(x) = \mathbb{Z}\langle\hat{x}\rangle = \{k_1\hat{x}_1 + \dots + k_s\hat{x}_s \mid k_i \in \mathbb{Z}\} \subset \mathbb{R}. \tag{82}$$

Furthermore, $\mathbb{Z}\langle\hat{x}\rangle$ is a complete invariant for $\mathrm{GL}(s; \mathbb{Z})$-orbits. See for example Cabrer and Mundici [10, Proposition 1]. □

This allows us to gain a good qualitative understanding of $\mathfrak{H}_x$.

Corollary 5.12 *The inclusion*

$$\mathfrak{H}_x \subset \{y \in \mathbb{R}^n_{>0} \mid d(y) = d(x),\ \#_d(y) = \#_d(x)\} \tag{83}$$

is dense if and only if $\operatorname{rank}\Gamma(x) \geqslant 2$.

Proof Let $\hat{x} \in \mathbb{R}^s_{>0}$ be the reduced vector as in (81). It follows from Corollary 5.11 that the inclusion (83) is dense if and only if the $\mathrm{GL}(s;\mathbb{Z})$-orbit of $\hat{x}$ is dense in $\mathbb{R}^s_{>0}$. The latter is equivalent to $\operatorname{rank}\Gamma(x) \geqslant 2$ by a classical theorem of Dani [17, Theorem 17]; see also [10]. □

Let us now have a look at the discrete case, ie the case where $\operatorname{rank}(\Gamma(x)) = 1$. In that case, the reduced vector $\hat{x} \in \mathbb{R}^s_{>0}$ is a real multiple of a lattice vector, $\hat{x} = \ell_{\mathrm{int}}(\hat{x})k$ with $k \in \mathbb{Z}^s$ a primitive vector and where $\ell_{\mathrm{int}}(\hat{x}) > 0$ denotes the integral affine length.

Corollary 5.13 *The product tori* $T(x), T(y) \subset \mathbb{C}^n$ *with* $d(x) = d(y)$, $\#_d(x) = \#_d(y)$ *and*

$$\operatorname{rank}(\Gamma(x)) = \operatorname{rank}(\Gamma(y)) = 1$$

are Hamiltonian isotopic if and only if their reduced vectors have the same integral affine length, $\ell_{\mathrm{int}}(\hat{x}) = \ell_{\mathrm{int}}(\hat{y})$.

Proof Write $\hat{x} = \ell_{\mathrm{int}}(\hat{x})k$ and $\hat{y} = \ell_{\mathrm{int}}(\hat{y})k'$, and note that $\mathrm{GL}(s;\mathbb{Z})$ acts transitively on the set of primitive lattice vectors and preserves integral affine length; thus the claim follows from Corollary 5.11. □

Let us now turn to the proof of Theorem 5.10. The following lemma is key.

Lemma 5.14 *Let* $x = (x_1, x_2, x_3) \in \mathbb{R}^3_{>0}$ *with* $x_1 < x_2, x_3$. *Then there is a symmetric probe showing*

$$T(x_1, x_2, x_3) \cong T(x_3, x_2 + x_3 - x_1, x_1) \subset \mathbb{C}^3. \tag{84}$$

Proof The directional vector $\eta = e_1^* + e_2^* - e_3^*$ defines a probe $\sigma = \mathbb{R}^3_{\geqslant 0} \cap \{x + t\eta \mid t \in \mathbb{R}\}$ which realizes the equivalence (84); see Figure 7. Indeed, since $x_1 < x_2$, the probe intersects the boundary $\partial\mathbb{R}^3_{\geqslant 0}$ in the points $(0, x_2 - x_1, x_3 + x_1)$ and $(x_1 + x_3, x_2 + x_3, 0)$ which lie in the interior of the facets $\{y_1 = 0\}$ and $\{y_3 = 0\}$ respectively. Since $\langle \eta, e_1 \rangle = 1$ and $\langle \eta, e_3 \rangle = -1$, both intersections are integrally transverse and hence the probe is admissible. Furthermore, since $x_1 \leqslant x_2, x_3$, the points (x_1, x_2, x_3) and $(x_3, x_2 + x_3 - x_1, x_1)$ both lie at integral distance x_1 to the boundary and hence the corresponding fibres are Hamiltonian isotopic. □

Proof of Theorem 5.10 Let $T(x), T(y) \subset \mathbb{C}^n$ be product tori whose Chekanov invariants agree. First note that coordinate permutations

$$(x_1, \dots, x_i, \dots, x_j, \dots, x_n) \mapsto (x_1, \dots, x_j, \dots, x_i, \dots, x_n) \tag{85}$$

can be realized by symmetric probes. Therefore, we may assume that both tori are given in the normal forms

$$x = (\underline{x}, \dots, \underline{x}, \underline{x} + \hat{x}_1, \dots, \underline{x} + \hat{x}_s), \quad y = (\underline{x}, \dots, \underline{x}, \underline{x} + \hat{y}_1, \dots, \underline{x} + \hat{y}_s), \tag{86}$$

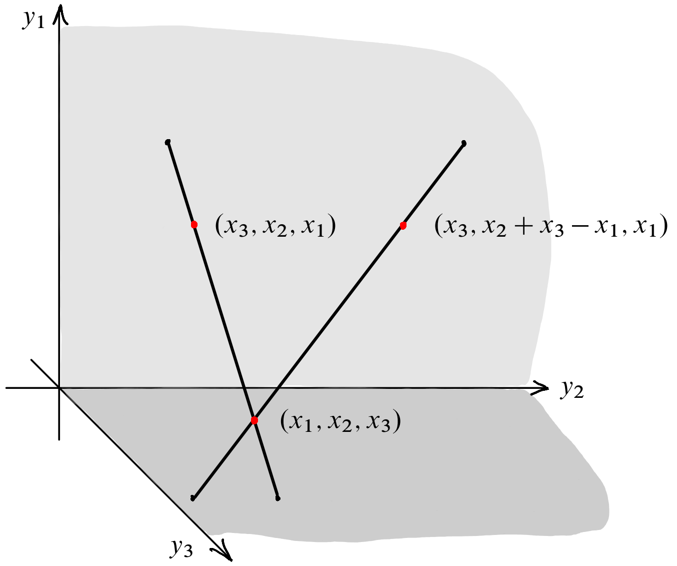


Figure 7: Two symmetric probes in $\mathbb{C}^3$, one realizes a coordinate permutation, the other is used in the proof of Lemma 5.14

where the reduced vectors $\hat{x}, \hat{y} \in \mathbb{R}^s_{>0}$ are $\mathrm{GL}(s;\mathbb{Z})$-equivalent by Corollary 5.11. We thus need to prove that all $\mathrm{GL}(s;\mathbb{Z})$-transformations on the reduced vectors can be realized by symmetric probes. The group $\mathrm{GL}(s;\mathbb{Z})$ is generated by coordinate permutations together with the transformation

$$(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_s) \mapsto (\hat{x}_1 + \hat{x}_2, \hat{x}_2, \dots, \hat{x}_s); \tag{87}$$

see for example [29]. Coordinate permutations can be realized by symmetric probes as we have just seen. By Lemma 5.14, the generator (87) can also be realized by a symmetric probe. Indeed, note that the reduced vectors associated to the tori $T(x_1,x_2,x_3) \subset \mathbb{C}^3$ and $T(x_3, x_2+x_3-x_1, x_1) \subset \mathbb{C}^3$ are given by $(x_2-x_1, x_3-x_1) \in \mathbb{R}^2_{>0}$ and $(x_2+x_3-2x_1, x_3-x_1) \in \mathbb{R}^2_{>0}$ which corresponds to $(\hat{x}_1,\hat{x}_2) \mapsto (\hat{x}_1+\hat{x}_2, \hat{x}_2)$ in terms of the reduced vectors. Hence (87) can be realized by a symmetric probe lying in an appropriately chosen coordinate subspace $\mathbb{C}^3 \subset \mathbb{C}^n$. □

Remark 5.15 Let us briefly discuss a quantitative version of Theorem 5.10. More specifically, For every $\varepsilon > 0$, we can find a Hamiltonian isotopy which has support in the ball $B^6(x_1+x_2+2x_3+\varepsilon)$ realizing the equivalence in Lemma 5.14. Indeed, note that the closure of $B^6(x_1+x_2+2x_3)$ is the smallest closed ball containing the symmetric probe σ in the proof of Lemma 5.14, and the support of the Hamiltonian isotopy can be chosen to lie in an arbitrarily small neighbourhood of σ. This yields a simple proof of [15, Lemma 4.1]. Furthermore, by the same argument as in the proof of [15, Theorem 1.1(ii)], this remark implies that for

$$\sum_{i=1}^n x_i + d(x) < a, \quad \sum_{i=1}^n y_i + d(y) < a, \tag{88}$$

the product tori $T(x)$ and $T(y)$ are Hamiltonian isotopic by iterated symmetric probes *inside the ball* $B^{2n}(a)$. In other words, given a ball $B^{2n}(a)$, there is a region $\mathcal{R}(a) \subset B^{2n}(a)$ in its (open) moment polytope

$$\mathcal{R}(a) = \mu_0^{-1}\left\{x \in \mathbb{R}^n_{\geqslant 0} \;\middle|\; \sum_{i=1}^{n} x_i + d(x) < a\right\} \tag{89}$$

in which the classification of product tori in $B^{2n}(a)$ coincides with the classification of product tori in all of $\mathbb{C}^n$ and all symmetric probes producing the equivalences are contained in $\mu_0(\mathcal{R}(a))$. Together with Chekanov's classification Theorem 4.3, this shows Corollary 5.16, which also follows from the methods in [15]. Let us point out that one cannot drop (88) as was shown in [15, Theorem 1.2]. A reasonable guess would be that in the complement of $\mathcal{R}(a)$, only coordinate permutations are allowed, since these are the only symmetric probes admissible in that region. However, we do not know the classification of product tori in the ball outside of $\mathcal{R}(a)$.

Corollary 5.16 *Let X be a toric manifold of dimension $\geqslant 6$ whose moment polytope has at least one vertex. Then X contains toric fibres such that $\mathfrak{H}_X$ has accumulation points.*

5.7 In arbitrary toric manifolds

The goal of this section is to illustrate that there are many symmetric probes in arbitrary toric manifolds. We focus on constructions near the boundary of the moment polytope Δ using normal forms of Delzant polytopes. For example each vertex $v \in \Delta$ yields an equivariant symplectic ball embedding $B^{2n}(a) \to X$, for each a smaller than the integral length of the shortest edge adjacent to v. Note that this ball embedding is unique up to coordinate permutations. Let us denote the corresponding subset by $B_v(a) \subset X$. Furthermore, denote the image of a product torus $T(x) \subset B^{2n}(a)$ under the equivariant embedding by $T_v(x) \subset B_v(a)$. From Remark 5.15, we deduce that there is a region $\mathcal{R}_v(a) \subset B_v(a)$ in which the same probes are admissible as those that are admissible in $\mathbb{C}^n$. Let us show the following result about the classification of toric fibres close to vertices.

Proposition 5.17 *Let $B^{2n}_v(a)$ and $B^{2n}_{v'}(a)$ be balls at vertices $v, v' \in \Delta$ with*

$$0 < a < \min\{\ell_{\mathrm{int}}(e) \mid e \text{ edge of } \Delta\}. \tag{90}$$

Then $T_v(x) \cong T_{v'}(x)$. In particular, we have $T_v(x) \cong T_v(y)$ if and only if $T_{v'}(x) \cong T_{v'}(y)$.

This means that in small enough neighbourhoods of vertices, the classification problem of toric fibres does not depend on the choice of vertex.

Proof The main idea of the proof is to use the edges of Δ to construct symmetric probes exchanging a pair of toric fibres sitting close to the vertices at the endpoints of the given edge.

Let $e \subset \Delta$ be an edge of the moment polytope with directional vector $v_e \in \Lambda^*$. Let $F, F' \subset \Delta$ be the two facets of Δ adjacent to the endpoints of e but not containing e. We note that the Delzant condition at the vertices adjacent to e implies that v_e intersects F and F' integrally transversely. This can be easily seen by using the corresponding normal form mapping e to the span of e_n^* and F (or F') to the span of $e_1^*, \dots, e_{n-1}^*$. In other words, we obtain symmetric probes parallel to e, as long as both endpoints intersect F and F'. If $a < \min\{\ell_{\mathrm{int}}(e) \mid e \text{ edge of } \Delta\}$, then every $T_v(x) \subset B_v(a)$ can be accessed by such a symmetric probe. Any two vertices $v, v' \in \Delta$ can be linked by a chain of edges and this proves the claim, up to performing coordinate permutations in one of the two ball embeddings (which can always be realized by symmetric probes). □

The technique used in the previous proof can be generalized to any symmetric probe in a face Δ' of a Delzant polytope. Recall that any face of a Delzant polytope is itself Delzant.

Proposition 5.18 *Let $\Delta' \subset \Delta$ be a face and $\sigma' \subset \Delta'$ a symmetric probe therein. Then there is a neighbourhood U of σ' such that any parallel translate of σ' in $U \cap \operatorname{int} \Delta$ is a symmetric probe.*

Proof Let $\sigma' \subset \Delta'$ be a symmetric probe with endpoints on the facets $f, f' \subset \Delta'$. We can write $f = \Delta' \cap F$ and similarly $f' = \Delta \cap F'$ for facets F and F' of Δ. Let us now show that any parallel translate of σ' with endpoints on F and F' is an admissible symmetric probe in Δ. At the face $f = \Delta' \cap F$, we can choose a normal form such that Δ' spans the coordinate subspace spanned by $e_1^*, \dots, e_k^*$ and F the one spanned by $e_2^*, \dots, e_n^*$. Integral transversality of the intersection of σ' and f implies that $v_{\sigma'}, e_2^*, \dots, e_k^*$ is a lattice basis for the sublattice spanned by $e_1^*, \dots, e_k^*$. Here, we have denoted the directional vector of the symmetric probe σ' by $v_{\sigma'}$. This implies that $v_{\sigma'}, e_2^*, \dots, e_n^*$ is a lattice basis of the full lattice in the ambient space, proving integral transversality. □

References

[1] **M Abreu**, **M S Borman**, **D McDuff**, *Displacing Lagrangian toric fibers by extended probes*, Algebr. Geom. Topol. 14 (2014) 687–752 MR Zbl

[2] **M Abreu**, **L Macarini**, *Remarks on Lagrangian intersections in toric manifolds*, Trans. Amer. Math. Soc. 365 (2013) 3851–3875 MR Zbl

[3] **M Audin**, *Torus actions on symplectic manifolds*, 2nd edition, Progr. Math. 93, Birkhäuser, Basel (2004) MR Zbl

[4] **M Augustynowicz**, **J Smith**, **J Wornbard**, *Homological Lagrangian monodromy for some monotone tori*, Quantum Topol. (online publication June 2024) Zbl

[5] **D Auroux**, *Infinitely many monotone Lagrangian tori in $\mathbb{R}^6$*, Invent. Math. 201 (2015) 909–924 MR Zbl

[6] **J Brendel**, *Real Lagrangian tori and versal deformations*, J. Symplectic Geom. 21 (2023) 463–507 MR Zbl

[7] **J Brendel**, **J Kim**, *On Lagrangian tori in $S^2 \times S^2$ and billiards*, preprint (2025) arXiv 2502.03324

[8] **J Brendel**, **J Kim**, **J Moon**, *On the topology of real Lagrangians in toric symplectic manifolds*, Israel J. Math. 253 (2023) 113–156 MR Zbl

[9] **J Brendel**, **G Mikhalkin**, **F Schlenk**, *Non-isotopic symplectic embeddings of cubes*, in preparation

[10] **L M Cabrer**, **D Mundici**, *Classifying* $\mathrm{GL}(n,\mathbb{Z})$*-orbits of points and rational subspaces*, Discrete Contin. Dyn. Syst. 36 (2016) 4723–4738 MR Zbl

[11] **A Cannas da Silva**, *Symplectic toric manifolds*, from "Symplectic geometry of integrable Hamiltonian systems", Birkhäuser, Basel (2003) 85–173 MR Zbl

[12] **Y V Chekanov**, *Lagrangian tori in a symplectic vector space and global symplectomorphisms*, Math. Z. 223 (1996) 547–559 MR Zbl

[13] **Y V Chekanov**, *Lagrangian intersections, symplectic energy, and areas of holomorphic curves*, Duke Math. J. 95 (1998) 213–226 MR Zbl

[14] **Y Chekanov**, **F Schlenk**, *Notes on monotone Lagrangian twist tori*, Electron. Res. Announc. Math. Sci. 17 (2010) 104–121 MR Zbl

[15] **Y Chekanov**, **F Schlenk**, *Lagrangian product tori in symplectic manifolds*, Comment. Math. Helv. 91 (2016) 445–475 MR Zbl

[16] **C-H Cho**, *Non-displaceable Lagrangian submanifolds and Floer cohomology with non-unitary line bundle*, J. Geom. Phys. 58 (2008) 1465–1476 MR Zbl

[17] **J S Dani**, *Density properties of orbits under discrete groups*, J. Indian Math. Soc. 39 (1975) 189–217 MR Zbl

[18] **T Delzant**, *Hamiltoniens périodiques et images convexes de l'application moment*, Bull. Soc. Math. France 116 (1988) 315–339 MR Zbl

[19] **K Fukaya**, **Y-G Oh**, **H Ohta**, **K Ono**, *Displacement of polydisks and Lagrangian Floer theory*, J. Symplectic Geom. 11 (2013) 231–268 MR Zbl

[20] **V Guillemin**, *Moment maps and combinatorial invariants of Hamiltonian* T^n*-spaces*, Progr. Math. 122, Birkhäuser, Boston, MA (1994) MR Zbl

[21] **S Hu**, **F Lalonde**, **R Leclercq**, *Homological Lagrangian monodromy*, Geom. Topol. 15 (2011) 1617–1650 MR Zbl

[22] **C Y Mak**, **I Smith**, *Non-displaceable Lagrangian links in four-manifolds*, Geom. Funct. Anal. 31 (2021) 438–481 MR Zbl

[23] **D McDuff**, *Displacing Lagrangian toric fibers via probes*, from "Low-dimensional and symplectic topology", Proc. Sympos. Pure Math. 82, Amer. Math. Soc., Providence, RI (2011) 131–160 MR Zbl

[24] **D McDuff**, **D Salamon**, *Introduction to symplectic topology*, 3rd edition, Oxford Univ. Press (2017) MR Zbl

[25] **L Polterovich**, *The geometry of the group of symplectic diffeomorphisms*, Birkhäuser, Basel (2001) MR Zbl

[26] **N W Porcelli**, *Families of relatively exact Lagrangians, free loop spaces and generalised homology*, Selecta Math. 30 (2024) art. id. 21 MR Zbl

[27] **E Shelukhin**, **D Tonkonog**, **R Vianna**, *Geometry of symplectic flux and Lagrangian torus fibrations*, preprint (2018) arXiv 1804.02044

[28] **J Smith**, *A monotone Lagrangian casebook*, Algebr. Geom. Topol. 21 (2021) 2273–2312 MR Zbl

[29] **S M Trott**, *A pair of generators for the unimodular group*, Canad. Math. Bull. 5 (1962) 245–252 MR Zbl

[30] **R Vianna**, *Infinitely many exotic monotone Lagrangian tori in* $\mathbb{CP}^2$, J. Topol. 9 (2016) 535–551 MR Zbl

[31] **R Vianna**, *Infinitely many monotone Lagrangian tori in del Pezzo surfaces*, Selecta Math. 23 (2017) 1955–1996 MR Zbl

[32] **M-L Yau**, *Monodromy and isotopy of monotone Lagrangian tori*, Math. Res. Lett. 16 (2009) 531–541 MR Zbl

Department of Mathematics, ETH Zürich
Zürich, Switzerland

joe.brendel@math.ethz.ch

Received: 10 July 2023

Algebraic & Geometric Topology **25**:3 (2025) 1877–1896
DOI: 10.2140/agt.2025.25.1877
Published: 20 June 2025

An example of higher-dimensional Heegaard Floer homology

YIN TIAN
TIANYU YUAN

We count pseudoholomorphic curves in the higher-dimensional Heegaard Floer homology of disjoint cotangent fibers of a two-dimensional disk. We show that the resulting algebra is isomorphic to the Hecke algebra associated to the symmetric group.

53D40; 53D10

1 Introduction

Many topological properties of a manifold M can be recovered from the symplectic geometry of its cotangent bundle T^*M. An example is the A_∞-equivalence between the wrapped Floer homology $\mathrm{CW}^*(T_q^*M)$ of a cotangent fiber and the space $C_{-*}(\Omega_q M)$ of chains on the based loop space of M, proved by Abbondandolo and Schwarz [1] and Abouzaid [2].

On the symplectic side, there is a generalization, the wrapped Floer homology $\mathrm{CW}^*\big(\bigsqcup_{i=1}^{\kappa} T_{q_i}^*M\big)$ of κ disjoint cotangent fibers in the framework of *higher-dimensional Heegaard Floer homology* (abbreviated HDHF) established by Colin, Honda and Tian [3]. It is related to the braid group of M on the topological side.

