AGT

Algebraic & Geometric
Topology

Comparing combinatorial models of moduli space and their compactifications

DANIELA EGAS SANTANDER

ALEXANDER KUPERS

We compare two combinatorial models for the moduli space of two-dimensional cobordisms (namely Bödigheimer's radial slit configurations and Godin's admissible fat graphs), using a "critical graph" map to produce an explicit homotopy equivalence. We also discuss natural compactifications of these two models, the unilevel harmonic compactification and Sullivan diagrams, respectively, and prove that the homotopy equivalence induces a cellular homeomorphism between these compactifications.

1 Introduction

In this paper we compare two combinatorial models of the moduli space of cobordisms. We start with an introduction to moduli space, giving a conformal description of it. After that, we describe various combinatorial models and how they relate to each other, which includes our main result, Theorem 1.1. Finally we describe two applications.

1.1 The moduli space of cobordisms

The study of families of surfaces, known as "moduli theory", goes back to the nineteenth century. One of the main points of this theory is the construction of a *moduli space*; informally, this is a space of all surfaces isomorphic to a given one, characterized by the property that equivalence classes of maps into it correspond to equivalence classes of families of surfaces. For applications to field theories, the surfaces

of interest are two-dimensional oriented cobordisms, an oriented surface S with parametrized boundary divided into an incoming and an outgoing part. More precisely, there is a pair of maps $\iota_{\text{in}} \colon \bigsqcup_{i=1}^{n} S^1 \to \partial S$ and $\iota_{\text{out}} \colon \bigsqcup_{j=1}^{m} S^1 \to \partial S$ such that $\iota_{\text{in}} \sqcup \iota_{\text{out}}$ is a diffeomorphism onto ∂S.

We will now give a conformal definition of the moduli space of these cobordisms, following work of Bödigheimer [3, Section 2] and Hamenstädt [24]. Let S be an isomorphism class of connected two-dimensional oriented cobordisms with nonempty incoming and outgoing boundary. As we will later endow S with a metric, a parametrization of its boundary is given by a point in each boundary component. So $S = S_{g,n+m}$ is a connected oriented surface of genus g with $n + m$ boundary components, each containing a single point p_i for $1 \leq i \leq n + m$. The marked points are ordered and divided into an incoming set (which contains the first $n \geq 1$ marked points) and an outgoing set (which contains the last $m \geq 1$ marked points).

To define the moduli space, we start by considering the set of metrics g on S. Two metrics are said to be conformally equivalent if they are equal, up to a pointwise rescaling by a continuous function. This is equivalent to having the same notion of angle. A diffeomorphism $f \colon S_1 \to S_2$ between two-dimensional manifolds $(S_1, [g]_1)$ and $(S_2, [g]_2)$ with conformal classes of metrics such that $f^*[g]_2 = [g]_1$ is said to be a conformal diffeomorphism. This is equivalent to each of its differentials $D_p f$ for $p \in S_1$ being a linear map that preserves angles.

We will restrict our attention to those conformal classes of metrics on S such that each incoming boundary component has a neighborhood that is conformally diffeomorphic to a neighborhood of the boundary of $\{z \in \mathbb{C} \mid \|z\| \geq 1\}$, and each outgoing boundary component has a neighborhood that is conformally diffeomorphic to a neighborhood of the boundary of $\{z \in \mathbb{C} \mid \|z\| \leq 1\}$. We say that these conformal classes have good boundary.

The moduli space $\mathcal{M}_g(n, m)$ will have as underlying set the conformal classes of metrics on S with good boundary, modulo the equivalence relation of conformal diffeomorphism fixing the points p_i. To topologize it, we introduce the Teichmüller metric. With respect to this metric, two equivalence classes of metrics on S are close if they are related by a homeomorphism that — away from a finite set — is not only differentiable, but also conformal up to a small error. To make this precise, note that a linear map $D \colon \mathbb{R}^2 \to \mathbb{R}^2$ is conformal if and only if $\max(\|Dv\|/\|v\|) = \min(\|Dv\|/\|v\|)$, with both the maximum and minimum taken over nonzero vectors. Hence we can quantify the deviation of a linear map from being conformal by its eccentricity

$$\text{Ecc}(D) := \frac{\max(\|Dv\|/\|v\|)}{\min(\|Dv\|/\|v\|)}.$$

If $f \colon (S, [g]_1) \to (S, [g]_2)$ is a homeomorphism that is continuously differentiable outside a finite set of points $\Sigma \subset S$, then its quasiconformal constant K_f is defined to be

$$K_f := \sup_{p \in S \setminus \Sigma} \text{Ecc}(D_p f),$$

and f is said to be quasiconformal if K_f is finite. If $QC([g]_1, [g]_2)$ denotes the set of all quasiconformal homeomorphisms between $(S, [g]_1)$ and $(S, [g]_2)$ fixing the points p_i, then we can define the Teichmüller distance between $[g]_1$ and $[g]_2$ as follows:

$$d_{\mathcal{T}}((S, [g]_1), (S, [g]_2)) := \log\inf\{K_f \mid f \in QC([g]_1, [g]_2)\}.$$

The moduli space of two-dimensional oriented cobordisms isomorphic to S is then defined to be the metric space

$$\mathcal{M}_g(n, m) := \left(\frac{\text{conformal classes of metrics on } S \text{ with good boundary}}{\text{conformal diffeomorphisms fixing the points } p_i}, d_{\mathcal{T}} \right).$$

For S that are not connected, we take the product of these spaces over all components. An alternative definition of these spaces is as the quotient of Teichmüller space (the space of quasiconformal maps modulo conformal equivalence) by the action of the mapping class group $\mathrm{Mod}(S, \partial S)$, ie the group of components of the diffeomorphism group $\mathrm{Diff}(S, \partial S)$. This is a free proper action on a contractible space, and hence $\mathcal{M}_g(n, m) \simeq B\,\mathrm{Mod}(S, \partial S)$. All connected components of $\mathrm{Diff}(S, \partial S)$ are contractible, and we can thus conclude that

$$\mathcal{M}_g(n, m) \simeq B\,\mathrm{Mod}(S, \partial S) \simeq B\mathrm{Diff}(S, \partial S).$$

This explains why $\mathcal{M}_g(n, m)$ is a model for the moduli space of two-dimensional oriented cobordisms; any bundle of cobordisms over a paracompact space B with transition functions given by diffeomorphisms can be obtained up to isomorphism by pulling back a universal bundle over $\mathcal{M}_g(n, m)$ along a map $B \to \mathcal{M}_g(n, m)$. This universal bundle is the quotient of the space consisting of pairs $([g], x)$ of a conformal class of metrics and a point $x \in S$, by conformal diffeomorphisms acting diagonally.

1.2 Combinatorial models of moduli space

We discuss several combinatorial models of $\mathcal{M}_g(n, m)$, as well as certain compactifications. The following diagram spells out the relations between them (we fix g, n and m and drop them from the notation):
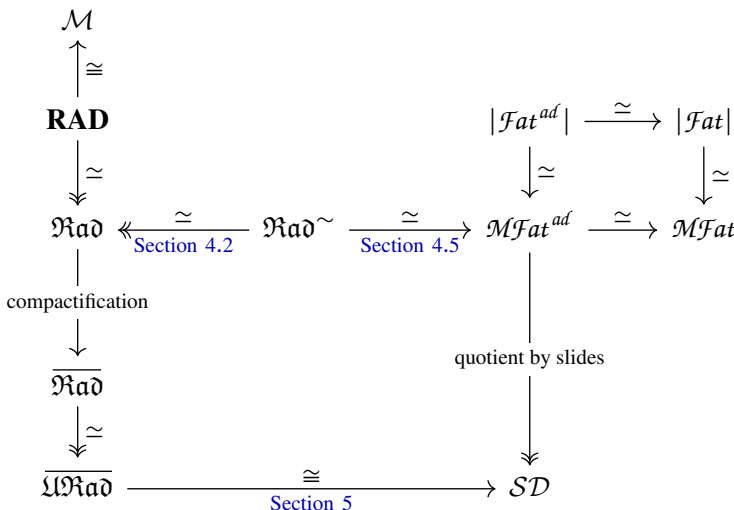
Each arrow is a continuous map; if decorated by $\simeq$ it is a homotopy equivalence, if it is double-headed it is a surjection, and if decorated by $\cong$ it is a homeomorphism. The objects that appear in this diagram are summarized below:

The moduli space $\mathcal{M}$ is the archetypical "space of cobordisms", a conformal model of which was discussed in Section 1.1. It consists of conformal classes of metrics modulo conformal diffeomorphisms, with the Teichmüller metric.

The radial slit configurations   The model **RAD**, due to Bödigheimer, consists of gluing data to construct a conformal class of metrics by gluing together annuli in $\mathbb{C}$. The main theorem of [3] is that there is a homeomorphism $\mathcal{M} \cong \mathbf{RAD}$. There is a deformation retraction of **RAD** onto $\mathfrak{Rad}$ by fixing the radii of the annuli. This and related models will be discussed in Section 2, and $\mathfrak{Rad}$ will be defined in Definition 2.15.

The fat graphs   Fat graphs are graphs with the additional structure of a cyclic ordering of the edges going into each vertex and data encoding the parametrization of its "boundary components". Taking as morphisms maps of fat graphs that collapse a disjoint union of trees defines a category of fat graphs, denoted by $\mathcal{Fat}$. The space $|\mathcal{Fat}|$ is the geometric realization of this category. This and related models will be discussed in Section 3, and $\mathcal{Fat}$ will be defined in Definition 3.7.

The admissible fat graphs   A fat graph is said to be admissible if its incoming boundary graph embeds in it, and the category of admissible fat graphs is denoted by $\mathcal{Fat}^{ad}$. The space $|\mathcal{Fat}^{ad}|$ is the geometric realization of the full subcategory on the admissible fat graphs. It is defined in Definition 3.7.

The metric fat graphs   Closely related to $\mathcal{Fat}$ is the space of metric fat graphs, denoted by $\mathcal{MFat}$. This is the space of fat graphs with the additional data of lengths of their edges. The topology is described in terms of these lengths, and it contains the realization of $\mathcal{Fat}$ as a deformation retract.

The admissible metric fat graphs   Just like $\mathcal{Fat}^{ad}$ is the subcategory of $\mathcal{Fat}$ consisting of fat graphs that are admissible, $\mathcal{MFat}^{ad}$ is the subspace of $\mathcal{MFat}$ consisting of metric fat graphs that are admissible. It is defined in Definition 3.11.

The fattening of the radial slit configurations   To discuss the relation between $\mathfrak{Rad}$ and $\mathcal{MFat}$, we introduce $\mathfrak{Rad}^{\sim}$ as a thicker version of $\mathfrak{Rad}$ by including resolutions of the critical graph for nongeneric radial slit configurations. This is done in Section 4.2.

The harmonic compactification   Naturally $\mathfrak{Rad}$ arises as an open subspace of a compact space $\overline{\mathfrak{Rad}}$. In this compactification we allow identifications of points on the outgoing boundary, and allow handles to degenerate to intervals. It is defined in Definition 2.15.

The unilevel harmonic compactification   The space $\overline{\mathfrak{URad}}$ is a deformation retract of $\overline{\mathfrak{Rad}}$ obtained by making all slits equal length. It is defined in Definition 2.21.

The Sullivan diagrams   The space of Sullivan diagrams, denoted by $\mathcal{SD}$, is the quotient of $\mathcal{MFat}^{ad}$ by the equivalence relation of slides away from the admissible boundary. It is defined in Definition 3.16.

We will focus on the bottom square, that is, the relations between radial slit configurations, admissible metric fat graphs and their compactifications. Our main result is:

Theorem 1.1 *The space $\mathfrak{Rad}^{\sim}$ and maps given in Corollaries 4.42 and 4.51, Proposition 5.1 and Lemma 2.22 form a commutative square*

$$
\begin{array}{ccc}
\mathfrak{Rad} & \xleftarrow[\simeq]{\text{Corollary 4.42}} \mathfrak{Rad}^{\sim} \xrightarrow[\simeq]{\text{Corollary 4.51}} & \mathcal{MFat}^{ad} \\
\downarrow & & \downarrow \\
\overline{\mathfrak{Rad}} & & \\
{\scriptstyle\text{Lemma 2.22}}\downarrow{\simeq} & & \downarrow \\
\overline{\mathfrak{URad}} & \xrightarrow[\text{Proposition 5.1}]{\cong} & \mathcal{SD}
\end{array}
$$

Furthermore, all maps that are decorated by $\simeq$ are homotopy equivalences and the map decorated by $\cong$ is a cellular homeomorphism.

There exist other combinatorial models related to the moduli space of cobordisms which are not discussed here. We will describe six such models in the following remarks.

Remark 1.2 To describe an action of the chains of the moduli space of surfaces on the Hochschild homology of $\mathcal{A}_{\infty}$–Frobenius algebras, Costello constructed a chain complex that models the homology of the moduli space [9; 10]. In [43], Wahl and Westerland described this chain complex in terms of fat graphs with two types of vertices, which they called *black and white fat graphs*. There is an equivalence relation of black and white graphs given by slides away from the white vertices. The quotient chain complex is the cellular chain complex of $\mathcal{SD}$. Furthermore, Egas Santander [14] showed that $\mathcal{MFat}^{ad}$ has a quasicell structure with *black and white fat graphs* as its cellular complex and where the quotient map to $\mathcal{SD}$ respects this cell structure.

Remark 1.3 In [8], Cohen and Godin defined *Sullivan chord diagrams* of genus g with p incoming and q outgoing boundary components, which were also used by Félix and Thomas [16]. These are fat graphs obtained from gluing trees to circles and comprise a space $\mathcal{CF}(g; p, q)$, which is a subspace of $\mathcal{MFat}^{ad}$. They are *not* the same as Sullivan diagrams as in Definition 3.16, though they do admit a map to $\mathcal{SD}$. The space of metric chord diagrams is not homotopy equivalent to moduli space; see Godin [21, Remark 3].

Remark 1.4 In [38], Poirier defined a space $\overline{SD}(g, k, l)/\sim$ of *string diagrams modulo slide equivalence* of genus g with k incoming and l outgoing boundary components, and more generally she defined *string diagrams with many levels modulo slide equivalence*, $\overline{LD}(g, k, l)/\sim$. Proposition 2.3 of [38] says that $\overline{SD}(g, k, l)/\sim \simeq \overline{LD}(g, k, l)/\sim$. She also defined a subspace $SD(g, k, l)$ of $\overline{SD}(g, k, l)$. Both $\overline{SD}(g, k, l)$ and $SD(g, k, l)$ are subspaces of $\mathcal{MFat}^{ad}$, and by counting components one can see that these inclusions cannot be homotopy equivalences. However, there is an induced map $\overline{SD}(g, k, l)/\sim \to \mathcal{SD}$ which is a homeomorphism.

Remark 1.5 In [11], Drummond-Cole, Poirier and Rounds defined a space of *string diagrams SD* which generalized the spaces of chord diagrams constructed in [38]. They conjectured that this space is homotopy equivalent to the moduli space of Riemann surfaces. There is an embedding $SD \hookrightarrow \mathcal{MFat}^{ad}$, but it is not clear this is a homotopy equivalence. Furthermore, there is an equivalence relation $\sim$ on SD, which is not discussed in their paper, and they conjectured that $SD/\sim$ is homotopy equivalent to the harmonic compactification.

Remark 1.6 Following the ideas of Wahl, Klamt constructed a chain complex of *looped diagrams*, denoted by $l\mathcal{D}$ in [31]. This complex gives operations on the Hochschild homology of commutative Frobenius algebras. Moreover, she gave a chain map from the cellular complex of the space of Sullivan diagrams to looped diagrams. However, a geometric interpretation of a space underlying the complex $l\mathcal{D}$ and its possible relation to moduli space are still unknown.

Remark 1.7 In [30], Kaufman described a space of open–closed Sullivan diagrams $\mathrm{Sull}_1^{c/o}$ in terms of arcs embedded in a surface. The closed part, Sull_1^c, is a space whose points correspond to weighted families of embedded arcs in the surface that flow from the incoming boundary to the outgoing boundary. This space has a natural cell structure, and there is a cellular homeomorphism $\mathrm{Sull}_1^c \xrightarrow{\cong} SD$ [43, Remark 2.12] .

1.3 Applications of these models

We will next explain two of the applications of combinatorial models for moduli spaces.

1.3.1 Explicit computations of the homology of moduli spaces
Combinatorial models provide cell decompositions for moduli spaces, allowing for explicit computations of the (co)homology groups of moduli spaces using cellular (co)homology. Instead of studying $\mathcal{M}_g(n, m)$, it is more convenient to study the closely related moduli space $\mathcal{M}_g^{1,n}$ of surfaces of genus g with one parametrized boundary component and n permutable punctures. There are variations of $\mathfrak{Rad}$ and $\mathcal{MFat}^{ad}$ that are models for $\mathcal{M}_g^{1,n}$.

Much is known about the homology of $\mathcal{M}_g^{1,n}$ and much is unknown about it. Harer stability tells us that $H_*(\mathcal{M}_g^{1,n})$ stabilizes as $g \to \infty$; see Harer [25] and Wahl [41]. As a consequence of homological stability for configuration spaces, it also stabilizes as $n \to \infty$. The Madsen–Weiss theorem gives the stable homology; see Galatius [19] and Madsen and Weiss [34]. (See Bödigheimer and Tillmann [5] for increasing the number of punctures.) Less is known outside of the stable range; explicit computations of $H_*(\mathcal{M}_g^{1,n})$ for low g and n can help inform and test conjectures about the homology of moduli spaces.

The computation of the homology of moduli spaces using radial slit configurations, or the closely related parallel slit configurations, is a long-term project of Bödigheimer and his students. The first example of this is Ehrenfried's thesis [15], where he computes $\mathcal{M}_2^{1,0}$. See Abhau, Bödigheimer and Ehrenfried [1] for computations of the integral homology of $\mathcal{M}_g^{1,n}$ for $2g + n \le 5$ using parallel slits. An example of an explicit computation using fat graphs is [22], in which Godin computes the integral homology of $\mathcal{M}_g^{1,0}$ for $g = 1, 2$ and $\mathcal{M}_g^{2,0}$ for $g = 1$.

1.3.2 Two-dimensional field theories, in particular string topology Combinatorial models of moduli spaces have been an important tool in the study of two-dimensional field theories. Two applications are Kontsevich's proof of the Witten conjecture [32], and Costello's classification of topological conformal field theories [10]. More concretely, combinatorial models for the moduli space of cobordisms play a role in the construction of string operations; these are operations $H_*(\mathcal{M}_g(n,m); \mathcal{L}^{\otimes d}) \otimes H_*(LM)^{\otimes n} \to H_*(LM)^{\otimes m}$ for compact oriented manifolds M. Chas and Sullivan thought of the pair of pants cobordism as a figure-eight graph [7], and many of the constructions of string operations since have used graphs. An important example is Godin's work [21], which uses $\mathcal{F}\!at^{ad}$. Using Costello's model for moduli space together with a Hochschild homology model for $H^*(LM)$, Wahl and Westerland [42; 43] not only constructed string operations, but showed that these factor through $\mathcal{SD}$. One can also use radial slit configurations to construct string operations.

A problem in string topology is that there are many constructions but few comparisons between them. The critical graph equivalence of Section 4 may help to compare constructions involving fat graphs and Sullivan diagrams to those involving radial slit configurations and the harmonic compactification.

Outline of paper

In Sections 2 and 3 we define radial slit configurations, fat graphs and their compactifications in detail. In Section 4 we use the critical graph of a radial slit configuration to construct a zigzag of homotopy equivalences between $\mathfrak{Rad}$ and $\mathcal{MF}\!at^{ad}$. In Section 5 we show that this descends to a homeomorphism between $\overline{\mathfrak{UR}\mathfrak{ad}}$ and $\mathcal{SD}$.

Acknowledgments

2 Radial slit configurations and the harmonic compactification

2.1 The definition

In this subsection we introduce Bödigheimer's radial slit configuration model for the moduli space of two-dimensional cobordisms with nonempty incoming and outgoing boundary. All material in this subsection is due to Bödigheimer, and references include [1; 2; 3; 12] and particularly [4] as it describes, in a related setting, an elegant alternative to the construction below, using subspaces of bar complexes

associated to symmetric groups. It leads, however, to a different compactification of moduli space than the harmonic compactification, so we use [3].

2.1.1 Spaces of radial slit configurations Before giving a definition of the radial slit configuration space $\mathfrak{Rad}$, we explain how to arrive at it from the perspective of building cobordisms by gluing annuli along cuts. The reader may prefer to skip this motivation and go directly to Definition 2.1.

The simplest cobordism with nonempty incoming and outgoing boundary is the cylinder, with one incoming and one outgoing boundary component. Using the theory of harmonic functions, one sees that each annulus is conformally equivalent to one of the following annuli for $R \in \left(\frac{1}{2\pi}, \infty\right)$ [24, Corollary 2.13]:

$$\mathbb{A}_R := \left\{ z \in \mathbb{C} \mid \frac{1}{2\pi} \leq |z| \leq R \right\}.$$

The reason for the choice of $\frac{1}{2\pi}$ is to facilitate comparison with fat graphs later on. We take these as our basic building blocks. Each of them has an inner boundary $\partial_{\text{in}}\mathbb{A}_R = \left\{ z \in \mathbb{C} \mid |z| = \frac{1}{2\pi} \right\}$ and an outer boundary $\partial_{\text{out}}\mathbb{A}_R = \{ z \in \mathbb{C} \mid |z| = R \}$. They come with a canonical metric, as subsets of the complex plane.

To construct a cobordism with n incoming boundary components, we start with an ordered disjoint union of n annuli $\mathbb{A}_{R_i}^{(i)}$, whose inner boundaries will be the incoming boundary of our cobordism. Next we make cuts radially inward from the outer boundaries of the annuli. Such cuts are uniquely specified by points $\zeta \in \bigsqcup_{i=1}^n \mathbb{A}_{R_i}^{(i)}$, which we will call slits. They need not be distinct. As will become clear, the number of slits must always be an even number $2h$, and we thus number them $\zeta_1, \ldots, \zeta_{2h}$. For a total genus g cobordism with n incoming and m outgoing boundary components we need $2h = 2(2g - 2 + n + m)$ slits.

We want to glue the different sides of the cuts back together. To get a metric on the surface from the metric on the cut annuli, the two cuts that we glue together must be of the same length. To get an orientation on the surface from the orientations on the cut annuli, we must glue a side clockwise from a cut to a side counterclockwise from a cut. To avoid singularities, if one side of the cut corresponding to ζ_i is glued to a side of the cut corresponding to ζ_j, the same must be true for the other two sides. Thus our gluing procedure is described by a pairing on $\{1, \ldots, 2h\}$, encoded by a permutation

$$\lambda \colon \{1, \ldots, 2h\} \to \{1, \ldots, 2h\}$$

consisting of h cycles of length 2. We should demand that if ζ_i lies on the annulus $\mathbb{A}_{R_j}^{(j)}$ and $\zeta_{\lambda(i)}$ lies on the annulus $\mathbb{A}_{R_{j'}}^{(j')}$, then $R_j - |\zeta_i| = R_{j'} - |\zeta_{\lambda(i)}|$. See Figure 1 for an example.

However, several problematic situations could occur. Firstly, if two slits ζ_i and ζ_j lie on the same *radial segment*, by definition a subset of the annulus $\mathbb{A}_{R_j}^{(j)}$ of the form

$$\{ z \in \mathbb{A}_{R_j}^{(j)} \mid \arg(z) = \theta \} \quad \text{for some } \theta,$$

then our cutting and gluing procedure is not well defined. We need to keep track of whether ζ_i lies clockwise or counterclockwise from ζ_j. To do this, we include the data of a successor permutation

$$\omega \colon \{1, \ldots, 2h\} \to \{1, \ldots, 2h\}.$$
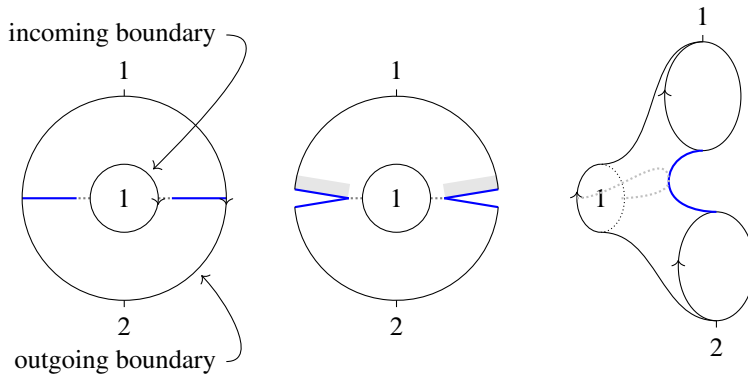
Figure 1: An example of constructing a cobordism by cutting and gluing slits in annuli. We start with the annulus on the left, cut along the blue lines to obtain the middle figure, and finally glue both the gray sides and the white sides of the cuts to get the cobordism on the right. In this simple example, the pairing λ and the successor permutation ω are uniquely determined.

This has n cycles, corresponding to the n annuli, and we should demand that each cycle contains the numbers of the slits in one of the annuli and is compatible with the weak cyclic ordering on these coming from the argument of the slits. The successor permutation keeps track of the fact that when two slits coincide, one actually lies "infinitesimally counterclockwise" from the other; see Figure 2.

This is not enough, because if all slits on an annulus lie on the same radial segment we can only deduce the ordering of the slits up to a cyclic permutation. To amend this, we add additional data: the angular distance $r_i \in [0, 2\pi]$ counterclockwise from ζ_i to $\zeta_{\omega(i)}$. In almost all cases one can deduce this from the
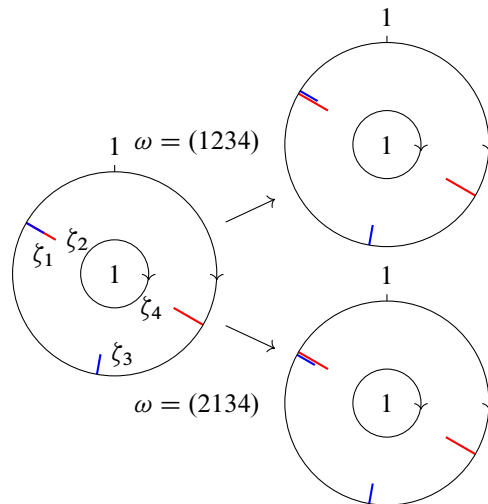
Figure 2: An example of a radial slit preconfiguration with two slits on the same radial segment; ζ_1 is the shorter blue slit and ζ_2 is the longer red slit. The successor permutation ω allows us to think of ζ_1 as either infinitesimally clockwise or counterclockwise from ζ_2.

locations of the ζ_i and ω, but in the case where all slits on an annulus lie on the same radial segment, one of them will have to be $r_i = 2\pi$, while the others will have to be $r_j = 0$. This allows one to determine the ordering of the slits, since the slit ζ_i with $r_i = 2\pi$ should be first in the clockwise direction from the angular gap between the slits.

We have almost described enough data to construct a cobordism. We can build a possibly degenerate surface, which has among its boundary components the inner boundaries of the annuli. Since we wanted m outgoing boundary components, we restrict to the subset of data that gives us m boundary components in addition to these inner boundaries of annuli. The inner boundaries of the annuli come with a canonical parametrization, but the outer ones do not. Because they already have a canonical orientation coming from the orientation of the outer boundary of the annuli, it suffices to add one point P_i in each of them, m in total. Thus we need to include these new parametrization points in ω and the r_i. To do this, we write $\xi_i = \zeta_i$ for $1 \leq i \leq 2h$ and $\xi_{2h+i} = P_i$ for $1 \leq i \leq m$, and expand our definition of ω to a permutation $\bar{\omega} \in \mathfrak{S}_{2h+m}$ and add additional $r_{2h+i} \in [0, 2\pi]$ for $1 \leq i \leq m$. It is also convenient to extend the definition of λ to a permutation $\bar{\lambda} \in \mathfrak{S}_{2h+m}$ by setting $\bar{\lambda}(2h + i) = 2h + i$ for $1 \leq i \leq m$.

Now we can state the definition of a radial slit configuration by collecting all the above data, identifying those configurations yielding the same conformal surface, and discarding those configurations yielding degenerate surfaces. Actually, it is only necessary to consider configurations with a fixed outer radius; we will say more on this towards the end of the section. Therefore, from now on we take $\vec{R} = (R, R, \dots, R)$ and $R = \frac{1}{2\pi} + \frac{1}{2}$ unless stated otherwise. This choice of outer radius is arbitrary, but it makes the connection with metric fat graphs cleanest.

Definition 2.1 The space of *possibly degenerate radial slit preconfigurations*, denoted by $\overline{\mathrm{PRad}}_h(n, m)$, is the subspace of

$$L = (\vec{\xi}, \bar{\lambda}, \bar{\omega}, \vec{r}) \in \left(\bigsqcup_{j=1}^{n} \mathbb{C} \right)^{2h+m} \times \mathfrak{S}_{2h+m} \times \mathfrak{S}_{2h+m} \times [0, 2\pi]^{2h+m}$$

with the following properties. For notation, let $\zeta_i := \xi_i$ for $1 \leq i \leq 2h$ and $P_i := \zeta_{2h+i}$ for $1 \leq i \leq m$. Then

- $\vec{\zeta} \in \left(\bigsqcup_{j=1}^{n} \mathbb{C} \right)^{2h}$ are the endpoints of the *slits*,
- $\vec{P} \in \left(\bigsqcup_{j=1}^{n} \mathbb{C} \right)^{m}$ are the *parametrization points*,
- $\bar{\lambda} \in \mathfrak{S}_{2h+m}$ is the *extended slit pairing*,
- $\bar{\omega} \in \mathfrak{S}_{2h+m}$ is the *extended successor permutation*,
- $\vec{r} \in [0, 2\pi]^{2h+m}$ are the *angular distances*.

These are subject to six conditions:

(i) Each slit ζ_i lies in $\bigsqcup_{j=1}^{n} \mathbb{A}_R^{(j)} \subset \bigsqcup_{j=1}^{n} \mathbb{C}$ and each parametrization point P_i lies in $\bigsqcup_{j=1}^{n} \partial_{\mathrm{out}} \mathbb{A}_R^{(j)}$.

(ii) The extended slit pairing $\bar{\lambda}$ consists of h 2–cycles and m 1–cycles. The latter are given by $2h + i$ for $1 \leq i \leq m$. We demand that $|\zeta_i| = |\zeta_{\bar{\lambda}(i)}|$ for all $1 \leq i \leq 2h$.

radial slits with pairing (12)

successor permutation $\omega = (12)$
angular distances $r_1 = r_2 = \pi$

outer boundary of annulus divided into
two outgoing boundary components
(here solid and dashed)

parametrization points in each
outgoing boundary component

labeled incoming boundary
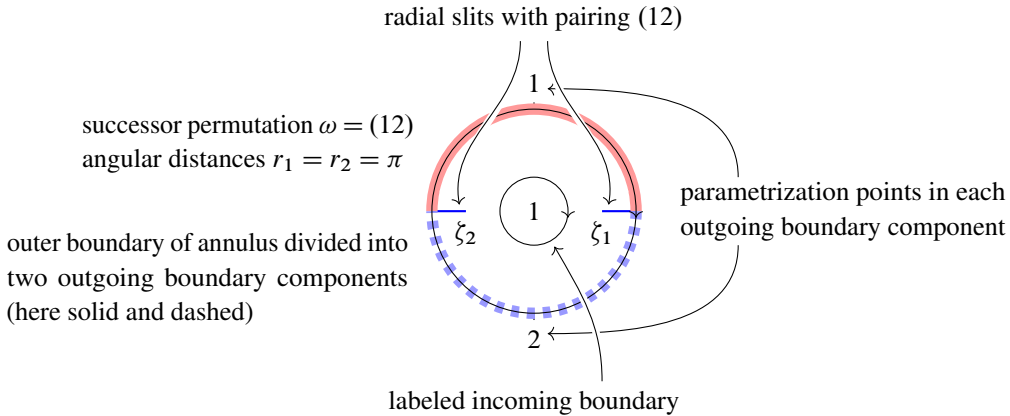
Figure 3: The configuration of Figure 1 with all its data pointed out.

(iii) The successor permutation $\bar\omega$ consists of a disjoint union of n cycles, and these cycles consist exactly of the indices of the ξ_i lying on each of the annuli. We demand that the permutation action of $\bar\omega$ on these ξ_i preserves the weakly cyclic ordering which comes from the argument (as usual taken in the counterclockwise direction).

(iv) The *boundary component permutation* $\bar\lambda \circ \bar\omega$ consists of m cycles. We will see that its cycles correspond to the outgoing boundary components.

(v) We demand that P_i lies in the subset O_i of $\bigsqcup_{j=1}^{n} \partial_{\mathrm{out}} \mathbb{A}_R^{(j)}$ which we will now define. The m cycles of $\bar\lambda \circ \bar\omega$ allow one to write the outer boundaries of the annuli as a union of m subsets, overlapping only in isolated points. We demand that each of these contains exactly one P_i, and denote that subset by O_i. To be precise, each O_i is the union of the parts in the outer boundary between the radial segments ξ_j and $\xi_{\bar\omega(j)}$ in the counterclockwise direction for all j in a cycle of $\bar\lambda \circ \bar\omega$.

(vi) The angular distances r_i must be compatible with the location of the ξ_i and the successor permutation $\bar\omega$ in the following sense. If ξ_i does not lie on an annulus with all slits and parametrization points coinciding, then r_i is equal to the angular distance in counterclockwise direction from ξ_i to $\xi_{\bar\omega(i)}$. If ξ_i lies on an annulus with all slits and parametrization points coinciding, then r_i is equal to either 0 or 2π and exactly one ξ_j on that annulus has $r_j = 2\pi$.

In terms of the previous notation, ω and λ are obtained from $\bar\omega$ and $\bar\lambda$ by deleting the elements $2h + i$ for $1 \le i \le m$ from the cycles.

We now give a construction of a possibly degenerate cobordism $S(L)$ for a preconfiguration L. To do so, we first define the sector space $\bar{\bar\Sigma}(L)$, the pieces used in the gluing construction. We slightly depart from our informal discussion by making cuts from the outer boundary to the inner boundary of the annuli and regluing these later. See Figure 4 for examples of the different types of sectors.

Definition 2.2 Let l be the number of annuli containing no elements of $\vec{\xi}$. Then $\bar\Sigma(L)$ will have $2h + m + l$ components F_i for $1 \le i \le 2h + m + l$. These come in four types:
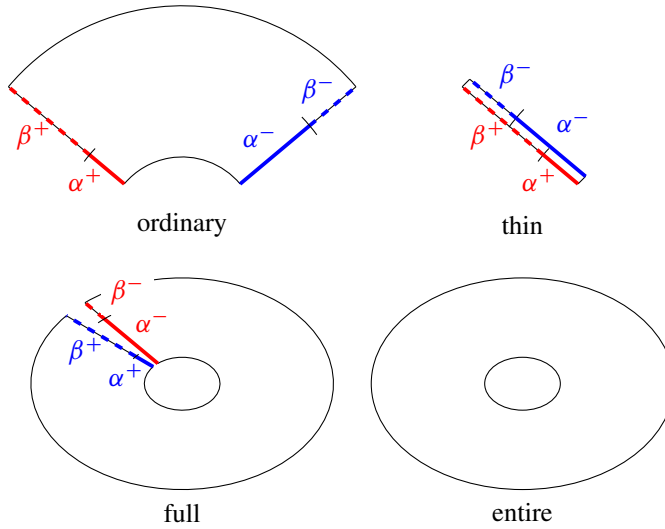
Figure 4: Examples of the different types of radial sectors with subsets $\alpha^{\pm}$ and $\beta^{\pm}$.

Ordinary sectors If $\arg(\xi_i) \neq \arg(\xi_{\overline{\omega}(i)})$ and ξ_i lies on the jth annulus $\mathbb{A}_R^{(j)}$, then we set

$$F_i = \{z \in \mathbb{A}_R^{(j)} \mid \arg(\xi_i) \leq \arg(z) \leq \arg(\xi_{\overline{\omega}(i)})\}.$$

Thin sectors If $\arg(\xi_i) = \arg(\xi_{\overline{\omega}(i)})$, $r_i = 0$ and ξ_i lies on the jth annulus $\mathbb{A}_R^{(j)}$, then we set

$$F_i = \{z \in \mathbb{A}_R^{(j)} \mid \arg(\xi_i) = \arg(z)\}.$$

Full sectors If $\arg(\xi_i) = \arg(\xi_{\overline{\omega}(i)})$, $r_i = 2\pi$ and ξ_i lies on the jth annulus $\mathbb{A}_R^{(j)}$, then we set F_i to be the annulus $\mathbb{A}_R^{(j)}$ cut open along the segment $\arg(z) = \arg(\xi_i)$, with that segment doubled so that it is homeomorphic to a closed rectangle.

Entire sectors If the jth annulus $\mathbb{A}_R^{(j)}$ does not contain any elements of $\vec{\xi}$ and is j'^{th} in the induced ordering on the r annuli that do not contain any slits, we set $F_{2h+m+j'} = \mathbb{A}_R^{(j)}$.

The surface $\Sigma(L)$ underlying the cobordism $S(L)$ will be obtained as a quotient space of the sector space by an equivalence relation that makes identifications on the boundary of the sectors. We next define the subsets involved in those identifications.

Definition 2.3 If F_i is an ordinary or thin sector corresponding to the element ξ_i on the jth annulus $\mathbb{A}_R^{(j)}$, then we define the following subspaces of F_i:

$$\alpha_i^+ := \{z \in \mathbb{A}_R^{(j)} \mid \arg(z) = \arg(\xi_{\overline{\omega}(i)}) \text{ and } |z| \leq |\xi_{\overline{\omega}(i)}|\},$$
$$\alpha_i^- := \{z \in \mathbb{A}_R^{(j)} \mid \arg(z) = \arg(\xi_i) \text{ and } |z| \leq |\xi_i|\},$$
$$\beta_i^+ := \{z \in \mathbb{A}_R^{(j)} \mid \arg(z) = \arg(\xi_{\overline{\omega}(i)}) \text{ and } |z| \geq |\xi_{\overline{\omega}(i)}|\},$$
$$\beta_i^- := \{z \in \mathbb{A}_R^{(j)} \mid \arg(z) = \arg(\xi_i) \text{ and } |z| \geq |\xi_i|\}.$$

If F_i is a full sector, then our definitions are different, because the two radial segments in the boundary have the same argument. Let S_i^+ be the radial segment bounding F_i in the counterclockwise direction and S_i^- be the radial segment bounding it in the clockwise direction. Then we define the following subspaces of F_i:

$$\alpha_i^+ := \{z \in S_i^+ \mid |z| \leq |\xi_{\overline{\omega}(i)}|\}, \quad \alpha_i^- := \{z \in S_i^- \mid |z| \leq |\xi_i|\},$$
$$\beta_i^+ := \{z \in S_i^+ \mid |z| \geq |\xi_{\overline{\omega}(i)}|\}, \quad \beta_i^- := \{z \in S_i^- \mid |z| \geq |\xi_{\overline{\omega}i}|\}.$$

These subspaces are empty for entire sectors.

Definition 2.4 The equivalence relation $\approx_L$ on $\overline{\Sigma}(L)$ is the one generated by identifying

(i) $z \in \alpha_i^+$ with $z \in \alpha_{\overline{\omega}(i)}^-$, and

(ii) $z \in \beta_i^+$ with $z \in \beta_{\overline{\lambda}(i)}^-$.

We define the surface $\Sigma(L)$ to be $\overline{\Sigma}(L)/\approx_L$.

Definition 2.5 The cobordism $S(L)$ has underlying surface $\Sigma(L)$. It has a map from each inner boundary $\partial_{\mathrm{in}}\mathbb{A}_R^{(j)}$

$$\iota_j^{\mathrm{in}} \colon S^1 \cong \partial_{\mathrm{in}}\mathbb{A}_R^{(j)} \to \Sigma(L),$$

and these are inclusions of subspaces if none of the slits lie on the inner boundary of an annulus. One can define the outgoing boundary components as a subspace of $\Sigma(L)$ by considering the intersection of the outer boundary of the annuli with the sectors. For each cycle in $\lambda \circ \omega$ these intersections form a circle with canonical orientation and starting point P_k. This yields, for the cycle $\lambda \circ \omega$ corresponding to P_k, a map

$$\iota_k^{\mathrm{out}} \colon S^1 \to \Sigma(L),$$

and these are inclusions of subspaces if none of the slits lie on the outer boundary of an annulus.

As mentioned before, this definition may result in a degenerate cobordism for some L. Moreover, two different preconfigurations might give the same conformal classes of cobordism. In fact, each conformal class of cobordisms occurs at least $(2h)!$ times, because the labeling on the slits does not matter. To see that degenerate surfaces can occur, consider the example in Figure 5. Now we explain how to resolve both issues.

We have already explained that one should identify configurations obtained by permuting the labels on the slits. We only need to make two additional identifications. For the first, instead of doing all the
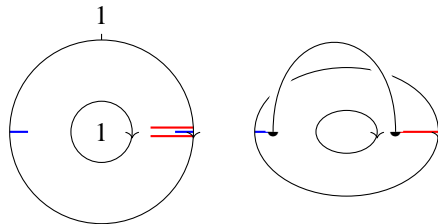
Figure 5: An example of a radial slit preconfiguration leading to a degenerate surface. The black arc connecting two points on the surface on the right was the line segment between the two red slits.
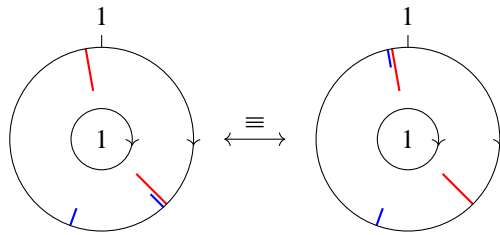
Figure 6: A jump of a slit. The pairing λ is given by the colors, but is uniquely determined by the configuration.

cutting and gluing simultaneously, do it in order of increasing modulus of the slits. This results in the same cobordism, but doing so makes clear it that if ζ_i lies on the same radial segment as ζ_j and satisfies $|\zeta_i| \geq |\zeta_j|$, it might as well be on the other side of $\zeta_{\lambda(j)}$. That is, it might as well have "jumped" over the slit ζ_j to $\zeta_{\lambda(j)}$. For the second, note that if a parametrization point similarly "jumps" over a slit, this does not change the parametrization of the outgoing boundary. These will turn out to be all required identifications, and we now use them to define equivalence relations on $\mathrm{PRad}_h(n,m)$.