When $M = \Sigma$ is a real oriented surface, the HDHF was recently studied by Honda, Tian and Yuan [7]. Pick κ basepoints $\boldsymbol{q} = \{q_1, \dots, q_\kappa\} \subset \Sigma$. By definition, $\mathrm{CW}^*\big(\bigsqcup_i T_{q_i}^*\Sigma\big)$ is an A_∞ algebra over $\mathbb{Z}[[\hbar]]$, where $\hbar$ keeps track of the Euler characteristic of the holomorphic curves that are counted in the definition of the A_∞-operations. If Σ is not a two sphere, then $\mathrm{CW}^*\big(\bigsqcup_i T_{q_i}^*\Sigma\big)$ is supported in degree zero. Hence, it is isomorphic to its homology $\mathrm{HW}^*\big(\bigsqcup_i T_{q_i}^*\Sigma\big)$ as an ordinary algebra. The main result of [7] is that the algebra $\mathrm{HW}^*\big(\bigsqcup_i T_{q_i}^*\Sigma\big)$ is isomorphic to the *braid skein algebra* $\mathrm{BSk}_\kappa(\Sigma)$ of Σ, which was defined by Morton and Samuelson [9]. Roughly speaking, $\mathrm{BSk}_\kappa(\Sigma)$ is a quotient of the group algebra of the braid group of Σ by the *HOMFLY skein relation*, which is expressed in terms of $\hbar$. The skein relation has an explanation as holomorphic curve counting due to Ekholm and Shende [4]. This is one of the keys to build the bridge between $\mathrm{HW}^*\big(\bigsqcup_i T_{q_i}^*\Sigma\big)$ and $\mathrm{BSk}_\kappa(\Sigma)$.

Morton and Samuelson showed that $\mathrm{BSk}_\kappa(\Sigma)$ is isomorphic to the *double affine Hecke algebra* associated to $\mathfrak{gl}_\kappa$ when Σ is a torus. Based on this result, Honda, Tian and Yuan proved the isomorphisms between $\mathrm{HW}^*\big(\bigsqcup_i T_{q_i}^*\Sigma\big)$ and various Hecke algebras of type A for Σ being a disk, a cylinder or a torus.

Here we focus on the local case: $\Sigma = D^2$ is a disk. Let $\mathrm{End}(L^{\otimes\kappa})$ denote the algebra $\mathrm{HW}^*\bigl(\bigsqcup_i T^*_{q_i} D^2\bigr)$ throughout the paper. It is isomorphic to the finite Hecke algebra associated to the symmetric group S_κ over $\mathbb{Z}[[\hbar]]$; see [7]. The main result of this paper is to show that $\mathrm{End}(L^{\otimes\kappa})$ can be defined over $\mathbb{Z}[\hbar]$, and the isomorphism to the finite Hecke algebra still holds.

The reduction from $\mathbb{Z}[[\hbar]]$ to $\mathbb{Z}[\hbar]$ has two advantages regarding connections to other fields. The first one is topological. The HOMFLYPT polynomial of links takes values in the ring of Laurent polynomials of $\hbar$. This polynomial can be obtained from a trace function on the Hecke algebra; see Jones [8]. Note that the HOMFLYPT polynomial has a Floer-theoretic interpretation due to Ekholm and Shende [4]. It is interesting to look for connections between our curve counting for the Hecke algebra and theirs for the link invariant.

The second one is algebraic. We expect that the HDHF Fukaya category of $T^*\Sigma$ is related to the category of modules over the algebra $\mathrm{HW}^*\bigl(\bigsqcup_i T^*_{q_i}\Sigma\bigr)$. Representation theory over $\mathbb{Z}[[\hbar]]$ and $\mathbb{Z}[\hbar]$ could be different. Modules over $\mathbb{Z}[\hbar]$ or $\mathbb{Z}[q,q^{-1}]$ are commonly used in representation theory of various Hecke algebras.

We explicitly describe our main result in the following. The Floer generators of $\mathrm{End}(L^{\otimes\kappa})$ are tuples of intersection points between the cotangent fibers $T^*_{q_i} D^2$. They are in one-to-one correspondence to elements of the symmetric group S_κ. Let $T_w \in \mathrm{End}(L^{\otimes\kappa})$ denote the corresponding Floer generator for $w \in S_\kappa$.

For the Hecke algebra, we change the variable from q to $\hbar$ via $\hbar = q - q^{-1}$ for our purpose.

Definition 1.1 The Hecke algebra H_κ is a unital $\mathbb{Z}[\hbar]$-algebra generated by $\widetilde{T}_1, \dots, \widetilde{T}_{\kappa-1}$, with relations

$$\widetilde{T}_i^2 = 1 + \hbar\widetilde{T}_i, \qquad \widetilde{T}_i\widetilde{T}_j = \widetilde{T}_j\widetilde{T}_i \quad \text{for } |i-j|>1 \qquad \text{and} \qquad \widetilde{T}_i\widetilde{T}_{i+1}\widetilde{T}_i = \widetilde{T}_{i+1}\widetilde{T}_i\widetilde{T}_{i+1}.$$

It is known that the Hecke algebra H_κ is a free $\mathbb{Z}[\hbar]$-module with a basis $\widetilde{T}_w$ for $w \in S_\kappa$, called the *standard basis*. Here $\widetilde{T}_i = \widetilde{T}_{s_i}$ for the transposition $s_i = (i, i+1)$. There is a length function on S_κ defined by $l(w) = \min\{l \mid w = s_{i_1}\cdots s_{i_l}\}$. The basis $\widetilde{T}_w = \widetilde{T}_{i_1}\cdots\widetilde{T}_{i_l}$ if $w = s_{i_1}\cdots s_{i_l}$ is an expression of minimal length. Moreover, the algebra structure on H_κ is uniquely determined by

$$\widetilde{T}_i\widetilde{T}_w = \begin{cases} \widetilde{T}_{s_i w} & \text{if } l(s_i w) > l(w), \\ \widetilde{T}_{s_i w} + \hbar\widetilde{T}_w & \text{if } l(s_i w) < l(w). \end{cases} \tag{1-1}$$

Our main result is the following.

Theorem 1.2 *The HDHF homology* $\mathrm{End}(L^{\otimes\kappa})$ *is defined over* $\mathbb{Z}[\hbar]$. *Moreover, there is an isomorphism of unital algebras* $\phi\colon H_\kappa \to \mathrm{End}(L^{\otimes\kappa})$ *such that* $\phi(\widetilde{T}_w) = T_w$ *for* $w \in S_\kappa$.

In other words, we give a Floer-theoretic explanation of the standard basis of the Hecke algebra H_κ.

Unlike the method presented in [7], our proof takes a different approach by directly counting holomorphic curves in HDHF. Curve counting is in general a challenging task unless the ambient symplectic manifold is of real dimension two. We arrange the Lagrangian boundary conditions in a split form such that the curve counting problem reduces to the case of two copies of $\mathbb{C}$. Our strategy consists of two main steps:

(1) When $\kappa = 2$, we are able to show that there exists a nontrivial holomorphic disk, which corresponds to the $\hbar$ term in the quadratic relation of the Hecke algebra; see Lemma 3.5. To see this curve, in each copy of $\mathbb{C}$, the counting is combinatorial and provides a relative homology class within the moduli space of a hexagon. The nontrivial intersection number of two relative classes from the two copies of $\mathbb{C}$ then shows the existence of such a curve.

(2) To prove the Hecke relation holds in HDHF in general, we proceed by induction on κ. By stretching the Lagrangians in a certain order, the corresponding family of holomorphic curves degenerates into several pieces due to Gromov compactness. The degenerated curves live in the moduli space associated to $\kappa' = \kappa - 1$ or $\kappa - 2$. Therefore we can do induction on κ.

Further directions (1) It is natural to ask whether the HDHF homology $\mathrm{End}\big(\bigsqcup_i T^*_{q_i}\Sigma\big)$ of disjoint fibers of $T^*\Sigma$ can be defined over $\mathbb{Z}[\hbar]$ for a general surface Σ. We hope to generalize our result from local to global by using some sheaf-theoretic techniques, for instance those of Ganatra, Pardon and Shende [6]. The idea is to establish a pushout diagram so that $\mathrm{End}\big(\bigsqcup_i T^*_{q_i}\Sigma\big)$ can be realized as a homotopy colimit of the local pieces which is defined over $\mathbb{Z}[\hbar]$.

(2) It is interesting to explain the geometric meaning of the change of variables $\hbar = q - q^{-1}$. Note that the *canonical basis* of the Hecke algebra is defined over $\mathbb{Z}[q, q^{-1}]$. We will express the canonical basis via HDHF in an upcoming paper.

(3) When the symplectic manifold is of dimension greater than four, a similar technique can be used to compute the local case T^*D^m for $m > 2$. In this case, we expect that the HDHF homology $\mathrm{HW}^*\big(\bigsqcup_i T^*_{q_i}D^m\big)$ is a nontrivial A_∞ algebra. Its nontrivial higher A_∞ relation can be viewed as a generalization of the quadratic relation in the Hecke algebra.

Acknowledgements The authors thank Ko Honda for numerous ideas and suggestions. Tian is supported by NSFC 11971256. Yuan is supported by China Postdoctoral Science Foundation 2023T160002.

2 Preliminaries

We first specify the ambient manifold and Lagrangian submanifolds of interest. For convenience of curve counting, we set $D^2 = I_1 \times I_2$ with $I_1 = I_2 = [0,1]$, which is topologically the same as the unit disk. Let $X = T^*D^2 = T^*I_1 \times T^*I_2$ be the total space of the cotangent bundle of D^2.

Consider the canonical Liouville form θ on T^*D^2, which induces a contact manifold structure at the infinity of (T^*D^2, θ). For a Lagrangian $L \subset T^*D^2$, denote its boundary at infinity by $\partial_\infty L$. An isotopy of Lagrangians L_t in T^*D^2 is called *positive* if $\alpha(\partial_t \partial_\infty L_t) > 0$ for all t. Let $T^*_v D^2 = T^*D^2|_{\partial D^2}$ be the vertical boundary of T^*D^2 over ∂D^2. We require that any isotopy L_t cannot cross $T^*_v D^2$. A positive isotopy is also called a "partially wrapping". For the details of partially wrapped Fukaya categories, we refer to [10] by Sylvan and [5] by Ganatra, Shende, and Pardon.

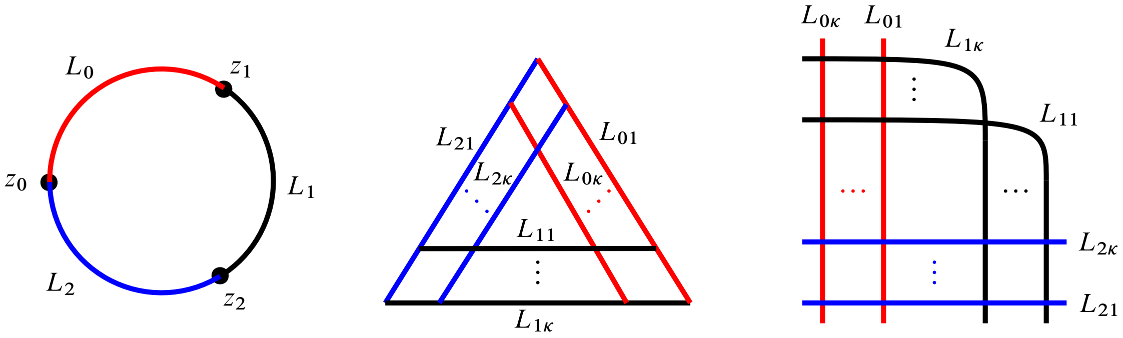


Figure 1: Left: D_3, the A_∞ base direction. Center: the Lagrangians in the T^*I_1 direction. Right: the Lagrangians in the T^*I_2 direction.

We next consider the generalization to wrapped HDHF. Pick κ disjoint basepoints $\boldsymbol{q}=\{q_1,\dots,q_\kappa\}\subset D^2\backslash\partial D^2$ and consider the κ cotangent fibers $L_{0i}=T^*_{q_i}D$ for $i=1,\dots,\kappa$. We define $\mathcal{L}_0=\{L_{01},\dots,L_{0\kappa}\}$. An isotopy of a κ-tuple of Lagrangians $\mathcal{L}_t=\{L_{t1},\dots,L_{t\kappa}\}$ is called *positive* if $\alpha(\partial_t\partial_\infty L_{ti})>0$ for all $i=1,\dots,\kappa$ and all t. For a pair of κ-tuples of Lagrangians $\mathcal{A}$ and $\mathcal{B}$, we write $\mathcal{A}\rightsquigarrow\mathcal{B}$ if there is an positive isotopy from $\mathcal{A}$ to $\mathcal{B}$.

We then perform positive wrapping on $\mathcal{L}_0$ to get $\mathcal{L}_j=\{L_{j1},\dots,L_{j\kappa}\}$ for $j=1,2$. Specifically, we put $\mathcal{L}_0$, $\mathcal{L}_1$ and $\mathcal{L}_2$ in the position as in Figure 1, center and right, which represent the T^*I_1-direction and T^*I_2-direction, respectively. It is easy to check that

$$\mathcal{L}_0\rightsquigarrow\mathcal{L}_1\rightsquigarrow\mathcal{L}_2.$$

The HDHF cochain complex $\mathrm{CF}^*(\mathcal{L}_i,\mathcal{L}_j)$ for $i<j$ is defined as the free abelian group generated by κ-tuples of intersection points between $\mathcal{L}_i$ and $\mathcal{L}_j$ over $\mathbb{Z}[[\hbar]]$. By definition, $\mathrm{CF}^*(\mathcal{L}_i,\mathcal{L}_j)$ is an A_∞-algebra. We refer the reader to [7] for details of the definition of HDHF in this case.

There is an absolute grading on $\mathrm{CF}^*(\mathcal{L}_i,\mathcal{L}_j)$ and the degree is supported at zero by [7, Proposition 2.9]. Hence, its homology $\mathrm{HF}^*(\mathcal{L}_i,\mathcal{L}_j)$ is an ordinary algebra over $\mathbb{Z}[[\hbar]]$. We denote it by $\mathrm{End}(L^{\otimes\kappa})$.

Remark 2.1 Strictly speaking, the algebra structure of $\mathrm{End}(L^{\otimes\kappa})$ is given by the composition map

$$\mu^2\colon \mathrm{HF}^*(\mathcal{L}_1,\mathcal{L}_2)\otimes\mathrm{HF}^*(\mathcal{L}_0,\mathcal{L}_1)\to\mathrm{HF}^*(\mathcal{L}_0,\mathcal{L}_2),$$

together with the continuation maps

$$c_{12}\circ-\colon\mathrm{HF}^*(\mathcal{L}_0,\mathcal{L}_1)\xrightarrow{\sim}\mathrm{HF}^*(\mathcal{L}_0,\mathcal{L}_2),\tag{2-1}$$

$$-\circ c_{01}\colon\mathrm{HF}^*(\mathcal{L}_1,\mathcal{L}_2)\xrightarrow{\sim}\mathrm{HF}^*(\mathcal{L}_0,\mathcal{L}_2),\tag{2-2}$$

where $c_{12}\in\mathrm{HF}^*(\mathcal{L}_1,\mathcal{L}_2)$ and $c_{01}\in\mathrm{HF}^*(\mathcal{L}_0,\mathcal{L}_1)$ are two specific generators, both denoted by T_{id} in Section 3. In this way, we have defined the algebra structure on $\mathrm{HF}^*(\mathcal{L}_0,\mathcal{L}_2)$.

In order to compute the algebra $\mathrm{End}(L^{\otimes\kappa})$, we need to explicitly compute the maps (2-1) and (2-2). It is then necessary that (2-1) and (2-2) are indeed identity maps (with respect to some choice of basis of

Floer generators) instead of just isomorphisms. With our specific choice of wrapping, we can show that the continuation maps preserve the Floer generators for different $\mathrm{HF}^*(\mathcal{L}_i, \mathcal{L}_j)$ with $i < j$. In other words, we show that T_{id} behaves as the identity of the algebra; see Proposition 3.1.

Remark 2.2 We fix the special wrapping of $\mathcal{L}_0$, $\mathcal{L}_1$ and $\mathcal{L}_2$, which is crucial for our counting of curves. We remark that the algebraic count is invariant under compactly supported perturbation of Lagrangians; see [3, Section 6]. Indeed, we can show that the specific choice of Figure 1, right, is only for computing convenience, and the count remains the same if we rearrange Figure 1, right, eg to be like Figure 1, center. However, currently we do not know whether the same holds if we perturb the Lagrangians at infinity without keeping the Lagrangians in a product form.

We describe the μ^2-composition map of $\mathrm{End}(L^{\otimes\kappa})$ in the following. We set $\hat{X} = D_3 \times X$ as the target manifold, where D_3 is the unit disk with three boundary punctures, and refer to it as the "A_∞ base direction". Let z_0, z_1 and z_2 be the boundary punctures of D_3 and let α_0, α_1 and α_2 be the boundary arcs. We extend $\mathcal{L}_i$ to the D_3 direction by setting $\hat{\mathcal{L}}_i = \alpha_i \times \mathcal{L}_i$ and $\hat{L}_{ij} = \alpha_i \times L_{ij}$ for $i = 0, 1, 2$ and $j = 1, \dots, \kappa$, which are Lagrangian submanifolds of $\hat{X}$.

For $i = 1, 2$, let $\boldsymbol{y}_i = \{y_{i1}, \dots, y_{i\kappa}\}$ be a tuple of intersection points $y_{ij} \in \hat{L}_{(i-1)j} \cap \hat{L}_{ij'}$, where $\{1', \dots, \kappa'\}$ is some permutation of $\{1, \dots, \kappa\}$. Let J be a small generic perturbation of $J_{D_3} \times J_1 \times J_2$, where J_{D_3}, J_1 and J_2 are the standard complex structures on D_3, T^*I_1 and T^*I_2, viewed as subsets of $\mathbb{C}$. Let $\mathcal{M}(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_0)$ be the moduli space of maps

$$u\colon (\dot{F}, j) \to (\hat{X}, J), \tag{2-3}$$

where (F, j) is a compact Riemann surface with boundary, $\boldsymbol{p}_i$ are disjoint tuples of boundary punctures of F for $i = 0, 1, 2$, and $\dot{F} = F \setminus \bigcup_i \boldsymbol{p}_i$. The map u satisfies

$$\begin{cases} du \circ j = J \circ du, \\ \text{each component of } \partial\dot{F} \text{ is mapped to a unique } \hat{L}_{ij}, \\ \pi_X \circ u \text{ tends to } \boldsymbol{y}_i \text{ as } s_i \to +\infty \text{ for } i = 1, \dots, m, \\ \pi_X \circ u \text{ tends to } \boldsymbol{y}_0 \text{ as } s_0 \to -\infty, \\ \pi_{D_3} \circ u \text{ is a } \kappa\text{-fold branched cover of } D_3, \end{cases} \tag{2-4}$$

where the third condition means that $\pi_X \circ u$ maps the neighborhoods of the punctures of $\boldsymbol{p}_i$ asymptotically to the Reeb chords of $\boldsymbol{y}_i$ for $i = 1, \dots, m$ at the positive ends. The fourth condition is similar.

The μ^2-composition map of $\mathrm{End}(L^{\otimes\kappa})$ is then defined as

$$\mu^2(\boldsymbol{y}_1, \boldsymbol{y}_2) = \sum_{\boldsymbol{y}_0, \chi \le \kappa} \#\mathcal{M}^{\mathrm{ind}=0,\chi}(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_0) \cdot \hbar^{\kappa-\chi} \cdot \boldsymbol{y}_0, \tag{2-5}$$

where the superscript "ind" denotes the Fredholm index and "χ" denotes the Euler characteristic of F; the symbol # denotes the signed count of the corresponding moduli space.

A choice of spin structures on the Lagrangians determines a canonical orientation of the moduli space. The Lagrangian in our case is the cotangent fiber, which is topologically $\mathbb{R}^2$. So there is a unique spin structure. We omit the details about the orientation, and refer the reader to [3, Section 3].

3 The case of $\kappa = 2$

In this section we compute $\mathrm{End}(L^{\otimes 2})$ as a model case. The general case will be discussed in Section 4.

For $0 \le i < j \le 2$, there are two Floer generators of $\mathrm{CF}^*(\mathcal{L}_i, \mathcal{L}_j)$: T_{id} and T_1, where $T_{\mathrm{id}} = (q_1, q_2)$ with $q_1 \in L_{i1} \cap L_{j1}$ and $q_2 \in L_{i2} \cap L_{j2}$, and $T_1 = (q_1, q_2)$ with $q_1 \in L_{i1} \cap L_{j2}$ and $q_2 \in L_{i2} \cap L_{j1}$. The main result of this section is the following:

Proposition 3.1 *The multiplication on* $\mathrm{End}(L^{\otimes 2})$ *is given by*

$$T_{\mathrm{id}} \cdot T_{\mathrm{id}} = T_{\mathrm{id}}, \quad T_{\mathrm{id}} \cdot T_1 = T_1, \quad T_1 \cdot T_{\mathrm{id}} = T_1 \quad \textit{and} \quad T_1 \cdot T_1 = 1 + \hbar T_1.$$

Hence Theorem 1.2 holds for $\kappa = 2$.

The proof of this proposition occupies the rest of the section. We directly compute the moduli spaces. There are trivial curves with $\chi = 2$ accounting for the $\hbar^0$ terms in the multiplication. We show that $\mathcal{M}_J^{\chi<2}(y_1, y_2, y_0) = \varnothing$ for almost all cases except that $\mathcal{M}_J^{\chi=1}(T_1, T_1, T_1) \neq \varnothing$ accounting for the $\hbar^1$ term in $T_1 \cdot T_1$. The main strategy to prove the nonexistence of curves is to stretch the Lagrangians in the T^*I_1-direction and apply Gromov compactness.

For later use, we make the following conventions:

- We denote the length of the line segment of $L_{1\kappa}$ in the I_1-direction by d; see Figure 1, center.
- For $q \in X$, we denote its projection in the T^*I_1 (resp. T^*I_2) direction by q' (resp. q'').
- We denote the line segment between q_1 and q_2 by (q_1q_2).
- When plotting figures, we denote the intersections q_i by i.
- When taking the limit, we denote the degenerated domain by $\dot{F}'$ and its irreducible component containing $\{p_1, p_2, \dots\}$ by $\dot{F}'_{(12\dots)}$.

Lemma 3.2 $$T_{\mathrm{id}} \cdot T_{\mathrm{id}} = T_{\mathrm{id}}.$$

Proof We first show that

$$\#\mathcal{M}^{\chi}(T_{\mathrm{id}}, T_{\mathrm{id}}, T_{\mathrm{id}}) = \begin{cases} 1 & \text{if } \chi = 2, \\ 0 & \text{if } \chi < 2. \end{cases} \tag{3-1}$$

The Floer generators are shown in Figure 2.

If $\chi = 2$, there is a unique trivial holomorphic curve consisting of two disks. So $\#\mathcal{M}^{\chi=2}(T_{\mathrm{id}}, T_{\mathrm{id}}, T_{\mathrm{id}}) = 1$.

If $\chi < 2$, let $d \to 0$, ie let q_1' and q_2' get closer. In the limit, since there are no slit or branch points separating q_1' and q_2', $\dot{F}'_{(12)}$ bubbles off as a triangle with vertices $\{p_1, p_2, p_a\}$, where $\{p_a\}$ is a boundary

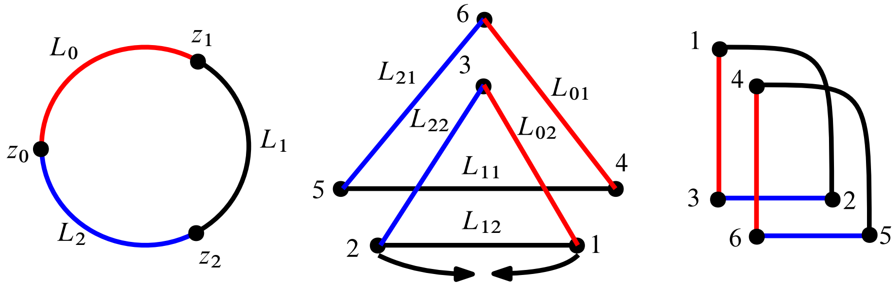


Figure 2: Generators for $\mathcal{M}(T_{\mathrm{id}}, T_{\mathrm{id}}, T_{\mathrm{id}})$.

nodal point. The projection of $\dot{F}'_{(12)}$ under $\pi_{T^*I_2} \circ u$ is a homeomorphism to the triangle with vertices $\{q''_1, q''_2, q''_3\}$. Hence the projection of $\dot{F}'_{(3)}$ under $\pi_{T^*I_2} \circ u$ is a constant map to q''_3. Since $\pi_{T^*I_2} \circ u$ is of degree zero or one near q''_3, the image $\pi_{T^*I_2} \circ u(\dot{F} \backslash (\dot{F}'_{(12)} \cup \dot{F}'_{(3)}))$ is disjoint from q''_3. So $\dot{F}'_{(12)} \cup \dot{F}'_{(3)}$ is a connected component of $\dot{F}'$. Therefore $\dot{F}$ consists of two components before the degeneration, which are homeomorphically mapped to the triangles $\{q''_1, q''_2, q''_3\}$ and $\{q''_4, q''_5, q''_6\}$ under $\pi_{T^*I_2} \circ u$, respectively. So $\chi = 2$, which is a contradiction. We conclude that $\#\mathcal{M}^\chi(T_{\mathrm{id}}, T_{\mathrm{id}}, T_{\mathrm{id}}) = 0$ if $\chi < 2$.

We next show that $\#\mathcal{M}(T_{\mathrm{id}}, T_{\mathrm{id}}, T_1) = 0$. The generators are shown in Figure 3. As $d \to 0$, $\dot{F}'_{(12)}$ bubbles off as a triangle with vertices $\{p_1, p_2, p_a\}$, where $p_a \in \dot{F}'$ is a nodal point. Denote the union of irreducible components of $\dot{F}'$ containing the preimage of the dashed lines in the T^*I_1-direction by $\dot{F}'_{\mathrm{dash}}$. Since p_3, p_6 and p_a are mapped to z_0 under $\pi_{D_3} \circ u$ in the limit, the preimages of the dashed lines are also mapped to z_0. Hence $\dot{F}'_{\mathrm{dash}}$ is mapped to the constant point z_0 under $\pi_{D_3} \circ u$. Since $(q'_5 q'_6)$ cannot be separated by slits, $q'_5 \in \dot{F}'_{\mathrm{dash}}$ and $\pi_{D_3} \circ u(q'_5) = z_0$. This contradicts with the fact that $\pi_{D_3} \circ u(q'_5) = z_2$. Therefore $\mathcal{M}(T_{\mathrm{id}}, T_{\mathrm{id}}, T_1) = \varnothing$. □

Lemma 3.3 $T_{\mathrm{id}} \cdot T_1 = T_1$.

Proof First we show that

$$\#\mathcal{M}^\chi(T_{\mathrm{id}}, T_1, T_1) = \begin{cases} 1 & \text{if } \chi = 2, \\ 0 & \text{if } \chi < 2. \end{cases} \tag{3-2}$$

The generators are shown in Figure 4.

If $\chi = 2$, there is a unique trivial holomorphic curve consisting of two disks. So $\#\mathcal{M}^{\chi=2}(T_{\mathrm{id}}, T_1, T_1) = 1$.

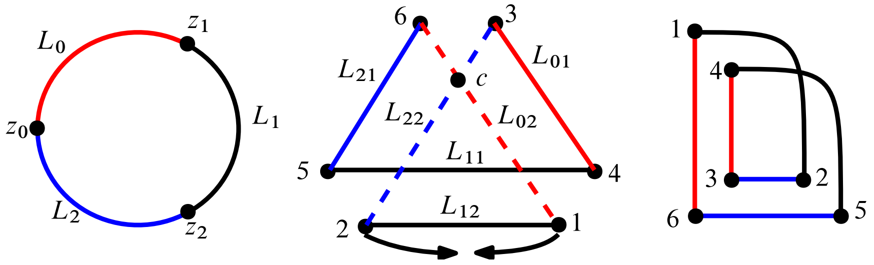


Figure 3: Generators for $\mathcal{M}(T_{\mathrm{id}}, T_{\mathrm{id}}, T_1)$.

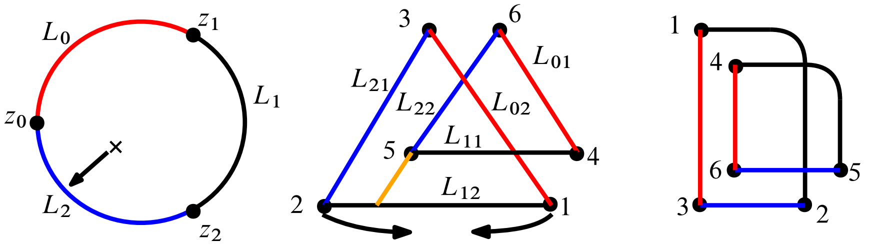


Figure 4: Generators for $\mathcal{M}(T_{\mathrm{id}}, T_1, T_1)$.

If $\chi < 2$, as $d \to 0$, there are two cases:

- If the orange slit extending $(q_6'q_5')$ is not long, then $\dot{F}'_{(12)}$ bubbles off as a triangle with vertices $\{p_1, p_2, p_a\}$. The remaining proof is the same as that of $\#\mathcal{M}^\chi(T_{\mathrm{id}}, T_{\mathrm{id}}, T_{\mathrm{id}}) = 0$ for $\chi < 2$ in Lemma 3.2. Hence, the limiting curve does not exist.

- If the orange slit extending $(q_6'q_5')$ is long, then there is a branch point approaching the interior of L_2 in the D_3-direction (as in the left of Figure 4). In the limit, the preimage of the branch point on the domain tends to some nodal point p_n such that $\pi_{D_3} \circ u(p_n) \in L_2$. This implies that $\pi_{T^*I_2} \circ u(p_n) \in L_{21} \cap L_{22}$. This contradicts with the fact that $L_{21} \cap L_{22} = \varnothing$ in the T^*I_2-direction.

Next we show that $\#\mathcal{M}^\chi(T_{\mathrm{id}}, T_1, T_{\mathrm{id}}) = 0$ for $\chi \leq 2$. The generators are shown in Figure 5. As $d \to 0$, $\dot{F}'_{(12)}$ bubbles off as a triangle $\{p_1, p_2, p_a\}$. Then p_a should be mapped to the intersection of the extension of the line segments $(q_1''q_6'')$ and $(q_2''q_3'')$. But this is impossible since the degree of the projection $\pi_{T^*I_2} \circ u$ is zero near the intersection. □

Similar arguments will be used in the proofs of Propositions 4.1 and 4.2. In general, $\dot{F}'_{(12)}$ always bubbles off as a triangle as $d \to 0$. Here q_1' and q_2' are on the bottom Lagrangian $L_{1\kappa}$ in the T^*I_1-direction. We then analyze the remaining irreducible components of $\dot{F}'$ and reduce the problem to simpler cases.

Lemma 3.4
$$T_1 \cdot T_{\mathrm{id}} = T_1.$$

Proof This is similar to the proof of Lemma 3.3. □

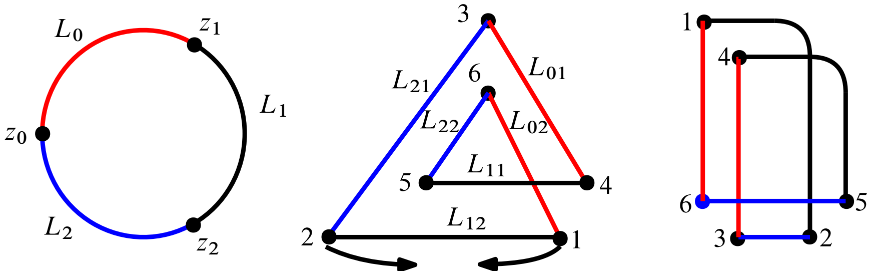


Figure 5: Generators for $\mathcal{M}(T_{\mathrm{id}}, T_1, T_{\mathrm{id}})$.

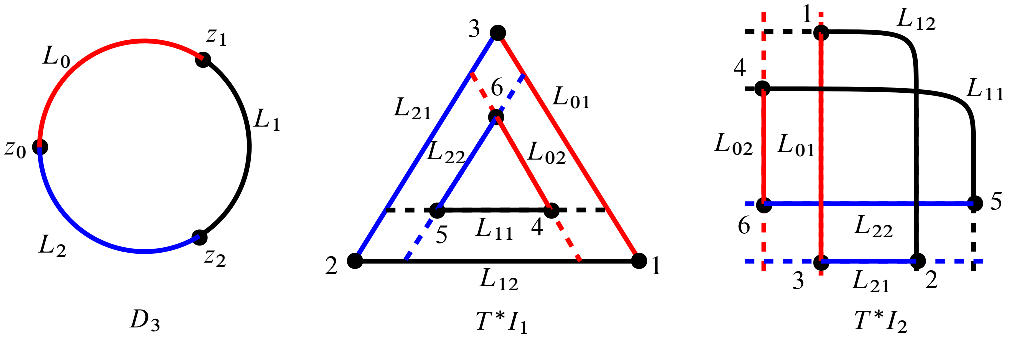


Figure 6: Generators for $\mathcal{M}(T_1, T_1, T_{\mathrm{id}})$.

Lemma 3.5 $T_1 \cdot T_1 = T_{\mathrm{id}} + \hbar T_1$.

Proof We first show that

$$\#\mathcal{M}^{\chi}(T_1, T_1, T_{\mathrm{id}}) = \begin{cases} 1 & \text{if } \chi = 2, \\ 0 & \text{if } \chi < 2. \end{cases} \tag{3-3}$$

The generators are shown in Figure 6.

If $\chi = 2$, there is a unique trivial holomorphic curve consisting of two disks, so $\#\mathcal{M}^{\chi=2}(T_1, T_1, T_{\mathrm{id}}) = 1$.

If $\chi < 2$, then $\dot{F}'_{(12)}$ bubbles off as a triangle as $d \to 0$. The projection of $\dot{F}'_{(12)}$ under $\pi_{T^*I_2} \circ u$ is a homeomorphism to the triangle $\{q''_1, q''_2, q''_3\}$. Hence the projection of $\dot{F}'_{(3)}$ under $\pi_{T^*I_2} \circ u$ is the constant map to q''_3. Since $\pi_{T^*I_2} \circ u$ is of degree zero or one near q''_3, the image of $\pi_{T^*I_2} \circ u(\dot{F} \backslash (\dot{F}'_{(12)} \cup \dot{F}'_{(3)}))$ is disjoint from q''_3. It follows that $\dot{F}'_{(12)} \cup \dot{F}'_{(3)}$ is a connected component of $\dot{F}'$. Therefore $\dot{F}$ consists of two components before the degeneration, which are homeomorphically mapped to the triangles $\{q''_1, q''_2, q''_3\}$ and $\{q''_4, q''_5, q''_6\}$ under $\pi_{T^*I_2} \circ u$, respectively. So $\chi = 2$, which is a contradiction.

We next show that

$$\#\mathcal{M}^{\chi}(T_1, T_1, T_1) = \begin{cases} 1 & \text{if } \chi = 1, \\ 0 & \text{if } \chi \neq 1. \end{cases} \tag{3-4}$$

The generators are shown in Figure 7.

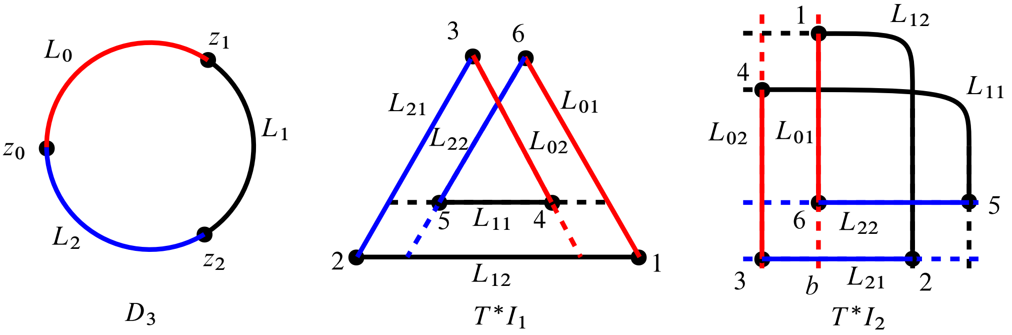


Figure 7: Generators for $\mathcal{M}(T_1, T_1, T_1)$.

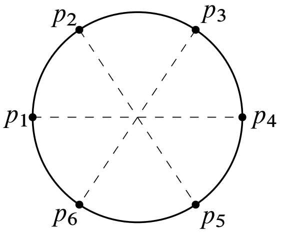


Figure 8: A disk $\dot{F} = D_6$ which satisfies the involution condition.

We denote the moduli space of domain $(\dot{F}, j)$ by $\mathcal{M}(\dot{F})$. By the Riemann–Roch formula, $\dim \mathcal{M}(\dot{F}) = 3(\kappa - \chi)$. Consider the moduli space of pseudoholomorphic maps from $(\dot{F}, j)$ to each direction D_3, $T^* I_1$ and $T^* I_2$, denoted by $\mathcal{M}(D_3)$, $\mathcal{M}(T^* I_1)$ and $\mathcal{M}(T^* I_2)$, respectively. The index formula says

$$\dim \mathcal{M}(D_3) = \dim \mathcal{M}(T^* I_1) = \dim \mathcal{M}(T^* I_2) = 2(\kappa - \chi),$$

for generic J. We have

$$\mathcal{M}(T_1, T_1, T_1) = \mathcal{M}(D_3) \cap \mathcal{M}(T^* I_1) \cap \mathcal{M}(T^* I_2). \tag{3-5}$$

Our main strategy to count curves in $\mathcal{M}(T_1, T_1, T_1)$ is computing the moduli space for each direction and then counting their intersection number.

The moduli space of curves restricted to each direction has an explicit parametrization. For example, $\pi_{D_3} \circ u$ from $\dot{F}$ to D_3 is a κ-fold branched cover, and its restriction to $\partial \dot{F}$ is a κ-fold cover over S^1. Generically, $\pi_{D_3} \circ u$ is parametrized by the positions of $\kappa - \chi$ double branch points on $\dot{F}$ over D_3.

In the case $\kappa = 2$ and $\chi = 1$, $\dot{F} = D_6$ is a disk with six boundary punctures. The moduli space of $(\dot{F}, j)$ is

$$\mathcal{M}(\dot{F}) \simeq \mathbb{R}^3.$$

Then we consider the cut-out moduli space $\mathcal{M}(D_3)$, viewed as a subset of $\mathcal{M}(\dot{F})$. The deck transformation of $\pi_{D_3} \circ u$ imposes an involution condition on $\dot{F}$. In other words, we require that $\{p_i, p_{i+3}\}$ lie on a diameter for $i = 1, 2, 3$ after some fractional linear transformation. Therefore,

$$\mathcal{M}(D_3) \simeq \mathbb{R}^2.$$

The moduli space $\mathcal{M}(D_3)$ admits a compactification $\overline{\mathcal{M}}(D_3)$, which is described in Figure 9.

We first consider $\partial \mathcal{M}(D_3) \cap \mathcal{M}(T^* I_1)$; see Figure 7, center. A map in $\mathcal{M}(T^* I_1)$ may have a double branch point inside the inner region with degree two. As the branch point touches the boundary of the inner region, it is replaced by a slit with two switch points along the Lagrangians. Since we are interested in $\partial \mathcal{M}(D_3)$, the bubbling behavior in Figure 9 requires the slit to be very long, so that some switch point meets another Lagrangian. The involution condition further requires that such switch points come in pairs. We conclude that $\partial \mathcal{M}(D_3) \cap \mathcal{M}(T^* I_1)$ consists of two points: one passes q'_5 to its left extending the line segment $(q'_4 q'_5)$ and downwards extending $(q'_6 q'_5)$; the other passes p_4 to its right extending $(q'_5 q'_4)$ and downwards extending $(q'_3 q'_4)$. The two points are depicted as the two circles on the boundary of the hexagon in Figure 10.

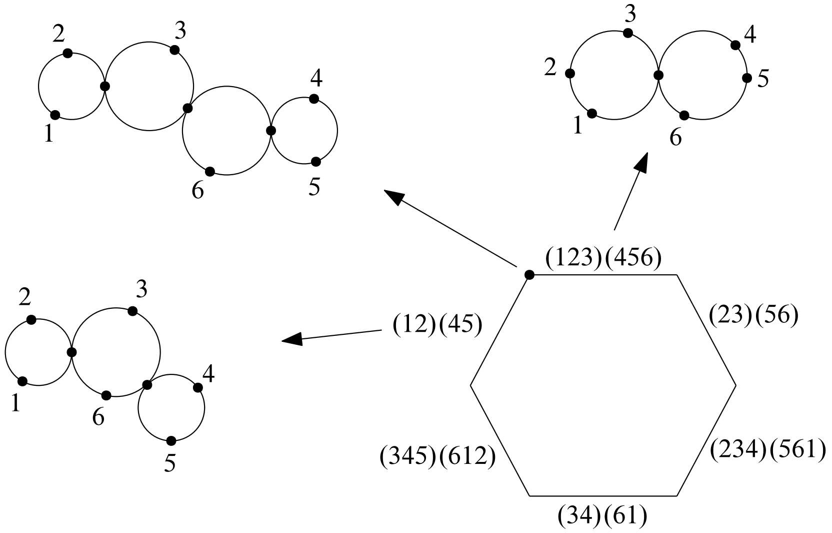


Figure 9: The compactified moduli space $\overline{\mathcal{M}}(D_3)$ is described by the hexagon. The index i stands for $p_i \in \partial \dot{F}$. The indices inside brackets describe the bubbling behavior, eg (12)(45) means p_1 is close to p_2 and p_4 is close to p_5. The involution condition is preserved on boundary strata, eg the cross ratio of the two bubble disks on the stratum (123)(456) are the same.

For $\partial \mathcal{M}(D_3) \cap \mathcal{M}(T^* I_2)$, see the Figure 7, right. Similar to the previous paragraph, the degeneration of D_6 requires the existence of long slits. There are two curves: one with a slit passing q''_6 to its left and downwards; the other lies on the Lagrangian $(q''_1 q''_2)$ or $(q''_4 q''_5)$ with one switch point meeting the intersection point $(q''_1 q''_2) \cap (q''_4 q''_5)$. The two curves are depicted as the dots on the hexagon in Figure 10.