Definition 2.6 Let $\equiv'$ be the equivalence relation on $\overline{\mathrm{PRad}}_h(n,m)$ generated by:

Relabeling of the slits We identify two preconfigurations if they can be obtained from each other by relabeling the slits. More precisely, for every permutation $\sigma \in \mathfrak{S}_{2h}$ extended by the identity to a permutation $\bar{\sigma} \in \mathfrak{S}_{2h+m}$ and $L = (\vec{\xi}, \bar{\lambda}, \bar{\omega}, \vec{r}) \in \mathrm{PRad}_h(n,m)$, we say that $L \equiv' \sigma(L)$ with

$$\sigma(L) = ((\vec{\xi})^{\bar{\sigma}}, (\bar{\lambda})^{\bar{\sigma}}, (\bar{\omega})^{\bar{\sigma}}, (\vec{r})^{\bar{\sigma}}),$$

where

- $(\vec{\xi})^{\bar{\sigma}}$ is given by $(\xi)_i^{\bar{\sigma}} = \xi_{\bar{\sigma}(i)}$,
- $(\bar{\lambda})^{\bar{\sigma}} = \bar{\sigma} \circ \bar{\lambda} \circ \bar{\sigma}^{-1}$,
- $(\bar{\omega})^{\bar{\sigma}} = \bar{\sigma} \circ \bar{\omega} \circ \bar{\sigma}^{-1}$,
- $(\vec{r})^{\bar{\sigma}}$ is given by $(r)_i^{\bar{\sigma}} = r_{\bar{\sigma}(i)}$.

Let $\equiv$ be the equivalence relation on $\overline{\mathrm{PRad}}_h(n,m)$ generated by relabeling of the slits (as above) and the following two identifications:

Slit jumps We say $L \equiv L'$ if L' can be obtained from L by a slit jump; see Figure 6. More precisely, if we are given a preconfiguration L and two indices i and j such that $j = \omega(i)$, $r_i = 0$ and $|\zeta_i| \geq |\zeta_j|$, then we can obtain a new preconfiguration L' as follows. We replace ζ_i by the point $\zeta_i' = (|\zeta_i|/|\zeta_{\lambda(j)}|)\zeta_{\lambda(j)}$ and keep all the other slits the same. We then put i after $\lambda(j)$ in $\bar{\omega}$ to obtain $\bar{\omega}'$, and set $r_i' = r_{\lambda(j)}$ and $r_{\lambda(j)}' = 0$. The rest of the data remains the same.

Parametrization point jumps We say $L \equiv L'$ if L' can be obtained from L by a jump of a parametrization point; see Figure 7. More precisely, if we are given a preconfiguration L in which there is a P_i such that $j = \bar{\omega}(i + 2h)$ for some j and $r_{i+2h} = 0$, then we can obtain a new preconfiguration L' by keeping
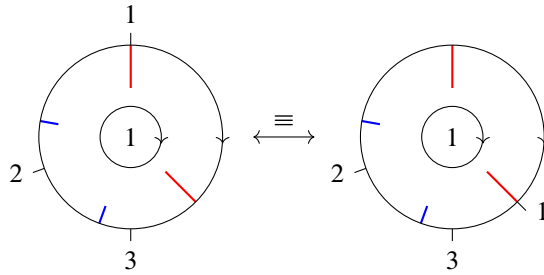
Figure 7: A jump of a parametrization point.

all the data the same, except replacing P_i with P_i' lying at the radial segment through $\zeta_{\lambda(j)}$ and setting $r_{i+2h}' = r_{\lambda(j)}$ and $r_{\lambda(j)}' = 0$.

Definition 2.7 We now define certain quotient spaces using these equivalence relations:

- the space $\overline{\mathrm{QRad}}_h(n,m)$ of *unlabeled possibly degenerate radial slit preconfigurations*, the quotient of $\overline{\mathrm{PRad}}_h(n,m)$ by $\equiv'$,
- the space $\overline{\mathrm{Rad}}_h(n,m)$ of *possibly degenerate radial slit configurations*, the quotient of $\overline{\mathrm{QRad}}_h(n,m)$ by $\equiv$.

We will denote by $[L]$ the radial slit configuration represented by a preconfiguration L. We are left to deal with the problem that certain preconfigurations give cobordisms whose underlying surface is degenerate. We call such preconfigurations *degenerate*. In [3], Bödigheimer gave a necessary and sufficient criterion for a (pre)configuration to lead to a degenerate surface:

Proposition 2.8 *The surface underlying the cobordism $\Sigma(L)$ constructed out of a preconfiguration L is degenerate if and only if it is equivalent under $\equiv$ to a preconfiguration satisfying at least one of the following three conditions:*

- **Slit hitting inner boundary** *There is a slit ζ_i with $|\zeta_i| = \frac{1}{2\pi}$.*
- **Slit hitting outer boundary** *There is a slit ζ_i on an annulus $\mathbb{A}_R^{(j)}$ with $|\zeta_i| = R_j$.*
- **Slits are "squeezed"** *There is a pair (i,j) such that $j = \lambda(i)$, ζ_i and ζ_j lie on the same annulus, $\zeta_i = \zeta_j$ and, for all k between i and j in the cyclic ordering coming from ω, we have that $|\zeta_k| \geq |\zeta_i| = |\zeta_j|$; see Figure 5 for an example. If all slits on the annulus containing ζ_i and ζ_j lie at the same point, we additionally require that $r_k = 0$ for all of the k between i and j.*

Definition 2.9 A radial slit preconfiguration is said to be *generic* if it is not equivalent to any other by slit or parametrization point jumps, ie all the slits are disjoint.

Definition 2.10 We define the following spaces:

- The space $\mathrm{PRad}_h(n,m)$ of *labeled radial slit preconfigurations* is the subspace of $\overline{\mathrm{PRad}}_h(n,m)$ consisting of nondegenerate preconfigurations.

- The space $\mathrm{QRad}_h(n,m)$ of *unlabeled radial slit preconfigurations* is the subspace of $\overline{\mathrm{QRad}}_h(n,m)$ consisting of equivalence classes with nondegenerate representatives.

- The space $\mathrm{Rad}_h(n,m)$ of *radial slit configurations* is the subspace of $\overline{\mathrm{Rad}}_h(n,m)$ consisting of equivalence classes with nondegenerate representatives.

2.1.2   Cell complexes of radial slit configurations Next we give CW–complexes $\overline{\mathfrak{Rad}}$ and $\mathfrak{Rad}$ homeomorphic to the spaces of radial slit configurations given before. On $\overline{\mathfrak{Rad}}$ this is the CW–structure given in [3, Section 8.2], and on the subspace $\mathfrak{Rad}$ it coincides with the radial analogue of [4]. The cells will be indexed by so-called combinatorial types, which we define first.

Definition 2.11 Fix an L in $\mathrm{PRad}_h(n,m)$.

- The radial segments of the slits, the parametrization points and the positive real lines divide the annuli of the preconfiguration L radially into different pieces, which we will call *radial chambers*; see Figure 8.

- Each slit ζ_i in L defines a circle of radius $|\zeta_i|$ on all of the n annuli. These circles divide the n annuli into different pieces, which we will call *annular chambers*; see Figure 8.

Remark 2.12 The orientation of the complex plane endows the radial chambers on each annulus with a natural ordering, and similarly the modulus endows the annular chambers with a natural ordering; see Figure 8.

Each of the annular chambers is homeomorphic to a disjoint union of n annuli, while each of the radial chambers is homeomorphic to a rectangle.

Definition 2.13 Two preconfigurations L and L' in $\mathrm{PRad}_h(n,m)$ are said to *have the same combinatorial data* if L' can be obtained from L by continuously moving the slits and parametrization points in each complex plane without collapsing any chamber. This defines an equivalence relation on $\mathrm{PRad}_h(n,m)$.

A *combinatorial type of preconfigurations* $\mathcal{L}$ is an equivalence class of preconfigurations under this relation. Informally, a combinatorial type is the data carried over by the picture of a preconfiguration without remembering the precise placement of the slits. Notice that this equivalence relation is also well defined on the sets of radial slit configurations $[L]$. Thus one can similarly define a *combinatorial type of configurations* $[\mathcal{L}]$ to be an equivalence class of configurations under this relation. We make a similar definition for the case of unlabeled radial slit configurations.

We will use Υ to denote the *set of all combinatorial types of configurations*.

Remark 2.14 If L is a degenerate (or nondegenerate) preconfiguration, then so is any preconfiguration of the same combinatorial type. Thus, we can talk about a degenerate or nondegenerate combinatorial type.
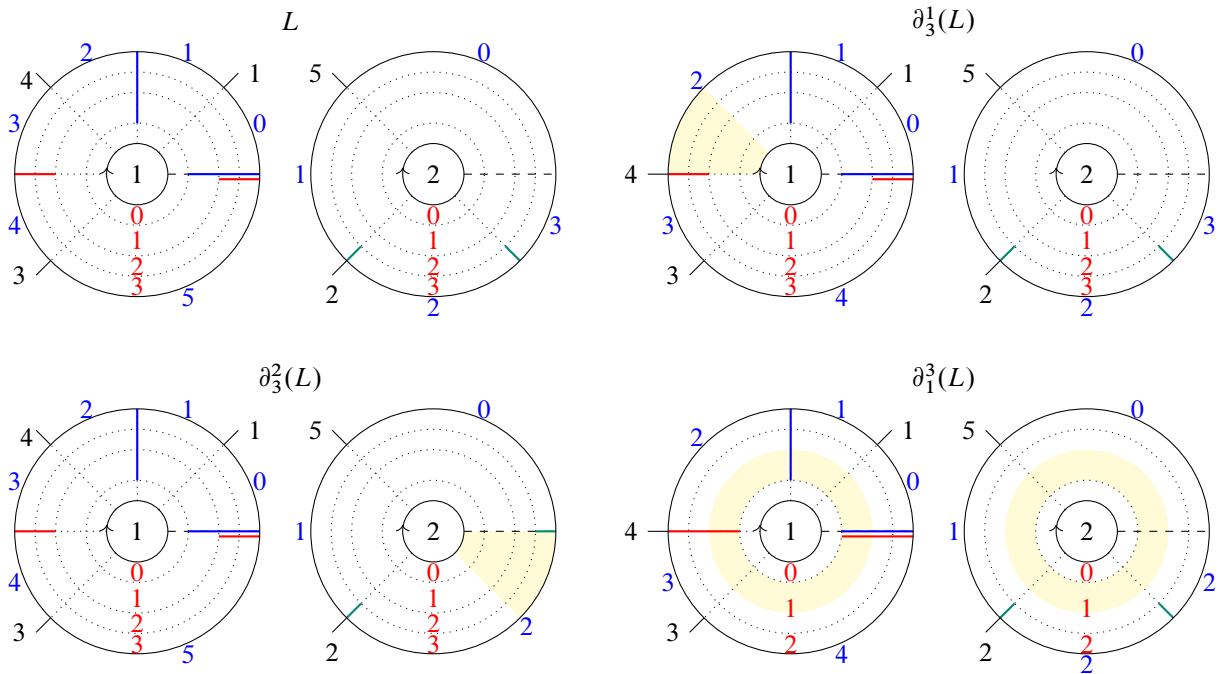
Figure 8: Top left: a configuration L and its radial and annular chambers divided by dotted lines. The radial chambers are numbered in blue (there are 6 radial chambers on the left annulus and 4 on the right annulus) and the annular chambers are numbered in red (there are 3 annular chambers consisting of a pair small annuli, one on each of the annuli). This combinatorial type gives an 11–cell in $\overline{\mathfrak{Rad}}$ given by $\Delta^5 \times \Delta^3 \times \Delta^3$. Top right and bottom: parts of the boundary of L and their chambers. The modified parts are marked in light yellow.

Now we give definitions of cell complexes of (pre)configurations and their compactifications. Note that the meaning of p and q is different from their meaning in [3].

Definition 2.15 The *multidegree of a combinatorial type* $[\mathcal{L}]$ on n annuli is the $(n+1)$–tuple of integers $(q_1, \ldots, q_n, p)$, where $q_i + 1$ is the number of radial chambers in the ith annulus and $p + 1$ is the number of annular chambers. For $0 \le j \le q_i$ and $0 \le i \le n$, we denote by $d_j^i([\mathcal{L}])$ the combinatorial type obtained by collapsing the jth radial chamber on the ith annulus; see Figure 8. For $0 \le j \le p$, we denote by $d_j^{n+1}([\mathcal{L}])$ the combinatorial type obtained by collapsing the jth annular chamber; see Figure 8.

The *cell complex of possibly degenerate radial slit configurations* $\overline{\mathfrak{Rad}}_h(n, m)$ is the realization of the multisimplicial set with

- $(q_1, \ldots, q_n, p)$–simplices given by

$$\{e_{[\mathcal{L}]} \mid [\mathcal{L}] \text{ is a combinatorial type of multidegree } (q_1, \ldots, q_n, p)\},$$

- the faces of $e_{[\mathcal{L}]}$ given by $d_j^i(\sigma_{[\mathcal{L}]}) := \sigma_{d_j^i([\mathcal{L}])}$.
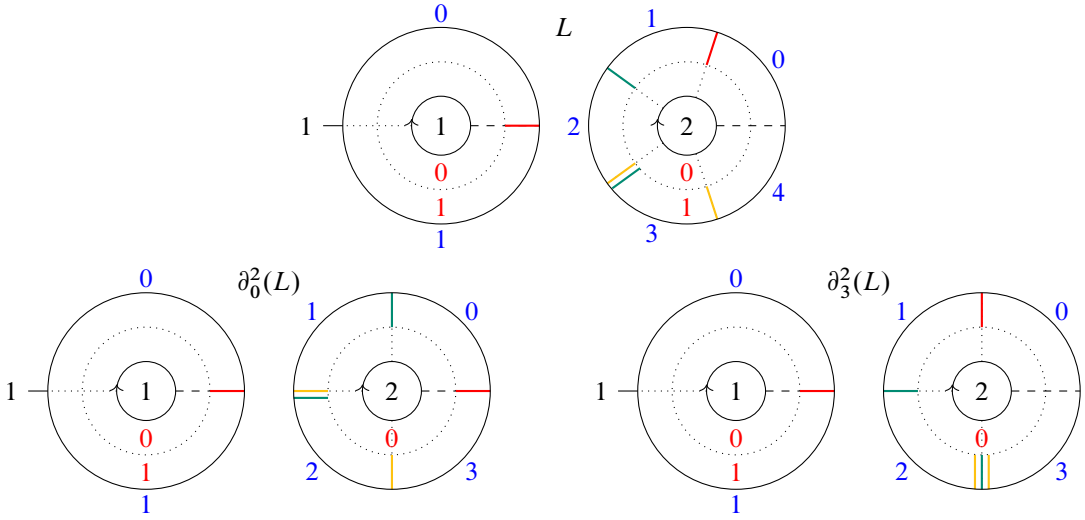
Figure 9: A second example of a cell and parts of its boundary. Here all slits have the same length.

That is, $\overline{\mathfrak{Rad}}_h(n,m)$ is a CW–complex with cells indexed by combinatorial types of radial slits configurations as follows. Let $e_{[\mathcal{L}]} := \Delta^{q_1} \times \cdots \times \Delta^{q_n} \times \Delta^p$. Then

$$\overline{\mathfrak{Rad}}_h(n,m) := \frac{\bigsqcup_{[\mathcal{L}] \in \Upsilon} e_{[\mathcal{L}]}}{\sim},$$

where the equivalence relation is generated by

$$(e_{[\mathcal{L}]}, (\vec{t}_1, \ldots, \delta^j(\vec{t}_i), \ldots, \vec{t}_{n+1})) \sim (e_{d_j^i([\mathcal{L}])}, (\vec{t}_1, \ldots, \vec{t}_i, \ldots, \vec{t}_{n+1})).$$

Here δ^j is the map $\Delta^{q_i-1} \to \Delta^{q_i}$ including 0 as the $(j+1)^{\text{st}}$ coordinate, and Υ is the set of combinatorial types of radial slit configurations.

The cell complexes of possibly degenerate radial slit preconfigurations $\overline{\mathfrak{PRad}}_h(n,m)$ and unlabeled preconfigurations $\overline{\mathfrak{QRad}}_h(n,m)$ are defined in similar ways.

Definition 2.16 If a combinatorial type $[\mathcal{L}]$ is degenerate, then $d_j^i([\mathcal{L}])$ is also degenerate. Thus, we define the *cell complex of degenerate radial slit configurations* as the subcomplex $\overline{\mathfrak{Rad}}_h(n,m)' \subset \overline{\mathfrak{Rad}}_h(n,m)$ obtained as the realization of the degenerate simplices. Finally, $\mathfrak{Rad}_h(n,m)$ is the complement. That is,

$$\mathfrak{Rad}_h(n,m) := \overline{\mathfrak{Rad}}_h(n,m) \setminus \overline{\mathfrak{Rad}}_h(n,m)'.$$

The spaces $\mathfrak{PRad}_h(n,m)$ and $\mathfrak{QRad}_h(n,m)$ are defined in a similar way.

We introduce notation for the image of $e_{[\mathcal{L}]}$ in $\mathfrak{Rad}$:

Definition 2.17 Letting $[\mathcal{L}]$ be a combinatorial type, we define the subspace $\mathfrak{Rad}_{[\mathcal{L}]}$ as the image of the interior of $e_{[\mathcal{L}]}$. We also let $\overline{\mathfrak{Rad}}_{[\mathcal{L}]}$ be the closure of $\mathfrak{Rad}_{[\mathcal{L}]}$ in $\overline{\mathfrak{Rad}}$, and define $\partial \overline{\mathfrak{Rad}}_{[\mathcal{L}]} = \mathfrak{Rad} \cap (\overline{\mathfrak{Rad}}_{[\mathcal{L}]} \setminus \mathfrak{Rad}_{[\mathcal{L}]})$.

2.1.3 Relationships Our final goal for this section is to explain the relationship between the spaces and cell complexes of radial slit configurations and the moduli space of cobordisms. The first relationship is straightforward, as there are continuous bijections

$$\text{Rad}_h(n,m) \to \mathfrak{Rad}_h(n,m), \qquad \overline{\text{Rad}}_h(n,m) \to \overline{\mathfrak{Rad}}_h(n,m),$$

$$\text{QRad}_h(n,m) \to \mathfrak{QRad}_h(n,m), \quad \overline{\text{QRad}}_h(n,m) \to \overline{\mathfrak{QRad}}_h(n,m),$$

$$\text{PRad}_h(n,m) \to \mathfrak{PRad}_h(n,m), \quad \overline{\text{PRad}}_h(n,m) \to \overline{\mathfrak{PRad}}_h(n,m),$$

compatible with the quotient maps and inclusions. These are given by sending a point to its combinatorial type and the simplicial coordinates obtained by rescaling the angles of the slits (for the first n coordinates) and their radii (for the last coordinate). The next lemma follows from [3], and we sketch a proof below.

Lemma 2.18 *These maps are homeomorphisms.*

Proof We start by noting that $\overline{\text{PRad}}_h(n,m)$ and $\overline{\mathfrak{PRad}}_h(n,m)$ are both compact Hausdorff spaces; the former is a closed subset of a compact Hausdorff space and the latter is a finite CW–complex. A continuous bijection between compact Hausdorff spaces is a homeomorphism. Next note that the maps $\overline{\text{Rad}}_h(n,m) \to \overline{\mathfrak{Rad}}_h(n,m)$ and $\overline{\text{QRad}}_h(n,m) \to \overline{\mathfrak{QRad}}_h(n,m)$ are induced by passing to quotients, as are their inverses, so they are also homeomorphisms.

Thus the maps on the right are homeomorphisms and the maps on the left are obtained by restricting these homeomorphisms to open subsets and replacing their codomain with their image. Hence they are also homeomorphisms. $\qquad\square$

The relationship to moduli space is less straightforward. In [3, Section 9], Bödigheimer defined a space $\overline{\mathbf{RAD}}_h(n,m)$ of all radial slit configurations with varying inner radii but fixed outer radii, and a subspace $\mathbf{RAD}_h(n,m)$ of all nondegenerate radial slit configurations. He also proved a version of the previous lemma.

Lemma 2.19 *There are homotopy equivalences*

$$\mathbf{RAD}_h(n,m) \simeq \text{Rad}_h(n,m) \quad and \quad \overline{\mathbf{RAD}}_h(n,m) \simeq \overline{\text{Rad}}_h(n,m).$$

Sketch of proof To explain the existence of these homotopy equivalences, we note that Bödigheimer's $\overline{\mathbf{RAD}}$ and $\mathbf{RAD}$ differ from $\overline{\text{Rad}}$ and Rad only in the following two ways:

   (i) In $\overline{\mathbf{RAD}}$ and $\mathbf{RAD}$ the inner radii are allowed to vary in $(0, R_0)$ for some choice of $R_0 > 0$, while in $\overline{\text{Rad}}$ and Rad they are fixed to $\frac{1}{2\pi}$.

  (ii) In $\overline{\mathbf{RAD}}$ and $\mathbf{RAD}$ an exceptional set Ω is used to remove ambiguity when all slits on an annulus lie on two segments, while in $\overline{\text{Rad}}$ and Rad this role is played by the angular distances $\vec{r}$.

The second of these encodes equivalent data; given the rest of the data of a radial slit configuration, Ω can be reconstructed from $\vec{r}$ and vice versa. The first says that the difference between the two spaces is in the choices of radii. More precisely, there is an inclusion $\overline{\text{Rad}} \hookrightarrow \overline{\mathbf{RAD}}$ with homotopy inverse

given by decreasing all radii to $\min(R_i)$ and changing the radial coordinates of all the data by an affine transformation that sends $\min(R_i)$ to $\frac{1}{2\pi}$ and fixes 1. This homotopy equivalence restricts to one between **RAD** and Rad. □

Bödigheimer proved in [3, Section 7.5], with additional details in [12], that a version of $\mathbf{RAD}_h(n,m)$ without parametrization points on the outgoing boundary is a model for the moduli space of cobordisms without parametrization of the outgoing boundary. This uses that $\Sigma(L)$ comes with a canonical conformal structure, being obtained by gluing subsets of $\mathbb{C}$. Adding in the parametrizations for the outer boundary, this result implies:

Theorem 2.20 (Bödigheimer) *The map that assigns to each $[L] \in \mathbf{RAD}_h(n,m)$ the conformal class of the cobordism $S(L)$ gives a homeomorphism*

$$\mathbf{RAD}_h(n,m) \cong \bigsqcup \mathcal{M}_g(n,m),$$

where the disjoint union is over triples (g,n,m) satisfying $h = 2g - 2 + n + m$.

By the remarks above,

$$\mathfrak{Rad}_h(n,m) \simeq \bigsqcup_{[\Sigma]} B\mathrm{Diff}(\Sigma, \partial\Sigma),$$

where the disjoint union is over two-dimensional cobordisms with $n \geq 1$ incoming boundary components, $m \geq 1$ outgoing boundary components and total genus $g \geq 0$.

Bödigheimer proved Theorem 2.20 for connected cobordisms with no parametrization of the outgoing boundary, but this version of the theorem is an easy consequence of his. His proof amounts to checking that $\mathbf{RAD}_h(n,m)$ is a manifold of dimension $3h + m + n$ (see also [13] for remarks on the real-analytic structure). It sits as a dense open subset in $\overline{\mathbf{RAD}}_h(n,m)$. In this way we can think of $\overline{\mathbf{RAD}}_h(n,m)$ as a "compactification" of $\mathbf{RAD}_h(n,m)$. Informally it is the compactification where handles or boundary components can degenerate to radius zero, as long as there is always a path from each incoming boundary component to an outgoing boundary component that does not pass through any degenerate handles or boundary components. Colloquially, "the water must always be able to leave the tap". Bödigheimer calls this the *harmonic compactification of moduli space*. We now describe a deformation retract of it:

Definition 2.21 The *unilevel harmonic compactification* $\overline{\mathfrak{URad}}_h(n,m)$ is the subspace of $\overline{\mathfrak{Rad}}_h(n,m)$ given by cells corresponding to configurations satisfying $|\zeta_i| = R$ for all $i \in \{1, \ldots, 2h\}$, ie all slits lie on the outer radius.

In addition to the inclusion $\iota : \overline{\mathfrak{URad}}_h(n,m) \hookrightarrow \overline{\mathfrak{Rad}}_h(n,m)$, there is also a projection $p : \overline{\mathfrak{Rad}}_h(n,m) \to \overline{\mathfrak{URad}}_h(n,m)$ which makes all slits have modulus R.

Lemma 2.22 *The maps ι and p are mutually inverse, up to homotopy.*

Proof The map $p \circ \iota$ is equal to the identity on $\overline{\mathfrak{UR}\mathfrak{ad}}$. For $\iota \circ p$, a homotopy from the identity on $\overline{\mathfrak{R}\mathfrak{ad}}$ to $\iota \circ p$ is given at time $t \in [0, 1]$ by sending each slit ζ_i to $\big(((1-t)|\zeta_i| + Rt)/|\zeta_i|\big)\zeta_i$ under the homeomorphism with Rad. $\qquad\square$

The spaces constructed in this section fit together in the diagram

$$
\begin{array}{ccc}
\mathfrak{PR}\mathfrak{ad}_h(n,m) & \xrightarrow{\text{compactification}} & \overline{\mathfrak{PR}\mathfrak{ad}}_h(n,m) \\
\downarrow & & \downarrow \\
\mathfrak{QR}\mathfrak{ad}_h(n,m) & \xrightarrow{\text{compactification}} & \overline{\mathfrak{QR}\mathfrak{ad}}_h(n,m) \\
\downarrow & & \downarrow \\
\mathfrak{R}\mathfrak{ad}_h(n,m) & \xrightarrow{\text{compactification}} & \overline{\mathfrak{R}\mathfrak{ad}}_h(n,m) \xrightleftharpoons{\;\simeq\;} \overline{\mathfrak{UR}\mathfrak{ad}}_h(n,m)
\end{array}
$$

where all the horizontal maps within the squares are inclusions.

Remark 2.23 One can make sense of gluing of cobordisms on the level of radial slits; see [3]. This construction gives $\mathbf{RAD}_h(n,m)$ the structure of a PROP in topological spaces. One of the advantages of the radial slit configurations over fat graphs is the ease with which one can describe the PROP structure.

2.2 The universal surface bundle

In the previous section, we motivated radial slit configurations by explaining that a preconfiguration consists of data to construct a cobordism $S(L)$. The topology on the collection of radial slit configurations was guided by the idea that this construction produces a conformal family of cobordisms. In this section we make this precise by defining a universal surface bundle over $\mathfrak{R}\mathfrak{ad}$ via its homeomorphism with Rad.

The equivalence relation $\equiv$ on $\mathrm{PRad}_h(n,m)$ is such that there is a canonical isomorphism of cobordisms with conformal structure between $S(L)$ and $S(L')$ if $L \equiv L'$. Thus we can make sense of the cobordism $S([L])$ for an equivalence class $[L]$. The idea for constructing the universal surface bundle over $\mathrm{Rad}_h(n,m)$ is to make the construction of $S([L])$ continuous in $[L]$. The result is a space over $\mathrm{Rad}_h(n,m)$, and we check it is a universal bundle by comparing it to the definition of the universal bundle in the conformal construction of moduli space.

We first make sense of the radial sectors $\overline{\Sigma}(L)$ as a space over $\overline{\mathrm{PRad}}_h(n,m)$. This seems obvious; we think of the sectors as a subspace of a disjoint union of annuli for each L, so one is tempted to just state that $\widetilde{\Sigma}(L)$ is the relevant subspace of

$$
\overline{\mathrm{PRad}}_h(n,m) \times \left(\coprod_{j=1}^{n} \mathbb{A}_R^{(j)} \right).
$$

Two minor problems arise:

(i) the full sectors are not actually subspaces of annuli, and

(ii) the number of entire sectors is not constant over $\overline{\mathrm{PRad}}_h(n,m)$.

Both are relatively harmless. Problem (ii) is solved by noting that the number of entire sectors is locally constant, so one can work separately over each of the subspaces of components with a fixed number of entire sectors. Problem (i) is solved by considering a version of $\overline{\mathrm{PRad}}_h(n,m)$ where the preconfigurations L are endowed with lifts of the slits to elements of $\bigsqcup_{i=1}^{n} \tilde{\mathbb{A}}_R$, the disjoint union of the universal covers of the annuli, under the condition that the distances between them are still equal to the angular distances. Over this version, one has a space with fibers given by $\bigsqcup_{i=1}^{n} \tilde{\mathbb{A}}_R$, which does contain the full sectors. One then notes that there is a canonical homeomorphism between the sectors over the same configurations with different choices of lifts. In the end, we conclude there exists a space $\tilde{\mathbb{A}}$ over $\overline{\mathrm{PRad}}_h(n,m)$ whose fibers consist of a disjoint union of annuli, and there is a subspace $\overline{\mathrm{PS}}_h(n,m) \subset \tilde{\mathbb{A}}$ whose fiber over L can be canonically identified with the sector space $\tilde{\Sigma}(L)$.

Recall that $\approx_L$ is the equivalence relation on $\overline{\Sigma}(L)$ used when gluing the sectors together to obtain a surface. Using it fiberwise defines an equivalence relation $\sim$:

Definition 2.24 Let $\sim$ be the equivalence relation on $\overline{\mathrm{PS}}_h(n,m)$ generated by $(L,z) \sim (L',z')$, where $L, L' \in \overline{\mathrm{PRad}}_h(n,m)$, $z \in \overline{\Sigma}(L) \subset \overline{\mathrm{PS}}_h(n,m)$ and $z' \in \overline{\Sigma}(L') \subset \overline{\mathrm{PS}}_h(n,m)$, if $L = L'$ and $z \approx_L z'$.

As mentioned before, there is a canonical isomorphism $\phi_{L,L'}$ between $\Sigma(L)$ and $\Sigma(L')$ if $L \equiv L'$. Using this, we can define a version of $\equiv$ for $\overline{\mathrm{PS}}_h(n,m)$:

Definition 2.25 Let $\cong$ be the equivalence relation on $\overline{\mathrm{PS}}_h(n,m)$ generated by $\sim$ and by saying that (L,z) and (L',z') are equivalent if $L \equiv L'$ and $z' = \phi_{L,L'}(z)$.

We can now define the surface bundle.

Definition 2.26 We define $\mathrm{PS}_h(n,m)$ to be the restriction of $\overline{\mathrm{PS}}_h(n,m)$ to $\mathrm{PRad}_h(n,m)$. We then define $\mathrm{S}_h(n,m)$ as $\mathrm{PS}_h(n,m)/\cong$, which is a space over $\mathrm{Rad}_h(n,m)$.

A priori this is a space over $\mathrm{Rad}_h(n,m)$ with fibers having the structure of cobordisms, but it is in fact a universal surface bundle. This is implicit in [3], but not explicitly stated there. We explain the reasoning below:

Proposition 2.27 *The space $\mathrm{S}_h(n,m)$ over $\mathrm{Rad}_h(n,m)$ is a universal surface bundle.*

Sketch of proof Varying radii allows one to extend $\mathrm{S}_h(n,m)$ to $\mathbf{RAD}_h(n,m)$. Theorem 2.20 tells us that the assignment $[L] \mapsto [S([L])]$ gives a homeomorphism $\mathbf{RAD}_h(n,m) \to \mathcal{M}_g(n,m)$. Pulling back the universal bundle over $\mathcal{M}_g(n,m)$ defined at the end of Section 1.1 exactly gives $\mathrm{S}_h(n,m)$.   □

There is a universal $\mathrm{Mod}(S_{g,n+m})$–bundle over $\mathrm{Rad}_h(n,m)$ given by the bundle with fiber over $[L]$ the isotopy classes of diffeomorphisms of $\Sigma(L)$ fixing the boundary. We give an alternative explicit construction of this bundle in Definition 4.46.

3 Admissible fat graphs and string diagrams

3.1 The definition

Following Strebel [39], Penner, Bowditch and Epstein gave a triangulation of Teichmüller space of surfaces with decorations, which is equivariant under the action of its corresponding mapping class group [6; 37]. In this triangulation, simplices correspond to equivalence classes of marked fat graphs and the quotient of this triangulation gives a combinatorial model of the moduli space of surfaces with decorations. These ideas were studied by Harer for surfaces with punctures and boundary components [26] and used by Igusa to construct a category of fat graphs that models the mapping class groups of punctured surfaces [29]. Godin extended Igusa's construction to surfaces with boundary and open–closed cobordisms [21; 22].

In this section we define a category of fat graphs, as well as specific subcategories of it, in the spirit of Godin. We also define the space of metric fat graphs in the spirit of Harer and Penner, as well as specific subspaces of these spaces, and show that these are the classifying spaces of these categories. Finally, we define the space of Sullivan diagrams as a quotient of a certain subspace of the space of metric fat graphs. It plays the role of a compactification.

3.1.1 Fat graphs We start with precise definitions of graphs and fat graphs:

Definition 3.1 A *combinatorial graph* G is a tuple $G = (V, H, s, i)$ with a finite set of *vertices* V, a finite set of *half-edges* H, a *source map* $s\colon H \to V$ and an *edge pairing* involution $i\colon H \to H$ without fixed points.

The source map s ties each half-edge to its source vertex, and the edge pairing involution i attaches half-edges together. The set E of *edges* of the graph is the set of orbits of i. The *valence* of a vertex $v \in V$ is the cardinality of the set $s^{-1}(v)$. A *leaf* of a graph is a univalent vertex and an *inner vertex* is a vertex that is not a leaf. The *geometric realization* of a combinatorial graph G is the CW–complex $|G|$ with one 0–cell for each vertex, one 1–cell for each edge and attaching maps given by s and $s \circ i$. A *tree* is a graph whose geometric realization is a contractible space and a *forest* is a disjoint union of trees.

Definition 3.2 A *fat graph* $\Gamma = (G, \sigma)$ is a combinatorial graph together with a cyclic ordering σ_v of the half-edges incident at each vertex v. The *fat structure* of the graph is given by the data $\sigma = (\sigma_v)$, which is a permutation of the half-edges.
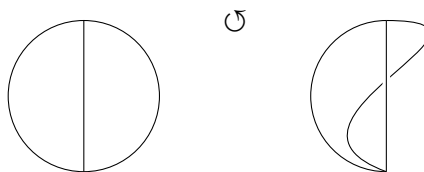
Figure 10: Two different fat graphs — the fat structure is given by the orientation of the plane, here denoted by the circular arrow — with the same underlying combinatorial graph.
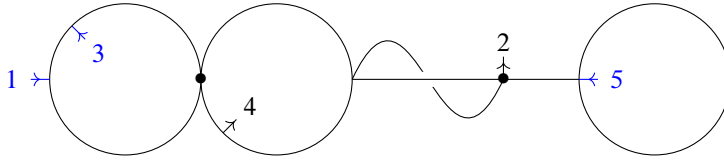
Figure 11: An example of a closed fat graph which is not admissible. The incoming and outgoing leaves are marked by incoming or outgoing arrows.

From a fat graph $\Gamma = (G, \sigma)$ one can construct a surface with boundary Σ_Γ by thickening the edges and the vertices. More explicitly, one can construct this surface by replacing each edge with a strip and gluing these strips to a disk at each vertex according to the fat structure. The cyclic ordering exactly gives the data required to do this. Notice that there is a strong deformation retraction of Σ_Γ onto $|G|$, so one can think of $|G|$ as the skeleton of the surface.

Definition 3.3 The *boundary cycles* of a fat graph are the cycles of the permutation of half-edges given by $\omega = \sigma \circ i$. Each cycle τ of ω gives a list of edges of the graph Γ and thus determines a subgraph $\Gamma_\tau \subset \Gamma$, which we call the *boundary graph* corresponding to τ.

Remark 3.4 The fat structure of Γ is completely determined by ω. Moreover, one can show that the boundary cycles of a fat graph $\Gamma = (G, \omega)$ correspond to the boundary components of Σ_Γ; see [22]. Therefore the surface Σ_Γ is completely determined up to topological type by the combinatorial graph and its fat structure.

A fat graph gives one a surface, but not yet a cobordism. The difference is that it does not distinguish between incoming and outgoing boundary components, nor do these come with canonical parametrizations. Note that after deciding whether a boundary component is incoming or outgoing, a parametrization is uniquely determined once we pick a marked point and edge lengths. Thus it suffices to add to each boundary component a leaf labeled either "incoming" or "outgoing".

Definition 3.5 A *closed fat graph* $\Gamma = (\Gamma, L_{\mathrm{in}}, L_{\mathrm{out}})$ is a fat graph with an ordered set of leaves and a partition of this set of leaves into two sets L_{in} and L_{out} such that:

   (i)   All inner vertices are at least trivalent.

  (ii)   There is exactly one leaf on each boundary cycle. Given a leaf l_i we denote its corresponding boundary graph by $\Gamma_{l_i} \subset \Gamma$.

Leaves in L_{in} or in L_{out}, are called *incoming* or *outgoing* respectively.

Note that the previous definition also removed unnecessary bivalent and univalent vertices. It turns out that one can consider an even more restricted type of fat graph, which reflects that (like in radial slits) we can decide to arrange the incoming boundary in a special way.
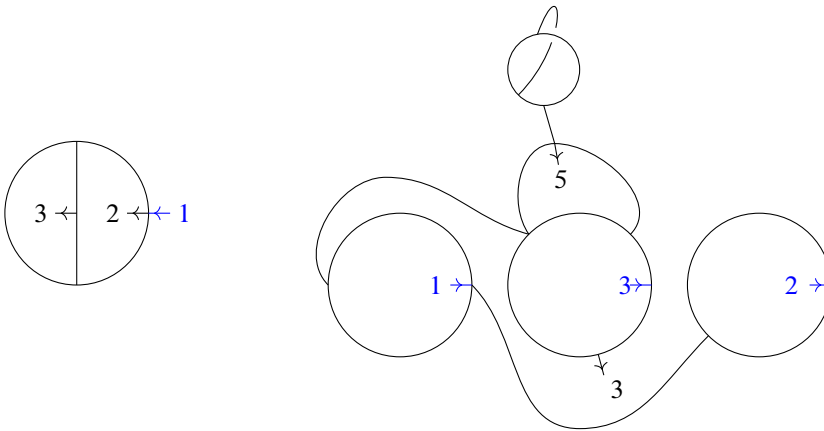
Figure 12: Two examples of admissible fat graphs. The first graph has the topological type of the pair of pants, and second graph that of a surface of genus 1 with 5 boundary components.

Definition 3.6 Let Γ be a closed fat graph. Let l_i denote a leaf of Γ and $\Gamma_{l_i} \subset \Gamma$ be its corresponding boundary graph. Γ is called *admissible* if the subgraphs $\Gamma_{l_i} - l_i$ for all incoming leaves l_i are disjoint embedded circles in Γ. We refer to these boundary cycles as *admissible cycles* (see Figure 12).

We organize fat graphs into a category. The idea is that when we use fat graphs to construct surfaces, we should be able to pick different lengths for the edges to obtain different conformal classes. Furthermore, if the length of an edge goes to zero, we expect the two disks corresponding to the vertices to be glued together. This makes sense as long as the edge is not a loop. The morphisms in the category of fat graphs encode this relationship between graphs. Recall that a tree is a graph whose geometric realization is contractible, and a forest is a disjoint union of trees.

Definition 3.7 We define two categories:

- The *category of closed fat graphs* $\mathcal{F}at$ is the category with objects isomorphism classes of closed fat graphs and morphisms $[\Gamma] \to [\Gamma/F]$ given by collapsing to a point in each tree in a subforest of Γ that does not contain any leaves.

- The *category of admissible fat graphs* $\mathcal{F}at^{ad}$ is the full subcategory of $\mathcal{F}at$ with objects isomorphism classes of admissible fat graphs.

The compositions in $\mathcal{F}at$ and $\mathcal{F}at^{ad}$, and hence the categories themselves, are well defined. The category $\mathcal{F}at$ was introduced by Godin in [22], and $\mathcal{F}at^{ad}$ is a slight variation, introduced by the same author in [21].

Note that the collapse of a subforest which does not contain any leaves induces a surjective homotopy equivalence upon geometric realizations and does not change the number of boundary components. Therefore, if there is a morphism $\varphi \colon [\Gamma] \to [\widetilde{\Gamma}]$ between isomorphism classes of fat graphs, then the surfaces $\Sigma_{[\Gamma]}$ and $\Sigma_{[\widetilde{\Gamma}]}$ are homeomorphic.

From a closed fat graph we can construct a two-dimensional cobordism. The underlying surface of the cobordism is the oriented surface Σ_Γ. This gives an orientation of the incoming and outgoing boundary component, so it is enough to give a labeled marked point in each boundary component. Note that each of the boundary components corresponds to exactly one leaf in the graph, which gives a marked point in the boundary component. We label this according to the labeling of its leaf. This gives a cobordism, well defined up to isomorphism.

3.1.2 Metric fat graphs

We motivated the morphisms in the category of fat graphs by thinking about lengths of edges. This is made more concrete in the space of metric fat graphs, which we describe now. This space has a deformation retraction onto the classifying space of the category of fat graphs, but we feel metric fat graphs are more intuitive and hence discuss them first. Several equivalent versions of this space and its dual concept (using weighted arc systems instead of fat graphs) have been studied by Harer, Penner, Igusa and Godin [20; 27; 29; 37].

The idea is simple: a metric fat graph is a fat graph with lengths assigned to its edges. We need a bit more care to make this interact well with the additional data and properties of admissible fat graphs.

Definition 3.8 A *metric admissible fat graph* is a pair (Γ, λ) where Γ is an admissible fat graph and λ is a *length function*, ie a function $\lambda \colon E_\Gamma \to [0, 1]$ where E_Γ is the set of edges of Γ and λ satisfies:

  (i) $\lambda(e) = 1$ if e is a leaf.
 (ii) $\lambda^{-1}(0)$ is a forest in Γ and $\Gamma/\lambda^{-1}(0)$ is admissible.
(iii)   For any admissible cycle C in Γ, we have $\sum_{e \in C} \lambda(e) = 1$.

We will call the value of λ on e the *length* of the edge e in Γ.

Definition 3.9 Suppose Γ is an admissible fat graph with p admissible cycles. Let $(n_1, n_2, \ldots, n_p)$ be the number of edges on each admissible cycle and set $n := \sum_i n_i$. The *space of length functions* on Γ is given as a set by

$$\mathcal{M}(\Gamma) := \{\lambda \colon E_\Gamma \to [0, 1] \mid \lambda \text{ is a length function}\}.$$

There is a natural inclusion

$$\mathcal{M}(\Gamma) \hookrightarrow \Delta^{n_1-1} \times \Delta^{n_2-1} \times \cdots \times \Delta^{n_p-1} \times ([0, 1])^{\#E_\Gamma - n}.$$

We give $\mathcal{M}(\Gamma)$ the subspace topology via this inclusion.

Definition 3.10 Two metric admissible fat graphs (Γ, λ) and $(\widetilde{\Gamma}, \widetilde{\lambda})$ are called *isomorphic* if there is an isomorphism of admissible fat graphs $\varphi \colon \Gamma \to \widetilde{\Gamma}$ such that $\lambda = \widetilde{\lambda} \circ \varphi_*$, where φ_* is the map induced by φ on E_Γ.