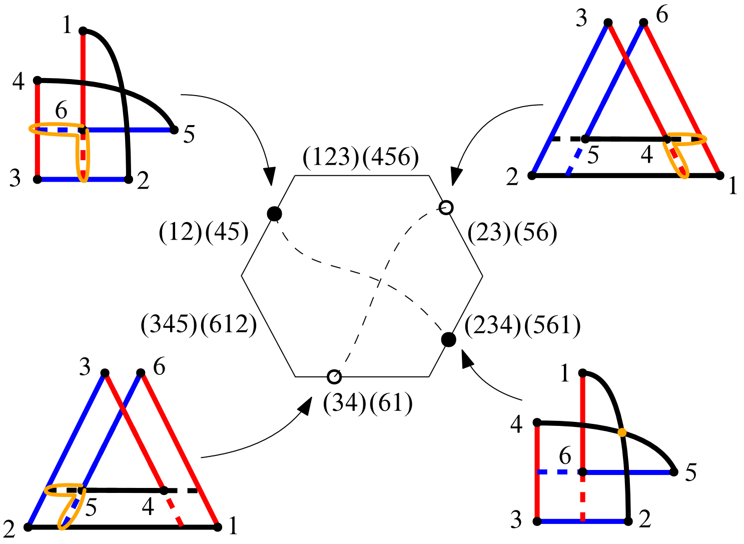


Figure 10: The orange arcs in the pictures outside the hexagon represent slits. The intersection of the two dashed arcs inside the hexagon represents a curve in $\mathcal{M}(D_3) \cap \mathcal{M}(T^* I_1) \cap \mathcal{M}(T^* I_2)$.

The relative position of dots and circles on $\partial\mathcal{M}(D_3)$ indicate $\mathcal{M}(D_3)\cap\mathcal{M}(T^*I_1)$ and $\mathcal{M}(D_3)\cap\mathcal{M}(T^*I_2)$ have intersection of algebraic count one inside $\mathcal{M}(D_3)$. Thus,

$$\#\mathcal{M}^{\chi=1}(T_1,T_1,T_1)=\#\mathcal{M}(D_3)\cap\mathcal{M}(T^*I_1)\cap\mathcal{M}(T^*I_2)=1. \tag{3-6}$$

If $\chi\neq 1$, we show that $\#\mathcal{M}^{\chi}(T_1,T_1,T_1)=0$. As $d\to 0$, $\dot{F}'_{(12)}$ bubbles off as a triangle. Since the projection of $\dot{F}'\backslash\dot{F}'_{(12)}$ to T^*I_2 is of degree one to its image (the polygon composed of $\{q''_3,q''_4,q''_5,q''_6,q''_b\}$ in Figure 7), the domain before degeneration has to be a disk. This contradicts with the fact that $\chi\neq 1$. □

The counting in (3-6) is essentially the only case where a nontrivial curve exists in our direct computation. It corresponds to the deformation $\widetilde{T}_i^2=1+\hbar\widetilde{T}_i$ from the symmetric group to the Hecke algebra.

4 The general case

In this section, we compute $\operatorname{End}(L^{\otimes\kappa})$ by induction on κ. Recall that $\operatorname{End}(L^{\otimes\kappa})$ is freely generated by $T_w=\{y_1,\dots,y_\kappa\}$, where $y_j\in L_{0j}\cap L_{1w(j)}$ and $w\in S_\kappa$ is viewed as a permutation. We compute $T_{w_1}\cdot T_{w_2}$ for $w_1,w_2\in S_\kappa$ by a case-by-case discussion depending on how w_1 acts on the last one or two elements of $\{1,\dots,\kappa\}$.

The first case is when w_1 fixes the last element. The schematic picture is shown in Figure 11. The following proposition is a generalization of Lemmas 3.2 and 3.3:

Proposition 4.1 *For $w_1,w_2,w_3\in S_\kappa$, suppose $w_1=w'_1$ and $w_2=w'_2s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-m}$, where $w'_1,w'_2\in S_{\kappa-1}$ and $m\geq 0$. We have*

$$\#\mathcal{M}^{\chi}(T_{w_1},T_{w_2},T_{w_3})=\begin{cases}\#\mathcal{M}^{\chi-1}(T_{w'_1},T_{w'_2},T_{w'_3}) & \text{if } w_3=w'_3s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-m},\\ 0 & \text{otherwise,}\end{cases}$$

where $w'_3\in S_{\kappa-1}$.

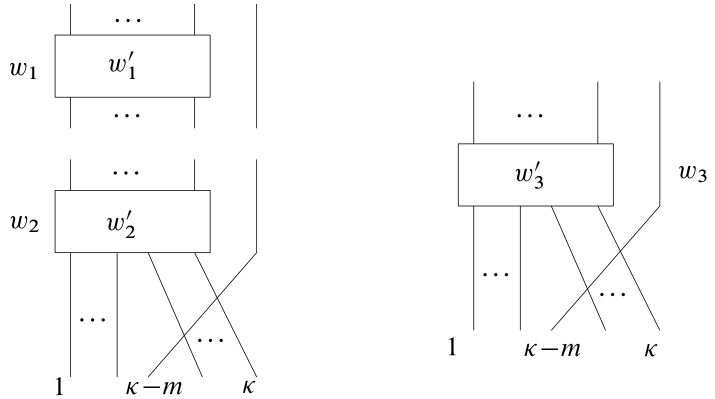


Figure 11: The case for Proposition 4.1.

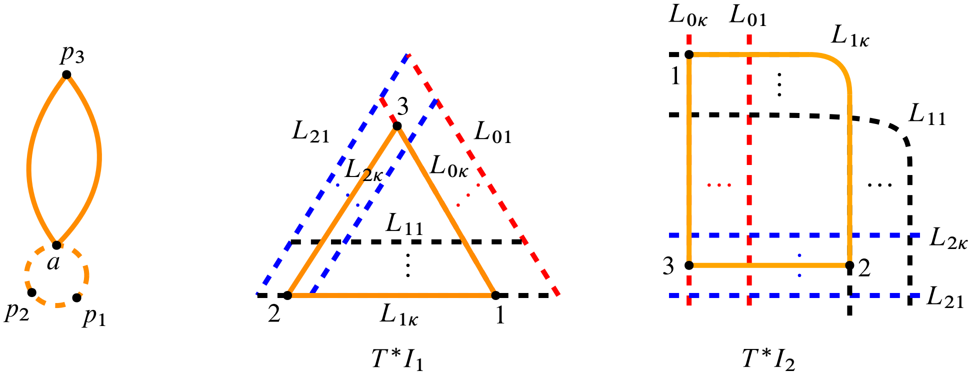


Figure 12: The case $t = \kappa - m$.

Proof Suppose that the strand of w_3 starting from the position κ ends on the position t; see Figure 11, right. Here, we are reading the picture for w_3 from top to bottom. We consider the following three cases depending on t:

(1) $(t = \kappa - m)$ Let $w_3 = w_3' s_{\kappa-1} s_{\kappa-2} \cdots s_{\kappa-m}$ for $w_3' \in S_{\kappa-1}$. This case is shown in Figure 12. The last vertical strand of w_1 in Figure 11 corresponds to p_1 in Figure 12. As $d \to 0$, $\dot F'_{(12)}$ bubbles off as a triangle with vertices $\{p_1, p_2, p_a\}$. The image of $\dot F'_{(12)}$ in the T^*I_2-direction is the orange triangle and the image of $\dot F'_{(3)}$ is the constant point q_3''. Since $\pi_{T^*I_2} \circ u$ is of degree zero or one near q_3'', the image $\pi_{T^*I_2} \circ u(\dot F \backslash (\dot F'_{(12)} \cup \dot F'_{(3)}))$ is disjoint from q_3''. This implies that $\dot F'_{(12)} \cup \dot F'_{(3)}$ is a connected component of $\dot F'$. Therefore $\dot F_{(123)}$ is a connected component of $\dot F$ before the degeneration, and it is mapped homeomorphically to the triangle $\{q_1'', q_2'', q_3''\}$ under $\pi_{T^*I_2} \circ u$. After removing the component $\dot F_{(123)}$, we see that $\#\mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3}) = \#\mathcal{M}^{\chi-1}(T_{w_1'}, T_{w_2'}, T_{w_3'})$.

(2) $(t > \kappa - m)$ This case is shown in Figure 13. As $d \to 0$, $\dot F'_{(12)}$ bubbles off as a triangle with vertices $\{p_1, p_2, p_a\}$. There is a vertex p_b in the component $\dot F'_{(3)}$ which is adjacent to p_a. Since $\pi_{T^*I_2} \circ u$ has degree zero near the intersection between the extensions of $(q_1'' q_b'')$ and $(q_2'' q_3'')$, $\dot F'_{(12)}$ cannot be a triangle. This leads to a contradiction. Therefore $\#\mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3}) = 0$.

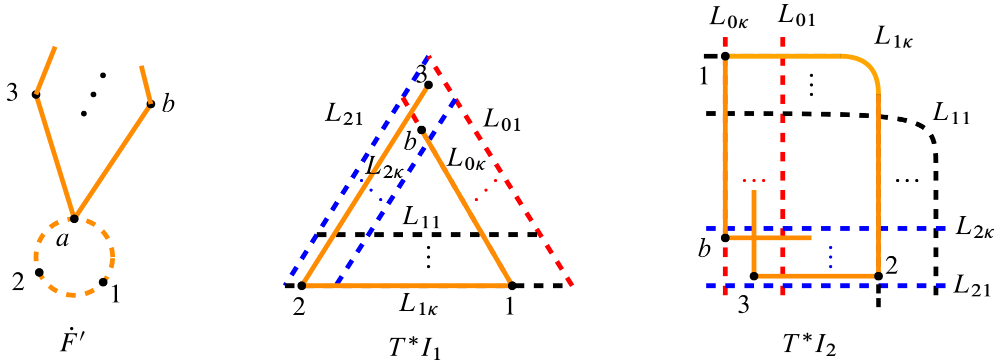


Figure 13: The case $t > \kappa - m$.

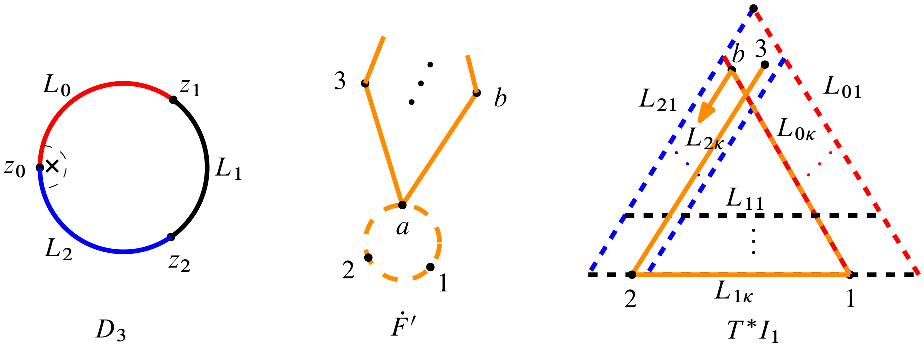


Figure 14: The case $t<\kappa-m$.

(3) $(t<\kappa-m)$ This case is shown in Figure 14. As $d\to 0$, $\dot F'_{(12)}$ bubbles off as a triangle with vertices $\{p_1,p_2,p_a\}$. On one hand, similar to the proof of $\#\mathcal{M}(T_{\mathrm{id}},T_{\mathrm{id}},T_1)=0$ of Lemma 3.2, the projection of $\dot F'_{(b)}$ under $\pi_{D_3}\circ u$ is the constant map to z_0. On the other hand, the line denoted by the orange arrow is disjoint from L_{0i} for $i=1,\dots,\kappa-1$ since $(q'_1q'_b)$ lies on $L_{0\kappa}$. So q'_b cannot be separated from the bottom left region. But the generators in this region are mapped to z_2 in the D_3-direction. We conclude that $\pi_{D_3}\circ u(\dot F'_{(b)})$ cannot be far from z_2. This is a contradiction. Therefore $\#\mathcal{M}^\chi(T_{w_1},T_{w_2},T_{w_3})=0$. □

The second case is when w_1 exchanges the last two elements. The schematic pictures are shown in Figures 15 and 16, which correspond to two subcases depending on the action of w_2 on the last two elements. The following proposition is a generalization of Lemmas 3.4 and 3.5:

Proposition 4.2 *For $w_1,w_2,w_3\in S_\kappa$, suppose that $w_1=w''_1 s_{\kappa-1}$, where $w''_1\in S_{\kappa-2}$.*

(1) *If $w_2=w''_2 s_{\kappa-2}\cdots s_{\kappa-m}s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-l}$, where $w''_2\in S_{\kappa-2}$ and $m>l\geq 0$, we have*

$$\#\mathcal{M}^\chi(T_{w_1},T_{w_2},T_{w_3})=\begin{cases}\#\mathcal{M}^{\chi-2}(T_{w''_1},T_{w''_2},T_{w''_3}) & \text{if } w_3=w''_3 s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-m}s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-l},\\ 0 & \text{otherwise,}\end{cases}$$

where $w''_3\in S_{\kappa-2}$.

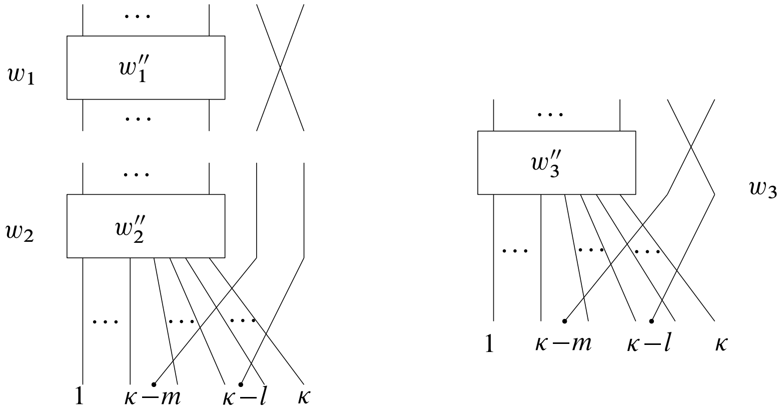


Figure 15: The case for Proposition 4.2(1).

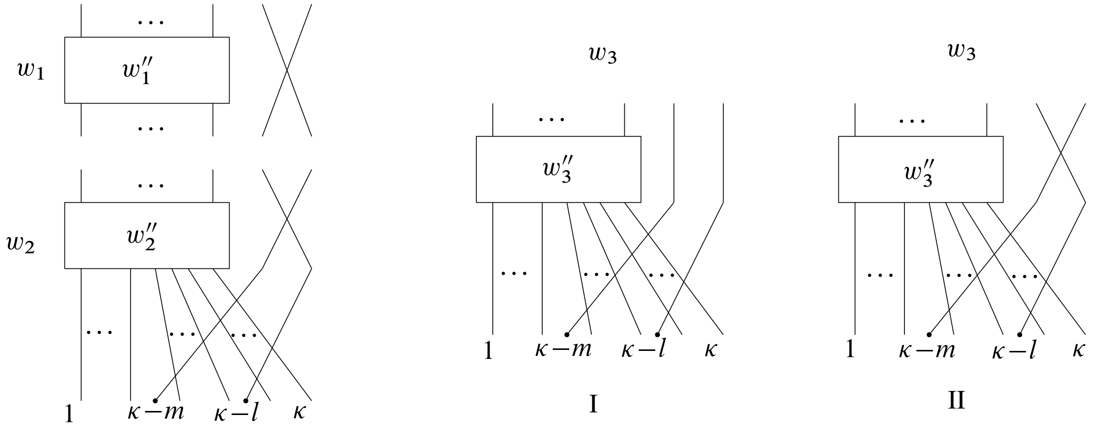


Figure 16: The case for Proposition 4.2(2).

(2) *If* $w_2 = w_2'' s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-m}s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-l}$, *where* $w_2''\in S_{\kappa-2}$ *and* $m>l\geq 0$, *we have*

$$\#\mathcal{M}^{\chi}(T_{w_1},T_{w_2},T_{w_3}) = \begin{cases} \#\mathcal{M}^{\chi-2}(T_{w_1''},T_{w_2''},T_{w_3''}) & \textit{if } w_3 = w_3'' s_{\kappa-2}\cdots s_{\kappa-m}s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-l}, \\ \#\mathcal{M}^{\chi-1}(T_{w_1''},T_{w_2''},T_{w_3''}) & \textit{if } w_3 = w_3'' s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-m}s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-l}, \\ 0 & \textit{otherwise}, \end{cases}$$

where $w_3''\in S_{\kappa-2}$.

Proof The proof is similar to but slightly longer than that of Proposition 4.1 since we need to discuss the last two strands of w_1 instead of one. We keep track of the following labels:

- the strands of w_3 which start from the positions κ and $\kappa-1$ end on positions t_1 and t_2, respectively,
- the strands of w_3 which end on the positions $\kappa-m$ and $\kappa-l$ start from positions r_1 and r_2, respectively.

Figure 17 describes the part of generators T_{w_1} and T_{w_2} corresponding to the last two strands of w_1, where the dashed circles describe the undetermined T_{w_3}. We discuss Cases 1 and 2 separately.

Case 1 ($t_1=\kappa-m$ and $t_2=\kappa-l$) This is equivalent to $w_3 = w_3'' s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-m}s_{\kappa-1}s_{\kappa-2}\cdots s_{\kappa-l}$, for some $w_3''\in S_{\kappa-2}$. Consider a holomorphic curve in $\mathcal{M}^{\chi}(T_{w_1},T_{w_2},T_{w_3})$ which contains two trivial

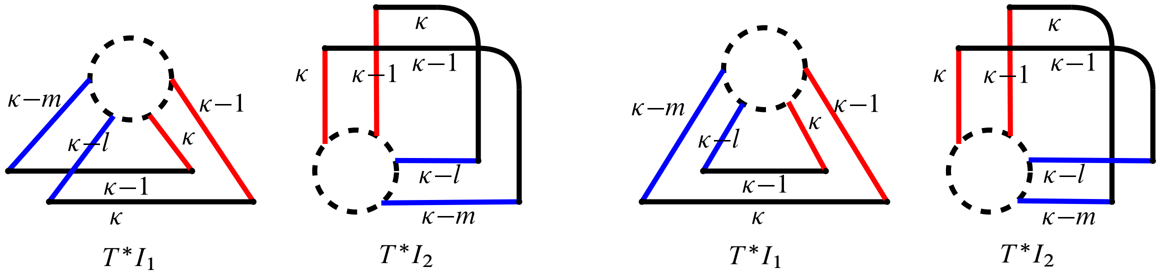
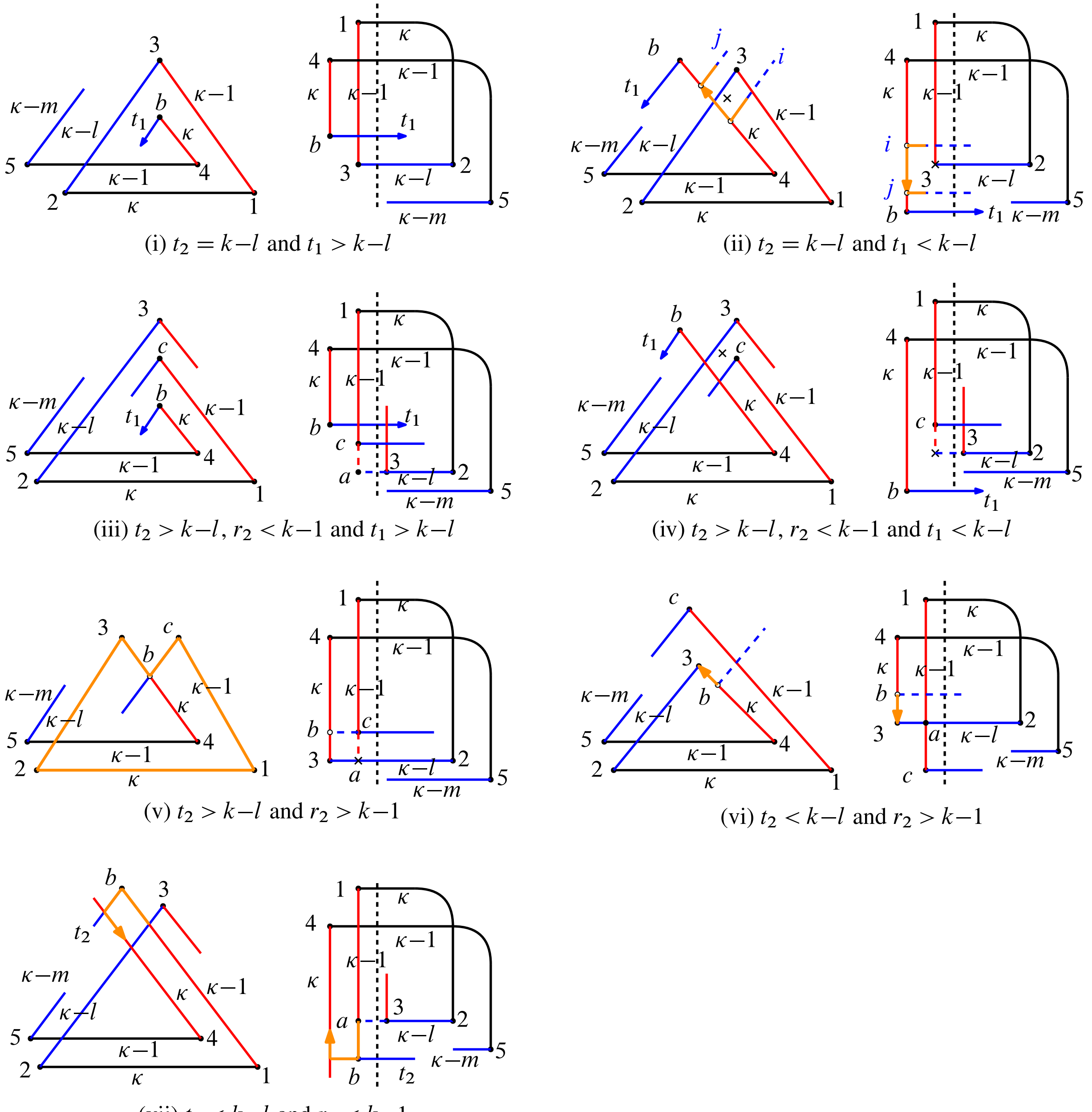


Figure 17: The part of generators T_{w_1} and T_{w_2} for Cases 1 (left) and 2 (right).