Definition 3.11 The space of *metric admissible fat graphs* is defined as

$$\mathcal{MFat}^{ad} := \frac{\bigsqcup_\Gamma \mathcal{M}(\Gamma)}{\sim},$$

where Γ runs over all admissible fat graphs and the equivalence relation $\sim$ is given by

$$(\Gamma, \lambda) \sim (\widetilde{\Gamma}, \widetilde{\lambda}) \iff (\Gamma/\lambda^{-1}(0), \lambda|_{E_\Gamma - \lambda^{-1}(0)}) \cong (\widetilde{\Gamma}/\widetilde{\lambda}^{-1}(0), \widetilde{\lambda}|_{E_{\widetilde{\Gamma}} - \widetilde{\lambda}^{-1}(0)}).$$

In other words, we identify isomorphic admissible fat graphs with the same metric, and we identify a metric admissible fat graph with some edges of length 0 with the metric fat graph in which these edges are collapsed and all other edge lengths remain unchanged.

Lemma 3.12 *There is a deformation retraction of the space of metric admissible fat graphs $\mathcal{MF}at^{ad}$ onto the geometric realization of the nerve of $\mathcal{F}at^{ad}$.*

Proof We will first give a continuous map $\iota \colon |\mathcal{F}at^{ad}| \to \mathcal{MF}at^{ad}$. A point $x \in |\mathcal{F}at^{ad}|$ is represented by $x = ([\Gamma_0] \to [\Gamma_1] \to \cdots \to [\Gamma_k], s_0, s_1, \ldots, s_k) \in N_k \mathcal{F}at^{ad} \times \Delta^k$, where N_k denotes the set of k–simplices of the nerve. Choose representatives Γ_i for $0 \le i \le k$, and for each i let C_j^i denote the jth admissible cycle of Γ_i, n_j^i denote the number of edges in C_j^i and m^i denote the number of edges that do not belong to the admissible cycles. Each graph Γ_i naturally defines a metric admissible fat graph (Γ_0, λ_i) where λ_i is given as follows:

$$\lambda_i \colon E_{\Gamma_0} \to [0, 1], \qquad e \mapsto \begin{cases} 0 & \text{if } e \text{ is collapsed in } \Gamma_i, \\ 1/n_j^i & \text{if } e \in C_j^i, \\ 1/m^i & \text{otherwise.} \end{cases}$$

Then define $\iota(x) := (\Gamma_0, \sum_{i=0}^k s_i \lambda_i)$. It is easy to show that this assignment is well defined and respects the simplicial relations of the geometric realization, and thus defines a continuous map. Moreover, it is an injective map between Hausdorff spaces with compact image, and so is a homeomorphism onto its image. Note that the image of ι is the subspace of metric graphs where the sum of the lengths of the edges that do not belong to the admissible cycles is 1.

We now construct a continuous map $r \colon \mathcal{MF}at^{ad} \times [0, 1] \to \mathcal{MF}at^{ad}$ which is a strong deformation retraction of $\mathcal{MF}at^{ad}$ onto the image of ι, by rescaling. Since all the graphs we are considering are finite, we can define a continuous function g by

$$g \colon \mathcal{MF}at^{ad} \to \mathbb{R}^{>0}, \qquad (\Gamma, \lambda) \mapsto \sum_{e \in \widetilde{E}_\Gamma} \lambda(e),$$

where $\widetilde{E}_\Gamma$ is the set of edges that do not belong to the admissible cycles. We then define r by linear interpolation as $r((\Gamma, \lambda), t) := (\Gamma, (1 - t)\lambda + t\lambda_g)$, where λ_g is the rescaled length function given by

$$\lambda_g \colon E_\Gamma \to \mathbb{R}^{\ge 0}, \qquad e \mapsto \begin{cases} \lambda(e) & \text{if } e \text{ belongs to an admissible cycle,} \\ \lambda(e)/g(\Gamma, \lambda) & \text{if } e \text{ does not belong to an admissible cycle.} \end{cases} \qquad \square$$

Remark 3.13 The space $\mathcal{MF}at^{ad}$ and the category $\mathcal{F}at^{ad}$ split into components indexed by the topological type of the graphs as two-dimensional cobordisms. That is,

$$\mathcal{MF}at^{ad} \cong \bigsqcup_{g,n,m} \mathcal{MF}at^{ad}_{g,n+m} \quad \text{and} \quad \mathcal{F}at^{ad} \cong \bigsqcup_{g,n,m} \mathcal{F}at^{ad}_{g,n+m},$$
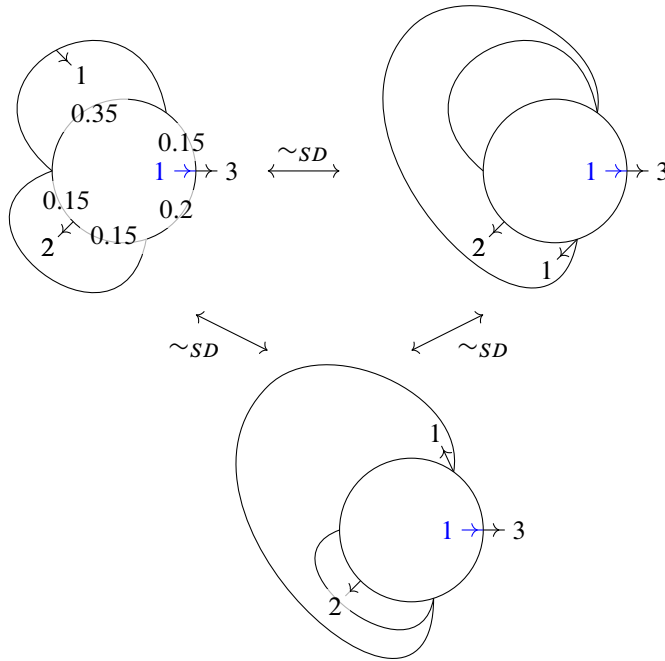
Figure 13: Three equivalent metric admissible fat graphs. On the last two graphs the lengths of the edges of the admissible cycle have been left out; they equal those of the first graph.

where $\mathcal{MF}at^{ad}_{g,n+m}$ and $\mathcal{F}at^{ad}_{g,n+m}$ are the connected components corresponding to admissible fat graphs with n admissible cycles which are homotopy equivalent to a surface of total genus g with $n+m$ boundary components.

3.1.3 Sullivan diagrams We now define a quotient space $\mathcal{SD}$ of $\mathcal{MF}at^{ad}$, which we will see in Section 5 is the analogue of the harmonic compactification for admissible fat graphs. To define it, we first describe an equivalence relation $\sim_{\mathcal{SD}}$ on metric admissible fat graphs.

Definition 3.14 We say $\Gamma_1 \sim_{\mathcal{SD}} \Gamma_2$ if Γ_2 can be obtained from Γ_1 by

- *slides*, ie sliding vertices along edges that do not belong to the admissible cycles, and
- *forgetting lengths of nonadmissible edges*, ie changing the lengths of the edges that do not belong to the admissible cycles.

Definition 3.15 A *metric Sullivan diagram* is an equivalence class of metric admissible fat graphs under the relation $\sim_{\mathcal{SD}}$.

We can informally think of a Sullivan diagram as an admissible fat graph where the edges not belonging to the admissible cycles are of length zero.

Definition 3.16 The space of *Sullivan diagrams* $\mathcal{SD}$ is the quotient space $\mathcal{SD} = \mathcal{MF}at^{ad}/\sim_{\mathcal{SD}}$.

Remark 3.17 A path in $\mathcal{SD}$ is given by continuously moving the vertices on the admissible cycles. This space splits into connected components given by topological type.

Remark 3.18 In Section 5 we show that $\mathcal{SD}$ has a canonical CW–complex structure. Its cellular chain complex is the complex of (cyclic) Sullivan chord diagrams introduced by Tradler and Zeinalian. They, and later Wahl and Westerland, used it to construct operations on the Hochschild chains of symmetric Frobenius algebras [40; 43].

3.2 The universal mapping class group bundle

In this section we describe the universal mapping class group bundles over $\mathcal{F}at^{ad}$ and $\mathcal{MF}at^{ad}$. Recall that from an admissible fat graph we can construct a cobordism which contains the graph as a deformation retract, though this depends on some choices. The idea for the construction of the universal mapping class group bundle is that its fiber over an admissible fat graph Γ consists of all ways that Γ can sit in a fixed standard cobordism.

For each topological type of cobordism fix a representative surface $S_{g,n+m}$ of total genus g with n incoming boundary components and m outgoing boundary components. Fix a marked point x_k in the kth incoming boundary for $1 \leq k \leq n$ and a marked point x_{k+n} in the kth outgoing boundary for $1 \leq k \leq m$.

Definition 3.19 Suppose Γ is an admissible fat graph of topological type $S_{g,n+m}$. Let $v_{\text{in},k}$ denote the kth incoming leaf and $v_{\text{out},k}$ denote the kth outgoing leaf. A *marking* of Γ is an isotopy class of embeddings $H: |\Gamma| \hookrightarrow S_{g,n+m}$ such that $H(v_{\text{in},k}) = x_k$, $H(v_{\text{out},k}) = x_{k+n}$ and the fat structure of Γ coincides with the one induced by the orientation of the surface. We will call a pair $(\Gamma, [H])$ a *marked fat graph* and denote by $\text{Mark}(\Gamma)$ the *set of markings of* Γ.

Lemma 3.20 *Any marking $H: |\Gamma| \hookrightarrow S_{g,n+m}$ is a homotopy equivalence, and the map on π_1 induced by H sends the ith boundary cycle of Γ to the ith boundary component of $S_{g,n+m}$.*

Proof Since the fat structure of Γ coincides with the one induced by the orientation of the surface, we can thicken Γ inside $S_{g,n+m}$ to a subsurface S_Γ of the same topological type as $S_{g,n+m}$. Moreover, by the definition of a marking, each boundary component of S_Γ meets a boundary component of $S_{g,n+m}$. Thus, there is a deformation retraction of $S_{g,n+m}$ onto this subsurface and onto Γ. $\qquad\square$

Lemma 3.21 *Let Γ be an admissible fat graph and F be a forest in Γ which does not contain any leaves of Γ. Then there is a bijection $\text{Mark}(\Gamma) \to \text{Mark}(\Gamma/F)$ denoted by $[H] \mapsto [H_F]$.*

This identification depends on the map connecting both graphs, ie given $[H]$ a marking of Γ, if $\widetilde{\Gamma} = \Gamma/F_1 = \Gamma/F_2$ then $[H_{F_1}]$ and $[H_{F_2}]$ can be different markings of $\widetilde{\Gamma}$. Figure 14 gives an example of this in the case of the cylinder.

Proof Let H be a representative of a marking $[H]$ of Γ. The image of $H|_F$ (the restriction of H to $|F|$) is contained in a disjoint union of disks away from the boundary. Therefore the marking H induces a
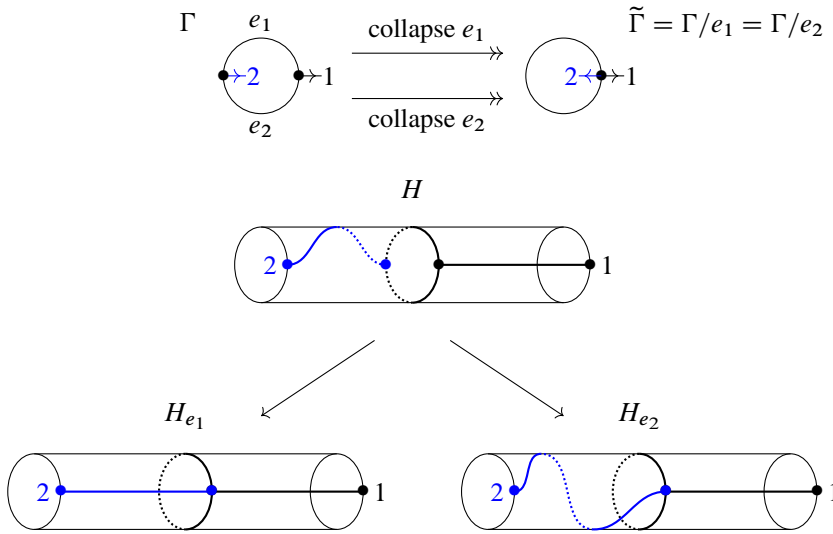
Figure 14: Two different embeddings of $\widetilde{\Gamma}$ in the cylinder differing by a Dehn twist and corresponding to the same marking of Γ.

marking $H_F : |\Gamma/F| \hookrightarrow S_{g,n+m}$ given by collapsing each of the trees of F to a point of the disk in which their image is contained. Note that H_F is well defined up to isotopy and it makes the following diagram commute up to homotopy:

$$
\begin{array}{ccc}
|\Gamma| & \longrightarrow\!\!\!\!\!\longrightarrow & |\Gamma/F| \\
 & H \searrow & \downarrow H_F \\
 & & S_{g,n+m}
\end{array}
$$

In fact, up to isotopy, there is a unique embedding of a tree with a fat structure into a disk, in which the fat structure of the tree coincides with the one induced by the orientation of the disk and the endpoints are fixed points on the boundary. This can proven by induction. Start with the case where F is a single edge. Up to homotopy, there is a unique embedding of an arc in a disk where the endpoints of the arc are fixed points on the boundary. Then by [17], there is also a unique embedding up to isotopy. For the induction step, let α be an arc embedded in the disk with its endpoints at the boundary and let a and b be fixed points in the boundary of a connected component of $D \setminus \alpha$. Then we have a map

$$\mathrm{Emb}^{a,b}(I, D \setminus \alpha) \to \mathrm{Emb}^{a,b}(I, D),$$

where $\mathrm{Emb}^{a,b}(I, D \setminus \alpha)$ is the space of embeddings of a path in $D \setminus \alpha$ which start at a and end at b, with the $\mathbb{C}^\infty$–topology, and similarly for $\mathrm{Emb}^{a,b}(I, D)$. By [23], this map induces injective maps in all homotopy groups, in particular in π_0, which gives the induction step.

It then follows that, given $[H_F]$ a marking of Γ/F, there is a unique marking $[H]$ of Γ such that the above diagram commutes up to homotopy. $\square$

Definition 3.22 Define the category $\mathcal{EFat}^{ad}$ to be the category with objects isomorphism classes of marked admissible fat graphs $([\Gamma], [H])$ (where two marked admissible fat graphs are isomorphic if their underlying fat graphs are isomorphic and they have the same marking) and morphisms given by morphisms in $\mathcal{Fat}^{ad}$ where the map acts on the marking as stated in the previous lemma. We denote by $\mathcal{EFat}^{ad}_{g,n+m}$ the full subcategory with objects marked admissible fat graphs whose thickening give a cobordism of topological type $S_{g,n+m}$.

Definition 3.23 The space of *marked metric admissible fat graphs* $\mathcal{EMFat}^{ad}$ is defined to be

$$\mathcal{EMFat}^{ad} := \frac{\bigsqcup_\Gamma \mathcal{M}(\Gamma) \times \mathrm{Mark}(\Gamma)}{\sim_E},$$

where Γ runs over all admissible fat graphs, and the equivalence relation is given by

$$(\Gamma, \lambda, [H]) \sim_E (\widetilde{\Gamma}, \widetilde{\lambda}, [\widetilde{H}]) \iff (\Gamma, \lambda) \cong (\widetilde{\Gamma}, \widetilde{\lambda}) \text{ and } [H_\lambda] = [\widetilde{H}_{\widetilde{\lambda}}].$$

Here $\cong$ denotes isomorphism of metric fat graphs, H_λ is the induced marking $H_\lambda : |\Gamma/F_\lambda| \hookrightarrow S_{g,n+m}$ where F_λ is the subforest of Γ of edges of length zero ie $F_\lambda = \lambda^{-1}(0)$ and $H_{\widetilde{\lambda}}$ is defined analogously.

The following result is proven in [14], in fact in more generality, for a category modeling open–closed cobordism and not only closed cobordisms.

Theorem 3.24 *The projection* $|\mathcal{EFat}^{ad}_{g,n+m}| \to |\mathcal{Fat}^{ad}_{g,n+m}|$ *is a universal* $\mathrm{Mod}(S_{g,n+m})$*–bundle.*

The proof follows the original ideas of Igusa [29] and Godin [22]. Since all spaces involved are CW–complexes, one first shows that $|\mathcal{EFat}^{ad}_{g,n+m}|$ is contractible, which follows from contractibility of the arc complex [28]. Second, one proves that the action of the mapping class group $\mathrm{Mod}(S_{g,n+m})$ on $\mathcal{EFat}^{ad}_{g,n+m}$ is free and transitive. That is, for any two markings $[H_1]$ and $[H_2]$, there is a unique $[\varphi] \in \mathrm{Mod}(S_{g,n+m})$ such that $[\varphi \circ H_1] = [H_2]$. This proof in particular gives rise to an abstract homotopy equivalence $\mathcal{M} \simeq \mathcal{Fat}^{ad}$.

By Lemma 3.21, as a set $\mathcal{EMFat}^{ad}$ is given by $\{([\Gamma, \lambda], [H]) \mid [\Gamma, \lambda] \in \mathcal{MFat}^{ad} \text{ and } [H] \in \mathrm{Mark}([\Gamma])\}$. As before, let $\mathcal{EMFat}^{ad}_{g,n+m}$ denote the subspace of marked metric admissible fat graphs whose thickenings give an open–closed cobordism of topological type $S_{g,n+m}$. Then $\mathrm{Mod}(S_{g,n+m})$ acts on $\mathcal{EMFat}^{ad}_{g,n+m}$ by composition with the marking, and it follows that:

Corollary 3.25 *The projection* $\mathcal{EMFat}^{ad}_{g,n+m} \to \mathcal{MFat}^{ad}_{g,n+m}$ *is a universal* $\mathrm{Mod}(S_{g,n+m})$*–bundle.*

Proof This is clear since we have a pullback diagram

$$
\begin{array}{ccc}
\mathcal{EMFat}^{ad}_{g,n+m} & \xrightarrow[r(-,1)\times\mathrm{id}]{\simeq} & |\mathcal{EFat}^{ad}_{g,n+m}| \\
\downarrow & & \downarrow \\
\mathcal{MFat}^{ad}_{g,n+m} & \xrightarrow[r(-,1)]{\simeq} & |\mathcal{Fat}^{ad}_{g,n+m}|
\end{array}
$$

The horizontal maps are the homotopy equivalences given by r, the map constructed in Lemma 3.12. $\square$

4   The critical graph equivalence

In this section we construct the space $\mathfrak{Rad}^{\sim}$ as well as the maps in Corollaries 4.42 and 4.51, and prove these are homotopy equivalences.

4.1   Lacher's theorem

The idea for proving that certain maps $f\colon X \to Y$ are homotopy equivalences is to show that they are nice enough maps between nice enough spaces with contractible fibers. This is made precise by [33, Theorem, page 510].

Definition 4.1   (i)   A subspace X of a space Y is a *neighborhood retract* if there exists an open subset U of Y containing X and a retraction $r\colon U \to X$.

  (ii)   A space X is an *ANR* if, whenever X is a closed subspace of a metric space Y, X is a neighborhood retract of Y.

Definition 4.2   (i)   A subset A of a manifold M is *cellular* if it is the intersection $\bigcap_n E_n$ of a nested sequence $E_1 \supset E_2 \supset \cdots$ of n–cells E_i in M, ie subsets homeomorphic to D^n.

  (ii)   A space X is *cell-like* if there is an embedding (ie a continuous map that is a homomorphism onto its image) $\phi\colon X \to M$ of X into a manifold M such that $\phi(X)$ is cellular.

 (iii)   A map $f\colon X \to Y$ is *cell-like* if for all $y \in Y$ the point inverse $f^{-1}(\{y\})$ is cell-like.

Theorem 4.3   (Lacher)   *A proper map $f\colon X \to Y$ between locally compact ANRs is cell-like if and only if, for all open $U \subset Y$, the restriction $f|_{f^{-1}(U)}\colon f^{-1}(U) \to U$ is a proper homotopy equivalence.*

The conditions in the above definitions are difficult to verify, so we will provide criteria which imply them. Our main reference are [35] for ANRs, [18, Chapter 3] for polyhedra and [33] for cell-like spaces.

Proposition 4.4   *The following are properties of ANRs*:

  (i)   *For all $n \geq 0$, the closed n–disk D^n is an ANR.*

  (ii)   *An open subset of an ANR is an ANR.*

 (iii)   *If X is a space with an open cover by ANRs, then X is an ANR.*

 (iv)   *If X and Y are compact ANRs, $A \subset X$ is a compact ANR and $f\colon A \to Y$ is continuous, then $X \cup_f Y$ is an ANR.*

  (v)   *Any locally finite CW–complex is an ANR.*

 (vi)   *Any locally finite polyhedron is an ANR.*

(vii)   *A product of finitely many ANRs is an ANR.*

(viii)   *A compact ANR is cell-like if and only if it is contractible.*

Proof   Properties (i), (ii), (iii) and (iv) follow from Corollary 5.4.6 and Theorems 5.4.1, 5.4.5 and 5.6.1 of [35], respectively. These combine to prove (v) by noting that by (ii) and (iii) one can reduce to the case

of finite CW–complexes, and since by definition these can be obtained by gluing closed n–disks together, (i) and (iv) prove that finite CW–complexes are ANRs. Property (vi) follows from (v), but is also [35, Theorem 3.6.11]. Property (vii) is [35, Proposition 1.5.7]. Finally, (viii) follows from Theorem 4.3 by considering the map to a point. □

4.2 The fattening of the radial slit configurations and the critical graph map

There is a natural admissible metric fat graph associated to a radial slit configuration: the unstable critical graph obtained by taking the inner boundaries of the annuli and the complements of the slit segments, and gluing these together according to the combinatorial data. The inner boundaries of the annuli give the admissible cycles of the graph and the incoming leaves are placed at the positive real line of each annulus. The outgoing leaves are obtained from marked points on the outgoing boundary components. This graph gets a canonical fat graph structure as a subspace of the surface $S(L)$.

We now make this definition precise. Because we fixed the outer radii of the annuli, we shorten $\mathbb{A}_{R_i}^{(i)}$ to $\mathbb{A}_i$. Recall the subsets $\alpha_i^{\pm}$ and $\beta_i^{\pm}$ in the sector F_i, from Definition 2.3. These lie in a pair of distinct radial segments of F_i, unless it is a thin sector in which case they lie in a single radial segment. To a radial slit configuration $L \in \mathfrak{QRad}$ we associate a space $\overline{E}_L$, defined as follows:

Definition 4.5 The space $\overline{E}_L$ is given by

$$\overline{E}_L := \left(\bigsqcup_{1 \leq j \leq n} \partial_{\mathrm{in}} \mathbb{A}_j \right) \sqcup \left(\bigsqcup_{1 \leq j \leq 2h+m} E_j \right) \sqcup \left(\bigsqcup_{1 \leq j \leq n} I_j \right),$$

where each of the terms is defined as follows:

- **Admissible boundaries** For each annulus $\mathbb{A}_j$ we take the inner boundary $\partial_{\mathrm{in}} \mathbb{A}_j$.
- **Radial segments for slits and outgoing leaves** For $1 \leq j \leq 2h + m$ with $\xi_j \in \mathbb{A}_k$ we take
  $E_j = \{z \in \mathbb{A}_k \mid \arg(z) = \arg(\xi_j) \text{ or } \arg(z) = \arg(\xi_{\overline{\omega}(j)})\}$.
- **Incoming leaves** For each annulus $\mathbb{A}_j$ we take $I_j = \left\{z \in \mathbb{C}_j \mid \arg(z) = 0 \text{ and } 0 \leq |z| \leq \frac{1}{2\pi}\right\}$.

The equivalence relation $\sim_L$ on $\overline{E}_L$ is that generated by:

- **Attaching incoming leaves** We set $\left(\frac{1}{2\pi} \in I_j\right) \sim_L \left(\frac{1}{2\pi} \in \partial_{\mathrm{in}} \mathbb{A}_j\right)$ for $j = 1, 2, \ldots, n$.
- **Attaching radial segments** For $r \in \partial_{\mathrm{in}} \mathbb{A}_k$ and $e \in E_j$, we set $r \sim_L e$ if $r = e$.
- **Identifying coinciding segments** Defining subsets $\alpha_i^{\pm}$ and $\beta_i^{\pm}$ of E_i as in Definition 2.3, we let $\sim_L$ identify $z \in \alpha_i^+$ with $z \in \alpha_{\overline{\omega}(i)}^-$ and $z \in \beta_i^+$ with $z \in \beta_{\overline{\lambda}(i)}^-$.

Note that each of the terms in $\overline{E}_L$ can be considered as a subspace of $\overline{\Sigma}(L)$; recalling Definition 2.4, one observes that $\sim_L$ identifies those points on $\overline{E}_L$ that are identified by $\approx_L$ on $\overline{\Sigma}(L)$. As a consequence, the quotient space $\overline{E}_L/\sim_L$ is invariant under the slit jump relation. Thus for a configuration $[L] \in \mathfrak{Rad}$, we obtain a well-defined graph $\Gamma_{[L]}$ if we demand it has no bivalent vertices. Some of its leaves are labeled by the incoming or outgoing boundary components; the remaining ones we will remove.
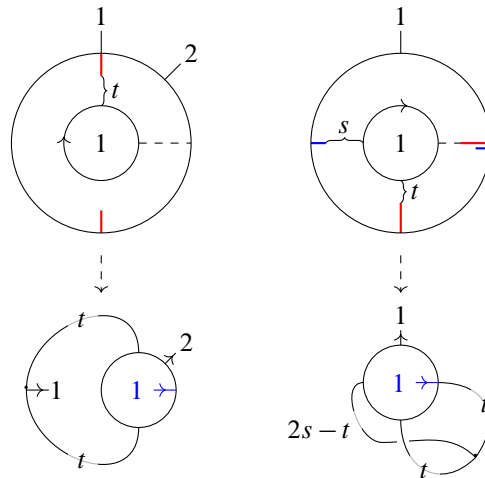
Figure 15: Critical graphs for different configurations. Edge lengths of the critical graphs are *not* to scale.

Definition 4.6 For $L \in \mathfrak{QRad}$, the corresponding *critical graph* Γ_L is the graph obtained from $\overline{E}_L/\sim_L$ by removing those leaves that do not correspond to incoming or outgoing boundary cycles; see Figure 15.

By construction, this graph comes embedded in the surface $\Sigma_{[L]}$ and thus inherits a fat structure. Moreover, it inherits a metric $\lambda_{[L]}$ from the standard metric in $\mathbb{C}$. In it, the incoming leaves have fixed length $\frac{1}{2\pi}$ and the outgoing leaves have strictly positive length. Because, for our purposes, the lengths of the outgoing leaves are superfluous information, we set $\lambda_{[L]}(e)$ to be given by the standard metric in $\mathbb{C}$ if e is not a leaf and $\lambda_{[L]}(e) = 1$ if e is a leaf. This makes $(\Gamma_{[L]}, \lambda_{[L]})$ a metric admissible fat graph.

Notation 4.7 We will just write Γ_L when it is clear from context that we consider it as a metric admissible fat graph.

The construction of the critical graph gives a function

$$\mathfrak{Rad} \to \mathcal{MFat}^{ad}, \quad [L] \mapsto (\Gamma_{[L]}, \lambda_{[L]}).$$

However, this function is *not* continuous at nongeneric configurations. For an example, consider the path in $\mathfrak{Rad}$ given by continuously varying the argument of a slit as in Figure 16; when the moving slit reaches a neighboring one, the associated metric graph jumps.

To solve this problem we enlarge $\mathfrak{Rad}$ at nongeneric configurations by a contractible space, by "opening up" the edges E_L. To do this, we first need to introduce some notation. We can think of the thin sector

$$F_i = \{z \in \mathbb{A}_j \mid \arg(\xi_i) = \arg(z)\}$$

as being obtained by identifying two copies of F_i, which we will denote by E_i^+ and E_i^-, along the equivalence relation that identifies $z \in E_i^+$ with $z \in E_i^-$. Let us extend this notation to ordinary and full sectors: if F_i is ordinary then

$$E_i^+ := \{z \in F_i \mid \arg(z) = \arg(\xi_{\overline{\omega}(i)})\} \quad \text{and} \quad E_i^- := \{z \in F_i \mid \arg(z) = \arg(\xi_{\omega(i)})\},$$
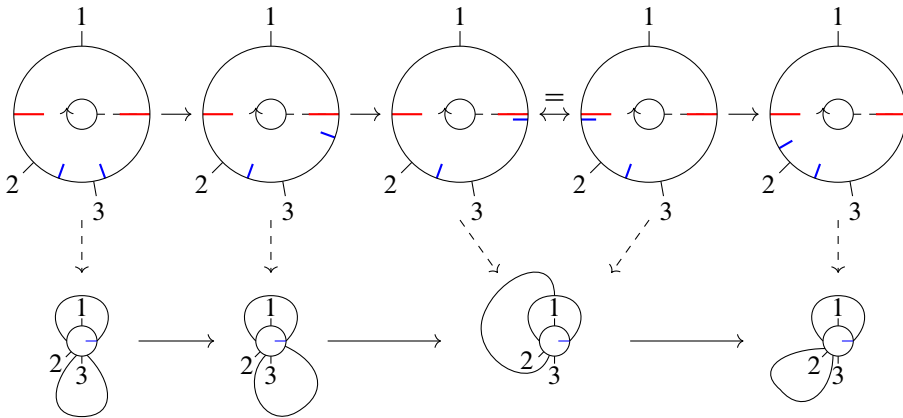
Figure 16: An example of a path in $\mathfrak{Rad}$ which leads to a path in $\mathcal{MFat}^{ad}$ that is not continuous. Labelings have been left out for the sake of clarity.

and if F_i is full then $E_i^+ = S_i^+$ and $E_i^- = S_i^-$. Let us also generalize Definition 2.3 to this section by taking $\alpha_i^+, \beta_i^+ \subset E_i^+$ and $\alpha_i^-, \beta_i^- \subset E_i^+$. Then we can also write $\bar{E}_L / \sim_L$ as $\bar{E}_L' / \sim_L'$ with

$$\bar{E}_L' := \left(\bigsqcup_{1 \leq j \leq n} \partial_{\mathrm{in}} \mathbb{A}_j \right) \sqcup \left(\bigsqcup_{1 \leq j \leq 2h+m} E_j^+ \sqcup E_j^- \right) \sqcup \left(\bigsqcup_{1 \leq j \leq n} I_j \right)$$

and $\sim_L'$ the equivalence relation on $\bar{E}_L'$ generated by replacing E_j with $E_j^{\pm}$ in the three operations generating $\sim_L$ and adding a fourth one:

- **Identifying thin sectors** If F_i is thin, we let $\sim_L'$ identify $z \in E_i^+$ with $z \in E_i^-$.

The idea is now to vary the extent to which we identify E_i^+ with E_i^- in the last of these:

Definition 4.8 Let $\mathrm{thin}(L)$ be the set of thin sectors of L and let $t: \mathrm{thin}(L) \to [0,1]$ be a function. The equivalence relation $\sim_t'$ on the space

$$\bar{E}_L' = \left(\bigsqcup_{1 \leq j \leq n} \partial_{\mathrm{in}} \mathbb{A}_j \right) \sqcup \left(\bigsqcup_{1 \leq j \leq 2h+m} E_j^+ \sqcup E_j^- \right) \sqcup \left(\bigsqcup_{1 \leq j \leq n} I_j \right)$$

is the one generated by:

- **Attaching incoming leaves** We set $\left(\frac{1}{2\pi} \in I_j \right) \sim_t' \left(\frac{1}{2\pi} \in \partial_{\mathrm{in}} \mathbb{A}_j \right)$ for $j = 1, 2, \ldots, n$.
- **Attaching radial segments** For $r \in \partial_{\mathrm{in}} \mathbb{A}_k$ and $e \in E_j^{\pm}$, we set $r \sim_t' e$ if $r = e$.
- **Identifying coinciding segments** With $\alpha_i^{\pm}$ and $\beta_i^{\pm}$ of the $E_j^{\pm}$ as above, we let $\sim_t'$ identify $z \in \alpha_i^+$ with $\alpha_{\bar{\omega}(i)}^-$ and $z \in \beta_i^+$ with $z \in \beta_{\bar{\lambda}(i)}^-$.
- **Partially identifying thin sectors** If F_i is thin, we let $\sim_t'$ identify $z \in E_i^+$ with $z \in E_i^-$ as long as $|z| \leq t(F_i) + \frac{1}{2\pi}$.

Definition 4.9 We define $\Gamma_{L,t}$ to be obtained from $\bar{E}_L' / \sim_t'$ by removing those leaves that do not correspond to incoming or outgoing boundary cycles.
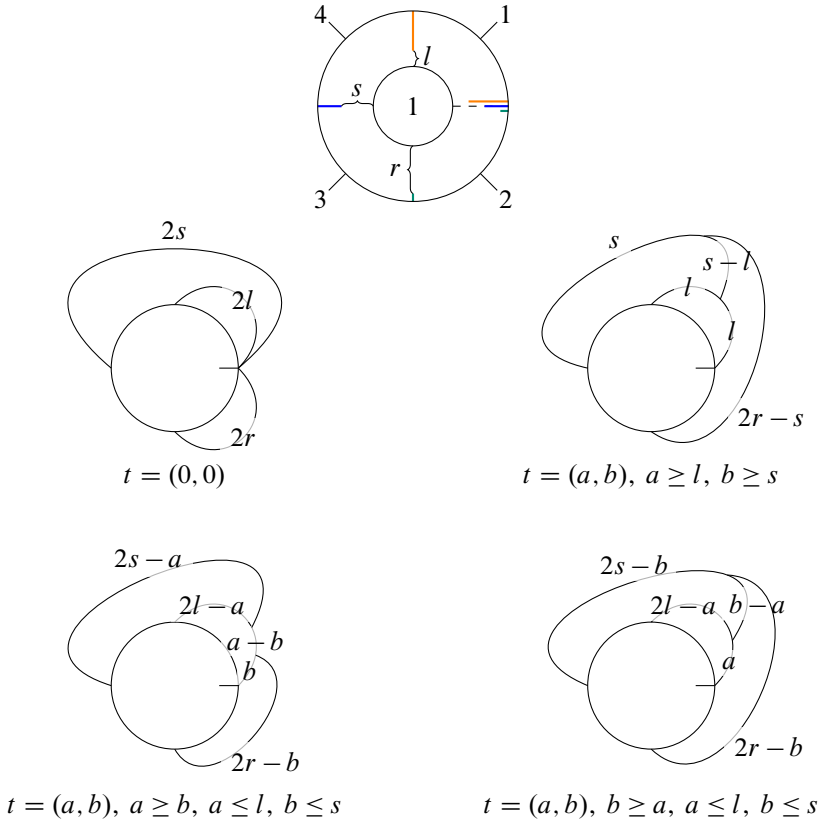
Figure 17: A configuration $[L]$ on the top and several graphs obtained from it using different functions $t: \text{thin}([L]) \to [0,1]$, here written as a pair of real numbers. The leaves have been omitted to make the graphs more readable, but they are all located along the admissible cycles according to the positions of the marked points in $[L]$. The edges are not to scale.

Example 4.10 When t is a constant function equal to 1, $\Gamma_{L,t}$ is the critical graph Γ_L, which is invariant under slit and parametrization points jumps. However, for most other t, the graph $\Gamma_{L,t}$ is *not* invariant under slit jumps.

Notation 4.11 If t is constant equal to 0, we will call this the *unfolded graph of L* and denote it by $\Gamma_{L,0}$; see Figure 17.

Just like the critical graph, the graph $\Gamma_{L,t}$ has a natural metric making $(\Gamma_{L,t}, \lambda_{L,t})$ an admissible metric fat graph. Figure 17 shows examples of unfolded and partially unfolded metric admissible fat graphs.

Remark 4.12 Two preconfigurations with the same combinatorial type have the same underlying admissible fat graphs, but with different metric. Thus it makes sense to talk about $\Gamma_{\mathcal{L},t}$ as an admissible fat graph. Similarly, it makes sense to talk about the critical graph of a combinatorial type, which we denote by $\Gamma_{[\mathcal{L}]}$.

Definition 4.13 Letting $[L] \in \mathfrak{Rad}$, we define a subspace of $\mathcal{MFat}^{ad}$

$$\mathcal{G}([L]) := \{[\Gamma_{L_i,t}, \lambda_{L_i,t}] \mid [L] = [L_i] \text{ and } t\colon \text{thin}(L_i) \to [0,1]\}.$$

We define the *fattening of* $\mathfrak{Rad}$ to be the space

$$\mathfrak{Rad}^{\sim} = \{([L], [\Gamma, \lambda]) \in \mathfrak{Rad} \times \mathcal{MFat}^{ad} \mid [\Gamma, \lambda] \in \mathcal{G}([L])\}.$$

For simplicity, we will just write $\Gamma_{L_i,t}$ or Γ when it is clear from the context that we are talking about metric graphs.

We will see that $\mathfrak{Rad}^{\sim}$ is constructed by replacing the point $[L] \in \mathfrak{Rad}$ by a contractible space $\mathcal{G}([L])$, which is a space of graphs which interpolate between the critical graph of $[L]$ and the unfolded graphs of the different representatives $L_1, L_2, \ldots, L_k$ of $[L]$ in $\mathfrak{QRad}$.

The fattening of $\mathfrak{Rad}$ splits into connected components given by the topological type of the cobordism they describe:

$$\mathfrak{Rad}^{\sim} := \bigsqcup_{h,n,m} \mathfrak{Rad}^{\sim}_h(n,m).$$

Moreover, it comes with two natural maps

$$\mathfrak{Rad} \xleftarrow{\pi_1} \mathfrak{Rad}^{\sim} \xrightarrow{\pi_2} \mathcal{MFat}^{ad}.$$

We call π_1 the *projection map* and π_2 the *critical graph map*. The goal of the remaining subsections is to prove that these are homotopy equivalences. The next section is the main input for proving π_1 is a homotopy equivalence.

4.3 The space $\mathcal{G}([L])$ is contractible

Proposition 4.14 $\mathcal{G}([L])$ *is contractible for any radial slit configuration* $[L]$.

We prove this inductively by removing parametrization points or slits. In particular, we allow radial slit configurations *without parametrization points*; all relevant definitions may be extended to this case in a straightforward manner.

Notation 4.15 For a radial slit configuration L^1, we denote by L the radial slit configuration obtained from L^1 by removing all parametrization points.

If L is not empty, then it has $m \geq 1$ shortest pairs of slits of L. That is, L has pairs of slits $(\zeta_{i_j}, \zeta_{\lambda(i_j)})$ for $1 \leq j \leq m$, which are all of the same length and are the shortest in the sense that

- $|\zeta_{i_j}| = |\zeta_{\lambda(i_j)}| = |\zeta_{i_l}| = |\zeta_{\lambda(i_l)}|$ for all $1 \leq j, l \leq m$, and
- $|\zeta_{i_j}| > |\zeta_s|$ for any $s \notin \{i_j, \lambda(i_j) \mid 1 \leq j \leq m\}$.

We denote by $\bar{L}$ the configuration obtained from L by forgetting the shortest slit pair(s).

Note that if L^1 is not degenerate, then L and $\bar{L}$ are also not degenerate. The induction step in the proof of Proposition 4.14 is provided by:

Lemma 4.16   *There are homotopy equivalences*

$$\pi_L^1 : \mathcal{G}([L^1]) \to \mathcal{G}([L]) \quad and \quad \pi_L : \mathcal{G}([L]) \to \mathcal{G}([\bar{L}]).$$

Informally, the map π_L^1 removes the leaves of $\Gamma^1 \in \mathcal{G}([L^1])$ corresponding to the outgoing boundary components. Similarly, the map π_L removes the edges of $\Gamma \in \mathcal{G}([L])$ corresponding to the shortest pair(s) of slits in $[L]$. Assuming Lemma 4.16, we now prove Proposition 4.14.

Proof of Proposition 4.14   By the first part of Lemma 4.16, it is enough to show that $\mathcal{G}([L])$ is contractible, where $[L]$ a radial slit configuration without parametrization points. We will prove this by induction on h, the number of pairs of slits of $[L]$. If $h = 0$, then $\mathcal{G}([L])$ is a point and therefore contractible. Assume that $\mathcal{G}([L])$ is contractible when $h < k$ for some fixed k. Now let $h = k$ and consider the map

$$\pi_L : \mathcal{G}([L]) \to \mathcal{G}([\bar{L}]).$$

Given that $[\bar{L}]$ has $\bar{h} < k$ pairs of slits, it is contractible by the induction hypothesis. Thus by the second part of Lemma 4.16, $\mathcal{G}([L])$ is also contractible. $\qquad\square$

4.3.1   Proof of Lemma 4.16   To prove Lemma 4.16 we will show that the spaces involved are compact ANRs and the maps involved are cell-like, and invoke Theorem 4.3. We start by considering the domain and target of the maps.

Lemma 4.17   *For all configurations $[L]$, with or without parametrization points, the space $\mathcal{G}([L])$ is a compact polyhedron and thus a compact ANR.*

Proof   We give the proof only when $[L]$ has parametrization points; the other case is similar.

The space $\mathcal{G}([L])$ is a subspace of $\mathcal{MFat}_{g,n+m}^{ad}$. The latter is contained in the larger compact polyhedron given by

$$P_{g,n+m} := \frac{\bigsqcup_\Gamma \Delta^{n_1-1} \times \Delta^{n_2-1} \times \cdots \times \Delta^{n_p-1} \times ([0,1])^{\#E_\Gamma - n}}{\sim},$$

with Γ indexed by the objects of $\mathcal{Fat}_{g,n+m}^{ad}$ and the equivalence relation $\sim$ given by Definition 3.7. This is compact because $\mathcal{Fat}_{g,n+m}^{ad}$ has finitely many objects.

The subspace $\mathcal{G}([L])$ can be characterized as the union of the images of maps from the cubes $[0,1]^{\mathrm{thin}(L_i)}$ to $\mathcal{MFat}_{g,n+m}^{ad}$ for all representatives L_i of $[L]$. Each map is piecewise linear between polyhedra, which implies that their image is a subpolyhedron. This is true because a piecewise linear map by definition can be made simplicial with respect to some triangulation, and the images of simplicial maps are polyhedra. Note that there are only finitely many representatives for $[L]$, so $\mathcal{G}([L])$ is a union of finitely many compact polyhedra, which implies it is a polyhedron by [18, Corollary 3.1.27]. The last claim then follows from Proposition 4.4(vi). $\qquad\square$

We now define the maps π_L^1 and π_L. We start with the former, which "removes leaves corresponding to the parametrization points".