(i) $t_2 = k-l$ and $t_1 > k-l$

(ii) $t_2 = k-l$ and $t_1 < k-l$

(iii) $t_2 > k-l$, $r_2 < k-1$ and $t_1 > k-l$

(iv) $t_2 > k-l$, $r_2 < k-1$ and $t_1 < k-l$

(v) $t_2 > k-l$ and $r_2 > k-1$

(vi) $t_2 < k-l$ and $r_2 > k-1$

(vii) $t_2 < k-l$ and $r_2 < k-1$

Figure 18: The subcases of Case 1.

disks corresponding to the last two strands of w_1. The remaining components represent a curve in $\mathcal{M}^{\chi-2}(T_{w_1''}, T_{w_2''}, T_{w_3''})$. Thus $\mathcal{M}^{\chi-2}(T_{w_1''}, T_{w_2''}, T_{w_3''})$ can be viewed as a subset of $\mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3})$. We show that no other curve exists in the rest of the proof. The subcases are shown in Figure 18.

(i) ($t_2 = k-l$ and $t_1 > k-l$) As $d \to 0$, $\dot{F}_{(12)}$ bubbles off as a triangle with vertices $\{p_1, p_2, p_a\}$, where $p_a \in \dot{F}'$ is the nodal point mapped to the limit of q_1' and q_2' in the T^*I_1-direction. Then the projection of $\dot F'_{(3)}$ to the T^*I_2-direction must be the constant map to q''_3. Since $\pi_{T^*I_2}\circ u$ is of degree zero or one near q_3, the image $\pi_{T^*I_2}\circ u(\dot F\backslash(\dot F'_{(12)}\cup \dot F'_{(3)}))$ is disjoint from q''_3. This implies that $\dot F'_{(12)}\cup \dot F'_{(3)}$ is a connected component of $\dot F'$. Therefore the triangle $\{p_1,p_2,p_3\}$ forms a connected component of $\dot F$ before the degeneration. By removing the triangle $\{p_1,p_2,p_3\}$, the problem reduces to case (2) of the proof of Proposition 4.1 with $\kappa-1$ strands. Hence $\#\mathcal{M}^\chi(T_{w_1},T_{w_2},T_{w_3})=0$.

(ii) $(t_2=k-l$ and $t_1<k-l)$ As $d\to 0$, $\dot F_{(12)}$ bubbles off as a triangle with vertices $\{p_1,p_2,p_a\}$ and the projection of $\dot F'_{(3)}$ to the T^*I_2-direction must be the constant map to q''_3. Moreover $\dot F'_{(3)}$ is a bigon with possible nodal degeneration points which are connected to other irreducible components of $\dot F'$. Denote one such nodal point on $\dot F'$ by p_n, whose images in T^*I_1 and T^*I_2 are drawn as the crossings in Figure 18(ii). We now remove the bigon $\dot F'_{(3)}$ from $\dot F'$ but keep p_n. We denote the irreducible component containing p_n in the remaining part of $\dot F'$ by $\dot F'_{p_n}$.

In the T^*I_2-direction, the projection of $u(\dot F'\backslash \dot F'_{(3)})$ to the left side of the vertical dotted line is of degree one. Let C be the boundary of the image $\pi_{T^*I_2}\circ(\dot F'_{p_n})$. Then the part of C near $L_{0\kappa}\cap L_{2(\kappa-l)}$ is locally drawn as the orange lines which go from L_{2i} to L_{2j} on $L_{0\kappa}$ for $i>\kappa-l$ and $j<\kappa-m$. We denote the preimage of the orange arrow from L_{2i} to L_{2j} by C_{arrow}. It has the positive boundary orientation.

In the T^*I_1-direction, the position of the crossing must be above $L_{0\kappa}$ since $\pi_{D_3}\circ u(p_n)=z_0$. However, the image of C_{arrow}, denoted by the orange arrow, has the negative boundary orientation. This leads to a contradiction. Therefore $\#\mathcal{M}^\chi(T_{w_1},T_{w_2},T_{w_3})=0$.

(iii) $(t_2>k-l$, $r_2<k-1$ and $t_1>k-l)$ As $d\to 0$, $\dot F_{(12)}$ bubbles off as a triangle with vertices $\{p_1,p_2,p_a\}$. Since on T^*I_2, $\pi_{T^*I_2}\circ u$ is of degree zero near the intersection of the extension of $(q''_1q''_c)$ and $(q''_2q''_3)$, $\{p_1,p_2,p_a\}$ cannot form a triangle. This leads to a contradiction. Therefore $\#\mathcal{M}^\chi(T_{w_1},T_{w_2},T_{w_3})=0$.

(iv) $(t_2>k-l$, $r_2<k-1$ and $t_1<k-l)$ As $d\to 0$, $\dot F_{(12)}$ bubbles off as a triangle with vertices $\{p_1,p_2,p_a\}$, where p_a is mapped to a point in T^*I_2, denoted by a crossing. We denote the preimage of this crossing in the irreducible component other than $\dot F_{(12)}$ and $\dot F_{(3)}$ by p_n. The image of p_n in T^*I_1 is also denoted by a crossing. It sits above $L_{0\kappa}$ for the same reason as in (ii). The remaining argument is the same as in (ii). We conclude that $\#\mathcal{M}^\chi(T_{w_1},T_{w_2},T_{w_3})=0$.

(v) $(t_2>k-l$ and $r_2>k-1)$ As $d\to 0$, $\dot F_{(12)}$ bubbles off as a triangle T with vertices $\{p_1,p_2,p_a\}$, where p_a is mapped to the crossing in T^*I_2. The other irreducible component of $\dot F'$ containing p_a is the quadrilateral Q with vertices $\{p_3,p_c,p_a,p_b\}$, which is the bottom-left part in the T^*I_2-direction. Figure 19 describes the degenerated domain $\dot F'$.

Removing T and Q from $\dot F'$ corresponds to removing the orange polygon in the T^*I_1-direction. As a result, the vertices $\{p_1,p_2,p_3,p_c\}$ are replaced by p_b. Then the problem is reduced to case (2) of the proof of Proposition 4.1 with $\kappa-1$ strands. Hence, $\#\mathcal{M}^\chi(T_{w_1},T_{w_2},T_{w_3})=0$.

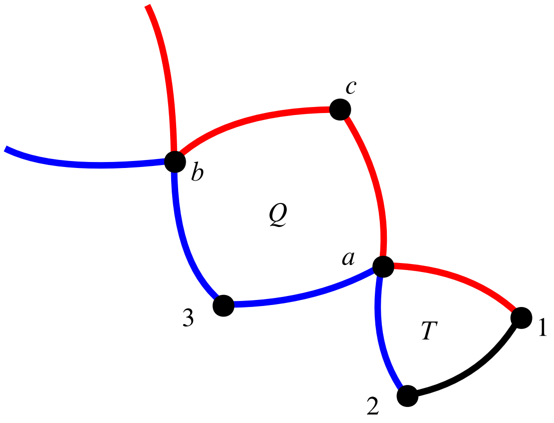


Figure 19: The subcase (v).

(vi) ($t_2 < k - l$ and $r_2 > k - 1$) This is similar to (ii). The orientation of the orange arrows leads to a contradiction. Hence $\#\mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3}) = 0$.

(vii) ($t_2 < k - l$ and $r_2 < k - 1$) This is similar to (ii). So $\#\mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3}) = 0$.

Case 2 The subcases are shown in Figure 20. The proofs of all subcases are similar to those in Case 1 except for (v). We discuss (v) only and omit the others.

(v) ($t_2 > k - m$ and $r_1 > k - 1$) The proof is similar to that of subcase (v) of Case 1. As $d \to 0$, $\dot{F}_{(12)}$ bubbles off as a triangle T with vertices $\{p_1, p_2, p_a\}$, where p_a is mapped to the crossing in T^*I_2, and $\{p_3, p_c, p_a, p_b\}$ forms a quadrilateral Q, as the bottom-left part in the T^*I_2-direction. Figure 19 describes the degenerated domain $\dot{F}'$. Removing T and Q from $\dot{F}'$ corresponds to removing the orange polygon in the T^*I_1-direction. As a result, the vertices $\{p_1, p_2, p_3, p_c\}$ are replaced by p_b. Then the problem is reduced to the case with $\kappa - 1$ strands. There are three possibilities:

(a) ($t_2 = \kappa - l$) This is similar to (1) in the proof of Proposition 4.1. If the limiting curve exists, then $\{p_1, p_2, p_3, p_4, p_5, p_c\}$ must forms a (hexagon) disk component H of $\dot{F}$. The count of $u \in \mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3})$ restricted to H is exactly the count of $\mathcal{M}^{\chi=1}(T_1, T_1, T_1)$ in Lemma 3.5, which is equals one. The count of u restricted to $\dot{F} \backslash H$ is the count of $\mathcal{M}^{\chi-1}(T_{w_1''}, T_{w_2''}, T_{w_3''})$. Therefore $\#\mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3}) = \#\mathcal{M}^{\chi-1}(T_{w_1''}, T_{w_2''}, T_{w_3''})$.

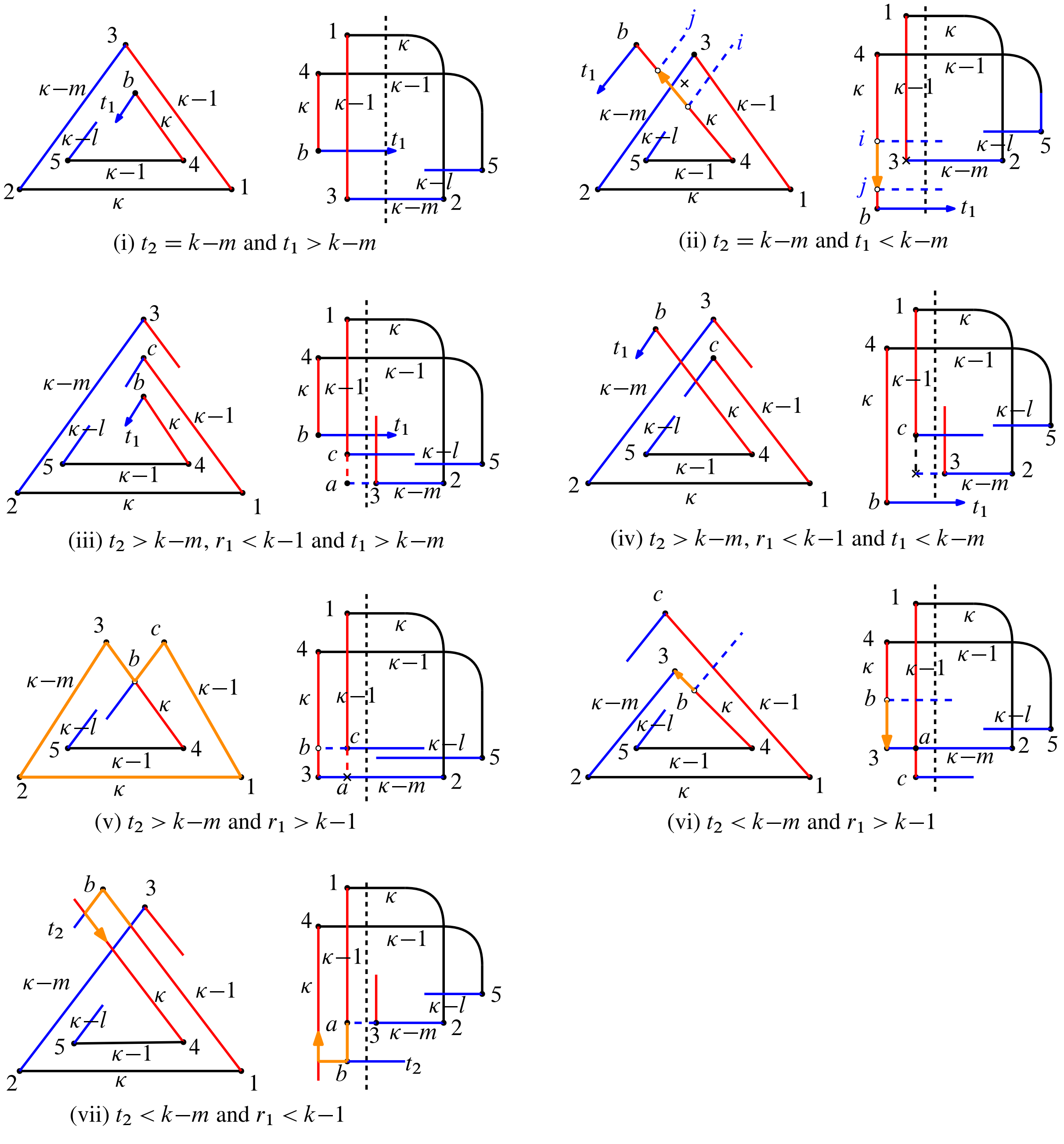
(b) ($t_2 > \kappa - l$) This is similar to (2) in the proof of Proposition 4.1. So $\#\mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3}) = 0$.

(c) ($t_2 < \kappa - l$) This is similar to (3) in the proof of Proposition 4.1. So $\#\mathcal{M}^{\chi}(T_{w_1}, T_{w_2}, T_{w_3}) = 0$. □

The following corollaries are direct consequences by inductively using the two propositions above.

Corollary 4.3 *The generator* T_{id} *is the identity in* $\mathrm{End}(L^{\otimes \kappa})$.

Corollary 4.4
$$T_i T_w = \begin{cases} T_{s_i w} & \text{if } l(s_i w) > l(w), \\ T_{s_i w} + \hbar T_w & \text{if } l(s_i w) < l(w). \end{cases}$$



(i) $t_2 = k-m$ and $t_1 > k-m$

(ii) $t_2 = k-m$ and $t_1 < k-m$

(iii) $t_2 > k-m$, $r_1 < k-1$ and $t_1 > k-m$

(iv) $t_2 > k-m$, $r_1 < k-1$ and $t_1 < k-m$

(v) $t_2 > k-m$ and $r_1 > k-1$

(vi) $t_2 < k-m$ and $r_1 > k-1$

(vii) $t_2 < k-m$ and $r_1 < k-1$

Figure 20: The subcases of Case 2.

Corollary 4.5 *The generators T_i satisfy the relations in the Hecke algebra*

(4-1) $$T_i^2 = 1 + \hbar T_i,$$

(4-2) $$T_i T_j = T_j T_i \quad \text{for } |i-j| > 1,$$

(4-3) $$T_i T_{i+1} T_i = T_{i+1} T_i T_{i+1}.$$

Proof of Theorem 1.2 Define a unital $\mathbb{Z}[\hbar]$-algebra map $\phi\colon H_\kappa \to \mathrm{End}(L^{\otimes\kappa})$ on the algebra generators by $\phi(\widetilde{T}_i) = T_i$. The map is well defined by Corollary 4.5. The multiplication rules on H_κ in (1-1) and that of $\mathrm{End}(L^{\otimes\kappa})$ in Corollary 4.4 are the same. So $\phi(\widetilde{T}_w) = T_w$ for all $w \in S_\kappa$. □

References

[1] **A Abbondandolo**, **M Schwarz**, *Floer homology of cotangent bundles and the loop product*, Geom. Topol. 14 (2010) 1569–1722 MR Zbl

[2] **M Abouzaid**, *On the wrapped Fukaya category and based loops*, J. Symplectic Geom. 10 (2012) 27–79 MR Zbl

[3] **V Colin**, **K Honda**, **Y Tian**, *Applications of higher-dimensional Heegaard Floer homology to contact topology*, J. Topol. 17 (2024) art. id. e12349 MR Zbl

[4] **T Ekholm**, **V Shende**, *Skeins on branes*, preprint (2019) arXiv 1901.08027

[5] **S Ganatra**, **J Pardon**, **V Shende**, *Covariantly functorial wrapped Floer theory on Liouville sectors*, Publ. Math. Inst. Hautes Études Sci. 131 (2020) 73–200 MR Zbl

[6] **S Ganatra**, **J Pardon**, **V Shende**, *Sectorial descent for wrapped Fukaya categories*, J. Amer. Math. Soc. 37 (2024) 499–635 MR Zbl

[7] **K Honda**, **Y Tian**, **T Yuan**, *Higher-dimensional Heegaard Floer homology and Hecke algebras*, J. Eur. Math. Soc. (online publication August 2024)

[8] **V F R Jones**, *Hecke algebra representations of braid groups and link polynomials*, Ann. of Math. 126 (1987) 335–388 MR Zbl

[9] **H R Morton**, **P Samuelson**, *DAHAs and skein theory*, Comm. Math. Phys. 385 (2021) 1655–1693 MR Zbl

[10] **Z Sylvan**, *On partially wrapped Fukaya categories*, J. Topol. 12 (2019) 372–441 MR Zbl

School of Mathematical Sciences, Beijing Normal University
Beijing, China

School of Mathematical Sciences, Eastern Institute of Technology, Ningbo
Zhejiang, China

yintian@bnu.edu.cn, tyyuan@eitech.edu.cn

Received: 24 August 2023 Revised: 24 November 2023

Algebraic & Geometric Topology **25**:3 (2025) 1897–1915
DOI: 10.2140/agt.2025.25.1897
Published: 20 June 2025

Fibered 3-manifolds and Veech groups

CHRISTOPHER J LEININGER
KASRA RAFI
NICHOLAS ROUSE
EMILY SHINKLE
YVON VERBERNE

We study Veech groups associated to the pseudo-Anosov monodromies of fibers and foliations of a fixed hyperbolic 3-manifold. Assuming Lehmer's conjecture, we prove that the Veech groups associated to fibers generically contain no parabolic elements. For foliations, we prove that the Veech groups are always elementary.

57K32; 57K20

1 Introduction

A pseudo-Anosov homeomorphism $f : S \to S$ on an orientable surface determines a complex structure and holomorphic quadratic differential, (X, q), up to Teichmüller deformation, for which the vertical and horizontal foliations are the stable and unstable foliations of f. The pseudo-Anosov generates an infinite cyclic subgroup of the full group of orientation preserving affine homeomorphisms, $\mathrm{Aff}_+(X, q)$.

For a finite type surface S, we say that the pseudo-Anosov homeomorphism f is *lonely* if $\langle f \rangle < \mathrm{Aff}_+(S, q)$ has finite index. The motivation for this paper is the following; see eg Hubert, Masur, Schmidt and Zorich [11] and Lanneau [15]

Conjecture 1.1 (lonely pseudo-Anosov) *There exist lonely pseudo-Anosov homeomorphisms. In fact, lonely pseudo-Anosov homeomorphisms are generic.*

There is not an agreed upon notion of "generic", and some care must be taken: work of Calta [2] and McMullen [19; 20] shows that *no* pseudo-Anosov homeomorphism on a surface of genus 2, with orientable stable/unstable foliation is lonely. In fact, in this case, not only are the pseudo-Anosov homeomorphisms not lonely, but their Veech groups always contain parabolic elements.

In this paper, we consider infinite families of pseudo-Anosov homeomorphisms arising as follows; see Section 2.1. Suppose $f : S \to S$ is a pseudo-Anosov homeomorphism of a finite type surface S and M_f is the mapping torus (which is hyperbolic by Thurston's hyperbolization theorem; see Otal [21]).

The connected cross sections of the suspension flow are organized by their cohomology classes (up to isotopy), which are primitive integral classes in the cone on the open fibered face $F \subset H^1(M, \mathbb{R})$ of the Thurston norm ball containing the Poincaré–Lefschetz dual of the fiber S. Given such an integral class α, the first return map to the cross section S_α is a pseudo-Anosov homeomorphism $f_\alpha\colon S_\alpha \to S_\alpha$. When $b_1(M) > 1$, there are infinitely many such pseudo-Anosov homeomorphisms; in fact, $|\chi(S_\alpha)|$ is a linear function of α, and hence tends to infinity with α.

We let $\bar{\alpha} \in F$ denote the projection of the primitive integral class α in the cone over F, and let $F_{\mathbb{Q}}$ be the set of all such projections, which is precisely the (dense) set of rational points in F.