Definition 4.18 Let $[L^1]$ be a radial slit configuration and let $[L]$ be the configuration obtained from $[L^1]$ by removing the parametrization points. We define the function

$$\pi_L^1 : \mathcal{G}([L^1]) \to \mathcal{G}([L])$$

by sending Γ to the metric fat graph obtained from Γ by

(1) removing all leaves corresponding to outgoing boundary components,

(2) removing all bivalent vertices, ie if there is a bivalent vertex we replace the two edges attached to it by a single edge whose length is the sum of the lengths of both.

Let $[L]$ be a radial slit configuration without parametrization points and assume it is nonempty, that is, $[L]$ has at least one pair of slits. We now define the function π_L, which "removes the edges corresponding to the longest slit pair(s) of $[L]$".

Definition 4.19 For any $\Gamma \in \mathcal{G}([L])$, the continuous function $d_{\mathrm{ad}} : \Gamma \to \mathbb{R}_{\geq 0}$ is defined by sending a point x in a leaf of Γ to 0 and any other point $x \in \Gamma$ to its path distance to the admissible cycles. By the extreme value theorem it attains a maximum $d_{\max}$. We denote by Γ' the fat graph with unlabeled leaves obtained by removing from Γ the preimage of $d_{\max}$. That is, we set $\Gamma' := \Gamma - d_{\mathrm{ad}}^{-1}(d_{\max}) \subset \Gamma$. We define the function

$$\pi_L : \mathcal{G}([L]) \to \mathcal{G}([\bar{L}])$$

by sending Γ to the metric fat graph $\overline{\Gamma}$ obtained from Γ' by recursively

(1) removing all unlabeled leaves of Γ',

(2) removing all bivalent vertices from to obtain a fat graph Γ'',

(3) repeating if Γ'' has unlabeled leaves.

Note that the only leaves of $\pi_L(\Gamma)$ are the ones corresponding to the admissible cycles.

We will focus on π_L first, leaving π_L^1 to the end of this subsection. We start with some properties of π_L:

Lemma 4.20   (i) *π_L is well defined.*

(ii) *π_L is continuous.*

(iii) *The fibers of π_L are compact ANRs.*

Proof Let $\Gamma \in \mathcal{G}([L])$, so that there is a representative L and function $t : \mathrm{thin}(L) \to [0, 1]$ such that $\Gamma = \Gamma_{L,t}$. Let $\bar{L}$ be the configuration obtained from L by removing the shortest pair(s) of slits. To prove that $\pi_L(\Gamma)$ is well defined, we exhibit a function $\bar{t} : \overline{\mathrm{thin}}(\bar{L}) \to [0, 1]$ such that $\pi_L(\Gamma) = \pi_L(\Gamma_{L,t}) = \Gamma_{\bar{L},\bar{t}}$. Note that any thin sector F of $\bar{L}$ is of one of two kinds:

(1) The sector F corresponds uniquely to a sector in L. In this case we define $\tilde{t}(F) := t(F)$.

(2) The sector F corresponds to several thin sectors $F_1, F_2, \ldots, F_s$ in L. This happens when, in between the slits defining the sector F in $\bar{L}$, there are one or more slits in L which have been removed. In this case, we define

$$\bar{t}(F) := \min\{t(F_1), t(F_2), \ldots, t(F_s)\}.$$

Then $\pi_L(\Gamma) = \Gamma_{\bar{L}, \bar{t}}$. This completes the proof of (i).

For (ii), it suffices to prove π_L is continuous on each of the finitely many closed subsets of the form $\{[\Gamma_{L_i, t}, \lambda_{L_i, t}] \mid t : \text{thin}(L_i) \to [0, 1]\}$, that is, fixing the representative L_i of $[L]$. This is clear from the construction of $\bar{t}$, and hence of $\Gamma_{\bar{L}, \bar{t}}$.

As in the proof of Lemma 4.17, for (iii) it suffices to prove the fibers are compact polyhedra by proving each fiber is the union of the images of finitely many piecewise linear maps with compact domain. But this follows once more from the construction of $\bar{t}$, and hence of $\Gamma_{\bar{L}, \bar{t}}$. $\square$

We now state the main ingredient for the proof of Lemma 4.16:

Lemma 4.21   For $\bar{\Gamma} \in \mathcal{G}([\bar{L}])$, the preimage $\pi_L^{-1}(\bar{\Gamma}) \subseteq \mathcal{G}([L])$ is contractible.

By construction, any $\Gamma \in \pi_L^{-1}(\bar{\Gamma})$ can be built from $\bar{\Gamma}$ by attaching to it a graph. We will show that the space of graphs that can be attached to Γ is contractible, and that there is a contractible space of ways to attach each of these. Before doing so, we give two illustrative examples:

Example 4.22   (single pair of shortest slits)   Consider the configurations L and $\bar{L}$ obtained by deleting the shortest pair of slits shown in Figure 18, top left. The other representatives L' of $[L]$ are given by letting the purple or green slit on the right jump; for any such representative, deleting its shortest pairs of slits also yields a representative of $\bar{L}$.

Figure 18, bottom left, shows two different graphs in $\mathcal{G}([L])$: Γ_1, the unfolded graph of L, and Γ_2, a partially folded graph of L. The map $\pi_L : \mathcal{G}([L]) \to \mathcal{G}([\bar{L}])$ is given by removing the point marked by an $\times$ in the green arc — which in the case of Γ_1 is the midpoint of the green arc — and deleting the resulting leaves. In particular, $\pi_L(\Gamma_i) = \bar{\Gamma}_i$ for $i = 1, 2$, where the graphs $\bar{\Gamma}_i$ are shown in Figure 18, top right. Note that $\bar{\Gamma}_1$ is the unfolded graph of $\bar{L}$, and $\bar{\Gamma}_2$ is the critical graph of $\bar{L}$. Therefore, in either case $\pi_L^{-1}(\bar{\Gamma}_i)$ is not empty.

The entire preimage $\pi_L^{-1}(\bar{\Gamma}_i)$ is given by the locations for attaching a chord to $\bar{\Gamma}_i$. This may be done along the dashed green segments for one end of the chord and the fixed point marked in green for the other, as in Figure 18, top right. Thus the preimages are homeomorphic to intervals. In either case, the endpoints of the interval correspond to the unfolded graphs of L and the radial slit configuration obtained from L by letting the shortest segment jump. In fact, the preimages $\pi_L^{-1}(\bar{\Gamma})$ are homeomorphic to an interval for all $\bar{\Gamma} \in \mathcal{G}([\bar{L}])$.

The reason the second end of the chord could only be attached to a single point is because its corresponding slit is isolated, ie it is the only slit on its radial segment. If this were not the case, then the other end of
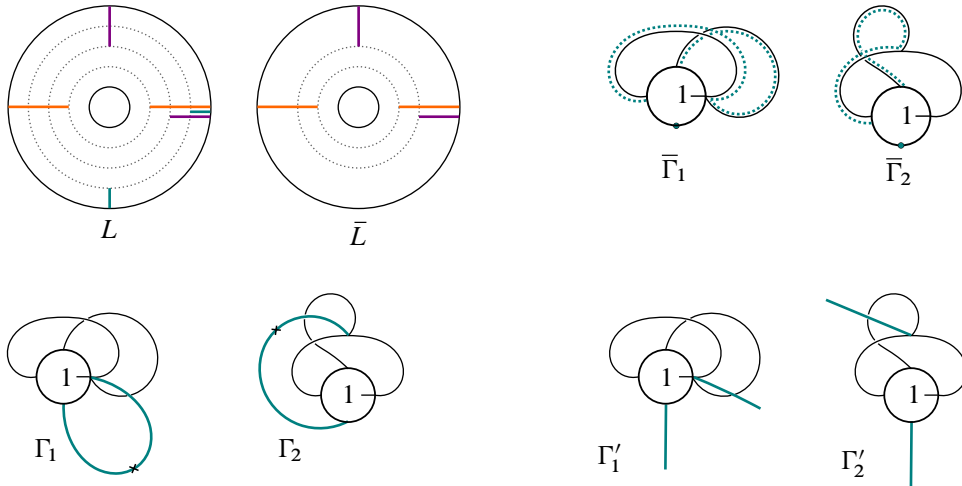
Figure 18: Top left: a configuration L and the configuration $\bar{L}$ obtained from it by deleting the shortest pair of slits (that is, those where $d_{\max}$ is attained). Top right: graphs in $\mathcal{G}([\bar{L}])$; $\bar{\Gamma}_1$ is the unfolded graph of $\bar{L}$ and $\bar{\Gamma}_2$ is the critical graph of $\bar{L}$. The green dotted lines trace the boundary interval defined by the open chord corresponding to the deleted green slit, and thus describe the places where one endpoint of the new chord can be attached. Bottom left: graphs in $\mathcal{G}([L])$ such that $\pi_L(\Gamma_i) = \bar{\Gamma}_i$. In both cases Γ_i is the maximally unfolded graph of L relative to $\bar{\Gamma}_i$. The points marked with an $\times$ denote the points in Γ at which the maximum of d_{ad} is attained. Bottom right: the open graphs of the maximally unfolded graphs relative to $\bar{\Gamma}$ given in the bottom left.

this chord could also be attached to an interval. The intervals at which both endpoints of the chord can be attached must be disjoint, otherwise there would be a sequence of jumps that would bring both slits together and thus L would be degenerate. So in this more generic case, $\pi_L^{-1}(\bar{\Gamma})$ is homeomorphic to a square. Finally, there is another simple generalization of this case: there are several pairs of shortest slits in L, but the intervals describing the endpoints where their corresponding chords can be attached are all disjoint. In this case, the preimage is homeomorphic to a higher-dimensional cube.

In the previous example we considered the case where there is exactly one pair of slits which is the shortest pair, as well as some simple generalizations. On the other end of the spectrum there is the case where all slits are of equal size:

Example 4.23 (all slits of equal size) In the following radial slit configuration L, the configuration $\bar{L}$ obtained by deleting all shortest slit pairs is empty:
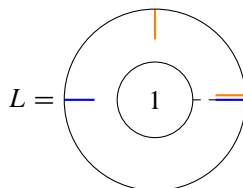
The configuration $[L]$ has three representatives, and $\mathcal{G}([L])$, which is the preimage over the unique point in $\mathcal{G}[\bar{L}]$, is homeomorphic to the cone on three points. These three points are represented by the unfolded graphs of the three representatives, and the cone point by the critical graph.

The general case is an amalgamation of these two cases. More precisely, in the first case — where there is exactly one pair of slits which is the shortest — the preimage is homeomorphic to an interval or to a cube arising from the choices of where to attach the endpoints of the attached chord. In the second case — where all slits are of the same length — the preimage is a cone on three points corresponding to the unfolded representatives. In general, the preimage is homeomorphic to a product of "cones on cubes". We will show this by going through an intermediary subspace of metric fat graphs corresponding to attaching trees on chords.

Definition 4.24 Let $\bar{\Gamma} \in \mathcal{G}([\bar{L}])$. By definition, there is a representative $\bar{L}$ and a function $\bar{t}$ such that $\bar{\Gamma} = \Gamma_{\bar{L},\bar{t}}$.

Let $L_1, L_2, \ldots, L_r$ be all the radial slit configurations that can obtained from $\bar{L}$ by adding slits such that each L_i is equivalent to L by slit jumps. For any i, there is at least one function $t : \text{thin}(L_i) \to [0, 1]$ such that $\pi_L(\Gamma_{L_i, t_i}) = \bar{\Gamma}$. Let t_i be the minimal one among such functions, ie the one that takes the smallest possible values for every element of $\text{thin}(L_i)$.

- The *maximally unfolded graph of L_i relative to $\bar{\Gamma}$* is the fat graph $\Gamma_i := \Gamma_{L_i, t_i}$.
- The *open graph of L_i relative to $\bar{\Gamma}$* is the fat graph with unlabeled leaves $\Gamma_i' := \Gamma_i - d_{\text{ad}}^{-1}(d_{\max})$, where $d_{\max}$ is the maximum of the distance from any point in Γ_i to the admissible cycles.

Examples of maximally unfolded graphs relative to some graph can be seen in Figure 18, bottom left. Their corresponding open graphs are given in Figure 18, bottom right.

Remark 4.25 Any maximally unfolded graph relative to $\bar{\Gamma}$, say Γ_i, is obtained from $\bar{\Gamma}$ by attaching a chord for each pair of slits deleted in L_i. In particular, if $\bar{\Gamma}$ is an unfolded graph then each Γ_i is an unfolded graph as well. Furthermore, the preimage $d_{\text{ad}}^{-1}(d_{\max})$ consists of exactly one point in each of these chords: that point at which the half-edges corresponding to each slit pair are glued to each other. Therefore, each leaf in the open graph of L_i relative to $\bar{\Gamma}$ corresponds precisely to a slit deleted from L_i.

Moreover, for any graph $\Gamma \in \mathcal{G}([L])$, there is at least one L_i and a function $t : \text{thin}(L_i) \to [0, 1]$ such that $\Gamma = \Gamma_{L_i, t}$ and $t \geq t_i$. Thus any graph in $\mathcal{G}([L])$ can be thought of as a "folding" of a maximally unfolded graph relative to $\bar{\Gamma}$, say Γ_i, where we only "fold" the chords that have been attached to $\bar{\Gamma}$ in the construction of Γ_i. In particular, this shows that any such Γ can be obtained from $\bar{\Gamma}$ by attaching to it a forest along its leaves.

A special example of this is the case of the critical graph $\Gamma_{\text{crit}} \in \mathcal{G}([L])$. It can be constructed from $\bar{\Gamma}$ by attaching corollas to $\bar{\Gamma}$. This graph can be obtained by "completely folding" any of the maximally

unfolded graphs relative to $\overline{\Gamma}$. Furthermore, the preimage $d_{\text{ad}}^{-1}(d_{\max})$ consist exactly of the central vertices of the corollas attached.

Informally, one can think of $\pi_L^{-1}(\overline{\Gamma})$ as a space of graphs that interpolates the maximally unfolded graphs relative to $\overline{\Gamma}$ with the critical graph. At one extreme we attach chords, at the other we attach corollas, and in between we attach forests that arise as all possible foldings of these chords on their way to the corollas.

We now show that these forests can be attached to boundary intervals (possibly of length 0, so points) in the outgoing boundary of the metric fat graph $\overline{\Gamma}$. Those boundary intervals that are not points are described combinatorially as follows:

Definition 4.26 Let Γ be a metric (admissible) fat graph and let τ be a boundary cycle of Γ. We can think of τ as a set of half-edges of Γ together with a cyclic order. A *boundary interval in τ*, denoted by $\mathcal{B}$, is a proper subset of the half-edges of τ which can be written as

$$\mathcal{B} = \{h_1, h_2 = \tau(h_1), h_3 = \tau^2(h_1), \ldots, h_n = \tau^{n-1}(h_1)\}$$

for some half-edge h_1 in τ. In particular, $\mathcal{B}$ is an ordered set.

A boundary interval determines an ordered list of edges in Γ in which an edge can appear at most twice. Consecutive edges in this list share a vertex and thus define a path in Γ between $s(h_1)$ and $s(\iota(h_n))$, where s and ι are the source and involution maps in the definition of the graph Γ. Up to scaling there is a canonical map from the unit interval to Γ which traces this path and sends 0 to $s(h_1)$ and 1 to $s(\iota(h_n))$. By scaling the unit interval, we can construct a canonical map which is an isometry when restricted to the edges of the path. We do this below.

Definition 4.27 Let $\mathcal{B}$ be a boundary interval in a boundary cycle τ. We denote by $I_{\mathcal{B}}$ an oriented interval whose length is the length of the path in Γ determined by $\mathcal{B}$. More precisely, $I_{\mathcal{B}}$ can be subdivided into consecutive subintervals I_i for $1 \leq i \leq |\mathcal{B}|$. The length of the ith subinterval I_i is the length of the ith edge $e_i = \{h_i, \iota(h_i)\}$ on the path determined by $\mathcal{B}$. We denote by x_i^- and x_i^+ the boundary points of I_i, using its orientation.

The *parametrization map of $\mathcal{B}$* is the unique map

$$f_{\mathcal{B}} \colon I_{\mathcal{B}} \to \Gamma$$

given by $x_1^- \mapsto s(h_1)$ and $x_n^+ \mapsto s(\iota(h_n))$ that, for all i, restricts to an isometry $f_{\mathcal{B}}| \colon I_i \to e_i := \{h_i, \iota(h_i)\}$ that sends x_i^- to $s(h_i)$.

The map $f_{\mathcal{B}}$ is a parametrization of an interval in the boundary component corresponding to τ. Thus a point in $x \in I_{\mathcal{B}}$ uniquely determines a way in which a leaf can be attached to Γ such that the leaf is in the boundary interval defined by $\mathcal{B}$.

We now describe the boundary intervals that will arise, given $\overline{\Gamma} \in \mathcal{G}([\overline{L}])$.

Definition 4.28 Let Γ_i' denote the open graph of L_i relative to $\overline{\Gamma}$ for $1 \le i \le r$. Let l be an unlabeled leaf of Γ_i'. This leaf defines a boundary cycle τ_l in Γ_i'. We define $\mathcal{B}_l$ to be the subset of the half-edges of τ_l given by

$$\mathcal{B}_l := \{\tau_l^j(l) \mid j \in \mathbb{Z}, \, j \neq 0 \text{ and } \tau_l^j(l) \text{ is not part of an edge in an admissible cycle}\}.$$

Note in particular that $\mathcal{B}_l$ could be empty, and this indeed happens when l is attached to a vertex v which is essentially trivalent in the sense that it has valence four if it is also attached to an admissible leaf but trivalent otherwise.

An example of this construction can be seen in Figure 18, top right, where the dotted lines in $\overline{\Gamma}_i$ for $i = 1, 2$ correspond precisely to the boundary intervals defined by the leaves of the open graph. The sets $\mathcal{B}_l$ have the following properties:

Lemma 4.29 For $1 \le i \le r$, let Γ_i' denote the open graphs of L_i relative to $\overline{\Gamma} \in \text{Im}(\pi_L)$, as in Definition 4.24. Recall that each unlabeled leaf of Γ_i, say l, corresponds precisely to a shortest slit of L_i, and thus it has a "pair" leaf which we denote by $\lambda(l)$. Then:

  (i) *For any unlabeled leaf l of Γ_i', the set $\mathcal{B}_l$ is either empty or it is a boundary interval in $\overline{\Gamma}$.*

 (ii) *For any unlabeled leaf l of Γ_i', the sets $\mathcal{B}_l$ and $\mathcal{B}_{\lambda(l)}$ are disjoint.*

(iii) *For any pair of unlabeled leaves l_1 and l_2 in Γ_i' such that $\mathcal{B}_{l_1} \neq \varnothing \neq \mathcal{B}_{l_2}$, either*

$$\mathcal{B}_{l_1} \cap \mathcal{B}_{l_2} = \varnothing \quad \text{or} \quad \mathcal{B}_{l_1} = \mathcal{B}_{l_2}.$$

 (iv) *For any open graphs relative to $\overline{\Gamma}$, say Γ_i' and Γ_j', the set of boundary intervals defined by their unlabeled leaves coincide.*

Proof We first show (i) holds. Let ζ_l denote the slit corresponding to the unlabeled leaf l in Γ_i'. Then $\mathcal{B}_l$ is the section of the outgoing boundary along which the leaf l can move around, given by slit jumps of ζ_l. In particular, if ζ_l is isolated, that is, it is the only slit on its radial segment, then this is a single point and $\mathcal{B}_l$ is empty. If $\mathcal{B}_l$ is not empty, it is enough to show that $\mathcal{B}_l$ is not the entire boundary cycle that corresponds to l. Assume, for contradiction, that $\mathcal{B}_l$ is the entire boundary cycle. Then there must be a set of slits $\{\zeta_1, \lambda(\zeta_1), \zeta_2, \lambda(\zeta_2), \dots, \zeta_s, \lambda(\zeta_s)\}$ in L_i for some $s \ge 1$ such that the following hold:

  (1) The slit ζ_l lies between ζ_1 and $\lambda(\zeta_s)$. More precisely, $\lambda(\zeta_s)$, ζ_l and ζ_1 all lie in the same radial segment, $\omega(\zeta_1) = \zeta_l$ and $\omega(\zeta_l) = \lambda(\zeta_s)$.

  (2) For each $1 \le i < s$, the slits $\lambda(\zeta_i)$ and ζ_{i+1} lie in the same radial segment and $\omega(\zeta_{i+1}) = \lambda(\zeta_i)$.

Let ζ_* be a slit in $\{\zeta_1, \dots, \zeta_s\}$ of largest modulus, ie a shortest slit in that set. Then ζ_* and $\lambda(\zeta_*)$ can jump along the other slits. In particular, L_i is equivalent via slit jumps to a configuration L_* where $\lambda(\zeta_*)$, ζ_* and λ_l lie in the same radial segment and

$$\omega(\zeta_*) = \zeta_l, \quad \omega(\zeta_l) = \lambda(\zeta_*) \quad \text{and} \quad |\zeta_l| \le |\zeta_*|.$$

So L_* and L_i are degenerate configurations, which is not possible.
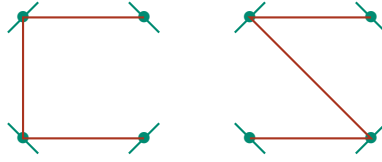
Figure 19: Two of the eight configurations of chords for $k = 4$. The green line segments are the intervals, the vertices are the marked points in these intervals and the red arcs are the chords.

Statement (ii) follows in a similar way. More precisely, if $\mathcal{B}_l$ and $\mathcal{B}_{\lambda(l)}$ are not disjoint, then L_i is equivalent via slit jumps to a configuration where ζ_l and $\lambda(\zeta_l)$ lie next to each other, and thus L_i is degenerate.

Statements (iii) and (iv) follow by construction. $\qquad\square$

Definition 4.30 (attaching intervals) Let $\overline{\Gamma} \in \mathcal{G}(\overline{L})$. We define $I_{L,\overline{\Gamma}}$ to be the set of oriented metric intervals (possibly of length zero) corresponding to the parametrization of the boundary intervals and isolated points in $\overline{\Gamma}$ along which a graph can be attached to obtain an element in its preimage.

That is, $I_{L,\overline{\Gamma}}$ is given by those $I_{\mathcal{B}_l}$ such that l is an unlabeled leaf of Γ', an open graph relative to $\overline{\Gamma}$, as in Definition 4.24. This interval is of length zero if its corresponding boundary interval is empty. Recall that this happens precisely when there is a leaf in $\overline{\Gamma}$ corresponding to an isolated slit, ie a slit that is the only one in its radial segment. Note in particular that, by Lemma 4.29(iv), this definition does not depend on the choice of Γ' but only on the class $[L]$ and the metric fat graph $\overline{\Gamma}$.

Any point in the preimage can be obtained by attaching a forest to $\overline{\Gamma}$ along the parametrization intervals in $I_{L,\overline{\Gamma}}$. To make this precise, we define certain spaces of forests attached to intervals, which will use the following combinatorial definition:

Definition 4.31 Let $\mathcal{I} := I_1 \sqcup I_2 \sqcup \cdots \sqcup I_k$ denote a disjoint union of k compact intervals of a given length. We allow intervals to have length zero. Let $\mathcal{D}$ denote a family of piecewise linear functions $\mathcal{D} := \{d_i \colon I_i \to \mathbb{R}_{>0} \mid 1 \leq i \leq k\}$ whose derivative is ± 1 outside a finite set. Then we define $\max \mathcal{D} := \max_{1 \leq i \leq k}\{\max_{x_i \in I_i} d_i(x_i)\}$.

Notation 4.32 (configurations of chords) We will consider the set of all possible configurations of $k - 1$ chords attached by their endpoints to the intervals in $\mathcal{I}$ such that the resulting graph is connected, planar and has no loops; we denote this set by $\mathrm{Conf}_{\mathcal{I}}$. See Figure 19 for examples of configuration of chords. We will construct a space of metric planar forests attached to these intervals and we will use the configurations above to restrict which metrics are allowed. For this, we will use the path distance function in a metric graph, which we denote by d_{path}.

Definition 4.33 Let $\mathcal{I}$ and $\mathcal{D}$ be as in the previous definition and $d \in \mathbb{R}_{>0}$ such that $2d > \max \mathcal{D}$. Denote by $\mathcal{F}_{\mathcal{I},\mathcal{D},d}$ those metric graphs obtained by attaching a metric forest F with at most $2(k-1)$ leaves to the intervals $\mathcal{I}$ such that:

- The graph obtained, denoted by G, is planar, connected and has no loops.

- There is a configuration $C \in \mathrm{Conf}_{\mathcal{I}}$ such that, for any pair of intervals I_i and I_j connected by a chord in C, the path distance in G from x_i to x_j (two attaching points of leaves of the forest F) is

$$d_{\mathrm{path}}(x_i, x_j) = 2d - d_i(x_i) - d_j(x_j).$$

Note that $\mathcal{F}_{\mathcal{I},\mathcal{D},d}$ is a subset of the space of metric fat graphs. We consider it as a space using the subspace topology.

Lemma 4.34 *The topological space $\mathcal{F}_{\mathcal{I},\mathcal{D},d}$ is contractible.*

Proof Fix a marked point $*_i \in I_i$ for all $1 \le i \le k$ such that $*_i$ is a local maximum for d_i. Let $\mathcal{F}_{\mathcal{I},\mathcal{D},d,*} \subset \mathcal{F}_{\mathcal{I},\mathcal{D},d}$ be the subspace where the forest is attached to the marked points in the intervals $\mathcal{I}$. We will construct a deformation retraction onto a point in two steps.

Step 1 We will construct a deformation retraction of $\mathcal{F}_{\mathcal{I},\mathcal{D},d}$ onto $\mathcal{F}_{\mathcal{I},\mathcal{D},d,*}$. Intuitively, we slide the endpoints along $\mathcal{I}$ towards the marked points, but some care is require to make sure the conditions on the metric remain satisfied. By definition, each I_i can be subdivided into finitely many intervals on which d_i is linear. Let N_i be the number of these in a uniquely minimal such subdivision. Our argument will be by induction over $N = N_1 + \cdots + N_k$.

In the initial case $N = 0$ there is nothing to prove. For the induction step, let $I' \subset I_j$ be an interval in the aforementioned minimal subdivision such that $I_j = I' \cup I'_j$ where $I' \cap I'_j$ is a point and $*_j \in I'_j$. Let $\mathcal{I}'$ be obtained from $\mathcal{I}$ by replacing I_j with I'_j and let $\mathcal{D}'$ be obtained by replacing d_j by $d'_j := d_j|_{I'_j}$. We will show that $\mathcal{F}_{\mathcal{I},\mathcal{D},d}$ deformation retracts onto a space homeomorphic to $\mathcal{F}_{\mathcal{I}',\mathcal{D}',d}$. There are two cases:

(A) **The point $I' \cap I'_j$ is a local minimum of d_j** In this case we "open" along the edge I' towards I'_j:

$$t = 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t = 1$$
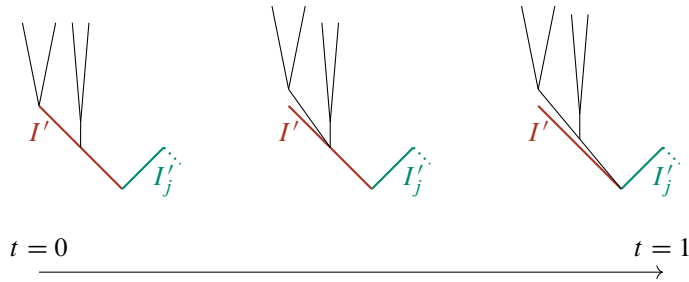
The precise construction is as follows. If I' has length ℓ, we linearly identify the interval I' with $[0, \ell]$, with $I' \cap I'_j$ corresponding to ℓ. Suppose that $s \in [0, \ell]$ is the unique smallest value at which an edge is attached to $I' \cong [0, \ell]$. Then on a metric graph G, the deformation retraction at time $t \in [0, 1]$ is the identity for $t\ell < s$, and for $t\ell \ge s$ replaces $I' \cong [0, \ell]$ with $[0, \ell] \cup_{t\ell} [s\ell, t\ell]$; note that we may identify $[t\ell, \ell] \cup_{t\ell} [s\ell, t\ell] \subset [0, \ell] \cup_{t\ell} [s\ell, t\ell]$ with $[s\ell, t\ell]$. We attach the edges originally attached to $[s\ell, t\ell] \subset I'$ to this new interval. The result has a canonical metric.

(B) **The point $I' \cap I'_j$ is a local maximum on d_j** In this case we "fold" along the edge I' towards I'_j:
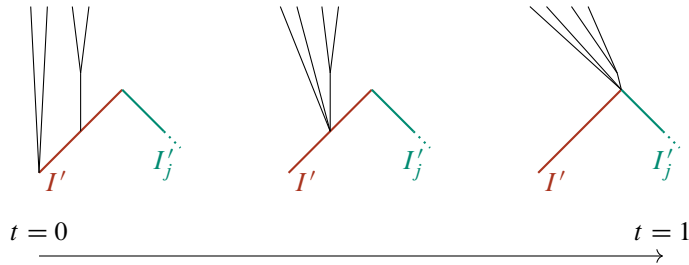
The precise construction is as follows. Let us linearly identify the interval I' with $[0, \ell]$, as in (A). Then consider the subtree of G given by points that are distance $t\ell$ from $0 \in I' \in [0, \ell]$. We identify this subtree with the interval $[0, t\ell]$ by identifying all points with distance s to $s \in [0, t\ell]$. The result has a canonical metric.

Step 2 We will prove that $\mathcal{F}_{\mathcal{I}, \mathcal{D}, d, *}$ is contractible by a variation of the Alexander trick. To do so, we replace the metric tree (T, d_T) attached to the marked points by $(T, (1-t)d_T)$ and add edges of length $t(d - d_i(*_i))$ connecting $*_i$ to the endpoint in this scaled tree originally attached to $*_i$ (the circles contain the rescaled graphs):
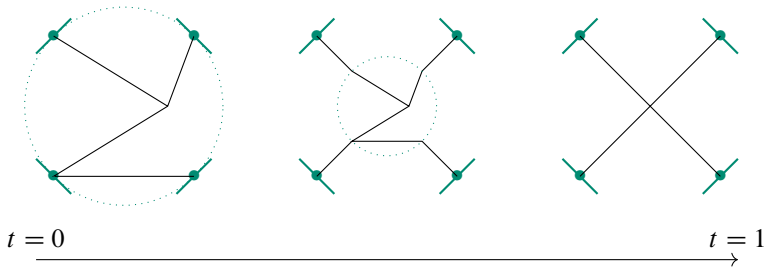
The resulting metric graphs are still planar, connected, without loops and satisfy the metric condition. At $t = 1$ we obtain the k–valent corolla attached to all intervals, with the edge between the vertex of the corolla and $*_i$ given by $d - d_i(*_i)$. □

Lemma 4.35 *Let $\overline{\Gamma} \in \mathcal{G}([\overline{L}])$. There is a positive real number $d \in \mathbb{R}_{>0}$ and a finite collection of sets of intervals $\mathcal{I}$ and sets of functions $\mathcal{D}$ such that there is a homeomorphism*

$$(4\text{-}1) \qquad \qquad \pi_L^{-1}(\overline{\Gamma}) \cong \mathcal{F}_{\mathcal{I}, \mathcal{D}, d} \times \cdots \times \mathcal{F}_{\mathcal{I}', \mathcal{D}', d}.$$

The intuition behind this homeomorphism is as follows. In the simplest scenario, there is only one term in the product of the right-hand side of (4-1). On the one hand, the critical graph corresponds to the unique point in $\mathcal{F}_{\mathcal{I}, \mathcal{D}, d}$ given by a single corolla. On the other hand, the maximally unfolded graphs relative to $\overline{\Gamma}$ correspond to elements in $\mathrm{Conf}_{\mathcal{I}}$, that is, to arrangements of $k - 1$ chords attached to the intervals (where $k - 1$ is the number of pairs of shortest slits of L). Finally, an arbitrary point in $\mathcal{F}_{\mathcal{I}, \mathcal{D}, d}$ is a "folding" of a configuration in $\mathrm{Conf}_{\mathcal{I}}$, and an arbitrary point in $\pi_L^{-1}(\overline{\Gamma})$ is a "folding" of a maximally unfolded graph relative to $\overline{\Gamma}$.

Proof   Given $[L]$ and $\overline{\Gamma}$, the set of intervals will be $\mathcal{I} = I_{L,\overline{\Gamma}}$; see Definition 4.30. Recall that there is a map

$$f : I_{L,\overline{\Gamma}} \to \overline{\Gamma},$$

which is an isometry when restricted to edges of $\overline{\Gamma}$ that are in the image. Moreover, we have a canonical embedding $\overline{\Gamma} \hookrightarrow \Gamma$ for which $\Gamma - \overline{\Gamma} = F$ is a forest and such that Γ is obtained from $\overline{\Gamma}$ by attaching the leaves of F to $I_{L,\overline{\Gamma}}$; see Remark 4.25.

For a choice of Γ in the preimage, we denote by G_Γ the subgraph of Γ that is given by the union of the forest F and the boundary intervals in $I_{L,\overline{\Gamma}}$ along which F is attached. The number of components of G_Γ is independent from the choice of Γ in the preimage of $\overline{\Gamma}$, and it corresponds to the number of elements in the product of the right-hand side of (4-1). An intuitive way to think about this is that the slits which are deleted from L to obtain $\bar{L}$ come in *clusters*, collections of slits which map to the same point in the glued surface $\Sigma([L])$, and each of these clusters contributes a single term in the product.

We will assume for the sake of simplicity that there is a single component in G_Γ or a single *cluster* of slits, thought the argument easily generalizes to the case of several components. The functions $d_i \in \mathcal{D}$ are induced by the modulus in $\mathbb{C}$. That is, they are determined by the path distance to the admissible cycles of $\overline{\Gamma}$. More precisely, for any $x \in I_i \in I_{L,\Gamma}$ we set $d_i(x) = d_{\mathrm{ad}}(x)$. This yields a well-defined piecewise-linear function on each I_i. The real number d is the common modulus of all slits which are deleted from L to obtain $\bar{L}$. Then there is a continuous map $\mathcal{F}_{\mathcal{I},\mathcal{D},d} \to \pi_L^{-1}(\bar{L})$ given by gluing the forest F into $\overline{\Gamma}$ according to the intervals $\mathcal{I}_i$. This has an inverse given by the continuous map that sends Γ to G_Γ. $\square$

Putting together these results, we prove that the preimages of π_L are contractible.

Proof of Lemma 4.21   Let $\overline{\Gamma} \in \mathcal{G}([\bar{L}])$. By Lemma 4.35, $\pi_L^{-1}(\overline{\Gamma})$ is homeomorphic to a product of spaces of forests attached at intervals. These are contractible by Lemma 4.34. $\square$

The proofs given above for π_L can be adapted to the simpler case of π_L^1, and we will spare the reader the technical details. The result is:

Lemma 4.36   (i) π_L *is well defined.*

  (ii) π_L *is continuous.*

  (iii)   *The fibers of π_L are compact contractible ANRs.*

We now finish the proof of Lemma 4.16, which said π_L and π_L^1 are homotopy equivalences:

Proof of Lemma 4.16   We apply Theorem 4.3. By Lemma 4.17, the domain and targets of the maps π_L and π_L^1 are compact ANRs, so it suffices to prove the fibers of both maps are cell-like. This follows by combining Proposition 4.4(viii) with Lemmas 4.20, 4.21 and 4.36. $\square$

4.4 The projection map is a homotopy equivalence

Our next goal is to check that the spaces $\mathfrak{Rad}$ and $\mathfrak{Rad}^\sim$ are ANRs, and that the map $\pi_1 \colon \mathfrak{Rad}^\sim \to \mathfrak{Rad}$ is proper and cell-like. For the remainder of this section we fix g, n and m.

Proposition 4.37 *The space $\mathfrak{Rad}$ is a locally compact ANR.*

Proof The space $\mathfrak{Rad}$ is a smooth manifold, so it is locally compact and has an open cover by copies of $\mathbb{R}^n$. These are ANRs by Proposition 4.4(v), so $\mathfrak{Rad}$ is an ANR by Proposition 4.4(iii). (Alternatively, one can argue that $\mathfrak{Rad}$ is an open subspace of the finite CW–complex $\overline{\mathfrak{Rad}}$ and use properties (ii) and (v) of Proposition 4.4.) $\qquad\square$

To prove that $\mathfrak{Rad}^\sim$ is an ANR and that π_1 is a proper cell-like map, we will write $\mathfrak{Rad}^\sim$ as an open subspace of a space $(\overline{\mathfrak{Rad}})^\sim$ obtained by gluing together finitely many compact ANRs. By Definition 2.16, $\overline{\mathfrak{Rad}} \setminus \mathfrak{Rad} = \mathfrak{Rad}'$ is a CW–complex, and in fact a subcomplex of $\overline{\mathfrak{Rad}}$. Then $(\overline{\mathfrak{Rad}})^\sim$ is defined by adding a boundary to the blowup $\mathfrak{Rad}^\sim$ in the most naive way. In the proof of Lemma 4.17, we saw that $\mathcal{MFat}^{ad}_{g,n+m}$ is a subspace of a compact polyhedron $P_{g,n+m}$, which we abbreviate to P here.

Definition 4.38 The space $(\overline{\mathfrak{Rad}})^\sim$ is the subspace of $\overline{\mathfrak{Rad}} \times P$ consisting of all $([L], \Gamma, \lambda)$ such that either

(i) $[L] \in \mathfrak{Rad}$ and $(\Gamma, \lambda) \in \mathcal{G}(L)$, or

(ii) $[L] \in \overline{\mathfrak{Rad}} \setminus \mathfrak{Rad}$ and $(\Gamma, \lambda) \in P$.

Lemma 4.39 *The topological space $(\overline{\mathfrak{Rad}})^\sim$ is a compact ANR.*

Proof Fix a representative $[L]$ for each combinatorial type $[\mathcal{L}]$, and note that, if $[L]$ and $[L']$ have the same combinatorial type, there is a canonical homeomorphism $\mathcal{G}([L]) \cong \mathcal{G}([L'])$. The space $\mathcal{G}([\mathcal{L}])$ is then by definition $\mathcal{G}([L])$ for the representative $[L]$ of $[\mathcal{L}]$. We remark that $(\overline{\mathfrak{Rad}})^\sim$ is obtained by gluing together $(\overline{\mathfrak{Rad}} \setminus \mathfrak{Rad}) \times P$ and $\overline{\mathfrak{Rad}}_{[\mathcal{L}]} \times \mathcal{G}([\mathcal{L}])$ for all combinatorial types $[\mathcal{L}]$ along $\partial\overline{\mathfrak{Rad}}_{[\mathcal{L}]} \times \mathcal{G}([\mathcal{L}])$. Note that $(\overline{\mathfrak{Rad}} \setminus \mathfrak{Rad}) \times P$ is the product of a subcomplex of the finite complex $\overline{\mathfrak{Rad}}$ with a compact polyhedron. Thus parts (v) and (vii) of Proposition 4.4 say it is a compact ANR. Similarly, by Lemma 4.17 $\overline{\mathfrak{Rad}}_{[\mathcal{L}]} \times \mathcal{G}([\mathcal{L}])$ and $\partial\overline{\mathfrak{Rad}}_{[\mathcal{L}]} \times \mathcal{G}([\mathcal{L}])$ are each a product of a finite CW–complex with a compact polyhedron, and thus compact ANRs by parts (v), (vi) and (vii) of Proposition 4.4. Attaching cells $\overline{\mathfrak{Rad}}_{[\mathcal{L}]}$ one at a time in order of dimension and repeatedly applying Proposition 4.4(iv), one proves inductively over k that

$$((\overline{\mathfrak{Rad}} \setminus \mathfrak{Rad}) \times P) \cup \left(\bigcup_{\dim \overline{\mathfrak{Rad}}_{[\mathcal{L}]} \leq k} \overline{\mathfrak{Rad}}_{[\mathcal{L}]} \times \mathcal{G}([\mathcal{L}]) \right)$$

is a compact ANR. This uses that $\overline{\mathfrak{Rad}}$ has finitely many cells after fixing g, n and m. In particular this process has to end at some $k \geq 0$, and hence $(\overline{\mathfrak{Rad}})^\sim$ is also a compact ANR. $\qquad\square$

Proposition 4.40 *The topological space $\mathfrak{Rad}^\sim$ is an ANR.*

Proof $\mathfrak{Rad}^\sim$ is an open subspace of $(\overline{\mathfrak{Rad}})^\sim$, so by Proposition 4.4(ii) we conclude it is an ANR. $\qquad\square$

Proposition 4.41 *The map $\pi_1 \colon \mathfrak{Rad}^\sim \to \mathfrak{Rad}$ is proper and cell-like.*

Proof Observe that π_1 extends to a continuous map $\overline{\pi}_1 \colon (\overline{\mathfrak{Rad}})^\sim \to \overline{\mathfrak{Rad}}$. If $K \subset \mathfrak{Rad}$ is compact, then it is also compact considered as a subset of $\overline{\mathfrak{Rad}}$, and thus closed. By continuity $\overline{\pi}_1^{-1}(K)$ is closed in $(\overline{\mathfrak{Rad}})^\sim$, and since the latter is a compact space it must be compact. But $\overline{\pi}_1^{-1}(K) \subset \mathfrak{Rad}^\sim$ and $\overline{\pi}_1^{-1}(K) \cap \mathfrak{Rad}^\sim = \pi_1^{-1}(K)$, so π_1 is proper.

That π_1 is cell-like is a consequence of Lemmas 4.14 and 4.17, which say that the point inverses of π_1 are contractible compact polyhedra, and Proposition 4.4(viii), which implies that contractible compact polyhedra are cell-like. □

Corollary 4.42 *The projection $\pi_1 \colon \mathfrak{Rad}^\sim \to \mathfrak{Rad}$ is a homotopy equivalence.*

Proof We may fix g, n and m. Then we can simply apply Theorem 4.3 to Propositions 4.37, 4.40 and 4.41. The domain is locally compact because it is an open subspace of a compact space by Lemma 4.39, and the target is locally compact by Proposition 4.37. □

4.5 The critical graph map is a homotopy equivalence

We now show that the critical graph map $\mathfrak{Rad}^\sim \to \mathcal{MFat}^{ad}$ is a homotopy equivalence using the relation between the universal bundles over $\mathfrak{Rad}$ and $\mathcal{MFat}^{ad}$. We start by recalling some well-known results regarding universal bundles:

Proposition 4.43 *Given a two-dimensional cobordism $S_{g,n+m}$ and a paracompact base space B, there are bijections natural in B between*

(i) *isomorphism classes of smooth $S_{g,n+m}$–bundles over B, that is, the transition functions lie in $\mathrm{Diff}(S_{g,n+m})$,*

(ii) *isomorphism classes of principal $\mathrm{Diff}(S_{g,n+m})$–bundles over B, and*

(iii) *isomorphism classes of principal $\mathrm{Mod}(S_{g,n+m})$–bundles over B.*

Sketch of proof For one direction of the first bijection, for a principal $\mathrm{Diff}(S_{g,n+m})$–bundle $p \colon W \to B$, its corresponding $S_{g,n+m}$–bundle is given by taking $S_{g,n+m} \times_{\mathrm{Diff}(S_{g,n+m})} W$.

For the other direction of the first bijection, suppose that $\pi \colon E \to B$ is a smooth $S_{g,n+m}$–bundle. Each fiber $E_b := \pi^{-1}(b)$ is a Riemann surface with boundary with a marked point in each boundary component. These marked points are ordered and labeled as incoming or outgoing. Let x_k^b denote the marked point in the kth incoming boundary component for $1 \le k \le n$ and x_{k+n}^b denote the marked point in the kth outgoing boundary for $1 \le k \le m$. Its corresponding $\mathrm{Diff}(S_{g,n+m})$–bundle is given by taking fiberwise orientation-preserving diffeomorphisms, ie it is the bundle $p \colon W \to B$ whose fibers are given by

$$W_b := p^{-1}(b) = \{\varphi \colon S_{g,n+m} \to E_b \mid \varphi \text{ is a diffeomorphism and } \varphi(x_i) = x_i^b\}.$$

These constructions are mutually inverse.

Because each connected component of $\mathrm{Diff}(S_{g,n+m})$ is contractible, taking π_0 gives a homotopy equivalence $\mathrm{Diff}(S_{g,n+m}) \to \mathrm{Mod}(S_{g,n+m})$. Thus there is a bijection between principal $\mathrm{Diff}(S_{g,n+m})$–bundles and principal $\mathrm{Mod}(S_{g,n+m})$–bundles, where one can obtain the $\mathrm{Mod}(S_{g,n+m})$–bundle corresponding to $p\colon W \to B$ by taking π_0 fiberwise. $\qquad\square$

We now construct a space $E\mathfrak{Rad}$ that maps to $\mathfrak{Rad}$ and use the previous proposition to show that $E\mathfrak{Rad} \to \mathfrak{Rad}$ is a universal $\mathrm{Mod}(S_{g,n+m})$–bundle. To construct this space we use the ideas of the construction of $\mathcal{EMFat}^{ad}$ in Definition 3.23. That is, as a set we define

$$E\mathfrak{Rad} := \{([L],[H]) \mid [L] \in \mathfrak{Rad} \text{ and } [H] \text{ is a marking of } \Gamma_{[L]}\}.$$

We will topologize $E\mathfrak{Rad}$ so that the map $E\mathfrak{Rad} \to \mathfrak{Rad}$ is a covering map. Then a path in $E\mathfrak{Rad}$ will be given by a path $\gamma\colon t \to [L(t)]$ in $\mathfrak{Rad}$ together with a marking $H_0\colon \Gamma_{[L(0)]} \hookrightarrow S_{g,n+m}$. Hence we must describe how H_0 and the path γ uniquely determine a sequence of markings $H_t\colon \Gamma_{[L(t)]} \hookrightarrow S_{g,n+m}$. To make this precise, we will give a procedure to obtain a well-defined marking of $\Gamma_{[\widetilde{\mathcal{L}}]}$ from a combinatorial type $[\mathcal{L}]$, a marking of $\Gamma_{[\mathcal{L}]}$ and a configuration $[\widetilde{L}] \in \partial\overline{\mathfrak{Rad}}_{[\mathcal{L}]}$, where $[\widetilde{\mathcal{L}}]$ is the combinatorial type of $[\widetilde{L}]$. To describe this procedure, notice that if $[\mathcal{L}]$ and $[\widetilde{\mathcal{L}}]$ are related in this manner, then $[\widetilde{\mathcal{L}}]$ must be obtained from $[\mathcal{L}]$ by collapsing radial and annular chambers. Hence we will start by analyzing these cases separately.

Definition 4.44 (annular chamber collapse map) Let $[\mathcal{L}]$ and $[\mathcal{L}']$ be two nondegenerate combinatorial types such that $[\mathcal{L}']$ can be obtained from $[\mathcal{L}]$ by collapsing the annular chambers $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ and let $A := \bigcup_i A_i$. We will define a map in $\mathcal{Fat}^{ad}$,

$$\rho\colon \Gamma_{[\mathcal{L}]} \to \Gamma_{[\mathcal{L}']},$$

which we will call the *annular chamber collapse map*; see Figure 20.

Choose a representative $[L]$ of $[\mathcal{L}]$. Then, following the construction of $\Gamma_{[L]}$, we can define a subgraph F_A which is given by the intersection of E_L and A. The subgraph F_A must be a forest inside $\Gamma_{[L]}$. To see this, assume there is a loop in F_A. Then there must be a loop in $\Gamma_{[L]}$, and hence there are two paired slits ζ_i and $\zeta_{\lambda(i)}$ which lie on the same radial segment. Since $[L]$ is nondegenerate there must be slits $\zeta_{i_1}, \zeta_{i_2}, \ldots, \zeta_{i_j}$ such that $i_j \geq 1$ and $|\zeta_{i_l}| < |\zeta_i|$ for all i_l. Finally, since the loop is in F_A, A must contain the radial segment between ζ_i and ζ_{i_l} for some i_l. But then collapsing A will give a degenerate configuration, and we assumed $[\mathcal{L}']$ is nondegenerate. Therefore F_A is a forest in $\Gamma_{[L]}$, and since $\Gamma_{[L]} = \Gamma_{[\mathcal{L}]}$ this description gives a well-defined subforest of $\Gamma_{[\mathcal{L}]}$, giving with a well-defined map on $\mathcal{Fat}^{ad}$.

Definition 4.45 (radial chamber collapse zigzag) Let $[\mathcal{L}]$ and $[\mathcal{L}'']$ be two nondegenerate combinatorial types such that $[\mathcal{L}'']$ can be obtained from $[\mathcal{L}]$ by collapsing radial chambers. We will define an admissible fat graph $\Gamma([\mathcal{L}], [\mathcal{L}''])$ together with a zigzag in $\mathcal{Fat}^{ad}$

$$\Gamma_{[\mathcal{L}]} \xrightarrow{\tau_1} \Gamma([\mathcal{L}], [\mathcal{L}'']) \xleftarrow{\tau_2} \Gamma_{[\mathcal{L}']},$$

which we will call the *radial chamber collapse zigzag*; see Figure 21.
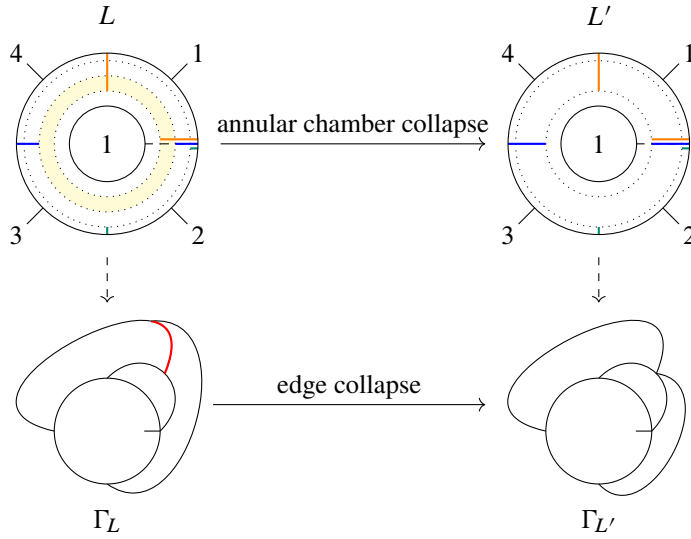
Figure 20: An example of the annular chamber collapse map. The leaves have been omitted from the graphs to make them more readable. The annular chambers are marked with dotted lines. The yellow radial sector is collapsed in L and the annular chamber collapse map is given by contracting the edge shown in red.

Choose a representative $L \in \mathfrak{QRad}$ of combinatorial type $[\mathcal{L}]$ and let $L'' \in \mathfrak{QRad}$ be the preconfiguration of combinatorial type $[\mathcal{L}'']$ obtained by collapsing radial chambers. We will call the radial segments
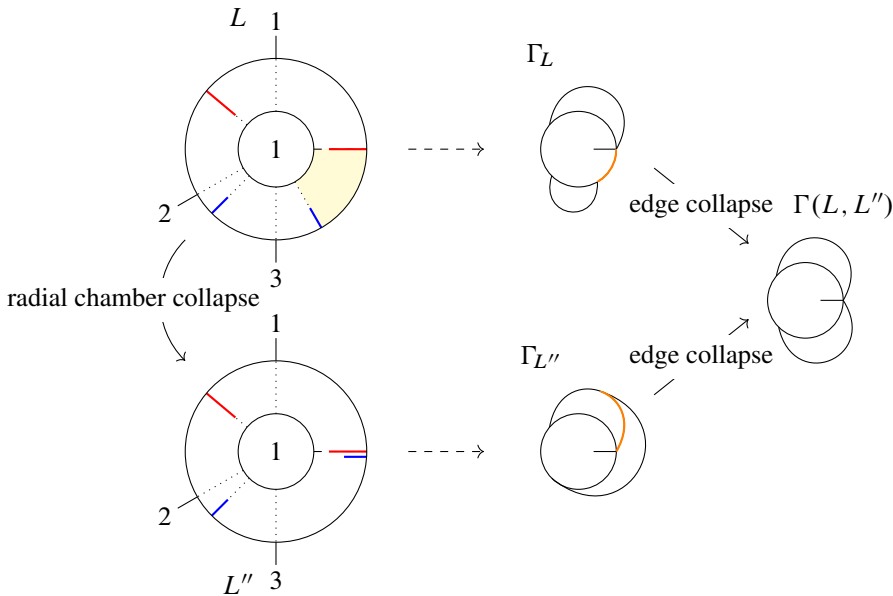
Figure 21: An example of the radial chamber collapse zigzag. The radial chambers are marked with dotted lines. The yellow radial chamber is collapsed in L and the radial chamber collapse zigzag is given by collapsing the edges shown in orange.

onto which the radial chambers have been collapsed the *special radial segments*. Notice that L'' is well defined up to a choice of L, and slit jumps and parametrization point jumps away from the special radial segments. Thus the idea is to define $\Gamma([\mathcal{L}], [\mathcal{L}''])$ as a partially unfolded graph of L'' which is unfolded at the special radial slit segments and folded everywhere else. This gives a well-defined isomorphism class of admissible fat graphs.

To make this precise, let $S_{k_1}, S_{k_2}, \ldots, S_{k_r}$ denote the special radial segments of L''. We define $\Gamma([\mathcal{L}], [\mathcal{L}'']) = \Gamma_{L'',t}$ where $t \in [0,1]^{d(L'')}$ is defined as follows:

$$t_\alpha := \begin{cases} 0 & \text{if } \alpha = k_i + j \text{ for } 1 \leq i \leq r \text{ and } 1 \leq j \leq s_{k_i} - 1, \\ 1 & \text{otherwise.} \end{cases}$$

This is a well-defined isomorphism class of admissible fat graphs, since the graph is folded in all radial segments in which jumps are allowed. Let F_L be the subgraph of Γ_L obtained by the intersection of E_L with the collapsing chambers. Then $\tau_1 \colon \Gamma_{[\mathcal{L}]} = \Gamma_L \to \Gamma_L / F_L = \Gamma([\mathcal{L}], [\mathcal{L}''])$ is a well-defined map in $\mathcal{F}at^{ad}$. Similarly, let $F_{L''}$ be the subgraph of $\Gamma_{L''}$ obtained from the intersection of $E_{L''}$ and the special radial segments. Then $\tau_2 \colon \Gamma_{[\mathcal{L}'']} = \Gamma_{L''} \to \Gamma_{L''} / F_{L''} = \Gamma([\mathcal{L}], [\mathcal{L}''])$ is a well-defined map in $\mathcal{F}at^{ad}$.

For the general case, consider any $[\widetilde{L}] \in \partial \overline{\mathfrak{Rad}}_{[\mathcal{L}]} \cap \mathfrak{Rad}_{[\widetilde{\mathcal{L}}]}$. Then $[\widetilde{\mathcal{L}}]$ is obtained from $[\mathcal{L}]$ by collapsing chambers. If we let $[\mathcal{L}']$ be the configuration obtained from collapsing only the annular chambers, then the previous construction gives a well-defined zigzag in $\mathcal{F}at^{ad}$:

(4-2) $$\Gamma_{[\mathcal{L}]} \xrightarrow{\rho} \Gamma_{[\mathcal{L}']} \xrightarrow{\tau_1} \Gamma([\mathcal{L}'], [\mathcal{L}]) \xleftarrow{\tau_2} \Gamma_{[\mathcal{L}']}.$$

Note that if $[\widetilde{\mathcal{L}}]$ is obtained by only collapsing annular chambers then $\tau_1 = \mathrm{id} = \tau_2$, and if $[\widetilde{\mathcal{L}}]$ is obtained by only collapsing radial chambers then $\rho = \mathrm{id}$.

Definition 4.46 We define the space $E\mathfrak{Rad}$ by

$$E\mathfrak{Rad} := \frac{\bigsqcup_{[\mathcal{L}]} \mathfrak{Rad}_{[\mathcal{L}]} \times \mathrm{Mark}(\Gamma_{[\mathcal{L}]})}{\sim},$$

where the disjoint union runs over all nondegenerate combinatorial types $[\mathcal{L}]$ and the equivalence relation $\sim$ is generated by saying that $([\widetilde{L}], [H]) \sim ([\widetilde{L}], [\widetilde{H}])$ if, given $[\widetilde{L}] \in \partial \overline{\mathfrak{Rad}}_{[\mathcal{L}]} \cap \mathfrak{Rad}_{[\widetilde{\mathcal{L}}]}$, $[H] \in \mathrm{Mark}(\Gamma_{[\mathcal{L}]})$ and $[\widetilde{H}] \in \mathrm{Mark}(\Gamma_{[\widetilde{\mathcal{L}}]})$, we have that $[\widetilde{H}] = (\tau_{2*})^{-1} \circ (\tau_{1*}) \circ \rho_*([H])$. Here ρ, τ_1 and τ_2 are given as in (4-2), and the induced maps are the ones constructed in Lemma 3.21.

Proposition 4.47 *The projection $E\mathfrak{Rad} \to \mathfrak{Rad}$ is a universal $\mathrm{Mod}(S_{g,n+m})$–bundle over $\mathfrak{Rad}$.*

Proof It is enough to show that $E\mathfrak{Rad} \to \mathfrak{Rad}$ is the $\mathrm{Mod}(S_{g,n+m})$–bundle corresponding to the universal surface bundle $p \colon S_h(n,m) \to \mathrm{Rad} \cong \mathfrak{Rad}$. Recall that the universal surface bundle has fibers $p_{[L]} = S([L])$, a surface with boundary with a marked point in each boundary component. These marked points are ordered, and labeled as incoming or outgoing.

Let x_k^L denote the marked point in the kth incoming boundary component for $1 \leq k \leq n$ and x_{k+n}^L denote the marked point in the kth outgoing boundary component for $1 \leq k \leq m$. Following the description in

the beginning of this subsection, the $\mathrm{Diff}(S_{g,n+m})$–bundle $W \to \mathfrak{Rad}$ corresponding to the universal surface bundle is given by taking fiberwise orientation-preserving diffeomorphisms. That is, we have

$$W_{[L]} := \{\varphi \colon S_{g,n+m} \to S([L]) \mid \varphi \text{ is an orientation-preserving diffeomorphism with } \varphi(x_i) = x_i^L\}.$$

Furthermore, its corresponding $\mathrm{Mod}(S_{g,n+m})$–bundle $Q \to \mathfrak{Rad}$, has fibers $Q_{[L]} := W_{[L]}/\text{isotopy}$. This amounts to passing to connected components of the group of diffeomorphisms.

Note that $Q_{[L]}$ is discrete, and thus by the description of $E\mathfrak{Rad}$ it is enough to show that there is a bijection between $\mathrm{Mark}(\Gamma_{[L]})$ and $Q_{[L]}$. We define inverse maps

$$\Phi \colon Q_{[L]} \leftrightarrows \mathrm{Mark}(\Gamma_{[L]}) : \Psi$$

By construction there is a canonical embedding $H_{[L]} \colon \Gamma_{[L]} \hookrightarrow S([L])$, and this embedding is a marking of $\Gamma_{[L]}$ in $S([L])$. Given $[\varphi] \in Q_{[L]}$ we define $\Phi([\varphi]) := [\varphi^{-1} \circ H_{[L]}]$; this is a well-defined map.

To go back, let $[H] \in \mathrm{Mark}(\Gamma_{[L]})$ and choose a representative $H \colon \Gamma_{[L]} \hookrightarrow S_{g,n+m}$. We will construct an orientation-preserving homeomorphism $f \colon S_{g,n+m} \to S([L])$ such that $[f \circ H] = [H_{[L]}]$, which we can approximate by a diffeomorphism φ using Nielsen's approximation theorem [36]. To do so, we use that the complements of the markings are disks: we construct the homeomorphism first on markings, and then extend it to the disks.

By Lemma 3.20, the complement $S_{g,n+m} \setminus H(\Gamma \setminus \text{leaves of } \Gamma)$ is a disjoint union of $n + m$ cylinders. For all $1 \le i \le n + m$, one of the boundary components of the ith cylinder consists of the ith boundary of $S_{g,n+m}$. The other boundary component consists of the image of the ith boundary cycles of Γ under H. The leaf corresponding to the ith boundary component is embedded in the cylinder and connects both boundary components. We conclude that $S_{g,n+m} \setminus H(\Gamma_{[L]}) \cong \bigsqcup_{i=1}^{n+m} D_i$, where each D_i is a disk.

Let x_i denote the marked point of the ith boundary component of $S_{g,n+m}$. The boundary of D_i has two copies of x_i. Connecting these on one side is the ith boundary component of $S_{g,n+m}$ and on the other side is the embedded image of the ith boundary cycle of $\Gamma_{[L]}$. The orientation of the ith boundary component of $S_{g,n+m}$ allows us to order the two copies of x_i and label them as $x_{i,1}$ and $x_{i,2}$. Similarly, $S([L]) \setminus H_{[L]}(\Gamma_{[L]}) \cong \bigsqcup_{i=1}^{n+m} \widetilde{D}_i$ where each $\widetilde{D}_i$ is a disk. Let $x_{i,j}^L$ for $j = 1, 2$ denote the two copies of the marked point on the ith boundary component of $S([L])$ that lie on the boundary of $\widetilde{D}_i$. Take $f_i|_{\partial D_i} \colon \partial D_i \to \partial \widetilde{D}_i$ to be an orientation-preserving homeomorphism satisfying $f(x_{i,j}) = x_{i,j}^L$ for $j = 1, 2$. Let f_i be an extension of $f_i|_{\partial D_i}$ to the entire disk. One can choose the maps $f_i|_{\partial D_i}$ consistently so that they glue together to a homeomorphism $f \colon S_{g,n+m} \to S([L])$. Since the maps f_i are unique up to homotopy, f is also unique up to homotopy.

We define $\Psi([H]) = [\varphi]$, where φ is a diffeomorphism approximating f. The map Ψ is well defined and by construction it is inverse to Φ. $\square$

We now extend this to $\mathfrak{Rad}^{\sim}$ by defining a fattening of $E\mathfrak{Rad}$ as follows:

Definition 4.48 The *fattening* $E\mathfrak{Rad}^\sim$ is defined as

$$E\mathfrak{Rad}^\sim := \{(([L],[H]),[\Gamma,\lambda,\tilde{H}]) \mid [\Gamma,\lambda] \in \mathcal{G}([L])\} \subset E\mathfrak{Rad} \times \mathcal{EMFat}^{ad},$$

where $\mathcal{G}([L])$ is the space given in Definition 4.13.

Recall that $E\mathfrak{Rad}$ consists of pairs $([L],[H])$ of a radial slit configuration and a marking, and that $\mathcal{EMFat}^{ad}$ consists of isomorphism classes of triples $[\Gamma,\lambda,H]$ of an admissible fat graph, a metric and a marking.

Corollary 4.49 *The projection* $E\mathfrak{Rad}^\sim \to \mathfrak{Rad}^\sim$ *is a universal* $\mathrm{Mod}(S_{g,n+m})$–*bundle over* $\mathfrak{Rad}^\sim$.

Proof Consider the diagram below, in which π_1 is a homotopy equivalence by Corollary 4.42:

$$
\begin{array}{ccc}
E\mathfrak{Rad}^\sim & \xrightarrow{\pi_1 \times \mathrm{id}} & E\mathfrak{Rad} \\
\downarrow & & \downarrow \\
\mathfrak{Rad}^\sim & \xrightarrow[\pi_1]{\cong} & \mathfrak{Rad}
\end{array}
$$

It suffices to prove this is a pullback diagram. To do so, observe that the path from $[\Gamma,\lambda] \in \mathcal{G}([L])$ to the critical graph $[\Gamma_{[L]}]$ described in Proposition 4.14 determines a zigzag in $|\mathcal{Fat}^{ad}|$ under the composite

$$\mathcal{G}([L]) \overset{\iota}{\hookrightarrow} \mathcal{MFat}^{ad} \xrightarrow{r(-,1)} |\mathcal{Fat}^{ad}|,$$

where ι is the inclusion and r is the map given in Lemma 3.12. Moreover, since $\mathcal{G}([L])$ is contractible, ι is an inclusion and $r(-,1)$ is a homotopy equivalence, there is a contractible choice of zigzags representing paths from $[\Gamma,\lambda]$ to $[\Gamma_{[L]}]$ in $\mathcal{G}([L])$. Therefore, by Lemma 3.21, a marking of $[\Gamma_{[L]}]$ uniquely determines a marking of $[\Gamma]$ and vice versa. Thus, for $[\Gamma,\lambda] \in \mathcal{G}([L])$, giving a tuple $(([L],[H]),[\Gamma,\lambda,\tilde{H}]) \in E\mathfrak{Rad} \times \mathcal{EMFat}^{ad}$ is equivalent to giving either a triple $(([L],[H]),[\Gamma,\lambda])$ or a triple $([L],[\Gamma,\lambda,\tilde{H}])$. $\square$

We now describe a general result on universal bundles, which we use to conclude that π_2 is a homotopy equivalence.

Proposition 4.50 *Let* $E \to B$ *and* $E' \to B'$ *be universal principal* G–*bundles with* B *and* B' *paracompact spaces. Let* $f : B \to B'$ *be a continuous map. If* $f^*(E')$ *is isomorphic to* E *as a bundle over* B, *then* f *is a homotopy equivalence.*

Proof For any paracompact space X, there is a diagram

$$
\begin{array}{ccc}
[X,B] & \xrightarrow{\cong} & \{\text{principal } G\text{–bundles over } X\} \\
{\scriptstyle f\circ-}\downarrow & \nearrow & \\
[X,B'] & {\scriptstyle \cong} &
\end{array}
$$

which commutes since $f^*(E') \cong E$. For $X = B'$, one finds there is a $[g] \in [B',B]$ such that $[f \circ g] = [\mathrm{id}_{B'}]$. Then $g^*(E) \cong g^*(f^*(E')) = E'$, so we can repeat the argument and obtain that there is an $h \in [B,B']$ such that $[g \circ h] = [\mathrm{id}_B]$. Finally, since $[h] = [f \circ g \circ h] = [f]$, f and g are mutually inverse homotopy equivalences. $\square$

Corollary 4.51 *The projection $\pi_2 \colon \mathfrak{Rad}^\sim \to \mathcal{MF}at^{ad}$ is a homotopy equivalence.*

Proof This follows from Proposition 4.50, as there is a pullback diagram

$$
\begin{array}{ccc}
E\mathfrak{Rad}^\sim & \xrightarrow{\ \pi_2 \times \mathrm{id}\ } & \mathcal{EMF}at^{ad} \\
\downarrow & & \downarrow \\
\mathfrak{Rad}^\sim & \xrightarrow{\ \pi_2\ } & \mathcal{MF}at^{ad}
\end{array}
$$
$\hfill\square$

5 Sullivan diagrams and the harmonic compactification

We now compare the harmonic compactification of radial slit configurations $\overline{\mathfrak{Rad}}$ and the space of Sullivan diagrams $\mathcal{SD}$, as in Definitions 2.15 and 3.16, respectively. To do this, we observe that $\overline{\mathfrak{URad}}$ is the subcomplex of $\overline{\mathfrak{Rad}}$ consisting of cells indexed by the subset $\Upsilon_{\mathfrak{U}}$ of Υ made up of all combinatorial types of unilevel radial slit configurations. As a consequence, the projection $p \colon \overline{\mathfrak{Rad}} \to \overline{\mathfrak{URad}}$ is cellular.

Proposition 5.1 *The space $\mathcal{SD}$ is homotopy equivalent to $\overline{\mathfrak{Rad}}$. In fact, there is a cellular homeomorphism between $\overline{\mathfrak{URad}}$ and $\mathcal{SD}$.*

Proof It is enough to show this for connected cobordisms. Recall that the harmonic compactification of the space of radial slit configurations $\overline{\mathfrak{Rad}}$ is homotopy equivalent to the space of unilevel radial slit configurations $\overline{\mathfrak{URad}}$ by Lemma 2.22, so it suffices to prove the second stronger statement.

Since in $\overline{\mathfrak{URad}}$ all annuli have the same outer and inner radius and all slits sit in the outer boundary, the annular chambers are superfluous information. Thus, the combinatorial type of a unilevel configuration is determined only by its radial chamber configuration. More precisely, two univalent configurations $[L]$ and $[L']$ have the same combinatorial type if and only if they differ from each other only by the size of the radial chambers. Finally, the orientations of the complex plane and the positive real line induce a total ordering of the radial chambers on each annulus.

Similarly, on a Sullivan diagram the leaves of the boundary cycles and the fat structure at the vertices where they are attached give a total ordering of the edges on the admissible cycles. We say two Sullivan diagrams $[\Gamma]$ and $[\Gamma']$ have the same combinatorial data if they differ from each other only on the lengths of the edges on the admissible cycles. A (*nonmetric*) *Sullivan diagram G* is an equivalence class of Sullivan diagrams under this relation. We will first show that a radial slit configuration and a Sullivan diagram are given by the same combinatorial data. That is, that there is a bijection

$\Upsilon_{\mathfrak{U}} := \{\text{combinatorial types of unilevel radial slit configurations}\} \leftrightarrow \Lambda := \{\text{nonmetric Sullivan diagrams}\}.$

We define a map $f \colon \Upsilon_{\mathfrak{U}} \to \Lambda$ by $[\mathcal{L}] \mapsto G_{[\mathcal{L}],0}$, where $G_{[\mathcal{L}],0}$ is the underlying (nonmetric) Sullivan diagram of a unfolded graph of $[\mathcal{L}]$. This map is well defined, since a slit or a parametrization point jumping along another slit corresponds to a slide of a vertex along an edge not belonging to the admissible cycle. For example, the configurations in Figure 9 are mapped to the graphs in Figure 22.
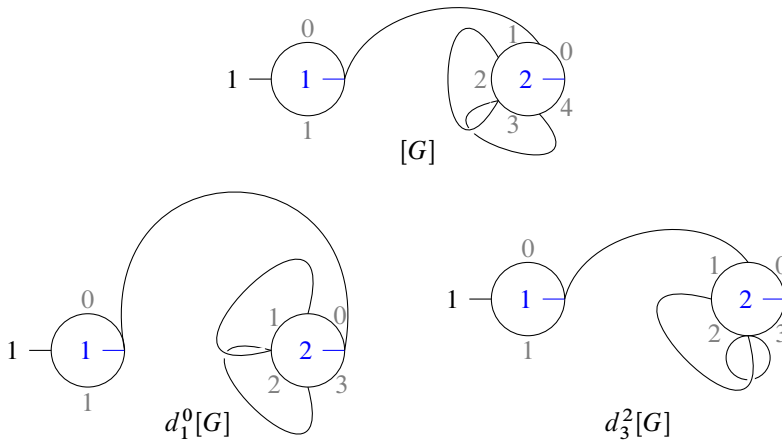
Figure 22: The top depicts a 5–cell which is a product $\Delta^1 \times \Delta^4$ of simplices in $\mathcal{SD}$, and the bottom two parts of its boundary. The edges are numbered in gray.

We next construct the inverse map $g: \Lambda \to \Upsilon_{\mathfrak{U}}$. Notice that any nonmetric Sullivan diagram has a canonically associated metric Sullivan diagram by assigning all the edges in an admissible cycle the same length. Moreover, any Sullivan diagram has a fat graph representative with all its vertices on the admissible cycles. A representative of a metric Sullivan diagram with all its vertices on the admissible cycles is given by the following data:

(i) $C_1, C_2, \ldots, C_n$ are parametrized circles which are disjoint, ordered and of length 1.

(ii) $l_1, l_2, \ldots, l_s$ are a finite number of chords, where a chord is a graph which consists of two vertices connected by an edge. Let V denote the set of vertices of such chords.

(iii) $\widetilde{V} \subset V$ is a subset such that $\widetilde{V}$ contains at least one vertex of each chord and $|V \setminus \widetilde{V}| = m$.

(iv) $\alpha: \widetilde{V} \to \bigsqcup_i C_i$ is an assignment which will indicate how to attach the chords onto the n circles. Two or more chords may be attached on the same circle and even on the same point. The assignment α should attach at least one chord on each circle.

(v) For each x in the image of α, we have an ordering of the subset of chords attached to x, that is, an ordering of the set $\alpha^{-1}(x)$.

From this data one can construct a metric fat graph with inner vertices of valence greater or equal to 3. The chords are attached onto the n circles using α. This gives the circles the structure of a graph by considering the attaching points as vertices and the intervals between them as edges. It just remains to give a fat structure at the attaching points. To do this, let x be in the image of α. The parametrization of the circles gives a notion of incoming and outgoing half-edges on x, say e_x^- and e_x^+, respectively. Moreover, there is an ordering of the chords attached on x, say $(l_{x,1}, l_{x,2}, \ldots, l_{x,s})$. The cyclic ordering at x is given by $(e_x^-, l_{x,1}, l_{x,2}, \ldots, l_{x,s}, e_x^+)$ as is shown in Figure 23. Informally, all chords are attached on the outside of the circles according to the order given by the data. The chords that are attached only at one vertex give the leaves of the Sullivan diagram.
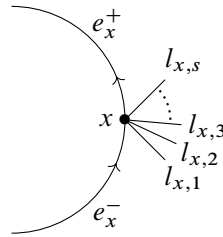
Figure 23: The fat structure induced at vertex x where the cyclic ordering is given by the orientation on the plane.

From this it is clear what the inverse map g should be. Given a Sullivan diagram G, its associated metric Sullivan diagram gives the data (i)–(v) listed above. Then $g(G) = (\zeta, \lambda, \tilde{\omega}, \vec{r}, \vec{P})$ where ζ is given by α on the chords attached at both ends, λ is given by those chords (ie $\lambda(i) = k$ if and only if there is a chord attached on both ends connecting i and k), $\vec{P}$ is given by α on the chords attached only at one vertex, and $\tilde{\omega}$ and $\vec{r}$ are completely determined by the ordering of the chords at each attaching point. This map is well defined since slides along chords correspond to jumps along slits, and it is an inverse to f.

We will show that $\overline{\mathfrak{URad}}$ and $\mathcal{SD}$ have homeomorphic CW–structures, where the cells are indexed by $\Upsilon_{\mathfrak{U}} \cong \Lambda$, by giving cellular homeomorphisms

$$\overline{\mathfrak{URad}} \xleftarrow{\varphi} \frac{\bigsqcup_{[\mathcal{L}] \in \Upsilon_{\mathfrak{U}}} e_{[\mathcal{L}]}}{\sim} \xrightarrow{\psi} \mathcal{SD}.$$

We already saw the map φ in Definition 2.15. To construct the map ψ, one first observes that any Sullivan diagram $[\Gamma]$ in $\mathcal{SD}$ is uniquely determined by its nonmetric underlying Sullivan diagram G and a tuple $(\vec{t}_1, \ldots, \vec{t}_{n_p})$, where t_{ij} is the length of the jth edge of the ith admissible cycle. Using this we can define

$$\psi(e_{[\mathcal{L}]}, (\vec{t}_1, \ldots, \vec{t}_{n_p})) = [\Gamma] = (f([\mathcal{L}]), (\vec{t}_1, \ldots, \vec{t}_{n_p})).$$

It is easy to show that the map ψ is continuous, and by construction the homeomorphism $\varphi \circ \psi^{-1}$ is cellular with respect to the CW–structures on $\overline{\mathfrak{URad}}$ and $\mathcal{SD}$. $\square$

References

[1]   **J Abhau**, **C-F Bödigheimer**, **R Ehrenfried**, *Homology of the mapping class group $\Gamma_{2,1}$ for surfaces of genus 2 with a boundary curve*, Geom. Topol. Monogr. 14, Geom. Topol. Publ., Coventry (2008)   MR   Zbl

[2]   **C-F Bödigheimer**, *On the topology of moduli spaces*, *I*: *Hilbert uniformization*, Math. Gottingensis 7–8, Sonderforschungsbereich Geom. Anal., Göttingen (1990)

[3]   **C-F Bödigheimer**, *Configuration models for moduli spaces of Riemann surfaces with boundary*, Abh. Math. Sem. Univ. Hamburg 76 (2006) 191–233   MR   Zbl

[4]   **C-F Bödigheimer**, *Hilbert uniformization of Riemann surfaces*, *I*: *Short version*, preprint (2007)   Available at http://www.math.uni-bonn.de/people/cfb/PUBLICATIONS/short-hilbert.pdf

[5] **C-F Bödigheimer**, **U Tillmann**, *Stripping and splitting decorated mapping class groups*, from "Cohomological methods in homotopy theory" (J Aguadé, C Broto, C Casacuberta, editors), Progr. Math. 196, Birkhäuser, Basel (2001) 47–57 MR Zbl

[6] **B H Bowditch**, **D B A Epstein**, *Natural triangulations associated to a surface*, Topology 27 (1988) 91–117 MR Zbl

[7] **M Chas**, **D Sullivan**, *String topology*, preprint (1999) arXiv math/9911159

[8] **R L Cohen**, **V Godin**, *A polarized view of string topology*, from "Topology, geometry and quantum field theory" (U Tillmann, editor), Lond. Math. Soc. Lecture Note Ser. 308, Cambridge Univ. Press (2004) 127–154 MR Zbl

[9] **K Costello**, *A dual version of the ribbon graph decomposition of moduli space*, Geom. Topol. 11 (2007) 1637–1652 MR Zbl

[10] **K Costello**, *Topological conformal field theories and gauge theories*, Geom. Topol. 11 (2007) 1539–1579 MR Zbl

[11] **G C Drummond-Cole**, **K Poirier**, **N Rounds**, *Chain-level string topology operations*, preprint (2015) arXiv 1506.02596

[12] **J Ebert**, *Hilbert-Uniformisierung Kleinscher Flächen*, Diplomarbeit in Mathematik, Universität Bonn (2003) Available at `https://ivv5hpp.uni-muenster.de/u/jeber_02/papers/Diplomarbeit.pdf`

[13] **J F Ebert**, **R M Friedrich**, *The Hilbert-uniformization is real-analytic*, preprint (2006) arXiv math/0601378

[14] **D Egas Santander**, *Comparing fat graph models of moduli space*, preprint (2015) arXiv 1508.03433

[15] **R Ehrenfried**, *Die Homologie der Modulräume berandeter Riemannscher Flächen von kleinem Geschlecht*, PhD thesis, Universität Bonn (1997)

[16] **Y Félix**, **J-C Thomas**, *String topology on Gorenstein spaces*, Math. Ann. 345 (2009) 417–452 MR Zbl

[17] **C D Feustel**, *Homotopic arcs are isotopic*, Proc. Amer. Math. Soc. 17 (1966) 891–896 MR Zbl

[18] **R Fritsch**, **R A Piccinini**, *Cellular structures in topology*, Cambridge Stud. Adv. Math. 19, Cambridge Univ. Press (1990) MR Zbl

[19] **S Galatius**, *Mod p homology of the stable mapping class group*, Topology 43 (2004) 1105–1132 MR Zbl

[20] **V Godin**, *A category of bordered fat graphs and the mapping class group of a bordered surface*, PhD thesis, Stanford University (2004) Available at `https://www.proquest.com/docview/305128598`

[21] **V Godin**, *Higher string topology operations*, preprint (2007) arXiv 0711.4859

[22] **V Godin**, *The unstable integral homology of the mapping class groups of a surface with boundary*, Math. Ann. 337 (2007) 15–60 MR Zbl

[23] **A Gramain**, *Le type d'homotopie du groupe des difféomorphismes d'une surface compacte*, Ann. Sci. École Norm. Sup. 6 (1973) 53–66 MR Zbl

[24] **U Hamenstädt**, *Teichmüller theory*, from "Moduli spaces of Riemann surfaces" (B Farb, R Hain, E Looijenga, editors), IAS/Park City Math. Ser. 20, Amer. Math. Soc., Providence, RI (2013) 45–108 MR Zbl

[25] **J L Harer**, *Stability of the homology of the mapping class groups of orientable surfaces*, Ann. of Math. 121 (1985) 215–249 MR Zbl

[26] **J L Harer**, *The virtual cohomological dimension of the mapping class group of an orientable surface*, Invent. Math. 84 (1986) 157–176 MR Zbl

[27] **J L Harer**, *The cohomology of the moduli space of curves*, from "Theory of moduli" (E Sernesi, editor), Lecture Notes in Math. 1337, Springer (1988) 138–221 MR Zbl

[28] **A Hatcher**, *On triangulations of surfaces*, Topology Appl. 40 (1991) 189–194 MR Zbl

[29] **K Igusa**, *Higher Franz–Reidemeister torsion*, AMS/IP Stud. Adv. Math. 31, Amer. Math. Soc., Providence, RI (2002) MR Zbl

[30] **R M Kaufmann**, *Open/closed string topology and moduli space actions via open/closed Hochschild actions*, Symmetry Integrability Geom. Methods Appl. 6 (2010) art. id. 036 MR Zbl

[31] **A Klamt**, *The complex of formal operations on the Hochschild chains of commutative algebras*, J. Lond. Math. Soc. 91 (2015) 266–290 MR Zbl

[32] **M Kontsevich**, *Intersection theory on the moduli space of curves and the matrix Airy function*, Comm. Math. Phys. 147 (1992) 1–23 MR Zbl

[33] **R C Lacher**, *Cell-like mappings and their generalizations*, Bull. Amer. Math. Soc. 83 (1977) 495–552 MR Zbl

[34] **I Madsen**, **M Weiss**, *The stable mapping class group and stable homotopy theory*, from "European Congress of Mathematics" (A Laptev, editor), Eur. Math. Soc., Zürich (2005) 283–307 MR Zbl

[35] **J van Mill**, *Infinite-dimensional topology*: *prerequisites and introduction*, North-Holland Math. Libr. 43, North-Holland, Amsterdam (1989) MR Zbl

[36] **J Nielsen**, *Die Isomorphismengruppe der freien Gruppen*, Math. Ann. 91 (1924) 169–209 MR Zbl

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

[38] **K Poirier**, *String topology and compactified moduli spaces*, PhD thesis, City University of New York (2010) Available at https://www.proquest.com/docview/763491239

[39] **K Strebel**, *Quadratic differentials*, Ergebnisse der Math. 5, Springer (1984) MR Zbl

[40] **T Tradler**, **M Zeinalian**, *On the cyclic Deligne conjecture*, J. Pure Appl. Algebra 204 (2006) 280–299 MR Zbl

[41] **N Wahl**, *Homological stability for mapping class groups of surfaces*, from "Handbook of moduli, III" (G Farkas, I Morrison, editors), Adv. Lect. Math. 26, International, Somerville, MA (2013) 547–583 MR Zbl

[42] **N Wahl**, *Universal operations in Hochschild homology*, J. Reine Angew. Math. 720 (2016) 81–127 MR Zbl

[43] **N Wahl**, **C Westerland**, *Hochschild homology of structured algebras*, Adv. Math. 288 (2016) 240–307 MR Zbl

Mathematical Institute, University of Bonn
Bonn, Germany
Department of Mathematics, University of Toronto
Toronto, ON, Canada
daniela.egassantander@epfl.ch, a.kupers@utoronto.ca

Towards a higher-dimensional construction of
stable/unstable Lagrangian laminations

Sᴀɴɢᴊɪɴ Lᴇᴇ

We generalize some properties of surface automorphisms of pseudo-Anosov type. First, we generalize the Penner construction of a pseudo-Anosov homeomorphism, and show that if a symplectic automorphism is constructed by our generalized Penner construction, then it has an invariant Lagrangian branched submanifold and an invariant Lagrangian lamination. These invariants are higher-dimensional generalizations of a train track and a geodesic lamination in the surface case. As an application, we compute the Lagrangian Floer homology of some Lagrangians on plumbings of cotangent bundles of spheres.

53D05, 53D40, 57R17

1 Introduction

By the Nielsen–Thurston classification of surface diffeomorphisms, an automorphism $\psi \colon S \xrightarrow{\sim} S$ of a compact oriented surface S is of one of three types: periodic, reducible, or pseudo-Anosov. We recommend Casson and Bleiler [2] or Thurston [14]. Maher [7] shows that, for a suitable notion of randomness, a random element of the mapping class group is pseudo-Anosov.

Let us assume that ψ is of pseudo-Anosov type. For any closed curve $C \subset S$, it is known that there is a sequence $\{L_m\}_{m \in \mathbb{N}}$ of closed geodesics such that L_m is isotopic to $\psi^m(C)$ for all $m \in \mathbb{N}$, and $\{L_m\}_{m \in \mathbb{N}}$ converges to a closed subset $\mathscr{L}$ with respect to the Hausdorff metric on closed subsets. Moreover, $\mathscr{L}$ is a geodesic lamination. The definitions of a lamination, a geodesic lamination, and a Lagrangian lamination are the following:

Definition 1.1   (1)   A k–*dimensional lamination* on an n–dimensional manifold M is a decomposition of a closed subset of M into k–dimensional submanifolds called *leaves* such that the closed subset is covered by charts of the form $I^k \times I^{n-k}$ where a leaf passing through a chart is a slice of the form $I^k \times \{\text{pt}\}$.

   (2)   An 1–dimensional lamination $\mathscr{L}$ on a Riemannian 2–manifold (S, g) is a *geodesic lamination* if every leaf of $\mathscr{L}$ is geodesic.

   (3)   An n–dimensional lamination $\mathscr{L}$ on a symplectic manifold (M^{2n}, ω) is a *Lagrangian lamination* if every leaf of $\mathscr{L}$ is a Lagrangian submanifold.

For more details, we refer the reader to Farb and Margalit [5, Chapter 15].