Question 1.2 Given a fibered hyperbolic 3-manifold and fibered face F, are the pseudo-Anosov homeomorphisms f_α for $\bar{\alpha} \in F_{\mathbb{Q}}$ generically lonely?

We will provide two pieces of evidence that the answer to this question is "yes". Write $\mathrm{Aff}_+(X_\alpha, q_\alpha)$ for the orientation preserving affine group containing f_α; see Section 2.3 for more details.

Theorem 1.3 *Suppose F is the fibered face of an orientable, fibered, hyperbolic 3-manifold. Assuming Lehmer's conjecture, the set of $\bar{\alpha} \in F_{\mathbb{Q}}$ such that $\mathrm{Aff}_+(X_\alpha, q_\alpha)$ contains a parabolic element is discrete in F.*

In certain examples, the set of classes whose associated Veech group contains parabolics is actually finite (again, assuming Lehmer's conjecture); see Theorem 4.2. In Section 3 we describe some explicit computations that illustrate this finite set. If M is the orientation cover of a nonorientable, fibered 3-manifold, then the conclusion of Theorem 1.3 holds on the invariant cohomology of the covering involution without assuming the validity of Lehmer's conjecture; see Theorem 4.3.

Much of the defining structure survives for nonintegral classes $\alpha \in F - F_{\mathbb{Q}}$; see Section 2.2 for details. Briefly, we first recall that every $\alpha \in F - F_{\mathbb{Q}}$ is represented by a closed 1-form ω_α which is positive on the vector field generating the suspension flow. The kernel of ω_α is tangent to a foliation $\mathscr{F}_\alpha$, and the flow can be reparameterized to send leaves of $\mathscr{F}_\alpha$ to other leaves. There is no longer a first return time, but rather a *higher rank abelian group* of return times, H_α, to any given leaf S_α of $\mathscr{F}_\alpha$. Work of McMullen [18] associates a *leafwise* complex structure and quadratic differential (X_α, q_α) to each $\alpha \in F - F_{\mathbb{Q}}$ such that the leaf-to-leaf maps of the flow are all Teichmüller maps. For every leaf S_α of $\mathscr{F}_\alpha$, the return maps to S_α thus determine an isomorphism from $H_\alpha < \mathbb{R}$ to a subgroup we denote by $H_\alpha^{\mathrm{Aff}} < \mathrm{Aff}_+(X_\alpha, q_\alpha)$, an abelian group of pseudo-Anosov elements. Our second piece of evidence for a positive answer to Question 1.2 is the following.

Theorem 1.4 *If F is a fibered face of a closed, orientable, fibered, hyperbolic 3-manifold, then for all $\alpha \in F - F_{\mathbb{Q}}$, and any leaf S_α of $\mathscr{F}_\alpha$, the abelian group $H_\alpha^{\mathrm{Aff}} < \mathrm{Aff}_+(X_\alpha, q_\alpha)$ has finite index.*

For $\alpha \in F - F_{\mathbb{Q}}$, the leaves S_α are infinite type surfaces. In general, there is much more flexibility in constructing affine groups for infinite type surfaces, and exotic groups abound. Indeed, work of Przytycki, Schmithüsen and Valdez [22] and Ramírez Maluendas and Valdez [23] proves that *any* countable subgroup of $\mathrm{GL}_2(\mathbb{R})$ without contractions is the derivative-image of some affine group. (See also Bowman [1] for a "naturally occurring" lonely pseudo-Anosov homeomorphism on an infinite type surface of finite area.) Theorem 1.4 says that for the leaves S_α of the foliations and their associated quadratic differentials, the situation is much more rigid.

Acknowledgements

The authors would like to thank Erwan Lanneau, Livio Liechti, Alan Reid and Ferrán Valdez for helpful conversations, and the referee for suggestions that improved the readability. We are particularly grateful to Liechti for suggesting Theorem 4.3. Leininger was partially supported by NSF grant DMS-2106419. Rafi was partially supported by NSERC Discovery grant RGPIN 06486. Verberne was partially supported by an NSERC-PDF fellowship.

2 Definitions and background

2.1 Fibered 3-manifolds

Here we explain the set up and background for our work in more detail. For a pseudo-Anosov homeomorphism $f\colon S \to S$ of an orientable, finite type surface S, let $\lambda(f)$ denote its *stretch factor* (also called its *dilatation*); see [3]. We write

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

to denote the mapping torus of the pseudo-Anosov homeomorphism f. The suspension flow ψ_s of f is generated by the vector field $\xi = \frac{\partial}{\partial t}$, where t is the coordinate on the $[0,1]$ factor. Alternatively, we have the local flow of the same name $\psi_s(x,t) = (x, t+s)$ on $S \times [0,1]$, defined for $t, s+t \in [0,1]$, which descends to the suspension flow.

A *cross section* (or just *section*) of the flow is a surface $S_\alpha \subset M$ transverse to ξ, such that for all $x \in S_\alpha$, $\psi_s(x) \in S_\alpha$ for some $s > 0$. If $s(x) > 0$ is the smallest such number, then the *first return map* of ψ_s is the map $f_\alpha\colon S_\alpha \to S_\alpha$ defined by $f_\alpha(x) = \psi_{s(x)}(x)$ for $x \in S_\alpha$. Note that $S(= S \times \{0\}) \subset M$ is a section, and the first return map to S is precisely the map $f = \psi_1|_S$.

Cutting open along an arbitrary section S_α we get a product $S_\alpha \times [0,1]$ where the slices $\{x\} \times [0,1]$ are arcs of flow lines. Thus, M can also be expressed as the mapping torus of f_α, or alternatively, M fibers over the circle with *monodromy* f_α. Up to isotopy, the fiber S_α is determined by its Poincaré–Lefschetz dual cohomology class $\alpha = [S_\alpha] \in H^1(M;\mathbb{Z}) \subset H^1(M;\mathbb{R}) = H^1(M)$. To see how these are organized, we first recall the following theorem of Thurston [27]

Theorem 2.1 *For $M = M_f$ as above, there is a finite union of open, convex, polyhedral cones $\mathscr{C}_1, \dots, \mathscr{C}_k \subset H^1(M)$ such that $\alpha \in H^1(M;\mathbb{Z})$ is dual to a fiber in a fibration over S^1 if and only* *if $\alpha \in \mathscr{C}_j$ for some j. Moreover, there is a norm $\|\cdot\|_T$ on $H^1(M)$ so that for each $\mathscr{C}_j$, $\|\cdot\|_T$ restricted to $\mathscr{C}_j$ is linear, and if $\alpha \in \mathscr{C}_j \cap H^1(M;\mathbb{Z})$ then $\|\alpha\|_T$ is the negative of the Euler characteristic of the fiber dual to α.*

The unit ball $\mathfrak{B}$ of $\|\cdot\|_T$ is a polyhedron, and each $\mathscr{C}_j$ is the cone over the interior of a top dimensional face F_j of $\mathfrak{B}$.

The cones in the theorem are called the *fibered cones* of M and the F_j the *fibered faces* of $\mathfrak{B}$. It follows from Thurston's proof of Theorem 2.1 that each of the sections S_α of (ψ_s) described above must lie in a single one of the fibered cones $\mathscr{C}$ over a fibered face F. The following theorem elaborates on this, combining results of Fried from [5; 6].

Theorem 2.2 *For $M = M_f$ as above, there is a fibered cone $\mathscr{C} \subset H^1(M)$ such that $\alpha \in H^1(M;\mathbb{Z})$ is dual to a section of (ψ_s) if and only if $\alpha \in \mathscr{C}$. Moreover, there is a function $\mathfrak{h}\colon \mathscr{C} \to \mathbb{R}_+$ which is continuous, convex, and homogenous of degree -1, with the following properties.*

- *For any $\alpha \in \mathscr{C} \cap H^1(M;\mathbb{Z})$, f_α is pseudo-Anosov and $\mathfrak{h}(\alpha) = \log(\lambda(f_\alpha))$.*
- *For any $\{\alpha_n\} \subset \mathscr{C}$ with $\alpha_n \to \partial\mathscr{C}$, we have $\mathfrak{h}(\alpha_n) \to \infty$.*

We let $\mathscr{C}_\mathbb{Z} \subset \mathscr{C}$ denote the primitive integral classes in the fibered cone $\mathscr{C}$; that is, the integral points which are not nontrivial multiples of another element of $H^1(M;\mathbb{Z})$. These correspond precisely to the connected sections of (ψ_s).

McMullen [18] refined the analysis of $\mathfrak{h}$, proving for example that it is actually real-analytic. For this, he computed the stretch factors using his *Teichmüller polynomial* $\Theta_\mathscr{C}$. This polynomial

$$\Theta_\mathscr{C} = \sum_{g \in G} a_g g$$

is an element of the group ring $\mathbb{Z}[G]$ where $G = H_1(M;\mathbb{Z})/\text{torsion}$. For $\alpha \in \mathscr{C}_\mathbb{Z}$, the *specialization* of the Teichmüller polynomial is

$$\Theta_\mathscr{C}^\alpha(t) = \sum_{g \in G} a_g t^{\alpha(g)} \in \mathbb{Z}[t^{\pm 1}]$$

where we view $\alpha \in H^1(M;\mathbb{Z}) \cong \mathrm{Hom}(G;\mathbb{Z})$. Further, $G \cong H \oplus \mathbb{Z}$ where $H = \mathrm{Hom}(H^1(S,\mathbb{Z})^f, \mathbb{Z}) \cong \mathbb{Z}^m$ and $H^1(S,\mathbb{Z})^f$ are the f-invariant cohomology classes. So we can regard $\Theta_\mathscr{C}$ as a Laurent polynomial on the generators $x_1, x_2, \ldots, x_m$ of H and the generator u of $\mathbb{Z}$. Then specialization to the dual of an element $(a_1, a_2, \ldots, a_m, b) \in \mathscr{C} \cap H^1(M;\mathbb{Z})$ amounts to setting $x_i = t^{a_i}$ for $1 \le i \le m$ and $u = t^b$. McMullen proves that the specializations and the pseudo-Anosov first return maps are related by the following.

Theorem 2.3 *For any $\alpha \in \mathscr{C}_\mathbb{Z}$, the stretch factor $\lambda(f_\alpha)$ is a root of $\Theta_\mathscr{C}^\alpha$ with the largest modulus.*

Combining the linearity of $\|\cdot\|_T$ on $\mathscr{C}$ together with the homogeneity of $\mathfrak{h}$, we have the following observation of McMullen; see [18].

Corollary 2.4 *The function $\alpha \mapsto \|\alpha\|_T \mathfrak{h}(\alpha)$ is continuous and constant on rays from 0. In particular, if $K \subset \mathscr{C}$ is any compact subset, then $\|\cdot\|_T \mathfrak{h}(\cdot)$ is bounded on $\mathbb{R}_+ K$.*

The key corollary for us is the following, also observed by McMullen in the same paper.

Corollary 2.5 *If $\{\alpha_n\}_n \subset \mathscr{C}_\mathbb{Z}$ is any infinite sequence of distinct elements, then $|\chi(S_{\alpha_n})| \to \infty$, and if the rays $\mathbb{R}_+\alpha_n$ do not accumulate on $\partial\mathscr{C}$, then*

$$\log(\lambda(f_{\alpha_n})) \asymp \frac{1}{|\chi(S_{\alpha_n})|}.$$

In particular, $\lambda(f_{\alpha_n}) \to 1$.

Remark 2.6 One can sometimes promote the final conclusion to *any* infinite sequence of distinct elements, without the assumption about nonaccumulation to $\partial\mathscr{C}$; see the examples in Section 3. This is not always the case, and the accumulation set of stretch factors can be fairly complicated, as described by work of Landry, Minsky and Taylor [14].

2.2 Foliations in the fibered cone

Fried's work described above [5; 6] implies that any $\alpha \in \mathscr{C}$ may be represented by a closed 1-form ω_α for which $\omega_\alpha(\xi) > 0$ at every point of M. For integral classes, ω_α is the pull-back of the volume form from the fibration over the circle $\mathbb{R}/\mathbb{Z}$, and in general, ω_α is a convex combination of such 1-forms. The kernel of ω_α defines a foliation $\mathscr{F}_\alpha$ transverse to ξ whose leaves are injectively immersed surfaces $S_\alpha \subset M$. We consider the reparameterized flow $\{\psi_s^\alpha\}$ defined by scaling the generating vector field ξ by $\xi/\omega_\alpha(\xi)$. Then for every leaf $S_\alpha \subset M$ of $\mathscr{F}_\alpha$ and for every $s \in \mathbb{R}$, the image by the flow $\psi_s^\alpha(S_\alpha)$ is another leaf of $\mathscr{F}_\alpha$. The subgroup $H_\alpha < \mathbb{R}$ mentioned in the introduction is precisely the set of return times of ψ_s^α to S_α. As such, H_α acts on S_α so that $s \in H_\alpha$ acts by $s \cdot x = \psi_s^\alpha(x)$, for all $x \in S_\alpha$.

The group $H_\alpha \cong \mathbb{Z}^n$ for some $n = n_\alpha \le b_1(M)$, and can alternatively be defined as the set of periods of α (ie the α-homomorphic image of $H_1(M;\mathbb{Z})$). A leaf S_α is a closed surface, and in fact a fiber as above if and only if $n_\alpha = 1$ in which case H_α is a discrete subgroup of $\mathbb{R}$ and $\bar{\alpha} \in F_\mathbb{Q}$. On the other hand, $n_\alpha \ge 2$ if and only if the group of return times H_α is indiscrete, and so S_α is *dense* in M.

2.3 Teichmüller flows and Veech groups

In [18], McMullen defines a conformal structure and quadratic differential, (X_α, q_α), on the leaves S_α of the foliation $\mathscr{F}_\alpha$, for all $\alpha \in \mathscr{C}$, with the following properties. For each $s \in \mathbb{R}$ and leaf S_α, the leaf-to-leaf map $\psi_s^\alpha \colon S_\alpha \to \psi_s^\alpha(S_\alpha)$ is a Teichmüller map with initial/terminal quadratic differentials given by q_α on the respective leaves. In fact, there exists some $K_\alpha > 1$ such that ψ_s^α is a $K_\alpha^{|s|}$-Teichmüller map, and hence $K_\alpha^{2|s|}$-quasiconformal.

Remark 2.7 The notation (X_α, q_α) is somewhat ambiguous: this really denotes a family of structures, one on every leaf, though we abuse notation and also use this same notation to denote the restriction to any given leaf.

The vertical and horizontal foliations of q_α on the leaves S_α of $\mathscr{F}_\alpha$ are obtained by intersecting with a *fixed* singular foliation on the 3-manifold; namely, the suspension of the unstable/stable foliations for the original pseudo-Anosov homeomorphism f. In particular, the cone points (ie zeros) of q_α are precisely the intersections of S_α with the ψ_s-flowlines through the cone points on the original surface S. Consequently, the cone points are isolated, and the cone angles are bounded by those of the original surface, and are hence bounded independent of α.

For $s \in H_\alpha$, $\psi_s^\alpha\colon S_\alpha \to S_\alpha$ is (a remarking) of the Teichmüller map, and thus an affine pseudo-Anosov homeomorphism with respect to q_α. In this way, we obtain an isomorphism from H_α to a subgroup $H_\alpha^{\mathrm{Aff}} < \mathrm{Aff}_+(X_\alpha, q_\alpha)$, the group of orientation preserving affine homeomorphisms of the leaf S_α with respect to (X_α, q_α). The derivative with respect to the preferred coordinates defines a map

$$D_\alpha\colon \mathrm{Aff}_+(X_\alpha, q_\alpha) \to \mathrm{GL}_2^+(\mathbb{R})/\pm I,$$

which is called the *Veech group* of (X_α, q_α). A *parabolic* element of $\mathrm{Aff}_+(X_\alpha, q_\alpha)$ is one whose image by D_α is parabolic.

Remark 2.8 The preferred coordinates for a quadratic differential are only defined up to translation and rotation through angle π, so the derivative is only defined up to sign. If all affine homeomorphisms are area preserving (eg if the surface has finite area) then the derivative maps to $\mathrm{PSL}_2(\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})/\pm I$.

Since the vertical/horizontal foliations are the stable/unstable foliations, the image of H_α^{Aff}, which we denote by $H_\alpha^D = D_\alpha(H_\alpha^{\mathrm{Aff}})$ is contained in the diagonal subgroup of $\mathrm{PSL}_2(\mathbb{R})$,

$$H_\alpha^D < \Delta = \left\{ \begin{pmatrix} a & 0 \\ 0 & \frac{1}{a} \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R}) \,\middle|\, a > 0 \right\} /\pm I.$$

Define $\mathrm{SAff}(X_\alpha, q_\alpha) < \mathrm{Aff}_+(X_\alpha, q_\alpha)$ to be the area preserving subgroup of orientation preserving affine homeomorphisms; this is the preimage of $\mathrm{PSL}_2(\mathbb{R})$ under D_α. In particular, $H_\alpha^{\mathrm{Aff}} < \mathrm{SAff}(X_\alpha, q_\alpha)$.

2.4 Trace fields

A number field is *totally real* if the image of every embedding into $\mathbb{C}$ lies in $\mathbb{R}$. Hubert and Lanneau [9] proved the following.

Theorem 2.9 *If a nonelementary Veech group contains a parabolic element, then the trace field is totally real.*

A pseudo-Anosov f being lonely implies that there are no parabolic elements in the Veech group, but not conversely; see [10].

McMullen [20, Corollary 9.6] proved the following fact about the trace field of a Veech group; see also Kenyon and Smillie [12].

Theorem 2.10 *The trace field of a Veech group containing a pseudo-Anosov is generated by the trace of that pseudo-Anosov. That is, the trace field is given by* $\mathbb{Q}(\lambda(f) + \lambda(f)^{-1})$.

Thus, this trace field is totally real precisely when the trace of the pseudo-Anosov has only real Galois conjugates.

Remark 2.11 Theorems 2.9 and 2.10 are proved for complex structures with an abelian differential, rather than a quadratic differential. The proof of Theorem 2.9 for the more general case of quadratic differentials follows verbatim since the key ingredient is the so-called Thurston–Veech construction, which works for both quadratic differentials and abelian differentials (see [28, Section 6]). Theorem 2.10 for quadratic differentials follows from the case of abelian differentials since every affine homeomorphism lifts to the canonical 2-fold cover where a quadratic differential pulls back to a square of an abelian differential, and thus the preimage of the Veech group of the original surface in $\mathrm{SL}_2(\mathbb{R})$ is contained in the Veech group for the abelian differential.

2.5 Lehmer's conjecture

Theorem 1.3 is dependent on the validity of what is known as Lehmer's conjecture [16] though Lehmer did not actually conjecture the statement we will use. See [26]. To state this conjecture, we need the following.

Definition 2.12 Let $p(x) \in \mathbb{C}[x]$ with factorization over $\mathbb{C}$,

$$p(x) = a_0 \prod_{i=1}^{m} (x - \gamma_i).$$

The *Mahler measure* of p is

$$\mathcal{M}(p) = |a_0| \prod_{i=1}^{m} (\max 1, |\gamma_i|).$$

With this definition, we state the conjecture we assume.

Conjecture 2.13 (Lehmer) *There is a constant $\mu > 1$ such that for every $p(x) \in \mathbb{Z}[x]$ with a root not equal to a root of unity, $\mathcal{M}(p) \geq \mu$.*

Lehmer's conjecture is known in some special cases, including the following result of Schinzel [25] which will be important in the proof of Theorem 4.3.

Theorem 2.14 *If $p(t)$ is the minimal polynomial for an algebraic integer not equal to 0 or ± 1, all of whose roots are real, then*

$$\mathcal{M}(p) \geq \left(\frac{1+\sqrt{5}}{2}\right)^{\deg(p)/2}.$$

3 Examples

Here we provide examples of fibered faces of fibered 3-manifolds and examine arithmetic features of the Veech groups of the corresponding pseudo-Anosov homeomorphisms.

3.1 Example 1

Let $\beta = \sigma_1\sigma_2^{-1}$ be an element of the braid group B_3 on three strands (viewed as the mapping class group of a four-punctured sphere, S), where σ_1 and σ_2 denote the standard generators. Let M denote the mapping torus of β. McMullen computes the Teichmüller polynomial for this manifold in detail in [18]. See also Hironaka [7].

Since β permutes the strands of the braid cyclically, $b_1(M) = 2$. Choosing appropriate bases, we obtain an isomorphism $H^1(M;\mathbb{Z}) \cong \mathbb{Z}^2$ such that the starting fiber surface S is dual to $(0,1)$, the fibered cone is

$$\mathscr{C} = \{(a,b) \in \mathbb{R}^2 \mid b > 0, -b < a < b\}$$

and the Teichmüller polynomial for this cone is

$$\Theta_{\mathscr{C}}(x,u) = u^2 - u(x + 1 + x^{-1}) - 1.$$

Specialization to an integral class $(a,b) \in \mathscr{C}_{\mathbb{Z}}$ equates to setting $x = t^a$ and $u = t^b$ and yields

$$\Theta_{\mathscr{C}}^{(a,b)}(t) = \Theta_{\mathscr{C}}(t^a, t^b) = t^{2b} - t^{b+a} - t^b - t^{b-a} + 1.$$

We used the mathematics software system SageMath [24] to factor $\Theta_{\mathscr{C}}^{(a,b)}(t)$ for all primitive integral pairs $(a,b) \in \mathscr{C}$ with $b < 50$, to determine the stretch factors $\lambda_{(a,b)}$ of the corresponding monodromies and their minimal polynomials. We then computed the conjugates of the corresponding traces, $\lambda_{(a,b)} + 1/\lambda_{(a,b)}$, to determine whether the trace field of each associated Veech group is totally real. The results are shown in Figure 1. Recall that by Theorem 2.9, when this trace field is not totally real, the Veech group has no parabolic elements.