In [3], Dimitrov, Haiden, Katzarkov and Kontsevich defined the notion of a *pseudo-Anosov functor* of a triangulated category, and they gave examples of it on the Fukaya categories: a pseudo-Anosov map ψ on a compact oriented surface S induces a functor, also called ψ, on the derived Fukaya category $D^\pi \operatorname{Fuk}(S, \omega)$, where ω is an area form of S. In [3], the authors showed that ψ is a pseudo-Anosov functor.

In [3, Section 4], the authors listed a number of open questions. One of them is to find a symplectic automorphism ψ on a symplectic manifold M of dimension greater than 2 which has invariant transversal stable/unstable Lagrangian measured foliations. A slightly weaker version of the question is to define a symplectic automorphism ψ with invariant stable/unstable Lagrangian laminations.

The goal of the present paper is to answer the latter question. First, we define symplectic automorphisms of generalized Penner type.

Definition 1.2 Let M be a symplectic manifold. A symplectic automorphism $\psi \colon M \overset{\sim}{\longrightarrow} M$ is of *generalized Penner type* if there are two collections $A = \{\alpha_1, \ldots, \alpha_m\}$ and $B = \{\beta_1, \ldots, \beta_l\}$ of Lagrangian spheres satisfying

- $\alpha_i \cap \alpha_j = \varnothing$, and $\beta_i \cap \beta_j = \varnothing$ for all $i \neq j$,
- $\alpha_i \pitchfork \beta_j$ for all i and j, and
- for each $\alpha_i \in A$ (resp. $\beta_j \in B$), there is at least one $\beta_j \in B$ (resp. $\alpha_i \in A$) such that $\alpha_i \cap \beta_j \neq \varnothing$,

so that ψ is a product of positive powers of Dehn twists τ_i along α_i and negative powers of Dehn twists σ_j along β_j, subject to the condition that every sphere appear in the product.

We will define a Dehn twist along a Lagrangian sphere in Section 2.2, partly to establish notation.

Then, we will define the notion of *Lagrangian branched submanifold* and *carried by*. These are higher-dimensional generalizations of the notion of train tracks and "carried by a train track" in surface theory. Roughly, in the surface theory, if a curve C is carried by a train track τ, then it is possible to encode C on τ with the extra data called "weights". We refer the reader to Farb and Margalit [5] for detail. Motivated by this, we will give the higher-dimensional generalizations of train tracks and the notion of "carried by" in Sections 3.1 and 3.3. Then, we prove Theorem 1.3 at the end of Section 3.

Theorem 1.3 *Let M be a symplectic manifold and let $\psi \colon M \overset{\sim}{\longrightarrow} M$ be a symplectic automorphism of generalized Penner type. Then there exists a Lagrangian branched submanifold $\mathcal{B}_\psi$ such that if L is a Lagrangian submanifold which is carried (resp. weakly carried) by $\mathcal{B}_\psi$, then $\psi^m(L)$ is carried (resp. weakly carried) by $\mathcal{B}_\psi$ for all $m \in \mathbb{N}$.*

Remark 1.4 Theorem 1.3 cares about symplectic automorphisms of generalized Penner type. However, there should be a generalized version of Theorem 1.3 for arbitrary symplectic automorphisms, which we do not prove in the current paper.

In Section 4, we will prove that if a Lagrangian L is carried by a Lagrangian branched submanifold $\mathcal{B}$, one can encode L on $\mathcal{B}$ with extra data called *braids*. The definition of braids will appear in Section 4.3. In Sections 5 and 6, by using the notion of braids, we prove our main theorem, ie Theorem 1.5.

Theorem 1.5 *Let M be a symplectic manifold, and let $\psi \colon M \stackrel{\sim}{\longrightarrow} M$ be a symplectic automorphism of generalized Penner type. Then there is a Lagrangian lamination $\mathcal{L}$ such that if L is a Lagrangian submanifold of M which is carried by $\mathcal{B}_\psi$, then there is a sequence of Lagrangian submanifolds L_m satisfying*

- *L_m is Hamiltonian isotopic to $\psi^m(L)$, and*
- *L_m converges to $\mathcal{L}$ as $m \to \infty$.*

Also, in Section 6.4, we will see how this generalizes to symplectic automorphisms which are not of generalized Penner type.

Finally, we will talk about Lagrangian Floer theory related to Theorems 1.3 and 1.5. The results will be written in Section 7.

Structure of the paper

This paper consists of 7 sections. In Section 2, we review plumbing spaces and generalized Dehn twists. We will prove Theorem 1.3 in Section 3 and Theorem 1.5 in Sections 4–6. In Section 7, we will discuss the relation of Theorems 1.3 and 1.5 to Lagrangian Floer theory and give a related calculation of Floer cohomology (Theorem 7.3).

Acknowledgments

2 Preliminaries

In this section, we will review plumbings of cotangent bundles and generalized Dehn twists, partly to establish notation.

2.1 Plumbing spaces

Let α and β be oriented spheres S^n. We describe how to plumb $T^*\alpha$ and $T^*\beta$ at $p \in \alpha$ and $q \in \beta$. Let $U \subset \alpha$ and $V \subset \beta$ be small disk neighborhoods of p and q. Then, we identify T^*U and T^*V so that the base U (resp. V) of T^*U (resp. T^*V) is identified with a fiber of T^*V (resp. T^*U).

To do this rigorously, we fix coordinate charts $\psi_1 \colon U \to \mathbb{R}^n$ and $\psi_2 \colon V \to \mathbb{R}^n$. Then we obtain a compositions of symplectomorphisms

$$T^*U \xrightarrow{(\psi_1^*)^{-1}} T^*\mathbb{R}^n \simeq \mathbb{R}^{2n} \xrightarrow{f} \mathbb{R}^{2n} \simeq T^*\mathbb{R}^n \xrightarrow{\psi_2^*} T^*V,$$

where $f(x_1, \ldots, x_n, y_1, \ldots, y_n) = (y_1, \ldots, y_n, -x_1, \ldots, -x_n)$.

A plumbing space $P(\alpha, \beta)$ of $T^*\alpha$ and $T^*\beta$ is defined by $T^*\alpha \sqcup T^*\beta/\sim$, where $x \sim (\psi_2^* \circ f \circ \psi_1^{*-1})(x)$ for all $x \in T^*U$. Since $\psi_2^* \circ f \circ \psi_1^{*-1}$ is a symplectomorphism, $P(\alpha, \beta)$ has a natural symplectic structure induced by the standard symplectic structures of cotangent bundles.

Since the plumbing procedure is a local procedure, we can plumb a finite collection of cotangent bundles of the same dimension at finitely many points. For convenience, we plumb cotangent bundles of oriented manifolds.

Note that we can replace f by

$$g(x_1, \ldots, x_n, y_1, \ldots, y_n) = (-y_1, y_2, \ldots, y_n, x_1, -x_2, \ldots, -x_n).$$

If we plumb $T^*\alpha$ and $T^*\beta$ at one point using g, this plumbing space is symplectomorphic to the previous plumbing space $P(\alpha, \beta)$, which is plumbed using f. However, if we plumb at more than one point, then by replacing f with g at a plumbing point, the plumbing space will change.

Definition 2.1 Let $\alpha_1, \ldots, \alpha_m$ be oriented manifolds of dimension n.

(1) A *plumbing datum* is a collection of pairs of nonnegative integers $(a_{i,j}, b_{i,j})$ for all $1 \le i \le j \le m$ and collections of distinct points

$$\{p_k^{i,j} \in \alpha_i \mid 1 \le i \le j \le m, 1 \le k \le a_{i,j} + b_{i,j}\},$$
$$\{q_k^{i,j} \in \alpha_j \mid 1 \le i \le j \le m, 1 \le k \le a_{i,j} + b_{i,j}\}.$$

(2) A *plumbing space* $P(\alpha_1, \ldots, \alpha_m)$, with the given plumbing datum, is given by

$$P(\alpha_1, \ldots, \alpha_m) = T^*\alpha_1 \sqcup \cdots \sqcup T^*\alpha_m/\sim,$$

where the equivalence relation $\sim$ is defined as follows: first, choose small disk neighborhoods $U_k^{i,j} \subset \alpha_i$ of $p_k^{i,j}$ and $V_k^{i,j} \subset \alpha_j$ of $q_k^{i,j}$ such that $U_{k_1}^{i_1,j_1} \cap U_{k_2}^{i_2,j_2} = \varnothing$ if $(i_1, j_1, k_1) \ne (i_2, j_2, k_2)$ and orientation-preserving coordinate charts $\psi_k^{i,j} \colon U_k^{i,j} \xrightarrow{\sim} \mathbb{R}^n$ and $\phi_k^{i,j} \colon V_k^{i,j} \xrightarrow{\sim} \mathbb{R}^n$; then for all $x \in T^*U_k^{i,j}$,

$$x \sim \begin{cases} (\phi_k^{i,j*} \circ f \circ (\psi_k^{i,j*})^{-1})(x) & \text{if } 1 \le k \le a_{i,j}, \\ (\phi_k^{i,j*} \circ g \circ (\psi_k^{i,j*})^{-1})(x) & \text{if } a_{i,j} + 1 \le k \le a_{i,j} + b_{i,j}. \end{cases}$$

(3) A *plumbing point* is an identified point $p_k^{i,j} \sim q_k^{i,j} \in P(\alpha_1, \ldots, \alpha_m)$.

[Figure 1](#) shows some examples of plumbing spaces.

If α_i is of dimension $n \ge 2$, then specific choices of plumbing points do not change the symplectic topology of $P(\alpha_1, \ldots, \alpha_m)$.
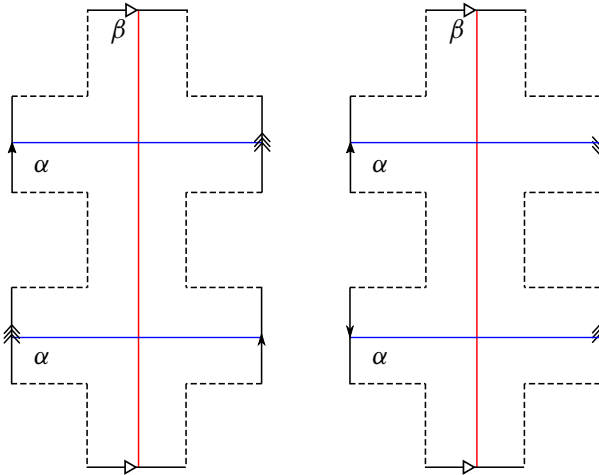
Figure 1: $P(\alpha \simeq S^1, \beta \simeq S^1)$ with plumbing datum $(2, 0)$ (left) and $(1, 1)$ (right).

2.2 Generalized Dehn twist

Let

$$T^*S^n = \{(u, v) \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \mid \|u\| = 1, \langle u, v \rangle = 0\},$$

$$S^n = \{(u, 0_{n+1}) \in T^*S^n\},$$

where $(u, v) \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$ and $\langle u, v \rangle$ is the standard inner product of u and v in $\mathbb{R}^{n+1}$. Moreover, let 0_k be the origin in $\mathbb{R}^k$.

We fix a Hamiltonian function $\mu(u, v) = \|v\|$ on $T^*S^n \setminus S^n$. Then μ induces a circle action on $T^*S^n \setminus S^n$ given by

$$\sigma(e^{it})(u, v) = \left(\left(\cos(t)u + \sin(t)\frac{v}{\|v\|} \right), (\cos(t)v - \sin(t)\|v\|u) \right).$$

Let $r: [0, \infty) \to \mathbb{R}$ be a smooth decreasing function such that $r(0) = \pi$ and $r(t) = 0$ for all $t \geq \epsilon$ for a small positive number ϵ. If ω_0 is the standard symplectic form of T^*S^n, we define a symplectic automorphism $\tau: (T^*S^n, \omega_0) \xrightarrow{\sim} (T^*S^n, \omega_0)$ by

$$(2\text{-}1) \qquad \tau(u, v) = \begin{cases} \sigma(e^{ir(\mu(u,v))})(u, v) & \text{if } v \neq 0_{n+1}, \\ (-u, 0_{n+1}) & \text{if } v = 0_{n+1}. \end{cases}$$

Let (M^{2n}, ω) be a symplectic manifold and let $L \simeq S^n$ be a Lagrangian sphere in M. By the Lagrangian neighborhood theorem — see Weinstein [16] — there is a neighborhood $N(L) \supset L$ and a symplecto-morphism $\phi: T^*S^n \xrightarrow{\sim} N(L)$. We define a generalized Dehn twist τ_L along L by

$$(2\text{-}2) \qquad \tau_L(x) = \begin{cases} (\phi \circ \tau \circ \phi^{-1})(x) & \text{if } x \in N(L), \\ x & \text{if } x \notin N(L). \end{cases}$$

Note that the support of τ_L is contained in $N(L)$. From now on, a generalized Dehn twist will just be called a Dehn twist.

Remark 2.2 We will use two specific Dehn twists $\tau, \tilde{\tau} : T^*S^n \xrightarrow{\sim} T^*S^n$ which are defined by (2-1) and two functions $r, \tilde{r} : [0, \infty) \to \mathbb{R}$. The function r (resp. $\tilde{r}$) defining τ (resp. $\tilde{\tau}$) satisfies the above conditions in addition to $r(t) = \pi$ for all $t \leq \frac{\epsilon}{2}$ (resp. $\tilde{r}'(0) < 0$). The two Dehn twists τ and $\tilde{\tau}$ are equivalent in the sense that $\tau \circ \tilde{\tau}^{-1}$ is a Hamiltonian isotopy.

Dehn twists have been studied extensively by Seidel. For example, Seidel [12] proved the following theorem.

Theorem 2.3 *Let α be a Lagrangian sphere and β be a Lagrangian submanifold of a symplectic manifold M. If α and β intersect transversally at only one point, $\beta \# \alpha$ is Lagrangian isotopic to $\tau_\alpha(\beta)$, where $\beta \# \alpha$ is a Lagrangian surgery of β and α.*

We prove Theorem 2.3 in the special case that β is also a sphere and $M = P(\alpha, \beta)$, as an illustration of the "spinning" procedure.

To define "spinning", we use the following notation. Let $y \in S^{n-1} \subset \mathbb{R}^n$. Then

$$\psi_y : T^*S^1 \simeq S^1 \times \mathbb{R} \to T^*S^n,$$

$$((\cos\theta, \sin\theta), t) \mapsto \big((\cos\theta(0_n, 1) + \sin\theta(y, 0)), (t\cos\theta(y, 0) - t\sin\theta(0_n, 1))\big)$$

is a symplectic embedding. Let W_y be the embedded symplectic surface $\psi_y(T^*S^1)$. We would like to note that $W_y = W_{-y}$.

Definition 2.4 Given a curve C in T^*S^1, its *spun image* $S(C)$ is $\bigcup_{y \in S^{n-1}} \psi_y(C)$.

Remark 2.5 A spun image $S(C)$ of a curve $C \subset T^*S^1$ is not an embedded submanifold of T^*S^n for all C. However, for some C, $S(C)$ is an embedded submanifold. For example, if C is invariant under the action $(\theta, t) \mapsto (-\theta, -t)$ on T^*S^1, then $S(C)$ is an embedded submanifold. Moreover, if $S(C)$ is a submanifold, then it is easy to prove that $S(C)$ is Lagrangian.

Proof of Theorem 2.3 We use $T^*\alpha$ and $T^*\beta$ to indicate neighborhoods of α and β inside $M = P(\alpha, \beta)$. Let p be the intersection point of α and β. Then, $T_p^*\alpha = \beta \cap T^*\alpha$. The closure of $T_p^*\alpha$ is denoted by D_p^-; we use D to indicate that this is a disk and the subscript p means that p is the center of D_p^-. The meaning of the negative sign in D_p^- will be explained in the next section. Since τ_α is supported on $T^*\alpha$,

$$\tau_\alpha(\beta) = \tau_\alpha(\beta \cap T^*\alpha) \cup \tau_\alpha(\beta \setminus T^*\alpha) = \tau_\alpha(D_p^-) \cup (\beta \setminus T^*\alpha).$$

There exists $\phi : T^*S^n \xrightarrow{\sim} T^*\alpha$ such that $\tau_\alpha = \phi \circ \tau \circ \phi^{-1}$. Without loss of generality, $\phi((0_n, 1), 0_{n+1}) = p$ and

$$D_p^- = \phi\big(\{(0_n, 1, ty, 0) \mid t \in \mathbb{R}, y \in S^{n-1} \subset \mathbb{R}^n\}\big).$$

Then

$$(\phi \circ \tau_\alpha \circ \phi^{-1})(D_p^-) = (\phi \circ \tau)\big(\{(0_n, 1, ty, 0) \mid t \in \mathbb{R}, y \in S^{n-1} \subset \mathbb{R}^n\}\big)$$

$$= \bigcup_{y \in S^{n-1}} \phi\big(\{\tau(0_n, 1, ty, 0) \mid t \in \mathbb{R}\}\big).$$
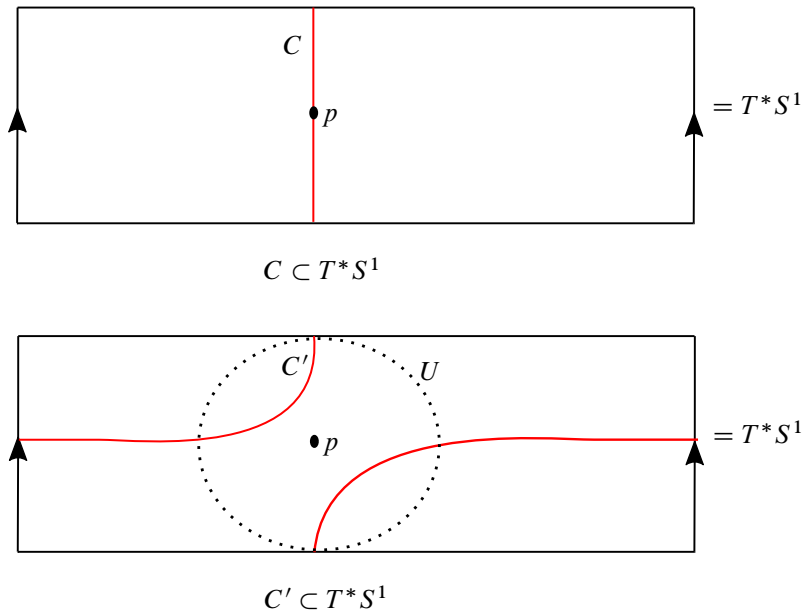
Figure 2: Curves C and C' in T^*S^1.

Since there is a curve C in T^*S^1 such that $\psi_y(C) = \{\tau(0_n, 1, ty, 0) \mid t \in \mathbb{R}\}$, $\tau_\alpha(D_p^-)$ is given by spinning with respect to p and ϕ.

Figure 2 represents $C \subset T^*S^1$. By Remark 2.5, it is easy to check that $\tau_\alpha(D_p^-)$ is Lagrangian.

One possible construction of $\beta \# \alpha$ is as follows. The Lagrangian surgery $\beta \# \alpha$ agrees with $\alpha \cup \beta$ outside of a small neighborhood U of p. On U, there is a Darboux chart ϕ satisfying

$$\phi(U) = \mathbb{C}^n, \quad \phi(\alpha) = \mathbb{R}^n, \quad \phi(\beta) = (i\mathbb{R})^n,$$

$$\phi(\beta \# \alpha) = \left\{ \left(x_1, \ldots, x_n, -\frac{\epsilon x_1}{\sqrt{x_1^2 + \cdots + x_n^2}}, \ldots, -\frac{\epsilon x_n}{\sqrt{x_1^2 + \cdots + x_n^2}} \right) \,\middle|\, x_i \in \mathbb{R} \right\}.$$

We refer the reader to Auroux [1]. Based on this construction, one could say that $\beta \# \alpha$ can be obtained by spinning a curve $C' \subset T^*S^1$ at p. Figure 2, bottom, represents $C' \subset T^*S^1$.

Similarly, we can construct a Lagrangian isotopy connecting $\tau_\alpha(\beta)$ and $\beta \# \alpha$ by spinning. $\qquad\square$

3 Lagrangian branched submanifolds

In Section 3.1, we will define Lagrangian branched submanifolds. In Section 3.2, we will introduce a construction of a fibered neighborhood of a Lagrangian branched submanifold. In Section 3.3, we will defined the notion of "carried by" by using a fibered neighborhood. In Section 3.4, we will introduce the generalized Penner construction. Finally, we will give a proof of Theorem 1.3 in Section 3.5.

3.1 Lagrangian branched submanifolds

Thurston [15] used train tracks, which are 1–dimensional branched submanifolds of surfaces, and defined the notion of "carried by a train track". In this subsection, we generalize train tracks.

The generalization of a train track is an n–dimensional branched submanifold of a $2n$–dimensional manifold. We define the n–dimensional branched submanifolds with local models, as Floyd and Oertel defined a branched surface in a 3–dimensional manifold in [6; 9]. For our definition, we need a smooth function $s \colon \mathbb{R} \to \mathbb{R}$ such that $s(t) = 0$ if $t \leq 0$ and $s(t) > 0$ if $t > 0$.

Definition 3.1 Let M^{2n} be a smooth manifold.

(1) A subset $\mathcal{B} \subset M$ is an *n–dimensional branched submanifold* if for every $p \in \mathcal{B}$, there exists a chart $\phi_p \colon U_p \xrightarrow{\sim} \mathbb{R}^{2n}$ about p such that $\phi_p(p) = 0$ and $\phi_p(\mathcal{B} \cap U_p)$ is a union of submanifolds $L_0, L_1, \ldots, L_k$ for some $k \in \{0, \ldots, n\}$, where

(3-1) $$L_i := \{(x_1, \ldots, x_n, s(x_1), s(x_2), \ldots, s(x_i), 0, \ldots, 0) \in \mathbb{R}^{2n} \mid x_j \in \mathbb{R}\}.$$

(2) A *sector* of $\mathcal{B}$ is a connected component of the set of all points in $\mathcal{B}$ that are locally modeled by L_0, ie $k = 0$.

(3) The *branch locus* $\mathrm{Locus}(\mathcal{B})$ of $\mathcal{B}$ is the complement of all the sectors.

(4) Let (M^{2n}, ω) be a symplectic manifold. A subset $\mathcal{B} \subset M$ is a *Lagrangian branched submanifold* if for every $p \in \mathcal{B}$, there exists a Darboux chart $\phi_p \colon (U_p, \omega|_{U_p}) \xrightarrow{\sim} (\mathbb{R}^{2n}, \omega_0)$ about p, satisfying that $\phi_p(\mathcal{B} \cap U_p)$ is a union of submanifolds $L_0, L_1, \ldots, L_k$ for some $k \in \{0, \ldots, n\}$ where L_i is defined in (3-1).

Remark 3.2 (1) At every point p of a branched submanifold $\mathcal{B}$, the tangent plane $T_p\mathcal{B}$ is well defined. Moreover, if $\mathcal{B}$ is Lagrangian, then $T_p\mathcal{B}$ is a Lagrangian subspace of T_pM.

(2) A point on the branch locus is (a smooth version of) an arboreal singularity in the sense of Nadler [8].

Example 3.3 (1) Every Lagrangian submanifold L is a Lagrangian branched submanifold. The branch locus $\mathrm{Locus}(L)$ is empty.

(2) Every train track of a surface equipped with an area form is a Lagrangian branched submanifold.

(3) Let (M, ω) be a symplectic manifold and let L_1 and L_2 be two Lagrangian submanifold of M such that

$$L_1 \pitchfork L_2 = L_1 \cap L_2 = \{p\}.$$

The Lagrangian surgery of L_1 and L_2 at p will be denoted by $L_2 \#_p L_1$. Then, $L_2 \#_p L_1 \cup L_1$ and $L_2 \#_p L_1 \cup L_2$ are examples of Lagrangian branched submanifolds.

In Section 3.3, we will define the notion of "carried by" which appears in Theorems 1.3 and 6.6. In order to define the notion of carried by, we will construct a fibered neighborhood first in Section 3.2.

3.2 Construction of fibered neighborhoods

Let $\mathcal{B}$ be a Lagrangian branched submanifold. A fibered neighborhood $N(\mathcal{B})$ of $\mathcal{B}$ is, roughly speaking, a codimension zero compact submanifold with boundary and corners of M, which is foliated by Lagrangian closed disks which are called *fibers*.

Definition 3.4 A *fibered neighborhood of* $\mathcal{B}$ is a union $\bigcup_{p \in \mathcal{B}} F_p$, where $\{F_p \mid p \in \mathcal{B}\}$ is a family of Lagrangian disks which are called *fibers* satisfying

(1) for any $p \in \mathcal{B}$, $F_p \pitchfork \mathcal{B}$,

(2) for any $p, q \in \mathcal{B}$, either $F_p = F_q$ or $F_p \cap F_q = \varnothing$,

(3) there exists a closed neighborhood $U \subset \mathcal{B}$ of $\mathrm{Locus}(\mathcal{B})$, such that $\{F_p \mid p \in U\}$ is a smooth family over each local sheet $L_i \cap U$,

(4) for each sector S of $\mathcal{B}$, $\{F_p \mid p \in S \setminus U\}$ is a smooth family,

(5) if $p \in S \cap \partial U$ where S is a sector of $\mathcal{B}$, then, for any sequence $\{q_n \in S \setminus U\}_{n \in \mathbb{N}}$, $\lim_{n \to \infty} F_{q_n}$ is a Lagrangian disk such that $\lim_{n \to \infty} F_{q_n} \subset F_p^\circ = F_p \setminus \partial F_p$.

Example 3.5 Let M be a symplectic manifold and let L be a Lagrangian submanifold of M. Then L is a Lagrangian branched submanifold of M. By the Lagrangian neighborhood theorem [16], for any Lagrangian submanifold L of M, there exists a small neighborhood $\mathcal{N}(L)$ of the zero section of T^*L such that a symplectic embedding $i_L \colon \mathcal{N}(L) \hookrightarrow M$ is defined on $\mathcal{N}(L)$. Without loss of generality, we assume that $\mathcal{N}(L)$ is a closed neighborhood. Then $\mathcal{N}(L)$ is foliated by closed Lagrangian disks $\mathcal{N}(L) \cap T_p^*L$. Thus, $\mathcal{N}(L)$ is a fibered neighborhood of L.

We will now give a specific construction of a fibered neighborhood $N(\mathcal{B})$. The rough sketch of the construction is as follows. If $p \in \mathcal{B}$ lies on a sector S of $\mathcal{B}$, by Example 3.5, there is a natural embedding $i_S \colon \mathcal{N}(S) \hookrightarrow M$. Then $i_s(\mathcal{N}(S) \cap T^*pS) \pitchfork \mathcal{B}$. Thus, it is natural to set $F_p := i_s(\mathcal{N}(S) \cap T^*pS) \pitchfork \mathcal{B}$. However, if one sets as above, the odds are that there are $p, q \in \mathcal{B} \setminus \mathrm{Locus}(\mathcal{B})$ near $\mathrm{Locus}(\mathcal{B})$ such that $F_p \cap F_q \neq \varnothing$, but $F_p \neq F_q$. See Figure 3 representing the case of $\dim M = 2$.

To handle this issue, we classify $p \in \mathcal{B}$ into three cases: "near the branch locus", "far from the branch locus", and "between the other two". Then, we construct a fiber F_p for p in each case.

Fibrations over near the branch locus First, we will construct fibers near the branch locus. For each connected component ℓ of $\mathrm{Locus}(\mathcal{B})$, we choose a small closed Lagrangian neighborhood L_ℓ of ℓ satisfying the following. Fix a Riemannian metric g or an almost complex structure J compatible with ω. Then, one can define a normal bundle for every Lagrangian submanifold. We choose any Lagrangian L_ℓ containing ℓ such that for any $x \in \ell$, $(T_x \phi_x^{-1}(L_i))^\perp \pitchfork T_x L_\ell$ for all i. Note that ϕ_x and L_i appeared in Definition 3.1.
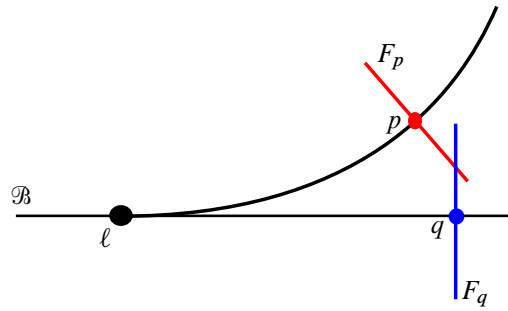
Figure 3: Black curves are part of a Lagrangian branched submanifold $\mathscr{B}$, the black point is a connected component ℓ of Locus($\mathscr{B}$), the red and blue points are $p, q \in \mathscr{B}$, and the red and blue lines are F_p and F_q.

Then, by Example 3.5, there exists a symplectic embedding

$$i_{L_\ell} : \mathcal{N}(L_\ell) \hookrightarrow M.$$

Let $U(\ell) = i_{L_\ell}(\mathcal{N}(\mathscr{L}_\ell))$.

Without loss of generality, we can choose a sufficiently small L_ℓ such that

$$i_{L_\ell}(\mathcal{N}(L_\ell) \cap T_x^* L_\ell) \cap \mathscr{B} \neq \varnothing \quad \text{for all } x \in L_\ell,$$
$$i_{L_\ell}(\mathcal{N}(L_\ell) \cap T_x^* L_\ell) \pitchfork \mathscr{B} \qquad \text{for all } x \in L_\ell,$$
$$U(\ell) \cap U(\ell') = \varnothing \qquad \text{if } \ell \neq \ell'.$$

If $p \in \mathscr{B}$ is "close" to the branch locus, ie there is a connected component ℓ of Locus($\mathscr{B}$) such that $p \in \mathscr{B} \cap U(\ell)$, then there exists $x \in L_\ell$ such that $p \in i_{L_\ell}(\mathcal{N}(L_\ell) \cap T_x^* L_\ell)$. Let $F_p := i_{L_\ell}(\mathcal{N}(L_\ell) \cap T_x^* L_\ell)$. Then F_p is a closed Lagrangian disk containing p.

By choosing a sufficiently small L_ℓ, for every $p \in \mathscr{B} \cap U(\ell)$,

(3-2) $$F_p \pitchfork \mathscr{B} \quad \text{and} \quad \partial F_p \cap \mathscr{B} = \varnothing.$$

After possibly renaming $U(\ell)$, from now we assume that

$$U(\ell) = \bigcup_{p \in L_\ell} F_p.$$

If $p \in \mathscr{B} \cap U(\ell)$, then there is a unique $q \in L_\ell$ such that $p \in F_q$. We define $F_p := F_q$. Thus, for $p \in \mathscr{B}$ which is close to Locus($\mathscr{B}$), ie $p \in U(\ell)$ for some connected component ℓ of Locus($\mathscr{B}$), we can define a fiber F_p at p.

Fibrations far from the branch locus   If $p \in \mathscr{B} \setminus \bigcup_\ell U(\ell)$, then there is a sector S of $\mathscr{B}$ containing p. Since S is Lagrangian, there is an embedding $i_S : \mathcal{N}(S) \hookrightarrow M$. We can assume $\mathcal{N}(S)$ is small enough that

$$F_q \cap i_S(\mathcal{N}(S)) \subset F_q^\circ = F_q \setminus \partial F_q \quad \text{for any } q \in \mathscr{B} \cap U(\ell),$$
$$\big(i_S(\mathcal{N}(S)) \setminus \cup U(\ell)\big) \cap \big(i_{S'}(\mathcal{N}(S')) \setminus \cup U(\ell)\big) = \varnothing \qquad \text{if } S \neq S'.$$

Figure 4, bottom right, represents examples of $\mathcal{N}(S)$.
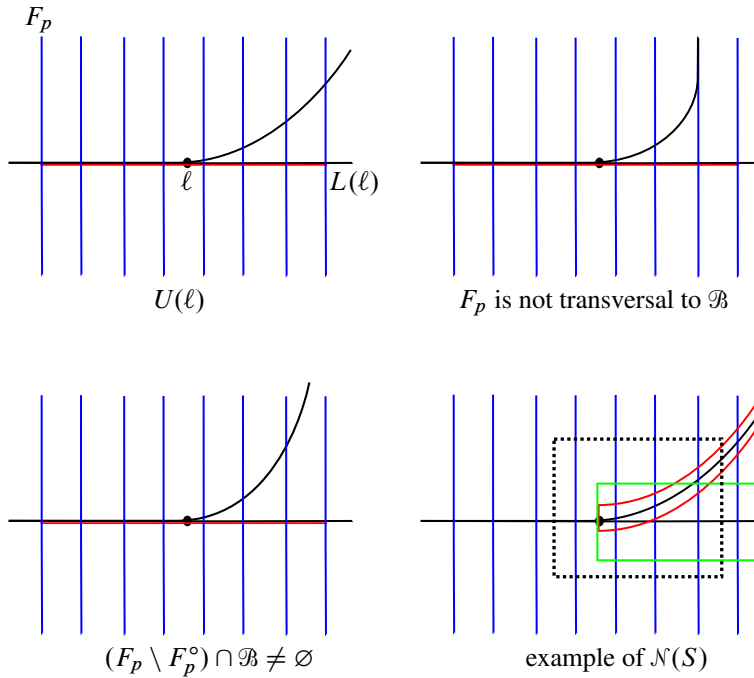
Figure 4: Black curves are part of a Lagrangian branched submanifold and the black marked points denote a connected component ℓ of $\mathrm{Locus}(\mathcal{B})$. In the top left, L_ℓ is in red, and the fibers F_p, for $p \in \mathcal{B} \cap U(\ell)$, are in blue; the top right and bottom left are not allowed by (3-2); and in the bottom right, the red and green boxes are examples of $\mathcal{N}(S)$ and the dotted box is an example of $U(\ell)$.

For any sector S, $S \setminus \bigcup_\ell \mathrm{Int}\, U(\ell)$ is a Lagrangian submanifold with boundary. The boundary of $S \setminus \bigcup_\ell \mathrm{Int}\, U(\ell)$ is a union of $S(\ell) := S \cap \partial(U(\ell))$. We fix a tubular neighborhood of $S(\ell)$, which is contained in $S \setminus \bigcup_\ell \mathrm{Int}\, U(\ell)$, and identify the tubular neighborhood with $S(\ell) \times [0, 1)$. For convenience, we will pretend that $S(\ell) \times [0, 1] \subset S$ and $S(\ell) \times \{0\} = S(\ell)$.

If $p \in S \setminus \bigcup_\ell \mathrm{Int}\, U(\ell)$ does not lie in any $S(\ell) \times (0, 1)$, then we set $F_p := i_S(\mathcal{N}(S) \cap T_p^* S)$. See Figure 5, top right.

Interpolation on $S(\ell) \times [0, 1]$ Let $p \in S(\ell)$. Then $F_{(p,0)}$ and $F_{(p,1)}$ are already constructed. We will construct $F_{(p,t)}$ from $F_{(p,0)}$ and $F_{(p,1)}$. The idea is to understand $F_{(p,0)}$ as a deformed $F_{(p,1)}$. In order to measure how much deformed $F_{(p,0)}$ is from $F_{(p,1)}$, we will construct a family of Lagrangian discs $B_{(p,t)}$ for all $t \in [0, 1]$, which are parallel to $F_{(p,1)}$. The family $B_{(p,t)}$ is defined by setting

$$B_{(p,t)} := i_S(\mathcal{N}(S) \cap T_{(p,t)}^* S).$$

We note that $B_{(p,t)}$ is parallel to $B_{(p,1)} = F_{(p,1)}$ so that there is a natural bijection map between $B_{(p,t)}$ and $B_{(p,1)}$.
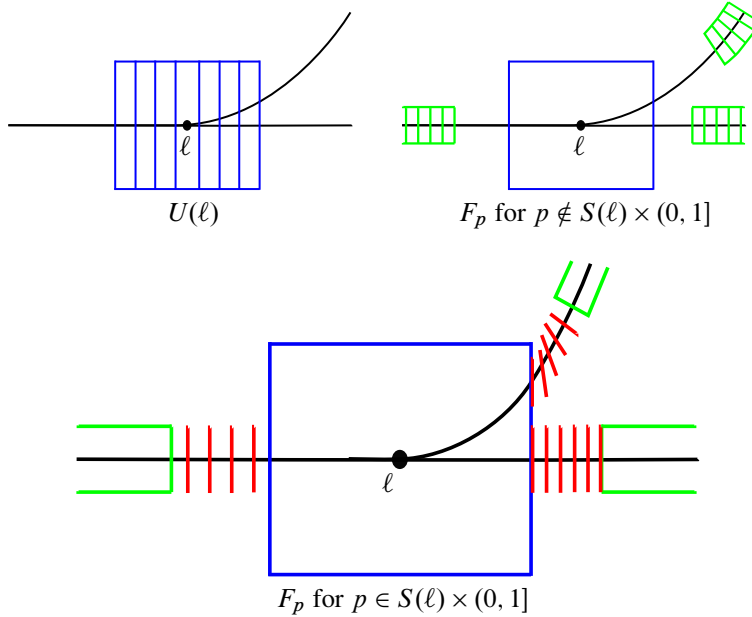
Figure 5: Black curves are part of a Lagrangian branched submanifold and marked points denote ℓ; in the top left, $U(\ell)$ is shaded blue, the vertical line segments are fibers. in the top right, the fiber F_p for $p \notin S(\ell) \times (0, 1]$ is in green; and in the bottom, the fiber F_p for $p \in S(\ell) \times (0, 1]$ is in red.

By applying the Lagrangian neighborhood theorem [16] to $B_{(p,0)}$,

$$F_{(p,0)} \cap i_S(\mathcal{N}(S)) = i_{B_{(p,0)}}(\text{the graph of a closed section in } T^* B_{(p,0)}).$$

Every closed section of $T^* B_{(p,0)}$ is an exact section because $B_{(p,0)}$ is a disk. Thus, there is a function $f_{(p,0)} : B_{(p,0)} \to \mathbb{R}$ such that

$$F_{(p,0)} \cap i_S(\mathcal{N}(S)) = i_{B_{(p,0)}}(\text{the graph of } df_{(p,0)}).$$

In other words, $F_{(p,0)}$ is obtained by deforming $B_{(p,0)}$. The deformation can be understood by using $f_{(p,0)}$.

Similarly, we will construct $F_{(p,t)}$ by deforming $B_{(p,t)}$. In order to deform, we define a function $f_{(p,t)} : B_{(p,t)} \to \mathbb{R}$ as

$$f_{(p,t)} : B_{(p,t)} \xrightarrow{\sim} B_{(p,0)} \xrightarrow{(1-t) f_{(p,0)}} \mathbb{R}.$$

The first arrow comes from the bijection between them. Then we set

$$F_{(p,t)} := i_{B_{(p,t)}}(\text{the graph of } df_{(p,t)}).$$

A fibered neighborhood $N(\mathcal{B})$ is given by the union of fibers, ie $N(\mathcal{B}) = \bigcup_{p \in \mathcal{B}} F_p$. Note that the construction of $N(\mathcal{B})$ is not unique.

3.3 Associated branched manifolds and the notion of "carried by"

We constructed a fibered neighborhood $N(\mathcal{B})$. In order to define what it means for a Lagrangian to be *carried by* $\mathcal{B}$, we introduce a projection map from $N(\mathcal{B})$ to an associated branched manifold $\mathcal{B}^*$.

Definition 3.6 Let $\mathcal{B}$ be a Lagrangian branched submanifold of M and let $N(\mathcal{B})$ be a fibered neighborhood of $\mathcal{B}$. Then, the *associated branched submanifold* $\mathcal{B}^*$ is defined by setting

$$\mathcal{B}^* := N(\mathcal{B})/\sim, \quad x \sim y \text{ if there exists an } F_p \text{ such that } x, y \in F_p.$$

Let $\pi \colon N(\mathcal{B}) \to \mathcal{B}^*$ denote the quotient map. We would like to remark that $\pi|_{\mathcal{B}}$ is not bijective, but $\mathcal{B}$ and $\mathcal{B}^*$ are equivalent as branched manifolds. We explain this with more detail in Remark 3.8.

We note that $\mathcal{B}^*$ is not contained in M. However, since $\mathcal{B}^*$ is a branched manifold, we can define the branch locus and sectors of $\mathcal{B}^*$ as follows:

Definition 3.7 (1) A *sector* of $\mathcal{B}^*$ is a connected component of

$$\{p \in \mathcal{B}^* \mid p \text{ has a neighborhood which is homeomorphic to } \mathbb{R}^n\}.$$

(2) A *branch locus* of $\mathcal{B}^*$ is the complement of all the sectors.

Remark 3.8 (1) Fibered neighborhoods $N(\mathcal{B})$ of $\mathcal{B}$ are not unique. However, if $N(\mathcal{B})$ is small enough, then $\mathcal{B}$ and $\mathcal{B}^*$ are equivalent as branched manifolds. For the equivalence between branched manifolds, we refer to Williams [17]. One can easily check their equivalence by using the Darboux chart that appeared in Definition 3.1. Thus, $\mathcal{B}^*$ is unique as a branched manifold under the assumption that $N(\mathcal{B})$ is small enough.

In the rest of this paper, when it comes to a Lagrangian branched submanifold $\mathcal{B}$, we will consider a triple $(\mathcal{B}, N(\mathcal{B}), \mathcal{B}^*)$ with an arbitrary choice of $N(\mathcal{B})$. Moreover, for any triple $(\mathcal{B}, N(\mathcal{B}), \mathcal{B}^*)$, the projection map is denoted by π for convenience.

(2) A fibered neighborhood $N(\mathcal{B})$ is a union of fibers, ie $N(\mathcal{B}) = \bigcup_{p \in \mathcal{B}} F_p$. In the equation, $\mathcal{B}$ is an index set. However, there is a possibility of having two distinct points $p, q \in \mathcal{B}$ such that $F_p = F_q$. From now on, we will use $\mathcal{B}^*$ as an index set and, by abuse of notation, F_x denotes $\pi^{-1}(x)$ for all $x \in \mathcal{B}^*$.

(3) Let x be a branch point of $\mathcal{B}^*$. Then there are sectors $S_0, S_1, \ldots, S_l$ of $\mathcal{B}^*$ for some $l \geq 2$ such that

$$x \in \overline{S}_i \text{ for every } i = 0, 1, \ldots, l,$$

$$F_x \cap \overline{\pi^{-1}(S_0)} = F_x \text{ and } F_x \cap \overline{\pi^{-1}(S_i)} \subset F_x^\circ = F_x \setminus \partial F_x \text{ for every } i = 1, 2, \ldots, l.$$

Figure 6, right, represents this.