These computations suggest that there are only finitely many pairs (a,b) where the trace field is not totally real. This is not a coincidence as we will see below. For this, we record the following improvement on Corollary 2.5 for the cone $\mathscr{C}$ for this example.

Lemma 3.1 *For any sequence $\alpha_n = (a_n, b_n) \in \mathscr{C}_{\mathbb{Z}}$ of distinct elements, we have $\lambda(f_{\alpha_n}) \to 1$.*

Proof Since $\mathfrak{h}$ is convex, the maximum value of $\mathfrak{h}(a,b) = \log(\lambda(f_{(a,b)}))$, for points $(a,b) \in \mathscr{C}_{\mathbb{Z}}$ and a fixed b, occurs at either $(b-1,b)$ or $(1-b,b)$.

First we consider the points of the form $(b-1,b)$. The specialization of $\Theta_{\mathscr{C}}$ in this case takes the form

$$\Theta_{\mathscr{C}}^{(b-1,b)}(t) = t^{2b} - t^{2b-1} - t^b - t + 1.$$

Recall that $\lambda_b = \lambda(f_{(b-1,b)}) > 1$. As $b \to \infty$, we claim that $\lambda_b \to 1$. Suppose instead that the sequence is bounded below by $1+\epsilon$, for $\epsilon > 0$ on some subsequence. Then in this subsequence we have

$$\begin{aligned}\Theta_{\mathscr{C}}^{(b-1,b)}(\lambda_b) &= \lambda_b^{2b}(1 - \lambda_b^{-1} - \lambda_b^{-b} - \lambda_b^{1-2b}) + 1 \\ &\geq (1+\epsilon)^{2b}\big(1 - (1+\epsilon)^{-1} - (1+\epsilon)^{-b} - (1+\epsilon)^{1-2b}\big).\end{aligned}$$

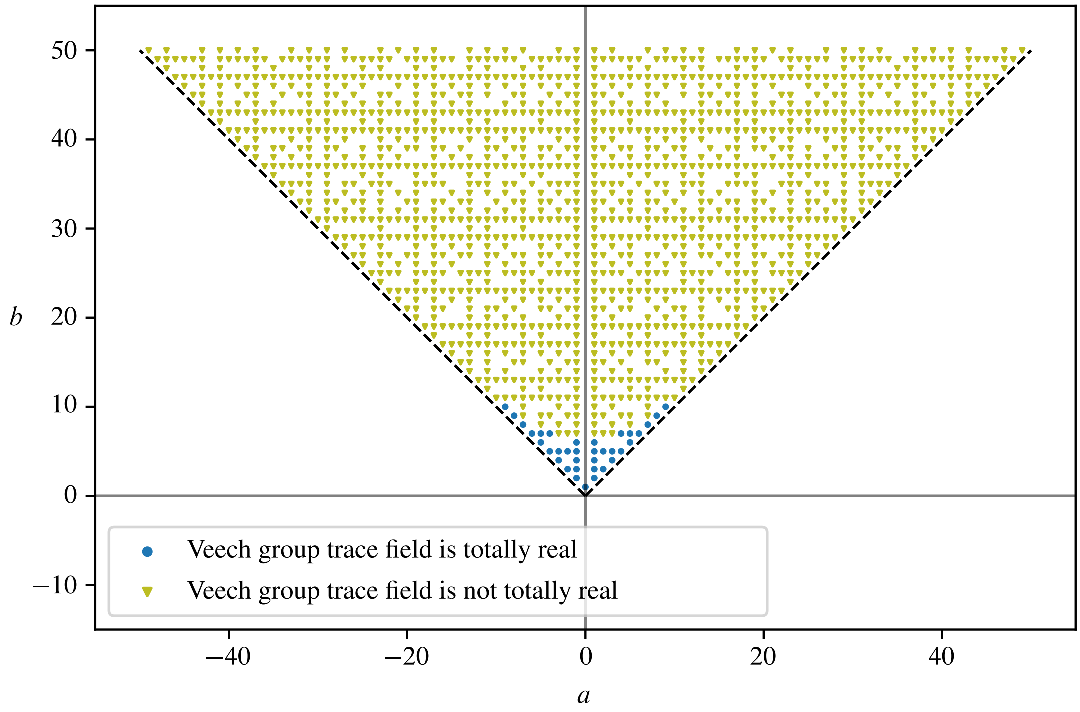


Figure 1: Primitive integral elements in a fibered cone for the mapping torus of the three-strand braid $\sigma_1\sigma_2^{-1}$. Elements marked with green triangles have corresponding Veech group with trace field that is not totally real.

The first factor on the right hand side tends to infinity when b does, while the second factor tends toward $1-(1+\epsilon)^{-1}=\epsilon/(1+\epsilon)>0$. This implies that $\Theta_{\mathscr{C}}^{(b-1,b)}(\lambda_b)$ approaches infinity, whereas instead it is identically equal to 0. This contradiction proves the claim.

For points of the form $(1-b,b)$, the specialization takes the form

$$\Theta_{\mathscr{C}}^{(1-b,b)}(t)=t^{2b}-t-t^b-t^{2b-1}+1=\Theta_{\mathscr{C}}^{(b-1,b)}(t).$$

Therefore, $\lambda(f_{(1-b,b)})=\lambda(f_{(b-1,b)})=\lambda_b$ and as $b\to\infty$; these both tend to 1. □

One of the difficulties in the proof of Theorem 1.3 is understanding the degrees of the trace field. This is complicated by the fact that the Teichmüller polynomial need not be irreducible in general. For example, when specialized to $(a,b)=(9,14)$, the Teichmüller polynomial in this example splits into the cyclotomic polynomials t^2-t+1 and t^4-t^2+1, plus the minimal polynomial of the corresponding stretch factor. However, in other cases, such as the specialization to $(a,b)=(5,14)$, the Teichmüller polynomial remains irreducible. We refer the reader to [4] for more on the factorizations of the specialized polynomials in the example above. As we will see in the example below, the Teichmüller polynomial also sometimes admits additional noncyclotomic factors aside from the minimal polynomial of the corresponding stretch factor.

3.2 Example 2

Let $\beta' = \beta^2$, for β from the preceding example. Let M' denote the mapping torus on β' and $\theta'_{\mathscr{C}'}$ the Teichmüller polynomial of the fibered cone $\mathscr{C}'$ containing the dual of β'. Here we will observe three different splitting behaviors of specializations of the Teichmüller polynomial. In particular, we see that certain specializations of $\theta'_{\mathscr{C}'}$ split into multiple noncyclotomic factors, limiting what information can be derived about conjugates of the corresponding stretch factors and their traces by looking at the collection of all roots of $\theta'_{\mathscr{C}'}$.

The Teichmüller polynomial here is

$$\theta'_{\mathscr{C}'}(x,u) = u^2 - u(x^2 + 2x + 1 + 2x^{-1} + x^{-2}) + 1$$

over the cone

$$\mathscr{C} = \{(a,b) \in \mathbb{R}^2 \mid b > 0, -\tfrac{1}{2}b < a < \tfrac{1}{2}b\}.$$

The specialization to $(a,b) = (6,17)$ is irreducible over $\mathbb{Z}$,

$$t^{34} - t^{29} - 2t^{23} - t^{17} - 2t^{11} - t^5 + 1,$$

while the specialization to $(a,b) = (7,17)$ splits as a cyclotomic and noncyclotomic factor,

$$(t^4 + t^3 + t^2 + t + 1)\big(t^{30} - t^{29} - t^{27} + t^{26} + t^{25} - t^{24} - t^{22} + t^{21} - t^{20} + t^{19} - t^{17} + t^{16} \\ - t^{15} + t^{14} - t^{13} + t^{11} - t^{10} + t^9 - t^8 - t^6 + t^5 + t^4 - t^3 - t + 1\big),$$

and the specialization to $(a,b) = (7,18)$ has multiple noncyclotomic factors,

$$(t^2 - t + 1)(t^4 + t^3 + t^2 + t + 1)(t^{12} - t^9 - t^8 + t^7 + t^6 + t^5 - t^4 - t^3 + 1)(t^{18} - t^{16} - t^9 - t^2 + 1).$$

Figure 2 shows whether the Veech groups corresponding to elements of $\mathscr{C}'$ have totally real trace field. For all three specializations described in this example, the corresponding Veech group trace field is not totally real.

The analog to Lemma 3.1 holds in this example as well. M' is a 2-fold cover of M so the stretch factors in $\mathscr{C}'_{\mathbb{Z}}$ are at most squares of the stretch factors in $\mathscr{C}_{\mathbb{Z}}$.

4 Most Veech groups have no parabolics

We are now ready for the proof of the first theorem from the introduction.

Theorem 1.3 *Suppose F is the fibered face of an orientable, fibered, hyperbolic 3-manifold. Assuming Lehmer's conjecture, the set of $\bar{\alpha} \in F_{\mathbb{Q}}$ such that $\mathrm{Aff}_+(X_\alpha, q_\alpha)$ contains a parabolic element is discrete in F.*

Proof Consider any sequence of distinct elements α_n in $\mathscr{C}_{\mathbb{Z}}$ such that $\bar{\alpha}_n$ does not accumulate on ∂F. We need to show that $\mathrm{Aff}(X_\alpha, q_{\alpha_n})$ contains a parabolic for at most finitely many n. According to

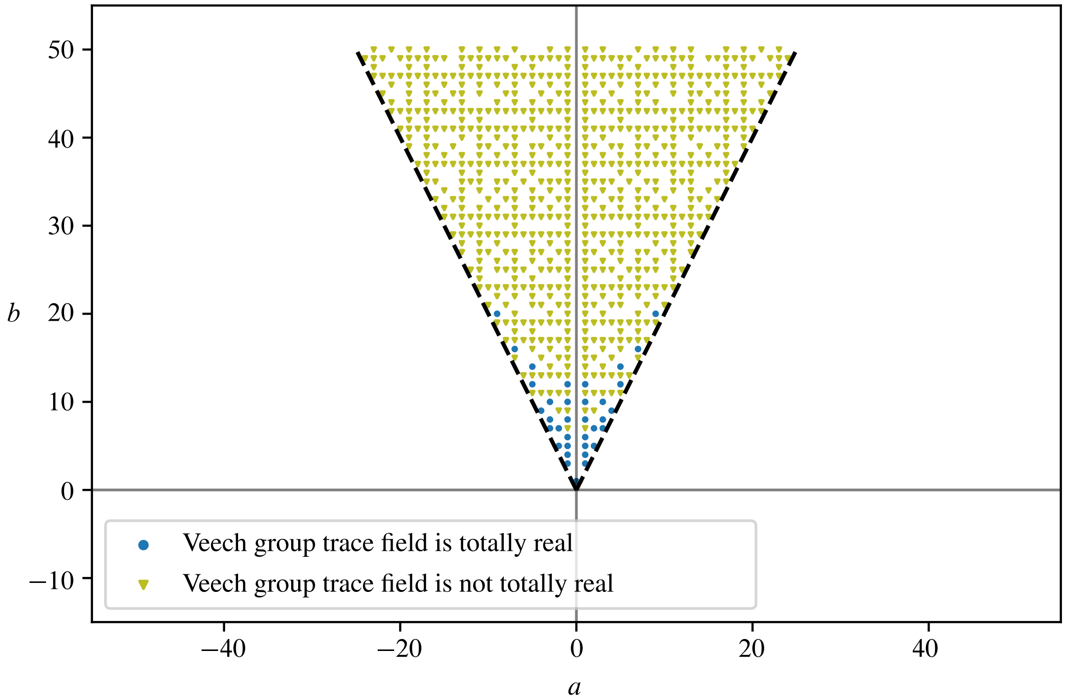


Figure 2: Primitive integral elements in a fibered cone for the mapping torus of the three-strand braid $(\sigma_1\sigma_2^{-1})^2$. Elements marked with green triangles have a not totally real corresponding Veech group.

Theorem 2.9, it suffices to prove that the trace field is totally real for at most finitely many n. Setting $\lambda_n = \lambda(f_{\alpha_n})$, Theorem 2.10 implies that the trace field of $\mathrm{Aff}(X_{\alpha_n}, q_{\alpha_n})$ is $\mathbb{Q}(\lambda_n + \lambda_n^{-1})$.

Next, let N be the number of terms of the Teichmüller polynomial, $\Theta_{\mathscr{C}}$ for $\mathscr{C}$. The stretch factor λ_n is the largest modulus root of the specialization $\Theta_{\mathscr{C}}^{\alpha_n}(t)$ by Theorem 2.3. We observe that this polynomial has no more nonzero terms than $\Theta_{\mathscr{C}}$, and thus has at most N terms. Descartes's rule of signs implies that the number of real roots of $\Theta_{\mathscr{C}}^{\alpha_n}$ is at most $2N-2$.

Suppose that $p_n(t)$ is the minimal polynomial of λ_n, which is thus a factor of $\Theta_{\mathscr{C}}^{\alpha_n}(t)$ (up to powers of t, which we will ignore). In particular, note that λ_n bounds the modulus of all other roots of $p_n(t)$. The stretch factors are always algebraic integers, and hence $p_n(t)$ is monic. The Mahler measure is therefore the product of the moduli of the roots outside the unit circle. There are at most $2N-2$ real roots of $\Theta_{\mathscr{C}}^{\alpha_n}(t)$, and hence the same is true of $p_n(t)$. Write

$$\mathcal{M}(p_n) = A_n B_n$$

where A_n is the product of the moduli of the *real roots* and B_n is the product of the moduli of the nonreal roots outside the unit circle (and 1 if there are none). Thus, we have

$$A_n \le \lambda_n^{2N-2}. \tag{1}$$

Now, as $n \to \infty$, we have $|\chi(S_{\alpha_n})| = \|\alpha_n\|_T \to \infty$ as $n \to \infty$. Since $\bar{\alpha}_n$ does not accumulate on ∂F, Corollary 2.5 implies $\lambda_n = \lambda(f_{\alpha_n}) \to 1$. By (1), it follows that $A_n \to 1$ as $n \to \infty$. Since we are assuming Lehmer's conjecture, it follows that $B_n > 1$ for all but finitely many n. That is, there is at least one nonreal root ζ_n of $p_n(t)$ outside the unit circle. (In fact, the number of such roots tends to infinity linearly with $|\chi(S_{\alpha_n})|$ since λ_n has the maximum modulus of any root of $p_n(t)$).

Therefore, for all but finitely many n, the embedding of $\mathbb{Q}(\lambda_n + \lambda_n^{-1})$ to $\mathbb{C}$ sending $\lambda_n + \lambda_n^{-1}$ to $\zeta_n + \zeta_n^{-1}$ has nonreal image, since ζ_n is nonreal and lies off the unit circle. Therefore, $\mathbb{Q}(\lambda_n + \lambda_n^{-1})$ is totally real for at most finitely many n, as required. □

Remark 4.1 The proof of Theorem 1.3 follows a strategy of Craig Hodgson [8] for understanding trace fields under hyperbolic Dehn filling.

The key ingredient is that for sequences $\{\alpha_n\}$ in $\mathscr{C}_{\mathbb{Z}}$, we have $\lambda(f_{\alpha_n}) \to 1$. Sometimes this happens for any sequence of distinct elements in the cone, and then one obtains the following stronger result.

Theorem 4.2 *Suppose F is the fibered face of an orientable, fibered, hyperbolic 3-manifold and that 1 is the only accumulation point of the set*

$$\{\lambda(f_\alpha) \mid \bar{\alpha} \in F_{\mathbb{Q}}\}.$$

Assuming Lehmer's conjecture, the set of $\bar{\alpha} \in F_{\mathbb{Q}}$ such that $\mathrm{Aff}_+(X_\alpha, q_\alpha)$ contains a parabolic element is finite.

Proof This is exactly the same as the proof of Theorem 1.3, except that the assumption that 1 is the only accumulation point of $\{\lambda(f_\alpha) \mid \bar{\alpha} \in F_{\mathbb{Q}}\}$ replaces the references to Corollary 2.5, and does away with the requirement that $\bar{\alpha}_n$ does not accumulate on ∂F. □

Returning to the examples from Section 3, Lemma 3.1 and the discussion in both examples implies that the hypotheses of Theorem 4.2 are satisfied. Thus only finitely many elements $\alpha \in \mathscr{C}_{\mathbb{Z}}$ are such that $\mathrm{Aff}_+(X_\alpha, q_\alpha)$ can contain parabolics. We refer the reader to [14] for more on the accumulation set of $\{\lambda(f_\alpha) \mid \alpha \in \mathscr{C}_{\mathbb{Z}}\}$

If $p\colon M \to N$ is the orientation double cover of a nonorientable fibered 3-manifold N with covering involution $\tau\colon M \to M$, then $p^*\colon H^1(N) \to H^1(M)$ is an isomorphism onto the τ^*-fixed subspace. There is a well-defined Thurston norm on $H^1(N)$, and the induced homomorphism $\pi_1 N \to \pi_1 S^1 = \mathbb{Z}$ determines an element $\alpha \in H^1(N)$ which lies in an open cone of a fibered face. Indeed, the p^*-image of this cone is the intersection of $p^*(H^1(N))$ with an open cone on a fibered face F for M, or equivalently, the cone over the τ^*-fixed set $F^\tau \subset F$; see [13, Theorem 2.11]. In this setting, and appealing to work of Liechti and Strenner [17] we can remove the assumption that Lehmer's conjecture holds, at the expense of restricting to F^τ.

Theorem 4.3 *With the assumptions above on $M \to N = M/\langle\tau\rangle$, the set of $\bar\alpha \in F^\tau_{\mathbb{Q}}$ such that $\mathrm{Aff}_+(X_\alpha, q_\alpha)$ contains a parabolic element is discrete in F^τ.*

Proof For every $\bar\alpha \subset F^\tau_{\mathbb{Q}}$, the associated monodromy $f_\alpha\colon S_\alpha \to S_\alpha$ is the lift of the monodromy for some fibration of N. Then either S_α covers a nonorientable surface S'_α and f_α is the lift of a pseudo-Anosov homeomorphism on S'_α, or else f_α is the square of an orientation reversing pseudo-Anosov homeomorphism. In either case, [17, Theorem 1.10] implies that if $p(t)$ is the minimal polynomial for $\lambda(f_\alpha)$, then $p(t)$ has no roots on the unit circle.

Now suppose $\{\bar\alpha_n\} \subset F^\tau_{\mathbb{Q}}$ is any infinite sequence of distinct elements not accumulating on the boundary of F and $\lambda_n = \lambda(f_{\alpha_n})$. As in the proof of Theorem 1.3, write $p_n(t)$ for the minimal polynomial and $\mathcal{M}(p_n) = A_n B_n$. Again, $A_n \to 1$, and thus by Theorem 2.14, there is a nonreal root ζ_n of $p_n(t)$ for all n sufficiently large (regardless of the behavior of B_n). By the previous paragraph ζ_n is not on the unit circle, and thus $\zeta_n + \zeta_n^{-1} \notin \mathbb{C}$; hence $\mathbb{Q}(\lambda_n + \lambda_n^{-1})$ is not totally real, proving our result. □

5 Veech groups of leaves

We now turn our attention to the nonintegral points in the cone and the second theorem from the introduction.

Theorem 1.4 *If F is a fibered face of a closed, orientable, fibered, hyperbolic 3-manifold, then for all $\alpha \in F - F_{\mathbb{Q}}$, and any leaf S_α of $\mathscr{F}_\alpha$, the abelian group $H^{\mathrm{Aff}}_\alpha < \mathrm{Aff}_+(X_\alpha, q_\alpha)$ has finite index.*

For the rest of the paper, we assume M is a closed, fibered, hyperbolic 3-manifold. The results of this section are only nontrivial if $b_1(M) > 1$, since otherwise $F - F_{\mathbb{Q}} = \varnothing$ for any fibered face F (since in that case $F = F_{\mathbb{Q}}$ is a point). Given $\alpha \in F$, we recall that ψ^α_s is the reparameterized flow as in Section 2.2, that sends leaves of $\mathscr{F}_\alpha$ to leaves. Furthermore, (X_α, q_α) is the leafwise conformal structure and quadratic differential, and there is $K_\alpha > 1$ such that ψ^α_s is the $K_\alpha^{|s|}$-Teichmüller map; hence $K_\alpha^{2|s|}$-quasiconformal and $K_\alpha^{|s|}$-bi-Lipschitz.

Lemma 5.1 *For any $\alpha \in F - F_{\mathbb{Q}}$ there exists a compact subsurface $Z \subset S_\alpha$ such that*

$$M = \bigcup_{s\in[0,1]} \psi^\alpha_s(Z).$$

Proof Choose an exhaustion of S_α by a sequence of compact subsurfaces,

$$Z_1 \subsetneq Z_2 \subsetneq Z_3 \subsetneq \cdots \subsetneq S_\alpha \quad \text{and} \quad \bigcup_{n=1}^{\infty} Z_n = S_\alpha,$$

and observe that

$$\Big\{ \bigcup_{s\in(0,1)} \psi^\alpha_s(\mathrm{int}(Z_n)) \Big\}_{n=1}^{\infty}$$

is an open cover of M since every leaf is dense. Since M is compact, the open cover admits a finite subcover of M. As the compact surfaces Z_n are nested, there exists an index N such that for $Z = Z_N$ we have

$$M = \bigcup_{s\in[0,1]} \psi_s^\alpha(Z). \qquad \square$$

The isomorphism $H_\alpha \cong H_\alpha^{\mathrm{Aff}}$ is given by $s \mapsto \psi_s^\alpha|_{S_\alpha}$. We write

$$H_\alpha^{\mathrm{Aff}}[0,1] \subset H_\alpha^{\mathrm{Aff}}$$

for the image of $H_\alpha \cap [0,1]$ under this isomorphism. Note that every element of H_α^{Aff} is K_α^2-quasiconformal and K_α-bi-Lipschitz since $s \le 1$. As a consequence of Lemma 5.1, we have the following.

Corollary 5.2 *For $\alpha \in F - F_{\mathbb{Q}}$ and $Z \subset S_\alpha$ as in Lemma 5.1 we have*

$$S_\alpha = \bigcup_{h\in H_\alpha^{\mathrm{Aff}}[0,1]} h(Z).$$

Proof Let $Z \subset S_\alpha$ be the compact subsurface from Lemma 5.1, so that for every $x \in S_\alpha \subseteq M$, we have $x \in \psi_s^\alpha(Z)$ for some $s \in [0,1]$. Since $x \in S_\alpha$, this implies that $s \in H_\alpha$. Therefore,

$$S_\alpha = \bigcup_{s\in H_\alpha\cap[0,1]} \psi_s^\alpha(Z) = \bigcup_{h\in H_\alpha^{\mathrm{Aff}}[0,1]} h(Z). \qquad \square$$

Corollary 5.3 *For any $\alpha \in F - F_{\mathbb{Q}}$ there exists $C > 0$ such that for any leaf S_α of $\mathscr{F}_\alpha$, the geometry of q_α is bounded. Specifically,*

(1) *there is a lower bound on the length of any saddle connection, in particular a lower bound on the distance between any two cone points,*

(2) *all cone points have finite (uniformly bounded) cone angle, and*

(3) *(X_α, q_α) is complete.*

Proof Let S_α be any leaf, and consider the compact surface Z from Corollary 5.2. By making Z slightly larger, we can assume that no singular points of q_α lie on the boundary of Z. Denote the set of all singularities of q_α by A. Let $d_{\partial Z}(a)$ denote the distance of a singularity $a \in A$ to the boundary of Z, and let $d_Z(a,b)$ denote the minimal length of a saddle connection in Z between two (not necessarily distinct) singularities $a, b \in A \cap Z$. Since Z is compact, we have that

$$\epsilon = \min\Bigl\{ \min_{a,b\in A\cap Z} d_Z(a,b), \min_{a\in A} d_{\partial Z}(a) \Bigr\} > 0.$$

Pick a saddle connection ω connecting any singularity a to any singularity b. There exists an $h \in H_\alpha^{\mathrm{Aff}}[0,1]$ such that $h(Z)$ contains a. Since h is K_α-bi-Lipschitz, either ω is contained in $h(Z)$ and has length at least ϵK_α^{-1}, or it leaves $h(Z)$ and we again deduce that ω has length at least the distance from a to $\partial h(Z)$, which is at least ϵK_α^{-1}. In either case, we obtain a uniform lower bound ϵK_α^{-1} to the length of ω, proving (1).

As was noted in Section 2.3, we have that all cone points have finite cone angle which proves (2). Since Z is compact, there is an ϵ' so that the ϵ'-neighborhood of Z also has compact closure, which is thus complete. Any Cauchy sequence has a tail that is contained in the h-image of the closure of this neighborhood for some $h \in H_\alpha^{\mathrm{Aff}}[0,1]$. Since this h-image is also complete, the Cauchy sequence converges, and we have that (X_α, q_α) is complete which proves (3). □

Remark 5.4 Note that Corollary 5.3 implies that our surfaces are tame in the sense of [22, Definition 2.1].

An important observation is the following: for any element of $g \in \mathrm{Aff}_+(X_\alpha, q_\alpha)$, we can choose some element $h \in H_\alpha^{\mathrm{Aff}}[0,1]$ so that $h \circ g(Z) \cap Z \neq \varnothing$, and furthermore, if g is K-quasiconformal, then $h \circ g$ is (KK_α^2)-quasiconformal.

Proposition 5.5 *Suppose $\alpha \in F - F_{\mathbb{Q}}$, $K_0 > 1$, and $\{g_n\}_{n=1}^\infty \subset \mathrm{Aff}_+(X_\alpha, q_\alpha)$ is a sequence of elements with $K(g_n) \le K_0$. Then there is a subsequence $\{g_{n_k}\}_{k=0}^\infty$ and $\{h_{n_k}\}_{k=0}^\infty \subset H_\alpha^{\mathrm{Aff}}[0,1]$ such that*

$$h_{n_k} \circ g_{n_k} = h_{n_0} \circ g_{n_0}$$

for all $k \ge 0$.

Proof From the observation before the statement, we can find $h_n \in H_\alpha^{\mathrm{Aff}}[0,1]$ such that $h_n \circ g_n(Z) \cap Z \neq \varnothing$. Next, observe that $h_n \circ g_n$ is $(K_0 K_\alpha^2)$-quasiconformal, so by compactness of quasiconformal maps, after passing to a subsequence, $h_{n_k} \circ g_{n_k}$ converges uniformly on compact sets to a map f. The maps $h_{n_k} \circ g_{n_k}$ are affine, so they must map cone points to cone points. Since the cone points are uniformly separated by Corollary 5.3, there is a pair of cone points a, b such that for k sufficiently large $h_{n_k} \circ g_{n_k}(a) = b$. Moreover, if we pick a pair of saddle connections in linearly independent directions emanating from a, then for n sufficiently large $h_{n_k} \circ g_{n_k}$ all agree on this pair, again by Corollary 5.3. But these conditions uniquely determines the affine homeomorphism, and hence $h_{n_k} \circ g_{n_k}$ is eventually constant, and passing to a tail-subsequence of this subsequence completes the proof. □

From this we can prove a special case of Theorem 1.4:

Proposition 5.6 *If $\alpha \in F - F_{\mathbb{Q}}$, then H_α^{Aff} has finite index in* $\mathrm{SAff}(X_\alpha, q_\alpha)$.

Proof Suppose H_α^{Aff} is not finite index and consider the closure of the D_α-image in $\mathrm{PSL}_2(\mathbb{R})$,

$$G = \overline{D_\alpha(\mathrm{SAff}(X_\alpha, q_\alpha))}.$$

Since $\alpha \in F - F_{\mathbb{Q}}$, every leaf S_α of $\mathscr{F}_\alpha$ is dense in M. Therefore $H_\alpha^D < \Delta \cong \mathbb{R}$ is an abelian subgroup with rank at least 2, and hence is dense in Δ. Consequently, $\Delta < G$.

By the classification of Lie subalgebras of $\mathfrak{sl}_2(\mathbb{R})$ (or a direct calculations) we observe that, after replacing G with a finite index subgroup, we must be in one of the following situations:

(1) $G = \mathrm{PSL}_2(\mathbb{R})$,

(2) G is the subgroup of upper triangular matrices, or

(3) $G = \Delta$.

In any case, we claim that there is a sequence of elements $\{g_n\} \subset \mathrm{SAff}(X_\alpha, q_\alpha)$ such that $D_\alpha(g_n) \to I$ in $\mathrm{PSL}_2(\mathbb{R})$ and so that $H_\alpha^{\mathrm{Aff}} g_n$ are distinct cosets of H_α^{Aff}. Assuming the claim, we prove the proposition. For this, we simply apply Proposition 5.5, pass to a subsequence (of the same name) so that $h_n \circ g_n = h_0 \circ g_0$ for all $n \geq 0$. This contradicts the fact that $\{H_\alpha^{\mathrm{Aff}} g_n\}$ are all distinct cosets.

To prove the claim, notice that in the first two cases, a finite index subgroup of $D_\alpha(\mathrm{SAff}(X_\alpha, q_\alpha))$ is dense in the Lie subgroup $G \leq \mathrm{PSL}_2(\mathbb{R})$, and $\Delta < G$ is a 1-dimensional submanifold of G, which itself has dimension 3 or 2 in cases (1) and (2), respectively. This implies that there exists a sequence $\{g_n\} \in \mathrm{SAff}(X_\alpha, q_\alpha)$ such that $D_\alpha(g_n) \to I$ as $n \to \infty$ but $D_\alpha(g_n) \notin \Delta$. By way of contradiction, suppose that there exists a subsequence $\{g_{n_i}\}$ such that g_{n_i} are in the same coset $H_\alpha^{\mathrm{Aff}} g$ where $D_\alpha(g) \notin \Delta$. This implies that $D_\alpha(g_{n_i}) \subset \Delta D_\alpha(g)$, which is a 1-manifold parallel to Δ and does not accumulate to I. This contradicts the fact that $D_\alpha(g_{n_i}) \to I$. Therefore, there exists a subsequence of $\{g_n\}$ such that $\{H_\alpha^{\mathrm{Aff}} g_n\}$ are all distinct cosets.

To prove the claim in the final case, we note that by assumption there exists a sequence of distinct cosets $H_\alpha^{\mathrm{Aff}} b_n^{\mathrm{Aff}}$ of H_α^{Aff} in $\mathrm{SAff}(X_\alpha, q_\alpha)$. Since both H_α^D and $D_\alpha(\mathrm{SAff}(X_\alpha, q_\alpha))$ are dense in Δ, so is every coset of H_α^D. Therefore, we can find a sequence $\{a_n^{\mathrm{Aff}}\} \subset H_\alpha^{\mathrm{Aff}}$ so that $D_\alpha(a_n^{\mathrm{Aff}}) D_\alpha(b_n^{\mathrm{Aff}}) \to I$ as $n \to \infty$. Let $g_n = a_n^{\mathrm{Aff}} b_n^{\mathrm{Aff}}$, so that $D_\alpha(a_n^{\mathrm{Aff}}) \to I$ and $H_\alpha^{\mathrm{Aff}} g_n$ are distinct cosets of H_α^{Aff}, as required. This completes the proof of the claim. Since we already proved the proposition assuming the claim, we are done. □

To complete the proof of Theorem 1.4, we need only prove the following.

Proposition 5.7 $\mathrm{Aff}_+(X_\alpha, q_\alpha) = \mathrm{SAff}(X_\alpha, q_\alpha)$.

Proof First, observe that $\mathrm{SAff}_+(X_\alpha, q_\alpha)$ is a normal subgroup of $\mathrm{Aff}_+(X_\alpha, q_\alpha)$ since it is precisely the kernel of the homomorphism given by the determinant of the derivative. In fact, from this homomorphism, either $\mathrm{Aff}_+(X_\alpha, q_\alpha) = \mathrm{SAff}(X_\alpha, q_\alpha)$ or else the index is infinite; $[\mathrm{Aff}_+(X_\alpha, q_\alpha) : \mathrm{SAff}(X_\alpha, q_\alpha)] = \infty$.

After passing to a finite index subgroup, $\Gamma < \mathrm{Aff}_+(X_\alpha, q_\alpha)$, if necessary, the conjugation action of Γ on $\mathrm{SAff}_+(X_\alpha, q_\alpha)$ preserves the finite index subgroup H_α^{Aff} (and without loss of generality, $H_\alpha^{\mathrm{Aff}} < \Gamma$). It thus suffices to prove $\Gamma < \mathrm{SAff}_+(X_\alpha, q_\alpha)$, or equivalently, $D_\alpha(\Gamma) < \mathrm{PSL}_2(\mathbb{R})$.

Consider any element

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in D_\alpha(\Gamma) \quad \text{and} \quad h = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \in H_\alpha^D,$$

with $\lambda \neq \pm 1$. Then $ghg^{-1} \in H_\alpha^D$, and is given by

$$ghg^{-1} = \frac{1}{\det(g)} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{\det(g)} \begin{pmatrix} ad\lambda - bc\lambda^{-1} & ab(\lambda - \lambda^{-1}) \\ cd(\lambda - \lambda^{-1}) & ad\lambda^{-1} - bc\lambda \end{pmatrix}.$$

In order for this element to be in H_α^D (hence diagonal), we must have that $ab = 0$ and $cd = 0$. Suppose that $a = 0$. If $c = 0$, then we have the zero matrix, so we must have that $c \neq 0$ and instead that $d = 0$. This gives us that g is a matrix of the form

$$g = \begin{pmatrix} 0 & b \\ c & 0 \end{pmatrix}.$$

We note that the square of a matrix of this form is a diagonal matrix. Similarly, if $b = 0$, we must have that $c = 0$ and we have that g is a matrix of the form

$$g = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}.$$

Together, these two conclusions imply that either g or g^2 is diagonal.

Now we show that $D_\alpha(\Gamma) < \mathrm{PSL}_2(\mathbb{R})$. If not, then there exists $g \in D_\alpha(\Gamma)$ with $0 < \det(g) \neq 1$. After squaring and inverting if necessary, we may assume that g is diagonal,

$$g = \begin{pmatrix} \lambda & 0 \\ 0 & \sigma \end{pmatrix},$$

and $0 < \det(g) = \lambda\sigma < 1$. Without loss of generality, suppose $\lambda < 1$. Notice that there exists an element $h \in H_\alpha^D$ such that

$$h = \begin{pmatrix} \mu & 0 \\ 0 & \mu^{-1} \end{pmatrix}$$

and there exist $n, k \in \mathbb{Z}$ such that

$$m = g^n h^k = \begin{pmatrix} r & 0 \\ 0 & s \end{pmatrix}$$

where $0 < r, s < 1$. Therefore, m^j is a contraction for all $j > 0$, which implies that it is contracting in both directions. Fixing a saddle connection ω of q_α, it follows that the length of $m^j(\omega)$ tends to 0 as $j \to \infty$. This contradicts Corollary 5.3, part (1), and thus proves that $g \in \mathrm{PSL}_2(\mathbb{R})$, as required. □

Remark 5.8 The final contradiction in the above proof also follows from [22, Theorem 1.1], since $D_\alpha(\mathrm{Aff}_+(X_\alpha, q_\alpha))$ is necessarily of type (i) in that theorem.

References

[1] **J P Bowman**, *The complete family of Arnoux–Yoccoz surfaces*, Geom. Dedicata 164 (2013) 113–130 MR Zbl

[2] **K Calta**, *Veech surfaces and complete periodicity in genus two*, J. Amer. Math. Soc. 17 (2004) 871–908 MR Zbl

[3] **A Fathi**, **F Laudenbach**, **V Poénaru** (editors), *Travaux de Thurston sur les surfaces*, Astérisque 66-67, Soc. Math. France, Paris (1979) MR Zbl

[4] **M Filaseta**, **S Garoufalidis**, *Factorization of polynomials in hyperbolic geometry and dynamics*, preprint (2022) arXiv 2209.08449

[5] **D Fried**, *Flow equivalence, hyperbolic systems and a new zeta function for flows*, Comment. Math. Helv. 57 (1982) 237–259 MR Zbl

[6] **D Fried**, *Transitive Anosov flows and pseudo-Anosov maps*, Topology 22 (1983) 299–303 MR Zbl

[7] **E Hironaka**, *Small dilatation mapping classes coming from the simplest hyperbolic braid*, Algebr. Geom. Topol. 10 (2010) 2041–2060 MR Zbl

[8] **C Hodgson**, *Commensurability*, *trace fields*, *and hyperbolic Dehn filling*, unpublished notes

[9] **P Hubert**, **E Lanneau**, *Veech groups without parabolic elements*, Duke Math. J. 133 (2006) 335–346 MR Zbl

[10] **P Hubert**, **E Lanneau**, **M Möller**, *The Arnoux–Yoccoz Teichmüller disc*, Geom. Funct. Anal. 18 (2009) 1988–2016 MR Zbl

[11] **P Hubert**, **H Masur**, **T Schmidt**, **A Zorich**, *Problems on billiards, flat surfaces and translation surfaces*, from "Problems on mapping class groups and related topics", Proc. Sympos. Pure Math. 74, Amer. Math. Soc., Providence, RI (2006) 233–243 MR Zbl

[12] **R Kenyon**, **J Smillie**, *Billiards on rational-angled triangles*, Comment. Math. Helv. 75 (2000) 65–108 MR Zbl

[13] **S Khan**, **C Partin**, **R R Winarski**, *Pseudo-Anosov homeomorphisms of punctured nonorientable surfaces with small stretch factor*, Algebr. Geom. Topol. 23 (2023) 2823–2856 MR Zbl

[14] **M P Landry**, **Y N Minsky**, **S J Taylor**, *Flows, growth rates, and the veering polynomial*, Ergodic Theory Dynam. Systems 43 (2023) 3026–3107 MR Zbl

[15] **E Lanneau**, *Raconte-moi … un pseudo-Anosov*, Gaz. Math. (2017) 52–57 MR Zbl Translated in Eur. Math. Soc. Newsl. 106 (2017) 12–16

[16] **D H Lehmer**, *Factorization of certain cyclotomic functions*, Ann. of Math. 34 (1933) 461–479 MR Zbl

[17] **L Liechti**, **B Strenner**, *Minimal pseudo-Anosov stretch factors on nonoriented surfaces*, Algebr. Geom. Topol. 20 (2020) 451–485 MR Zbl

[18] **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

[19] **C T McMullen**, *Billiards and Teichmüller curves on Hilbert modular surfaces*, J. Amer. Math. Soc. 16 (2003) 857–885 MR Zbl

[20] **C T McMullen**, *Teichmüller geodesics of infinite complexity*, Acta Math. 191 (2003) 191–223 MR Zbl

[21] **J-P Otal**, *Thurston's hyperbolization of Haken manifolds*, from "Surveys in differential geometry, III", International, Boston, MA (1998) 77–194 MR Zbl

[22] **P Przytycki**, **G Schmithüsen**, **F Valdez**, *Veech groups of Loch Ness monsters*, Ann. Inst. Fourier (Grenoble) 61 (2011) 673–687 MR Zbl

[23] **C Ramírez Maluendas**, **F Valdez**, *Veech groups of infinite-genus surfaces*, Algebr. Geom. Topol. 17 (2017) 529–560 MR Zbl

[24] *SageMath*, *version* 9.3 (2021) Available at `https://www.sagemath.org`

[25] **A Schinzel**, *Addendum to 'On the product of the conjugates outside the unit circle of an algebraic number', 24 (1973) 385–399*, Acta Arith. 26 (1974/75) 329–331 MR Zbl

[26] **C Smyth**, *The Mahler measure of algebraic numbers: a survey*, from "Number theory and polynomials", Lond. Math. Soc. Lect. Note Ser. 352, Cambridge Univ. Press (2008) 322–349 MR Zbl

[27] **W P Thurston**, *A norm for the homology of* 3*-manifolds*, Mem. Amer. Math. Soc. 339, Amer. Math. Soc., Providence, RI (1986) MR Zbl

[28] **W P Thurston**, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. 19 (1988) 417–431 MR Zbl

CJL: *Department of Mathematics, Rice University*
Houston, TX, United States

KR: *Department of Mathematics, University of Toronto*
Toronto, ON, Canada

NR: *Department of Mathematics, University of British Columbia*
Vancouver, BC, Canada

ES: *Department of Mathematics, University of Illinois at Urbana-Champaign*
Urbana, IL, United States

YV: *Department of Mathematics, Western University*
London, ON, Canada

`cjl12@rice.edu, rafi@math.toronto.edu, rouse@math.ubc.ca, esshinkle@gmail.com, verberne.math@gmail.com`

Received: 14 October 2023 Revised: 10 January 2024

Guidelines for Authors

Submitting a paper to Algebraic & Geometric Topology

Papers must be submitted using the upload page at the AGT website. You will need to choose a suitable editor from the list of editors' interests and to supply MSC codes.

The normal language used by the journal is English. Articles written in other languages are acceptable, provided your chosen editor is comfortable with the language and you supply an additional English version of the abstract.

Preparing your article for Algebraic & Geometric Topology

At the time of submission you need only supply a PDF file. Once accepted for publication, the paper must be supplied in LaTeX, preferably using the journal's class file. More information on preparing articles in LaTeX for publication in AGT is available on the AGT website.

`arXiv` papers

If your paper has previously been deposited on the `arXiv`, we will need its `arXiv` number at acceptance time. This allows us to deposit the DOI of the published version on the paper's `arXiv` page.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Fuzzy or sloppily drawn figures will not be accepted. For labeling figure elements consider the pinlabel LaTeX package, but other methods are fine if the result is editable. If you're not sure whether your figures are acceptable, check with production by sending an email to graphics@msp.org.

Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

ALGEBRAIC & GEOMETRIC TOPOLOGY

Volume 25 Issue 3 (pages 1265–1915) 2025