If a Lagrangian submanifold L (resp. Lagrangian branched submanifold $\mathcal{L}$) is contained in $N(\mathcal{B})$, there is a restriction of π to L (resp. $\mathcal{L}$). For convenience, we will simply use π instead of $\pi|_L \colon L \to \mathcal{B}^*$ (resp. $\pi|_{\mathcal{L}} \colon \mathcal{L} \to \mathcal{B}^*$).
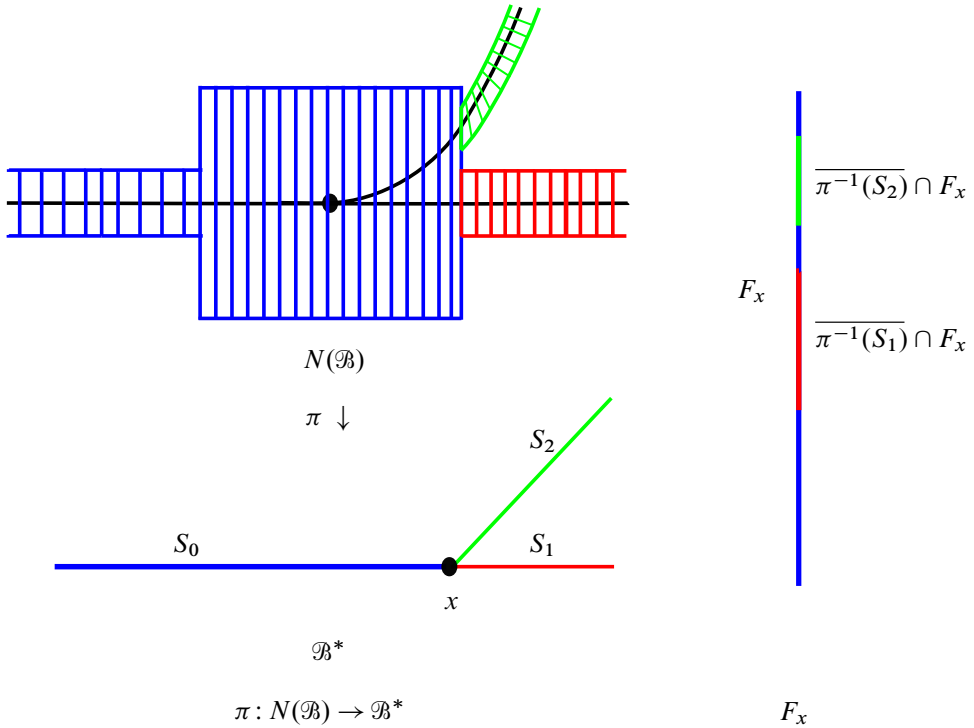
Figure 6: The left represents $\pi\colon N(\mathcal{B}) \to \mathcal{B}^*$. In $N(\mathcal{B})$, the blue, red, and green represent $\pi^{-1}(S_0)$, $\pi^{-1}(S_1)$, and $\pi^{-1}(S_2)$, where S_i is the corresponding sector of $\mathcal{B}^*$. The right represents F_x where x is in the branch locus of $\mathcal{B}^*$ to the left.

Definition 3.9 Let L be a Lagrangian submanifold (resp. $\mathcal{L}$ be a Lagrangian branched submanifold) of $N(\mathcal{B})$.

(1) A point x of L (resp. $\mathcal{L}$) is a *regular point* of π if $L \pitchfork F_{\pi(x)}$ (resp. $\mathcal{L} \pitchfork F_{\pi(x)}$) at x.

(2) A point x of L (resp. $\mathcal{L}$) is a *singular point* of π if x is not regular point of π. Moreover, $y \in \mathcal{B}^*$ is a *singular value* of π if there is a singular point x of π such that $\pi(x) = y$.

(3) L is *minimally singular with respect to* $\mathcal{B}$ if $\pi\colon L \to \mathcal{B}^*$ has no singular value on the branch locus of $\mathcal{B}^*$ and $|F_x \cap L| = |F_y \cap L|$, for any nonsingular value x and y which lie in the same sector of $\mathcal{B}^*$, where $|\cdot|$ means the cardinality of a set.

We recall that by definition, branched manifolds have tangent spaces even along the branch locus, so Definition 3.9 makes sense.

Definition 3.10 Let $\mathcal{B}$ and $\mathcal{L}$ be branched Lagrangian submanifolds.

(1) $\mathcal{L}$ is *strongly carried by* $\mathcal{B}$ if L (resp. $\mathcal{L}$) is Hamiltonian isotopic to $\mathcal{L}'$ such that $\mathcal{L}' \subset N(\mathcal{B})$ and $\pi\colon \mathcal{L}' \to \mathcal{B}^*$ has no singular value.

(2) $\mathscr{L}$ is *weakly carried by* $\mathscr{B}$ if $\mathscr{L}$ is Hamiltonian isotopic to $\mathscr{L}'$ such that $\mathscr{L}' \subset N(\mathscr{B})$, $\mathscr{L}'$ is minimally singular, and $\pi \colon \mathscr{L}' \to \mathscr{B}^*$ has countably many singular values.

We would like to remark that Lagrangian submanifolds are branched Lagrangian submanifold with empty branch locus. In the rest of this paper, if L is weakly carried by $\mathscr{B}$, then we will assume that $L \subset N(\mathscr{B})$ and L is minimally singular with respect to $\mathscr{B}$.

Note that the notion of "carried by" used by Thurston in [14] is our notion of "strongly carried by". For the case of surfaces, singularities of π can be easily resolved. However, for the case of higher-dimensional symplectic manifolds, there exists singularities which cannot be resolved. Thus, we defined the notion of "weakly carried by". We will give more detail in Section 3.4 with examples.

Thurston showed that for a pseudo-Anosov surface automorphism $\psi \colon S \xrightarrow{\sim} S$, there is a 1–dimensional branched submanifold τ which is called a train track such that $\psi(\tau)$ is strongly carried by τ. Our higher-dimensional generalization is slightly weaker, ie for some symplectic automorphism $\psi \colon (M, \omega) \xrightarrow{\sim} (M, \omega)$, we construct a Lagrangian branched submanifold $\mathscr{B}_\psi$ such that $\psi(\mathscr{B}_\psi)$ is weakly carried by $\mathscr{B}_\psi$. In other words, we allow nontransversality at countably many point $p \in \mathscr{B}_\psi$. However, we allow only one type of nontransversality. In the rest of the present subsection, we will describe the unique type of nontransversality.

Definition 3.11 Let L be weakly carried by $\mathscr{B}$. A *singular component* V of $\pi \colon L \to \mathscr{B}$ is a connected component of the set of all singular points of π.

Example 3.12 Let M be the symplectic manifold $T^*\mathbb{R}^n \simeq \mathbb{R}^{2n}$ equipped with the canonical symplectic form. The zero section $\mathscr{Z} := \mathbb{R}^n \times 0 \subset \mathbb{R}^{2n}$ is a Lagrangian branched submanifold. The fibered neighborhood $N(\mathscr{Z})$ is M with fibers $F_p := T_p^*\mathbb{R}^n$ for all $p \in \mathbb{R}^n = \mathscr{Z}$. Then, a Lagrangian submanifold

$$L_* := \{(tx, x) \in \mathbb{R}^n \times \mathbb{R}^n \mid t \in \mathbb{R}, x \in S^{n-1} \subset \mathbb{R}^n\}$$

is weakly carried by $\mathscr{Z}$, and π_* has only one singular component

$$V_* := \{(0, x) \mid x \in S^{n-1}\},$$

where π_* is the projection map.

In order to understand the singularity, we would like to restrict π_* on L_*. By definition L_* is $\mathbb{R} \times S^{n-1}$, and the restriction is the map described as follows. First, the map collapses the center sphere $\{0\} \times S^{n-1}$ to a point and get two cones of S^{n-1} glued at the vertex. Then, second, the map projects each cone of S^{n-1} to a disk $\mathbb{D}^n$. Figure 7 describes the case of $n = 2$.

Definition 3.13 A singular component V of $\pi \colon L \to \mathscr{B}$ is of *real blow-up type* if there exists an open neighborhood U of V and a symplectomorphism $\phi \colon U \xrightarrow{\sim} \mathbb{R}^{2n}$ such that $\phi(U \cap \mathscr{B}) = \mathscr{Z}, \phi(V) = V_*$, and $\phi^{-1} \circ \pi_* \circ \phi = \pi$, where $\mathscr{Z}$, V_*, and π_* are defined in Example 3.12.
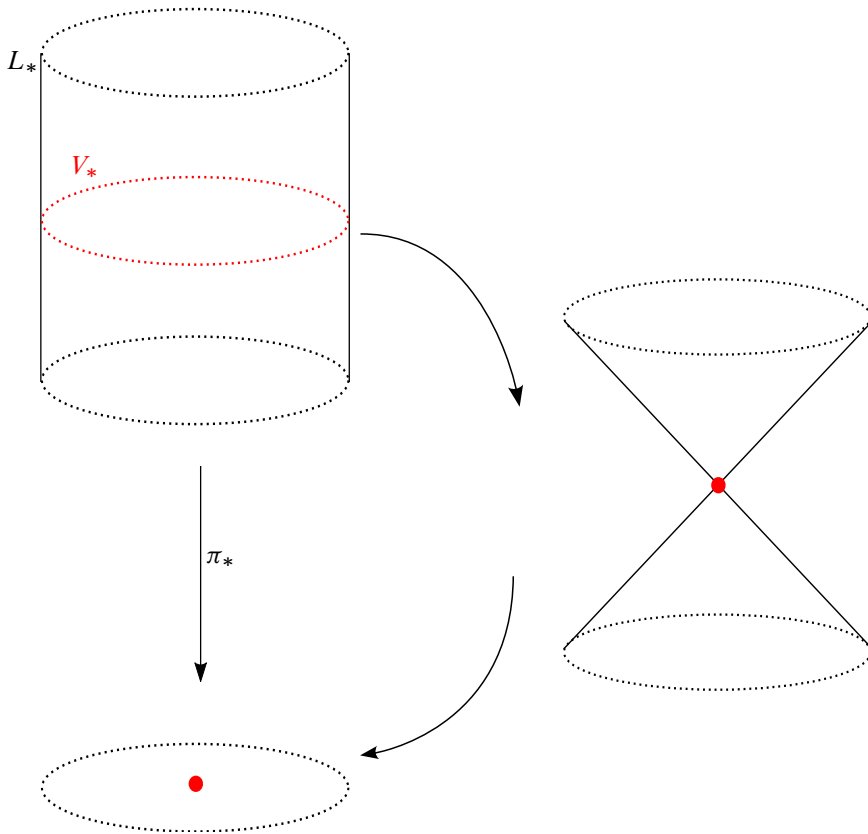
Figure 7: In the upper left, the Lagrangian L_* is shaded in black and the set of singular points V_* is shaded in red; the vertical arrow means the projection map π_*. In the lower left, the disk is a target of π_*; the red marked point is the singular value. In the middle right, the picture describes two cones glued at the vertex(red marked point), which is obtained by collapsing V_*.

Definition 3.14 A Lagrangian submanifold L (resp. Lagrangian branched submanifold $\mathcal{L}$) is *carried by* a Lagrangian branched submanifold $\mathcal{B}$ if L (resp. $\mathcal{L}$) is weakly carried by $\mathcal{B}$ and every singular component of π is a singular component of real blow-up type.

3.4 Examples of "weakly carried by"

In Section 3.4, we will give three examples of Lagrangians which are weakly carried by Lagrangian branched submanifolds. The first example is the lowest dimensional example, ie a 1–dimensional Lagrangian in a 2–dimensional symplectic manifold. The second example is a Lagrangian torus in T^*S^2. We will introduce these two examples in order to help the reader's understanding on the notion of "weakly carried by". The third example is a Lagrangian sphere in an A_3–surface singularity. With the example, we will explain why singular components occur naturally by iterating Dehn twists, which we will consider in the present paper.
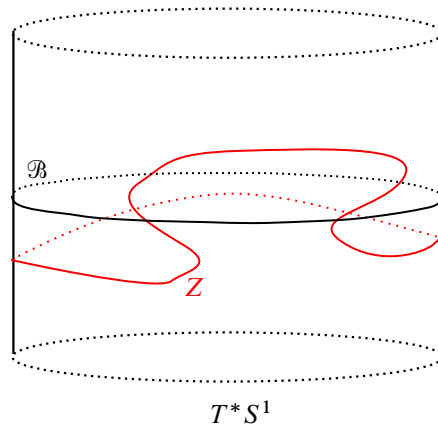
Figure 8: T^*S^1 together with the zero section $\mathcal{B}$ (black) and a Lagrangian Z (red) Hamiltonian isotopic to $\mathcal{B}$.

An example in T^*S^1   We consider the cotangent bundle of S^1. Let $\mathcal{B}$ denote the zero section of T^*S^1. Figure 8 describes T^*S^1 and $\mathcal{B}$. Let Z denote the red curve in Figure 8. Then Z is a Lagrangian which is Hamiltonian isotopic to $\mathcal{B}$.

By restricting a cotangent bundle map π on Z, Z is weakly carried by $\mathcal{B}$. However, by Hamiltonian isotoping Z, one obtains $\mathcal{B}$ and one can resolve the singularities of $\pi\colon Z \to \mathcal{B}$. In other words, Z is strongly carried by $\mathcal{B}$.

In [14], Thurston proved that on a surface, if a Lagrangian L is carried by a branched submanifold $\mathcal{B}$, then by isotoping L, one can resolve the singularities. Thus, Thurston used the notion of "carried by" without defining the notion of "weakly carried by" and his notion of "carried by" is the same to the notion of "strongly carried by".

Remark 3.15   Thurston resolved the singularities by isotoping, not Hamiltonian isotoping. Thus, for a 1–dimensional Lagrangian L which is weakly carried by a branched submanifold $\mathcal{B}$, it is possible that one cannot resolve the singularities of $\pi\colon L \to \mathcal{B}$, ie L is not strongly carried by. However, we do not discuss the existence of such examples in the current paper.

A torus in T^*S^2   We will introduce an example of a torus T in T^*S^2 such that T is weakly carried by, but not strongly carried by, the zero section $\mathcal{B}$. In order to describe the example, let assume that

$$T^*S^2 = \left\{ (x_1, x_2, x_3, y_1, y_2, y_3) \in \mathbb{R}^6 \,\Big|\, \sum_{i=1}^{3} x_i^2 = 1,\ \sum_{i=1}^{3} x_i y_i = 0 \right\} \subset \mathbb{R}^6 \simeq T^*\mathbb{R}^3.$$

Then it is easy to check that $\omega|_{T^*S^2}$ is a symplectic form on T^*S^2, where $\omega = \sum_{i=1}^{3} dx_i \wedge dy_i$.

Let T be given by

$$T = \{ (\cos\theta(0,0,1) + \sin\theta(\cos\phi, \sin\phi, 0), -\sin\theta(0,0,1) + \cos\theta(\cos\phi, \sin\phi, 0)) \mid \theta, \phi \in \mathbb{R} \}.$$

Then it is easy to check that T is a Lagrangian submanifold of T^*S^2. By restricting the cotangent bundle map π on T, T is weakly carried by $\mathcal{B}$. However, L cannot be strongly carried by $\mathcal{B}$. If L is strongly carried by $\mathcal{B}$, then L should be a covering space of $\mathcal{B}$. However, since $\mathcal{B}$ is S^2, a torus T cannot be a covering space.

This example shows the reason why we need to define the notion of "weakly carried by" in a symplectic manifold of dimension greater than or equal to 4.

Singularity arising from iterating a Dehn twist   We will give an exact Lagrangian sphere in A_3–surface singularity. By definition, A_3–surface singularity M is symplectically identified with

$$M := \{(x, y, z) \mid x^2 + y^2 + z^4 = 1\} \subset (\mathbb{C}^3, \omega_{\text{std}}).$$

We will use well-known properties of M without proof. For details, we refer the reader to Wu [18].

The first property is that M is symplectically equivalent to the plumbing of two copies of T^*S^2 at one point, ie

$$M \simeq P(\alpha, \beta).$$

We defined $P(\alpha, \beta)$ in Section 2.1. The second property of M is that M is equipped with a Weinstein Lefschetz fibration $f(x, y, z) = z$. The Lefschetz fibration has three singular points. Fibers at regular points are T^*S^1.

The Lagrangian sphere which we will consider is $\tau^2(\beta)$, where τ is a Dehn twist along α. We will encode $\tau^2(\beta)$ on the base of the Lefschetz fibration. Figure 9 describes the base of the Lefschetz fibration $f : M \to \mathbb{C}$. Then α (resp. β) is a union of vanishing cycles over a curve connecting two singular points on the base, which is shaded red (resp. blue) in Figure 9, top. Similarly, $\tau^2(\beta)$ is a union of vanishing cycles over a curve shaded green in Figure 9, top.

Let $\mathcal{B}$ be the union of vanishing cycles over a curve shaded red in Figure 9, bottom. Then $\tau^2(\beta)$ is carried by $\mathcal{B}$. The projection map from $\tau^2(\beta)$ to $\mathcal{B}$ could be drawn as arrows on the base of f; see Figure 9, bottom.

One can observe that, in Figure 9, bottom, there is a arrow from a regular point x to a singular point y. On $\tau^2(\beta)$, the point x corresponds to the vanishing cycle on $f^{-1}(x)$. The vanishing cycle is projected to a point on $f^{-1}(y)$ by $\pi : \tau^2(\beta) \to \mathcal{B}$. Moreover, one can observe that the singular component is of real blow-up type. Thus, $\tau^2(\beta)$ is carried by $\mathcal{B}$.

Remark 3.16   The last example shows that a singular component could occur when we iterate a Dehn twist. We will consider the natural occurrence in later sections.

3.5   The generalized Penner construction

In this subsection, we give a higher-dimensional generalization of Penner construction [10] of pseudo-Anosov surface automorphisms. The generalization replaces Dehn twists by generalized Dehn twists along Lagrangian spheres.
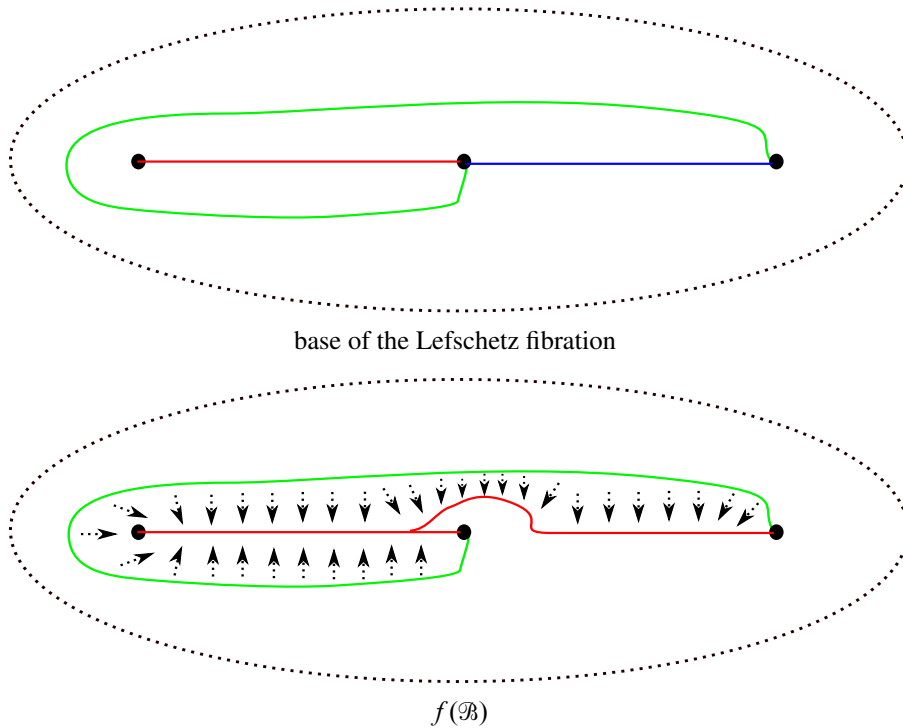
Figure 9: Top: base of the Lefschetz fibration f; the red, blue and green curves are images of α, β and $\tau^2(\beta)$. Bottom: the red curves are the image of $\mathcal{B}$ under f and the arrows are describing the projection maps from $\tau^2(\beta)$ to $\mathcal{B}$.

Generalized Penner construction Let M be a symplectic manifold. A symplectic automorphism $\psi : M \xrightarrow{\sim} M$ is of *generalized Penner type* if there are two collections,

$$A = \{\alpha_1, \ldots, \alpha_m\}, \quad B = \{\beta_1, \ldots, \beta_l\},$$

of Lagrangian spheres satisfying

$$\alpha_i \cap \alpha_j = \varnothing, \quad \beta_i \cap \beta_j = \varnothing \quad \text{for all } i \neq j,$$
$$\alpha_i \pitchfork \beta_j \qquad \qquad \text{for all } i, j$$

such that ψ is a product of positive powers of Dehn twists τ_i along α_i and negative powers of Dehn twists σ_j along β_j, subject to the condition that every sphere appear in the product.

A Lagrangian sphere α_i (resp. β_j) is called a *positive* (resp. *negative*) sphere since only positive powers of τ_i (resp. negative powers of σ_j) appear in ψ.

Remark 3.17 (1) In Theorems 1.3 and 1.5, we can assume that the symplectic manifold M is a plumbing space. Every τ_i (resp. σ_j) is supported on a neighborhood of α_i (resp. β_j), which is denoted
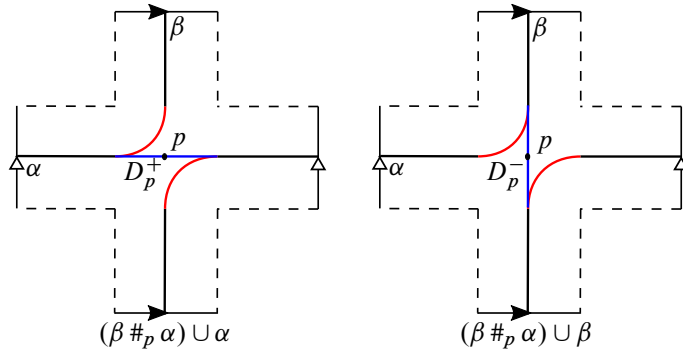
Figure 10: The blue curves represent D_p^+ in the left-hand picture and D_p^- in the right-hand picture; the red curves represent N_p in both.

by $T^*\alpha_i$ (resp. $T^*\beta_j$). Thus, ψ is supported on the union of $T^*\alpha_i$ and $T^*\beta_j$. By the transversality condition $\alpha_i \pitchfork \beta_j$, we can identify the union with a plumbing space

$$P = P(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l).$$

Thus, it suffices to prove Theorems 1.3 and 1.5 on the plumbing space P, which we take to be connected.

(2) In [10], the Penner construction required that A and B fill the surface S; ie the complement of $A \cup B$ is a union of disks and annuli, one of whose boundary components is a component of ∂S. In the current paper, we do not require the analogue of the filling condition since we only construct an invariant Lagrangian branched submanifold and an invariant Lagrangian lamination, not an invariant singular foliation on all of M.

In the rest of this subsection, we define a set of Lagrangian branched submanifolds in a plumbing space $P(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l)$. We start from the simplest plumbing space, having one positive and one negative sphere intersecting at only one point.

Example 3.18 Let α and β be n–dimensional spheres and let M be a plumbing $P(\alpha, \beta)$ which is plumbed at only one point p. Let $\beta \#_p \alpha$ be the Lagrangian surgery of α and β at p such that $\beta \#_p \alpha \simeq \tau_\alpha(\beta) \simeq \sigma_\beta^{-1}(\alpha)$. See Figure 10, which represents the case $n = 1$. The cross-shape is the plumbing space $P(\alpha, \beta)$, where α is the horizontal line and β is the vertical line.

The *neck* N_p at p connecting α and β is the closure of $(\beta \#_p \alpha) - (\alpha \cup \beta)$. In Figure 10, N_p is drawn in red. The *positive disk* D_p^+ at p is the closure of $\alpha - (\beta \#_p \alpha)$ and the *negative disk* D_p^- at p is the closure of $\beta - (\beta \#_p \alpha)$. The disks $D_p^\pm$ are drawn in blue in Figure 10. Then, by attaching D_p^+ or D_p^- to $\beta \#_p \alpha$, we obtain Lagrangian branched submanifolds $(\beta \#_p \alpha) \cup \alpha$ and $(\beta \#_p \alpha) \cup \beta$.

On a general plumbing space $M = P(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l)$ with positive spheres α_i and negative spheres β_j, we similarly construct Lagrangian branched submanifolds. More precisely, given a plumbing point p, N_p, D_p^+ and D_p^- are the closures of $(\beta_j \#_p \alpha_i) - (\alpha_i \cup \beta_j)$, $\alpha_i - (\beta_j \#_p \alpha_i)$ and $\beta_j - (\beta_j \#_p \alpha_i)$,

respectively. In order to construct a Lagrangian branched submanifold $\mathfrak{B}$, let $D_p(\mathfrak{B})$ be either D_p^+ or D_p^-. In other words, to construct $\mathfrak{B}$, one choose either D_p^+ or D_p^- for each plumbing points p. Then we construct a Lagrangian branched submanifold $\mathfrak{B}$ by setting

$$(3\text{-}3) \qquad \mathfrak{B} := \bigcup_i \left(\alpha_i - \bigcup_{p \in \alpha_i} D_p^+ \right) \cup \bigcup_j \left(\beta_j - \bigcup_{p \in \beta_j} D_p^- \right) \cup \bigcup_p N_p \cup \bigcup_p D_p(\mathfrak{B}).$$

There are 2^N possible choices of $\mathfrak{B}$, where N is the number of plumbing points. Let $\mathbb{B}$ be the set of all 2^N Lagrangian branched submanifolds constructed above.

3.6 Proof of Theorem 1.3

In this subsection, let $M = P(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l)$, let τ_i (resp. σ_j) be a Dehn twist along α_i (resp. β_j), and let ψ be of generalized Penner type.

In the rest of the paper, we assume that every Dehn twist, τ_i and σ_j, satisfies that

(1) τ_i (resp. σ_j) is supported on a small neighborhood $T^*\alpha_i$ (resp. $T^*\beta_j$) of α_i (resp. β_j);

(2) τ_i (resp. σ_j) agrees with the antipodal map on α_i (resp. β_j).

We define

$$(3\text{-}4) \qquad \begin{aligned} A_p &:= \tau_i(D_p^+), \quad B_p := \sigma_j^{-1}(D_p^-) \quad \text{if } p \in \alpha_i \cap \beta_j, \\ \alpha_i' &:= \alpha_i - \bigcup_{p \in \alpha_i} (D_p^+ \cup A_p), \quad \beta_j' := \beta_j - \bigcup_{p \in \beta_j} (D_p^- \cup B_p). \end{aligned}$$

In words, A_p (resp. B_p) is a neighborhood of an antipodal point of p in α_i (resp. β_j). We are assuming that $D_p^\pm$, A_p and B_p are sufficiently small that they are disjoint to each other.

Recall that $\mathbb{B}$ is the set of Lagrangian branched submanifolds defined in Section 3.5; see the last sentence of that subsection.

Lemma 3.19 *For all k, there exists a function $F_{\tau_k} : \mathbb{B} \to \mathbb{B}$ such that $\tau_k(\mathfrak{B})$ is carried by $F_{\tau_k}(\mathfrak{B})$ for all $\mathfrak{B} \in \mathbb{B}$. Similarly, there is a function $F_{\sigma_j^{-1}} : \mathbb{B} \to \mathbb{B}$ for all j such that $\sigma_j^{-1}(\mathfrak{B})$ is carried by $F_{\sigma_j^{-1}}(\mathfrak{B})$.*

Proof In this proof, τ_k is given by (2-2) and $\tilde{\tau} : T^*S^n \xrightarrow{\sim} T^*S^n$ defined in Section 2.2; ie $\tau_k = \phi \circ \tilde{\tau} \circ \phi^{-1}$ in a neighborhood of α_k, where ϕ is an identification of T^*S^n and a neighborhood of α_k.

Given $\mathfrak{B} \in \mathbb{B}$, $\mathfrak{B}$ admits the decomposition

$$(3\text{-}5) \qquad \mathfrak{B} = \bigcup_i \alpha_i' \cup \bigcup_j \beta_j' \cup \bigcup_p N_p \cup \bigcup_p A_p \cup \bigcup_p B_p \cup \bigcup_p D_p(\mathfrak{B}),$$

where $D_p(\mathfrak{B})$ is either D_p^+ or D_p^-. This follows from (3-3) and (3-4).

We prove the first statement for τ_k; the proof for σ_j^{-1} is analogous. Our strategy is to apply τ_k to α_i', β_j', N_p, A_p, B_p, and $D_p^\pm$. We claim:

(i) $\tau_k(\alpha_i') = \alpha_i'$ and $\tau_k(\beta_j') = \beta_j'$, and they are strongly carried by α_i' and β_j'.

(ii) If $p \notin \alpha_k$, then $\tau_k(N_p) = N_p$, $\tau_k(D_p^\pm) = D_p^\pm$, $\tau_k(A_p) = A_p$ and $\tau_k(B_p) = B_p$, and they are strongly carried by N_p, $D_p^\pm$, A_p and B_p.

(iii) If $p \in \alpha_k$, then $\tau_k(D_p^+) = A_p$, $\tau_k(A_p) = D_p^+$ and $\tau_k(B_p) = B_p$, and they are strongly carried by A_p, D_p^+ and B_p.

(iv) If $p \in \alpha_k$, then $\tau_k(D_p^-)$ and $\tau_k(N_p)$ are obtained by spinning with respect to p. Moreover, $\tau_k(D_p^-)$ is strongly carried by $N_p \cup (\alpha_k - D_p^+)$ and $\tau_k(N_p)$ is carried by $N_p \cup (\alpha_k - D_p^+)$.

By (3-5) and (i)–(iv), $\tau_k(\mathcal{B})$ is carried by $\mathcal{B}'$ such that

$$(3\text{-}6) \qquad \mathcal{B}' = \bigcup_i \alpha_i' \cup \bigcup_j \beta_j' \cup \bigcup_p N_p \cup \bigcup_p A_p \cup \bigcup_p B_p \cup \bigcup_p D_p(\mathcal{B}'),$$

where $D_P(\mathcal{B}')$ is $D_p(\mathcal{B})$ if $p \notin \alpha_k$ and D_p^+ if $p \in \alpha_k$. Then $F_{\tau_k} : \mathbb{B} \to \mathbb{B}$ is defined by $F_{\tau_k}(\mathcal{B}) = \mathcal{B}'$.

For (i), since τ_k agrees with the antipodal map on α_k, $\tau_k(\alpha_k') = \alpha_k'$ and $\tau_k(\alpha_k')$ is strongly carried by α_k'. Moreover, since τ_k is supported on $T^*\alpha_k$, α_i' does not intersect the support of τ_k for all $i \neq k$. Thus, $\tau_k(\alpha_i')$ agrees with α_i' and $\tau_k(\alpha_i')$ is strongly carried by itself. The same proof applies to $\tau_k(\beta_j')$.

Statements (ii) and (iii) are proved in the same way.

For (iv), we compute $\tau_k(D_p^-)$ and $\tau_k(N_p)$ by spinning with respect to p and ϕ. We assume

$$\phi\big(((1, 0_n), 0_{n+1})\big) = p$$

without loss of generality. Using the notation from Section 2, D_p^- and N_p are contained in $\bigcup_{y \in S^{n-1}} \phi(W_y)$. Thus,

$$(3\text{-}7) \qquad \tau_k(D_p^-) = \bigcup_{y \in S^{n-1}} (\phi \circ \tilde{\tau} \circ \phi^{-1})(D_p^- \cap \phi(W_y))$$

$$= \bigcup_{y \in S^{n-1}} \big(\phi\big(\tilde{\tau}|_{W_y}(\phi^{-1}(D_p^-) \cap W_y)\big)\big) = \bigcup_{y \in S^{n-1}} \tau_k(D_p^-) \cap \phi(W_y),$$

$$(3\text{-}8) \qquad \tau_k(N_p) = \bigcup_{y \in S^{n-1}} (\phi \circ \tilde{\tau} \circ \phi^{-1})(N_p \cap \phi(W_y))$$

$$= \bigcup_{y \in S^{n-1}} \phi\big(\tilde{\tau}|_{W_y}(\phi^{-1}(N_p) \cap W_y)\big) = \bigcup_{y \in S^{n-1}} \tau_k(N_p) \cap \phi(W_y).$$

The restriction $\tilde{\tau}|_{W_y}$ is a Dehn twist on $W_y \simeq T^*S^1$ along the zero section. Thus, we obtain Figure 11 which represents intersections $\phi(W_y) \cap D_p^-$, $\phi(W_y) \cap N_p$, $\phi(W_y) \cap \tau_k(D_p^-)$, and $\phi(W_y) \cap \tau_k(N_p)$. Equation (3-8) and Figure 11 imply that $\tau_k(N_p)$ is carried by $N_p \cup (\alpha_k - D_p^+)$ This is because in each W_y, the vertical projection has no critical values. Thus, if there is a singular value, then the singular value is created when one takes the union in (3-8). One can easily check that $\tau_k(p)$ is the only singular value when one takes the union. Similarly, $\tau_k(D_p^-)$ is strongly carried by $N_p \cup (\alpha_k - D_p^+)$.

Then (i)–(iv) and (3-5) prove that $\tau_k(\mathcal{B})$ is carried by $F_{\tau_k}(\mathcal{B})$. $\square$
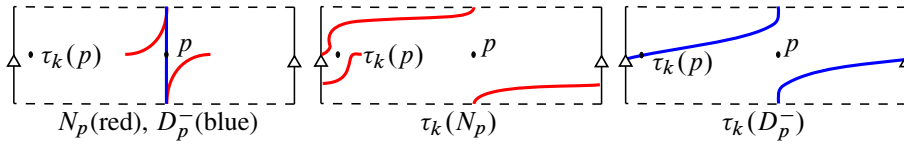
Figure 11: In the left picture, the blue curve represents D_p^- and the red curve represents N_p; in the middle picture, the red curve represents $\tau_k(N_p)$; and in the right picture, the blue curve represents $\tau_k(D_p^-)$.

Lemma 3.20 *If L is a Lagrangian submanifold which is carried (resp. weakly carried) by $\mathcal{B} \in \mathbb{B}$, then $\tau_k(L)$ is carried (resp. weakly carried) by $F_{\tau_k}(\mathcal{B})$. The case of σ_j^{-1} is analogous.*

Proof We can assume that L is contained in an arbitrary small neighborhood of $\mathcal{B}$. Then we apply a Dehn twist τ_k as we did in the proof of Lemma 3.19. The details are similar to the proof of Lemma 3.19. □

Proof of Theorem 1.3 Let $\psi: M \xrightarrow{\sim} M$ be a symplectic automorphism of generalized Penner type. Then we can write $\psi = \delta_1 \circ \cdots \circ \delta_l$, where δ_k is a Dehn twist τ_i or σ_j^{-1}. By Lemma 3.19, we have specific functions F_{τ_i} and $F_{\sigma_j^{-1}}$ acting on $\mathbb{B}$. We then define $F_\psi = F_{\delta_1} \circ \cdots \circ F_{\delta_l}: \mathbb{B} \to \mathbb{B}$.

We claim that F_ψ is a constant map, ie there is a unique $\mathcal{B}_\psi \in \mathbb{B}$ such that $F_\psi(\mathcal{B}) = \mathcal{B}_\psi$ for all $\mathcal{B} \in \mathbb{B}$, which we define as follows: in (3-3), for $p \in \alpha_i \cap \beta_j$, we set $D_p(\mathcal{B}_\psi) = D_p^+$ if the last τ_i in ψ appears later than the last σ_j^{-1}, and $D_p(\mathcal{B}_\psi) = D_p^-$ otherwise. Note that every Dehn twist τ_i and σ_j^{-1} appears in ψ; thus $\mathcal{B}_\psi$ is well defined. By (3-6), $F_\psi(\mathcal{B}) = \mathcal{B}_\psi$ for all $\mathcal{B} \in \mathbb{B}$. Lemma 3.20 completes the proof. □

Remark 3.21 (1) A singular value of $\pi: \psi^m(L) \to \mathcal{B}^*$ can be moved by isotoping $\psi^m(L)$.

(2) Every singular value of $\pi: \psi^m(\mathcal{B}_\psi) \to \mathcal{B}^*$ lies near $\pi(p)$, $\pi(\tau_i(p))$, or $\pi(\sigma_j^{-1}(p))$ by isotoping, where p is a plumbing point.

4 Encoding a Lagrangian on a Lagrangian branched submanifold

In the previous section, we generalized the notion of "carried by" for higher-dimensional symplectic manifolds. It is well known that on a surface, if a curve is carried by a train track, then one can encode the isotopy class of curve on the train track with an extra data. The extra data is called *weight*. We briefly review the notion of weight in Section 4.1, then generalize this for higher-dimensional case in Section 4.

4.1 Weights on a train track

We will briefly review the notion of weights on a train track with a simple example, and how one can construct a stable lamination of a surface automorphism of generalized Penner type from them in Section 4.1. We will introduce some well-known facts without proofs. For more detail, we refer the reader to Penner and Harer [11], or Farb and Margalit [5].
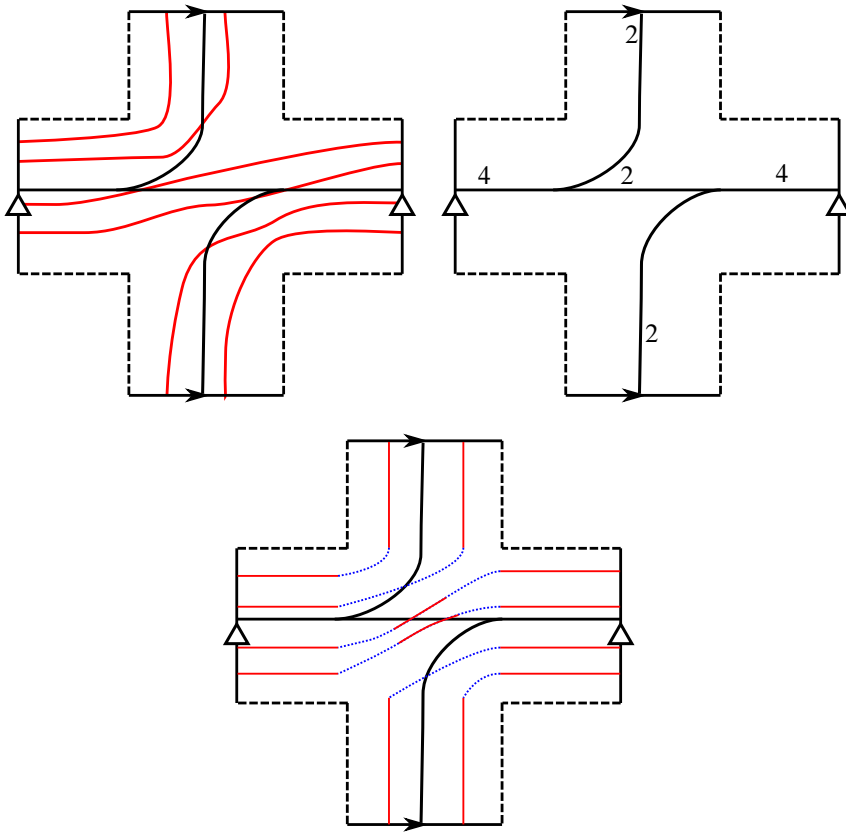
Figure 12: Top left: the cross shape is a surface S, the black graph is a Lagrangian branched submanifold $\mathcal{B}_\psi$, and the red curves are carried by $\mathcal{B}_\psi$. Top right: the numbers are the weights corresponding to the red curve in the diagram to the left. Bottom: the red curves are parallel copies of each edges and the blue dotted curves are the unique way to connecting the parallel copies.

At the end of Section 4.1, we will explain why the construction on surfaces does not work on the cases of a higher-dimensional symplectic manifold. Then, we will give a detailed organization of Section 4.

The notion of weights Let S be a surface obtained by plumbing two copies of T^*S^1 at one point. Two zero sections of each copies of T^*S^1 will be denoted by α and β as we did in previous sections. Similarly, let τ and σ denote Dehn twists along α and β respectively. We will fix a surface automorphism $\psi := \tau \circ \sigma^{-1}$. Then, by Section 3, there is a branched submanifold $\mathcal{B}_\psi$ such that if a curve $C \subset S$ is carried by $\mathcal{B}_\psi$, then $\psi(C)$ is also carried by $\mathcal{B}_\psi$. Moreover, as mentioned in Section 3.4, one can assume that there is no singular value of $\pi: C \to \mathcal{B}_\psi$ by isotoping. Figure 12, top left, describes the surface S and $\mathcal{B}_\psi$ together with an example of a curve C which is carried by $\mathcal{B}_\psi$.

Weights on a train track are collection of nonnegative numbers assigned on each edges of the train track. If a curve C is carried by a train track $\mathcal{B}$, then C gives weights on $\mathcal{B}$ by assigning the number of connected

components of $\pi^{-1}(e)$ for each edge e of $\mathcal{B}_\psi$. Figure 12, top right, is an example of weights, which are induced from the curve C drawn in Figure 12, top left.

Conversely, one can recover the isotopy class of a curve C from a train track $\mathcal{B}$ which carries C and the weights induced from C. In order to recover, one can considers parallel copies of each edges. The numbers of copies are the weights on each edges. Then, it is known that there is a unique way to connect each copies to construct an isotopy class of the curve C. Figure 12, bottom, is the example of the recovering process.

Linear algebra on weights   By Theorem 1.3, for a surface automorphism ψ of generalized Penner type, if a curve C is carried by a train track $\mathcal{B}_\psi$, then $\psi(C)$ is carried by $\mathcal{B}_\psi$. Since C and $\psi(C)$ are carried by $\mathcal{B}_\psi$, they induce weights on $\mathcal{B}_\psi$. Moreover, it is well known that the weights for $\psi(C)$ is obtained from the weights for C by doing linear algebra. We will review this with the example which we used above, ie S is the plumbing of $T^*\alpha$ and $T^*\beta$ and $\psi = \tau \circ \sigma^{-1}$.

Let C be a curve carried by $\mathcal{B}_\psi$ such that the induced weights on $\mathcal{B}_\psi$ are a, b and c, as drawn in Figure 13, top left. For simplicity, we write the weight for C in a vector

$$\vec{w}_C = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

Figure 13, top right and bottom left, are $\tau(\mathcal{B}_\psi)$ and $\psi(\mathcal{B}_\psi)$. One can observe that $\psi(\mathcal{B}_\psi)$ is carried by $\mathcal{B}_\psi$ and induces weights $3b + 2c$, $2b + c$ and a on $\mathcal{B}_\psi$. Thus, the weights for C and $\psi(C)$ satisfy

(4-1)
$$\vec{w}_{\psi(C)} = \begin{pmatrix} 0 & 3 & 2 \\ 0 & 2 & 1 \\ 1 & 0 & 0 \end{pmatrix} \cdot \vec{w}_C.$$

Remark 4.1   In (4-1), a 3×3 matrix appears. One can replace this matrix with a 2×2 matrix. Since the weight assigned on the blue edge in Figure 13 should be the same to the sum of weights assigned on the red and black edges in Figure 13. This condition is called the *switch* condition. For the detail, see Farb and Margalit [5].

Stable lamination of ψ   For a surface automorphism ψ of generalized Penner type, it is well known that the stable lamination of ψ is easily constructed from $\mathcal{B}_\psi$ and the linear algebra which we did above. For a rigorous treatment, we should define the notion of measured lamination and should explain how a measured lamination $\mathcal{L}$ can be encoded onto a pair $(\mathcal{B}, \vec{w}_\mathcal{L})$ of a train track $\mathcal{B}_\psi$ and weights $\vec{w}_\mathcal{L}$. However, for simplicity, we skip this excepts that the weight vector $\vec{w}_\mathcal{L}$ is an eigenvector of A_ψ corresponding to an eigenvalue $\lambda > 1$, where A_ψ is the matrix appearing in (4-1).

For more details including the notion of measured laminations and the existence of an eigenvalue $\lambda > 1$ of A_ψ, we refer the reader to Farb and Margalit [5].
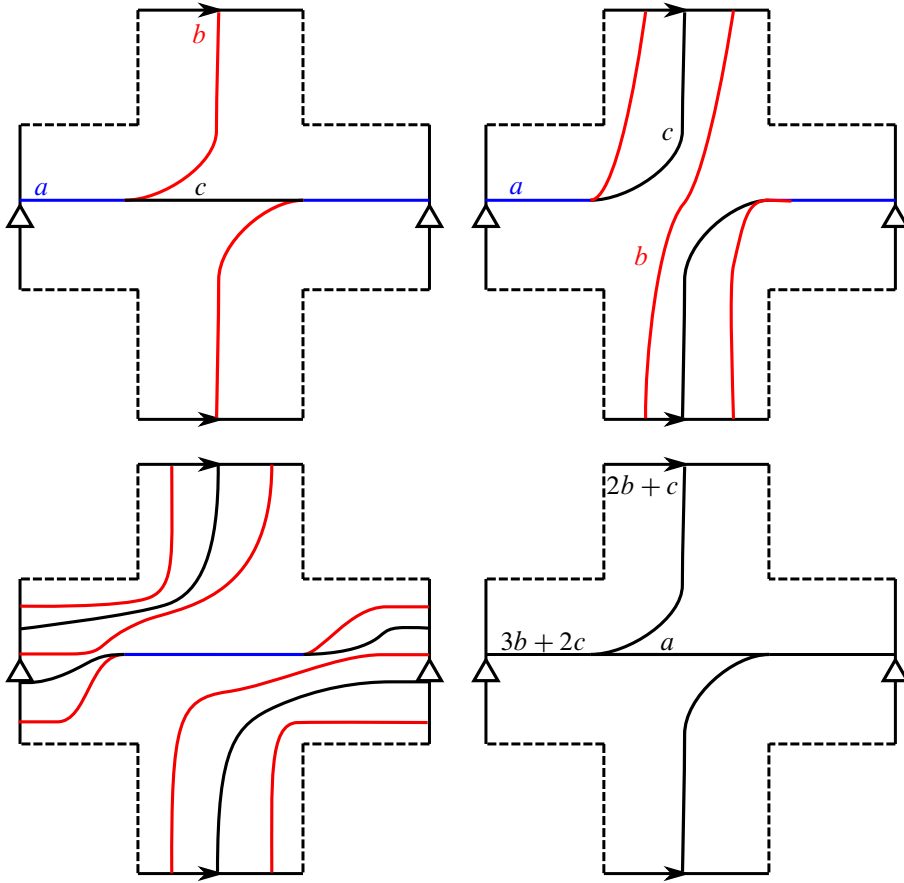
Figure 13: Top left: the Lagrangian branched submanifold $\mathcal{B}_\psi$ has three edges shaded blue, red and black, and a, b and c are weights on the edges, respectively. Top right describes $\sigma^{-1}(\mathcal{B}_\psi)$; blue, red and black edges are assigned the same weights a, b and c. Bottom left describes $\tau(\sigma^{-1}(\mathcal{B}_\psi)) = \psi(\mathcal{B}_\psi)$; blue, red and black edges are assigned the same weights a, b and c. Bottom right describes the projection of $\psi(\mathcal{B}_\psi)$ onto $\mathcal{B}_\psi$, and one finds new weights on each edge of $\mathcal{B}_\psi$.

A difficulty on higher-dimensional symplectic manifolds For a surface automorphism ψ of generalized Penner type, one can construct the stable lamination of ψ by doing some linear algebra on weights on a train track $\mathcal{B}_\psi$. This is because, in the case of a surface, the notion of carried by is the notion of strongly carried by, ie there is no singular component. However, in the case of a higher-dimensional symplectic manifold, the construction of laminations on surfaces does not work, because of singularities.

In Section 4.2, we will decompose $\mathcal{B}_\psi^*$ into a union of disks. The disks are of two types, one with singularities and one without singularities. Then, in Section 4.3, we will generalize the notion of weights. Since the generalization should have information on singularities, it will be defined by using the disks with singularities. In Section 5.1, we will generalize the linear algebra on weights. In Section 6.2, we

will construct a stable Lagrangian lamination on a disc with singularities, and in Section 6.3, we will construct on a disc without singularities.

4.2 Singular and regular disks

As mentioned in the previous subsection, the construction of laminations on surfaces does not work because of singularities. In Section 4.2, for a symplectic automorphism ψ of generalized Penner type, we decompose $\mathcal{B}_\psi$ into a union of disks. The disks are classified into two types, with and without singularities.

Definition 4.2 Let assume that there is a pair $(\psi : M \xrightarrow{\sim} M, \mathcal{B}_\psi)$ of a symplectic automorphism ψ and a Lagrangian branched submanifold $\mathcal{B}$ such that $\psi^n(\mathcal{B}_\psi)$ is carried by $\mathcal{B}_\psi$ for all $n \in \mathbb{N}$. Let $\mathcal{B}_\psi^*$ be the associated branched manifold of $\mathcal{B}_\psi$. Then the triple $(\psi, \mathcal{B}_\psi, \mathcal{B}_\psi^*)$ *admits a decomposition into singular and regular discs* if $\mathcal{B}_\psi^*$ can be decomposed into the union of a finite number of disks $S_i \simeq \mathbb{D}^n$, which are called *singular disks*, and $R_j \simeq \mathbb{D}^n$, which are called *regular disks*, ie

$$(4\text{-}2) \qquad \mathcal{B}_\psi^* = \bigcup_i S_i \cup \bigcup_j R_j$$

such that

(1) each singular disk S_i is a closed disk contained in a closure of a sector of $\mathcal{B}^*$;

(2) $S_i \cap S_j = \varnothing$ for any $i \neq j$;

(3) every singular value of $\pi : \psi^m(\mathcal{B}_\psi) \to \mathcal{B}_\psi$ after weakly fibered isotopy lies in $\bigcup_i S_i^\circ$ for all $m \in \mathbb{N}$, where S_i° is the interior of S_i;

(4) each regular disk R_j is a closed disk contained in a closure of a sector minus $\bigcup_i S_i^\circ$;

(5) S_i and R_j (resp. R_i and R_j for $i \neq j$) meet only along their boundaries.

For convenience, we simply say that $\mathcal{B}_\psi^*$, instead of a triple $(\psi, \mathcal{B}_\psi, \mathcal{B}_\psi^*)$, admits a decomposition into singular and regular discs.

Definition 4.3 Let a triple $(\psi, \mathcal{B}_\psi, \mathcal{B}_\psi^*)$ admit a decomposition into singular and regular discs. A Lagrangian L which is carried by $\mathcal{B}_\psi$ is *compatible with the decomposition* if L is Hamiltonian isotopic to L' such that every singular value of $\pi : L' \to \mathcal{B}_\psi$ lies on a singular disc.

Remark 4.4 In Section 3, we used a decomposition of $\mathcal{B}$ with notation $D_p^\pm$, A_p, B_p and so on. However, the decomposition introduced in Definition 4.3 is a decomposition of the associated branched manifold $\mathcal{B}^*$, not $\mathcal{B}$.

In the rest of Section 4.2, we will introduce and use a specific decomposition of $\mathcal{B}_\psi^*$ for ψ of generalized Penner type. Since the specific decomposition of $\mathcal{B}_\psi^*$, together with the decomposition of $\mathcal{B}_\psi$ in (3-5), is likely to confuse the reader, we remark that here.

If $\mathscr{B}_\psi^*$ admits a decomposition into singular and regular discs, then one obtains a decomposition of $N(\mathscr{B}_\psi)$ as

$$N(\mathscr{B}_\psi) = \bigcup_i \pi^{-1}(S_i) \cup \bigcup_j \pi^{-1}(R_j).$$

Remark 4.5 In Section 4.3 (resp. Section 6.3), we will construct a Lagrangian lamination on $\overline{\pi^{-1}(S_i^\circ)}$ (resp. $\overline{\pi^{-1}(R_j^\circ)}$) which is the closure of $\pi^{-1}(S_i^\circ)$, not on $\pi^{-1}(S_i)$ (resp. $\subset \pi^{-1}(R_j)$). This is because $\pi^{-1}(S_i)$ (resp. $\pi^{-1}(R_j)$) is not a (closed) submanifold of M if S_i (resp. R_j) intersects the branch locus of $\mathscr{B}^*$.

Figure 6 is an example. If S_1 in Figure 6 is a singular disk, then $\pi^{-1}(S_1)$ is the union of the red box in Figure 6, left, and F_x.

Decomposition of $\mathscr{B}_\psi^*$ for ψ of generalized Penner type   Let us assume that a symplectic automorphism $\psi \colon M \xrightarrow{\sim} M$ is of generalized Penner type. Then, in Section 3, we constructed a Lagrangian branched submanifold $\mathscr{B}_\psi$. We will now give a specific decomposition of $\mathscr{B}_\psi^*$ into singular and regular discs, which we will call *the standard decomposition of $\mathscr{B}_\psi^*$*.

By Remark 3.21, after weakly fiber isotoping, every singular value of $\pi \colon \psi^m(\mathscr{B}_\psi) \to \mathscr{B}_\psi^*$ lies in the interior of $S_p(\mathscr{B}_\psi)$ or $S_p^\pm$, where $S_p(\mathscr{B}_\psi) := \pi(D_p(\mathscr{B}_\psi))$, $S_p^+ := \pi(A_p)$ and $S_p^- := \pi(B_P)$. We note that as the notation suggests, $S_p(\mathscr{B})$ depends on $\mathscr{B}$, but $S_p^\pm$ does not. In the specific decomposition, $S_p(\mathscr{B}_\psi)$ and $S_p^\pm$ are singular disks of $\mathscr{B}_\psi^*$ and there is no other singular discs.

Remark 4.6 As mentioned in Remark 4.4, $S_p(\mathscr{B}_\psi)$ and $S_p^\pm$ are subsets of $\mathscr{B}_\psi^*$, not $\mathscr{B}_\psi$. However, in the rest of the current paper, if there is no chance of misunderstanding, we will abuse notation and will identify the singular disks with $D_p(\mathscr{B}_\psi)$, A_p and B_p. This is for notational convenience.

We will divide the complement of singular disks from $\mathscr{B}_\psi^*$, ie

(4-3)
$$\mathscr{B}_\psi^* \setminus \left(\bigcup_p S_p(\mathscr{B}_\psi) \sqcup \bigcup_p S_p^+ \sqcup \bigcup_p S_p^- \right),$$

into regular disks. In order to do this, we cut out a symplectic submanifold $W^{2n-2} \subset M^{2n}$, which is defined as follows: for each α_i (resp. β_j), there is an equator C_{α_i} (resp. C_{β_j}) $\simeq S^{n-1}$ such that

(1)   for any plumbing point $p \in \alpha_i$ (resp. β_j), p lies on C_{α_i} (resp. C_{β_j});

(2)   if $p \in \alpha_i \cap \beta_j$, then $T^*C_{\alpha_i} \equiv T^*C_{\beta_j}$ near p.

Note that the equators on Lagrangian spheres α_i and β_j are defined using identifications $\phi_{\alpha_i} \colon \alpha_i \xrightarrow{\sim} S^n$ and $\phi_{\beta_j} \colon \beta_j \xrightarrow{\sim} S^n$. Thus, by choosing proper identification ϕ_{α_i} and ϕ_{β_j}, we can assume the existence of C_{α_i} and C_{β_j}. Then

$$W := \bigcup_i T^*C_{\alpha_i} \cup \bigcup_j T^*C_{\beta_j}$$

is a $(2n-2)$–dimensional symplectic submanifold of M.
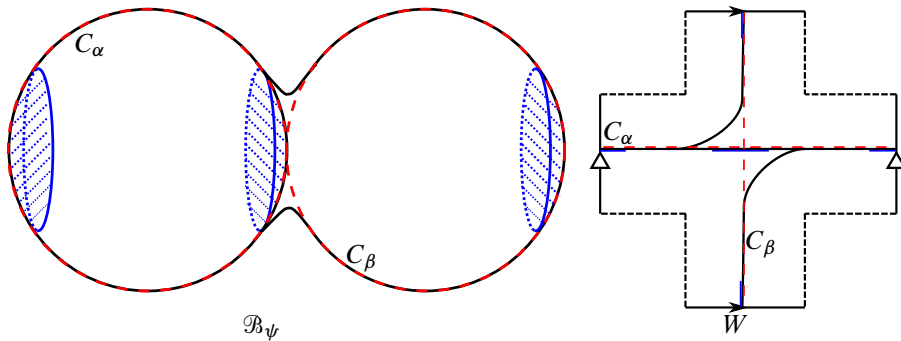
Figure 14: Left: the black curves represent $\mathcal{B}_\psi$ and the red dotted circles are C_α (left) and C_β (right). The blue shaded regions are singular disks. Right: W as a symplectic submanifold, black curves are $W \cap \mathcal{B}_\psi$, red dotted lines are C_α (horizontal) and C_β (vertical), blue shaded regions are intersections of W and singular disks.

We cut (4-3) along $\pi(W)$. The components of the complement of W are the regular discs R_j in the specific decomposition of $\mathcal{B}_\psi$. Each R_k is a manifold with corners, where the corners are at $R_k \cap \pi(W) \cap S_l$. Then the proof of Theorem 1.3 shows that this decomposition of $\mathcal{B}_\psi^*$ is a decomposition into singular and regular discs. More precisely, there are two types of singularities of $\pi \colon \psi^m(\mathcal{B}_\psi) \to \mathcal{B}_\psi$, one coming from a singularity of $\psi^{m-1}(\mathcal{B}_\psi)$ and the other occurring when one applies ψ. The proof of Theorem 1.3 shows two things; first, ψ sends a singular value of $\psi^{m-1}(\mathcal{B}_\psi)$ onto a singular disk, and second, a new born singular value lies on a singular disk.

Remark 4.7 (1) If $\mathcal{B}_{\psi_1} = \mathcal{B}_{\psi_2}$, then it is easy to check that the standard decomposition with respect to ψ_i are the same.

(2) In Section 3.5, we defined a set $\mathbb{B}$ of Lagrangian branched submanifolds in M. For all $\mathcal{B} \in \mathbb{B}$, one can find a symplectic automorphism ψ such that $\mathcal{B} = \mathcal{B}_\psi$. Together with the above argument, for all $\mathcal{B} \in \mathbb{B}$, $\mathcal{B}^*$ admits the standard decomposition.

Example 4.8 Let M be the plumbing of $T^*\alpha$ and $T^*\beta$ at one point p, where $\alpha, \beta \simeq S^2$. Let $\psi = \tau \circ \rho^{-1}$ where τ (resp. β) is a Dehn twist along α (resp. β). Then $\mathcal{B}_\psi = (\beta \#_p \alpha) \cup \alpha$ and $(\psi, \mathcal{B}_\psi, \mathcal{B}_\psi^*)$ admits the standard decomposition.

Figure 14, left, is a schematic picture of $\mathcal{B}_\psi$. The regions shaded blue are singular disks of the standard decomposition. The red dotted circles are C_α and C_β. Figure 14, right, is the symplectic submanifold W of codimension 2.

Remark 4.9 For a given symplectic manifold M and a given triple $(\psi, \mathcal{B}_\psi, \mathcal{B}_\psi^*)$, it is natural to ask which Lagrangians L are compatible with the standard decomposition of $\mathcal{B}_\psi^*$. One can easily check that if L is one of zero sections α_i or β_j, or if L is obtained by applying a series of Dehn twist to one of zero sections, then L is compatible with the standard decomposition. See Remark 3.17 for the notation α_i

and β_j and see the proof of Lemma 5.1. Also we note that by Wu [18], if M is an A_n–surface singularity, then every exact Lagrangian L is compatible with the standard decomposition of $\mathcal{B}^*_\psi$ if L is carried by $\mathcal{B}_\psi$. In the current paper, we simply assume that a Lagrangian M is compatible with the standard decomposition for convenience.

4.3 Braids

In Section 4.3, we will generalize the notion of weights on higher-dimensional symplectic manifolds. We will assume that a given triple $(\psi, \mathcal{B}_\psi, \mathcal{B}^*_\psi)$ admits a decomposition into singular and regular disks.

Let S be a singular disk in $\mathcal{B}^*_\psi$ and let S° be the interior of S. Then $\pi^{-1}(S^\circ) = \bigcup_{p \in S^\circ} F_p$ is symplectomorphic to $DT^*(\mathbb{D}^n)^\circ$. Thus, the closure $\overline{\pi^{-1}(S^\circ)}$ is symplectomorphic to $DT^*\mathbb{D}^n$ and there is a natural symplectomorphism between them. The boundary $\partial\overline{\pi^{-1}(S^\circ)}$ is a $\mathbb{D}^n$–bundle over $\partial S \simeq S^{n-1}$ and the natural symplectomorphism induces $\varphi \colon \partial\overline{\pi^{-1}(S^\circ)} \xrightarrow{\sim} S^{n-1} \times \mathbb{D}^n$.

Definition 4.10 $D(S)$ (resp. $D(\partial S)$) is the $\mathbb{D}^n$–bundle $\overline{\pi^{-1}(S^\circ)}$ (resp. $\partial\overline{\pi^{-1}(S^\circ)}$) over S_i (resp. ∂S).

Definition 4.10 is for notational convenience.

Remark 4.11 Since $D(S)$ is symplectomorphic to a disk cotangent bundle of $\mathbb{D}^n$, coordinate charts on the base will induce a natural identification between $D(S)$ and $\mathbb{D}^n \times \mathbb{D}^n$. By restricting the identification on the boundary, $D(\partial S)$ is identified with $S^{n-1} \times \mathbb{D}^n$.

If L is a Lagrangian submanifold which is carried by $\mathcal{B}_\psi$ and if L is compatible with the decomposition of $\mathcal{B}^*_\psi$, then, for all $p \in \partial S$, $\varphi(L \cap F_p)$ is a finite collection of isolated points in $F_p \simeq \mathbb{D}^n$; recall that $\pi \colon L \to \mathcal{B}^*_\psi$ has no singular value on ∂S. Thus, $\varphi(L \cap D(\partial S))$ can be identified with a map from $\partial S \simeq S^{n-1}$ to the configuration space $\mathrm{Conf}_l(\mathbb{D}^n)$ of l points on $\mathbb{D}^n$ where $l = l(L, S)$, ie a braid. Since L is Lagrangian, $(\varphi^{-1})^*\omega$ vanishes on $\phi(L \cap D(\partial S))$.

From now on, we will define the braids on the boundary of a singular disk S. Let $f \colon S^{n-1} \to \mathrm{Conf}_l(\mathbb{D}^n)$ for some l. In other words, there are maps

$$f_1, \dots, f_l \colon S^{n-1} \to \mathbb{D}^n$$

such that $f(p) = \{f_1(p), \dots, f_l(p)\}$ with $f_i(p) \neq f_j(p)$ for all $i \neq j$. We define

(4-4)
$$B(f) := \{(p, f_i(p)) \in S^{n-1} \times \mathbb{D}^n \mid p \in S^{n-1}, i \in \{1, \dots, \ell\}\},$$
$$\widetilde{\mathrm{Br}}_{\partial S} := \{\varphi^{-1}(B(f)) \mid f \colon S^{n-1} \to \mathrm{Conf}_l(\mathbb{D}^n) \text{ such that } (\varphi^{-1})^*(\omega) \text{ is zero on } B(f) \text{ for some } l\}.$$

Note that $\widetilde{\mathrm{Br}}_{\partial S}$ is a set of closed subsets of $D(\partial S)$ and independent of φ.

We define an equivalence relation on $\widetilde{\mathrm{Br}}_{\partial S}$ as follows: $b_0 \sim b_1$ for $b_i \in \widetilde{\mathrm{Br}}_{\partial S}$ if there exists a smooth 1–parameter family $b_t \in \widetilde{\mathrm{Br}}_{\partial S}$ connecting b_0 and b_1. Let $\mathrm{Br}_{\partial S} := \widetilde{\mathrm{Br}}_{\partial S}/\sim$.

Definition 4.12 Let $(\psi, \mathcal{B}_\psi, \mathcal{B}_\psi^*)$ admit a decomposition into singular and regular discs. If L is a Lagrangian submanifold which is carried by $\mathcal{B}$ and is compatible with the decomposition of $\mathcal{B}_\psi^*$, then the *braid* $b(L, S)$ of L on a singular disk S is the braid isotopy class of $\text{Br}_{\partial S}$ which is given by

$$b(L, S) = [L \cap D(\partial S)] \in \text{Br}_{\partial S}.$$

Remark 4.13 The word "braid" comes from the case of $\frac{1}{2} \dim M = n = 2$. If $n = 2$, then f in (4-4) is an element of $\pi_1(\text{Conf}_l(\mathbb{D}^2))$, ie a braid. For a general n, we consider an element of $\pi_n(\text{Conf}_l(\mathbb{D}^n))$.

The notion of braid is defined as an equivalence class in Definition 4.12. However, in the rest of the present paper, if it is not likely to be misunderstood, then we use the word "braid $b(L, S)$" to indicate a representative of the class. This is for the notational convenience. By considering a representative of a braid, we can consider $b(L, S)$ as a subset in $D(\partial S)$. For the case of $\frac{1}{2} \dim M = n = 2$ (resp. general n), a braid $b(L, S)$ is a union of circles (resp. S^{n-1}) embedded in $D(\partial S)$.

Definition 4.14 A *strand* of a braid $b(L, S)$ is a connected component of $b(L, S) \subset D(\partial S)$.

As similar to Remark 4.13, the word "strand" comes from the case of $\frac{1}{2} \dim M = n = 2$.

5 Action of a symplectomorphism

In Section 5.1, we briefly review how one can keep track of the action of a surface automorphism changing the isotopy classes of curves. The action can be written as a linear map acting on the set of weights. Also, we generalized the notion of weights in Section 5.

In Section 5, we generalize the "linear algebra on weights" for higher-dimensional cases.

5.1 Linear algebra on braids

We would like to generalize the linear algebra on weights, which we reviewed in Section 4.1. More precisely, we claim the following:

Claim (⋆) *If L is carried by $\mathcal{B}_\psi$ and L is compatible with the standard decomposition of $\mathcal{B}_\psi^*$ for a symplectic automorphism ψ of generalized Penner type, then there is a systematic way to obtain*

$$\{b(\psi(L), S) \mid S \text{ is a singular disk of the standard decomposition of } \mathcal{B}_\psi^*\}$$

from

$$\{b(L, S) \mid S \text{ is a singular disk of the standard decomposition of } \mathcal{B}_\psi^*\}.$$

Moreover, the systematic way depends only on ψ, independent of L, as one has a matrix A_ψ for ψ of generalized Penner type as in Section 4.1.

Instead of proving ($\star$), we will prove Lemma 5.1, which considers Dehn twists τ_k and σ_j^{-1} instead of ψ. Recall Remark 3.17 saying that the symplectic manifold M is a plumbing space of copies of T^*S^n and τ_i (resp. σ_j) is a Dehn twist along one of the zero sections of T^*S^n. We also recall Lemma 3.20 saying the following: there exists a set $\mathbb{B}$ of Lagrangian branched submanifolds and functions $F_{\tau_k} : \mathbb{B} \to \mathbb{B}$ (resp. $F_{\sigma_j^{-1}} : \mathbb{B} \to \mathbb{B}$), such that if L is carried by $\mathcal{B} \in \mathbb{B}$, then $\tau_k(L)$ (resp. $\sigma_j^{-1}(L)$) is carried by $F_{\tau_k}(\mathcal{B})$ (resp. $F_{\sigma_j^{-1}}(\mathcal{B})$).

Lemma 5.1 *Let L be a Lagrangian submanifold of M such that L is carried by $\mathcal{B} \in \mathbb{B}$ and L is compatible with the standard decomposition of $\mathcal{B}^*$. Then $\tau_k(L)$ is compatible with the standard decomposition of $F_{\tau_k}(\mathcal{B})$. Moreover, there exists a systematic way to obtain*

$$\{b(\tau_k(L), S) \mid S \text{ is a singular disk of the standard decomposition of } F_{\tau_k}(\mathcal{B})^*\}$$

from

$$\{b(L, S) \mid S \text{ is a singular disk of the standard decomposition of } \mathcal{B}^*\}.$$

The case of σ_j^{-1} is analogous.

Remark 5.2 Since a symplectic automorphism ψ of generalized Penner type is a product of Dehn twists τ_k and σ_j^{-1}, Lemma 5.1 is enough to prove ($\star$).

We will prove Lemma 5.1 in Sections 5.2 and 5.3. The proof will be given for an example case. In the rest of Section 5.1, we will introduce the main idea of the proof. Also, we will introduce the example case which we will consider in Sections 5.2 and 5.3.

The main idea The main idea is to consider $\tau_k(N(\mathcal{B}))$ instead of $\tau_k(L)$. More precisely, for a given singular disk S' of $F_{\tau_k}(\mathcal{B}) = \mathcal{B}'$, we consider $\tau_k(N(\mathcal{B})) \cap D(\partial S')$. One can check that every connected component of $\tau_k(N(\mathcal{B})) \cap D(\partial S')$ is homeomorphic to $S^{n-1} \times \mathbb{D}^n$. For an arbitrary component, there is a map $f_{S \to S', i} : D(\partial S) \to D(\partial S')$, where S is a singular disk of $\mathcal{B}^*$ and a natural number i, such that the image of $f_{S \to S', i}$ is the connected component. In other words, every connected component of $\tau_k(N(\mathcal{B})) \cap D(\partial S')$ is given as the image of a function defined on $D(\partial S)$ where S is a singular disk of $\mathcal{B}$.

The subscription $(S \to S', i)$ of $f_{S \to S', i}$ means that it is a function explaining the contribution of $b(L, S)$ on $b(\tau_k(L), S')$. Since it is possible that there are multiple connected components of $\tau_k(N(\mathcal{B})) \cap D(\partial S')$, which are induced from the same singular disk S, one needs multiple functions, which are labeled by natural numbers i in the subscription.

Since L is carried by $\mathcal{B}$, $L \subset N(\mathcal{B})$. Thus, $\tau_k(L) \subset \tau_k(N(\mathcal{B}))$. By definition,

$$b(\tau_k(L), S') = \tau_k(L) \cap D(\partial S') \subset \tau_k(N(\mathcal{B})) \cap D(\partial S').$$

We consider the intersection of $b(\tau_k(L), S')$ with each connected components of $\tau_k(N(\mathcal{B})) \cap D(\partial S')$. In the connected component, which is the image of $F_{S \to S', i}$, $b(\tau_k(L), S')$ is given by

$$f_{S \to S', i}(b(L, S)).$$

Thus, the set $\{f_{S \to S', i}\}$ of functions gives the systematic way to construct

$$\{b(\tau_k(L), S) \mid S \text{ is a singular disk of the standard decomposition of } F_{\tau_k}(\mathcal{B})^*\}$$

from

$$\{b(L, S) \mid S \text{ is a singular disk of the standard decomposition of } \mathcal{B}^*\}.$$

The example case The symplectic manifold we consider is $M = P(\alpha, \beta_1, \beta_2)$, where α and β_j are spheres such that $\alpha \cap \beta_1 = \{p\}$ and $\alpha \cap \beta_2 = \{q\}$, ie M is a plumbing space of three copies of $T^* S^n$. Let τ_0 and σ_j be Dehn twists along α and β_j, and $\psi = \tau_0 \circ \sigma_1^{-1} \circ \sigma_2^{-1}$. Then Theorem 1.3 gives a Lagrangian branched submanifold $\mathcal{B}_\psi$. For the case of dim $M = 2n = 2$, Figure 15 describes the example symplectic manifold M. In the example, we will consider the effects of σ_2^{-1} on $\mathcal{B}_\psi$ in Section 5.2 and τ_0 in Section 5.3.

For convenience, we establish notation here. The standard decomposition of $\mathcal{B}_\psi$ has 6 singular disks which are centered at p, $\tau_0(p)$, $\sigma_1^{-1}(p)$, q, $\tau_0(q)$ and $\sigma_2^{-1}(q)$. As mentioned in Remark 4.6, we are abusing notation and pretending that the singular disks are in $\mathcal{B}_\psi$, not in $\mathcal{B}_\psi^*$. We also note that $\tau_0(p)$ and $\tau_0(q)$ are antipodal points of p and q on α. Similarly, $\sigma_1^{-1}(p)$ (resp. $\sigma_2^{-1}(q)$) is the antipodal point of p (resp. q) on β_1 (resp. β_2). Let $S_1, \ldots, S_6$ denote the singular disks centered at p, $\tau_0(p)$, $\sigma_1^{-1}(p)$, q, $\tau_0(q)$ and $\sigma_2^{-1}(q)$ respectively. Moreover, let b_i denote $b(L, S_i)$ for $i = 1, \ldots, 6$.

Similarly, the Lagrangian branched submanifolds $\mathcal{B}' := F_{\sigma_2^{-1}}(\mathcal{B})$ and $\mathcal{B}'' := F_{\tau_0}(\mathcal{B})$ each have 6 singular disks. By definition of the standard decomposition, those singular disks are also centered at p, $\tau_0(p)$, $\sigma_1^{-1}(p)$, q, $\tau_0(q)$ and $\sigma_2^{-1}(q)$. As we did for $\mathcal{B}$, let $S_1', \ldots, S_6'$ (resp. $S_1'', \ldots, S_6''$) denote the singular disks of $\mathcal{B}'$ (resp. $\mathcal{B}''$) centered at p, $\tau_0(p)$, $\sigma_1^{-1}(p)$, q, $\tau_0(q)$ and $\sigma_2^{-1}(q)$. Moreover, we label

$$b_i = b(L, S_i), \quad b_i' = b(\sigma_2^{-1}(L), S_i'), \quad b_i'' = b(\tau_0(L), S_i'').$$

In the rest of this paper, we make specific choices of τ_0 and σ_j, given by (2-2) and $\tau : T^* S^n \xrightarrow{\sim} T^* S^n$, which is defined in Remark 2.2. In other words, $\tau_0 = \phi_\alpha \circ \tau \circ \phi_\alpha^{-1}$ and $\sigma_j = \phi_{\beta_j} \circ \tau \circ \phi_{\beta_j}^{-1}$, where ϕ_α (resp. ϕ_{β_j}) is a symplectomorphism from $T^* S^n$ to a neighborhood of α (resp. β_j). The neighborhood of α (resp. β_j) will be denoted by $T^* \alpha$ (resp. $T^* \beta_j$).

Remark 5.3 Recall that τ is a Dehn twist on $T^* S^n$ which agrees with the antipodal map

$$T^* S^n \xrightarrow{\sim} T^* S^n, \quad (u, v) \mapsto (-u, -v),$$

on a neighborhood of the zero section S^n.
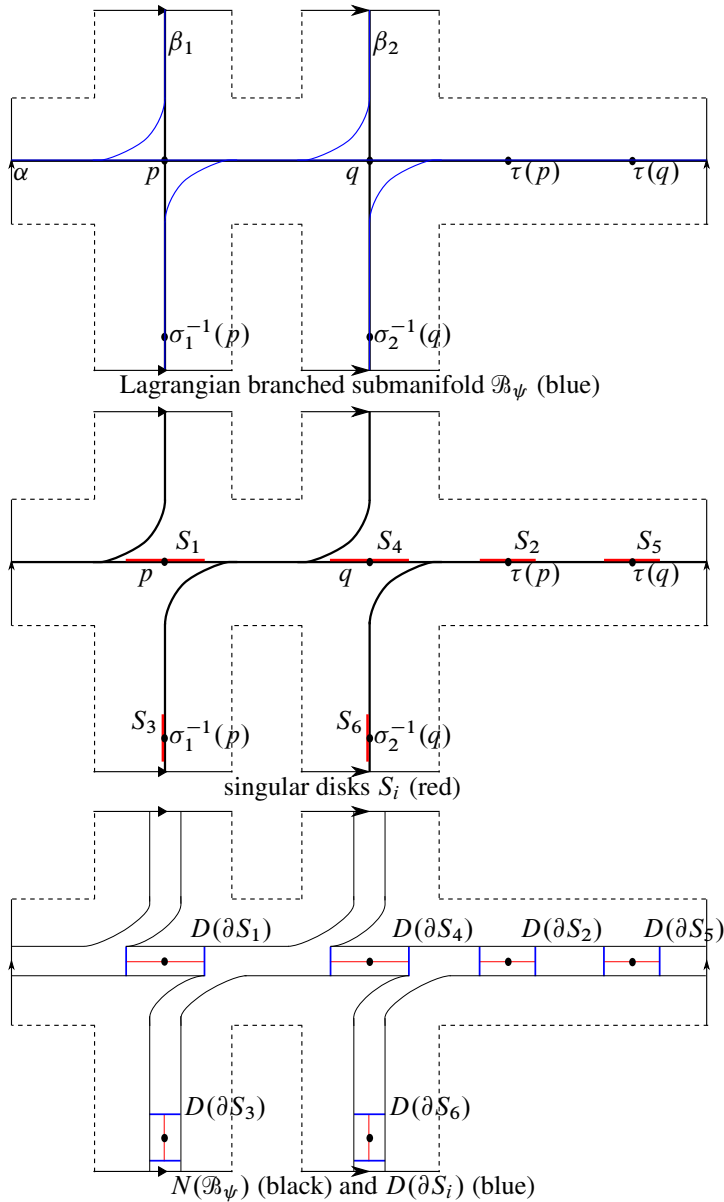
Figure 15: Top: the black curves represent α, β_1 and β_2 in $M = P(\alpha, \beta_1, \beta_2)$, and the blue curve is $\mathcal{B}_\psi$. Middle: the red curves are singular disks S_i. Bottom: the fibered neighborhood $N(\mathcal{B}_\psi)$ and a disk bundle $D(\partial S_i) \simeq \mathbb{D}^1 \times S^0$, ie two intervals attached at ∂S_i.

Remark 5.4 In the next sections, we will consider the example which we specified in the present subsection. Moreover, for convenience, we will assume that the dimension $2n$ of the symplectic manifold M is 4. For the case of $n = 2$, we specify identifications φ_i, φ_i' and φ_i'' from $D(\partial S_i)$, $D(\partial S_i')$ and $D(\partial S_i'')$ to $S^1 \times \mathbb{D}^2$. We would like to point out that there is no reason to choose these specific identifications, this is only for the notational convenience.

In order to construct $\varphi_1 \colon D(\partial S_1) \xrightarrow{\sim} S^1 \times \mathbb{D}^2$, we remark that

$$D(\partial S_1) = \partial \overline{\pi^{-1}(S_1^\circ)}, \quad D(S_1) = \overline{\pi^{-1}(S_1^\circ)}$$

by definition. Thus, in order to specify φ_1, it is enough to identify $D(S_1)$ and $\mathbb{D}^2 \times \mathbb{D}^2$. We remark that $D(S_1)$ is a disk bundle over $S_1 \simeq \mathbb{D}^2$.

By abuse of notation, let's assume that $S_1 \subset \mathbb{B}_\psi$, not $\mathbb{B}'_\psi$. Then S_1 is a Lagrangian disk in M. Thus, $D(S_1)$ is a small neighborhood of a Lagrangian disk S_1. By the Lagrangian neighborhood theorem [16], it is enough to choose coordinate charts on S_1. Similarly, it is enough to choose coordinate charts for S_i, S_i' and S_i''.

In order to choose specific coordinate charts, we use the symplectic submanifold $W \subset M$ defined in Section 4.2.

Let (x_1, x_2) be a coordinate chart on $S_1 \subset \alpha$ such that the x_1–axis agrees with $W \cap S_1$. There are two choices for the positive x_1–direction corresponding to the two orientations of $W \cap S_1$, or equivalently orientations of C_α. We can choose either of them. Then, let (y_1, y_2) be an oriented chart on S_2 such that the y_1–axis agrees with $W \cap \beta_1$ and $\omega(\partial_{x_1}, \partial_{y_1}) > 0$. The positive y_1–direction determines an orientation of C_{β_1}. On S_3, there exists an oriented chart (x_1, x_2) such that the positive x_1–direction agrees with the orientation of C_α. For the other singular disks, we obtain oriented coordinate charts from the orientations of C_α, C_{β_i}, α and β_i in the same way.

5.2 Effect of σ_2^{-1}

In Section 5.2, we discuss how $\{b_i' \mid i = 1, \dots, 6\}$ are obtained from $\{b_i \mid i = 1, \dots, 6\}$. Since σ_2^{-1} is supported on $T^*\beta_2$, a small neighborhood of β_2, b_i and b_i' are the same braid in $\mathrm{Br}_{\partial S_i}$ for $i = 1, 2, 3$ and 5. We will explain how b_6' is constructed.

We can obtain $\sigma_2^{-1}(\mathbb{B}_\psi)$ by spinning with respect to q in $T^*\beta_2$, ie $\sigma_2^{-1}(\mathbb{B}_\psi)$ is the union of curves in a 2–dimensional submanifold $\phi_{\beta_2}(W_y)$ over $y \in S^{n-1}$. Recall that the spinning and W_y are defined in Section 2.2.

Figure 16 represents a support of σ_2^{-1} in M, ie a small neighborhood of $\beta_2 \subset M$ where M is a symplectic manifold of dimension 2 given in Figure 15. Similarly, in Section 5.2, the rectangles in Figures 15–19 are the support of σ_2^{-1}.

By spinning blue, red, and green points in Figure 16, we obtain $\sigma_2^{-1}(\mathbb{B}_\psi) \cap D(\partial S_6')$. Let B, R and G be obtained by spinning constant curves drawn blue, red and green points in Figure 16, respectively.

Since $N(\mathbb{B}_\psi) \supset \mathbb{B}_\psi$, $\sigma_2^{-1}(N(\mathbb{B}_\psi)) \cap D(\partial S_6')$ is a neighborhood of $\sigma_2^{-1}(\mathbb{B}_\psi) \cap D(\partial S_6')$. By assuming that $N(\mathbb{B}_\psi)$ is a sufficiently small neighborhood of $\mathbb{B}_\psi$, $\sigma_2^{-1}(N(\mathbb{B}_\psi)) \cap D(\partial S_6')$ consists of three connected components, which are neighborhoods of B, R and G. Each connected component will be called $N(B)$, $N(R)$ and $N(G)$.
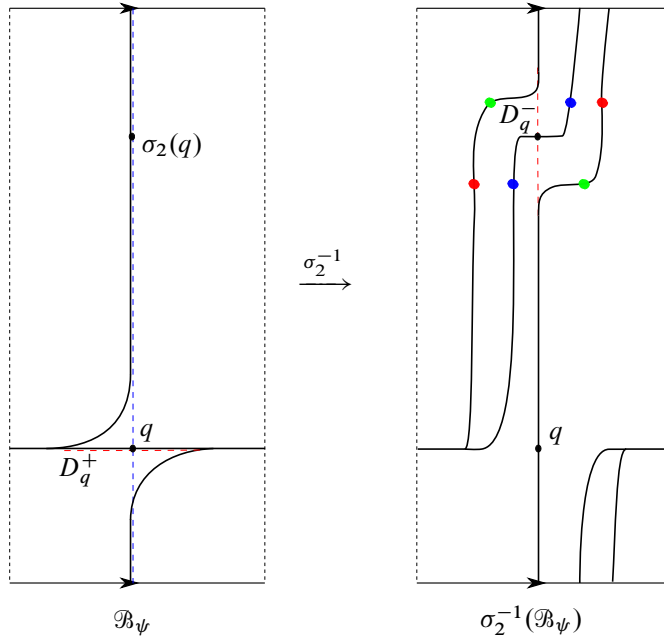
Figure 16: The left picture represents $\mathcal{B}_\psi \cap \phi_{\beta_2}(W_y)$ and the right picture represents $\sigma_2^{-1}(\mathcal{B}_\psi) \cap \phi_{\beta_2}(W_y)$.

Since $b_6' = \sigma_2^{-1}(L) \cap D(\partial S_6') \subset N(B) \sqcup N(R) \sqcup N(G)$, b_6' is divided into three groups, which are contained in $N(B)$, $N(R)$ and $N(G)$ respectively. We argue the group which is contained in $N(B)$ first.

Let assume that $\sigma_2^{-1}(S_4) = S_6'$. Then $\sigma_2^{-1}(\partial S_4) = \partial S_6'$. Moreover, if $\sigma_2^{-1}(D(\partial S_4))) \subset D(\partial S_6')$, then $N(B) = \sigma_2^{-1}(D(\partial S_4)) \subset D(\partial S_6')$. Also, one concludes that $\sigma_2^{-1}|_{D(\partial S_4)} \colon D(\partial S_4) \xrightarrow{\sim} N(B)$. If one can assume that b_4 is a subset of $D(\partial S_4)$ by definition of braids, the set of braids of b_6' inside $N(B)$ is $\sigma_4^{-1}(b_4)$.

However, $\sigma_2^{-1}(S_4)$ is not s_6'. Thus, we will construct a Hamiltonian isotopy Φ_t so that there exists a slightly smaller disk D_B is S_4 satisfying

$$(\Phi_1 \circ \sigma_2^{-1})(D_B) = S_6'.$$

Note that "slightly smaller" means that there is no singular value on $S_4 \setminus D_B$. Then

$$(\Phi_1 \circ \sigma_2^{-1})(D(\partial D_B)) = N(B),$$

where $D(\partial D_B)$ is defined as similar to Definition 4.10. The strands of b_6' in $N(B)$ will be given by $(\Phi_1 \circ \sigma_2^{-1})(D(\partial D_B) \cap L)$. Moreover, $D(\partial D_B)$ (resp. $D(\partial D_B) \cap L$) and $D(\partial S_4)$ (resp. $b_4 = D(\partial S_4) \cap L$) are naturally isotopic. Under the isotopic relation, there is a function $f_1 \colon D(\partial S_4) \to D(\partial S_6')$ such that the strands of b_6' in $N(B)$ are $f_1(b_4)$.

From now on, we will construct a specific Φ_t. For notational simplicity, we assume that $\dim(M) = 4$, but the construction of Φ_t is easily generalized for the case of higher dimensions.
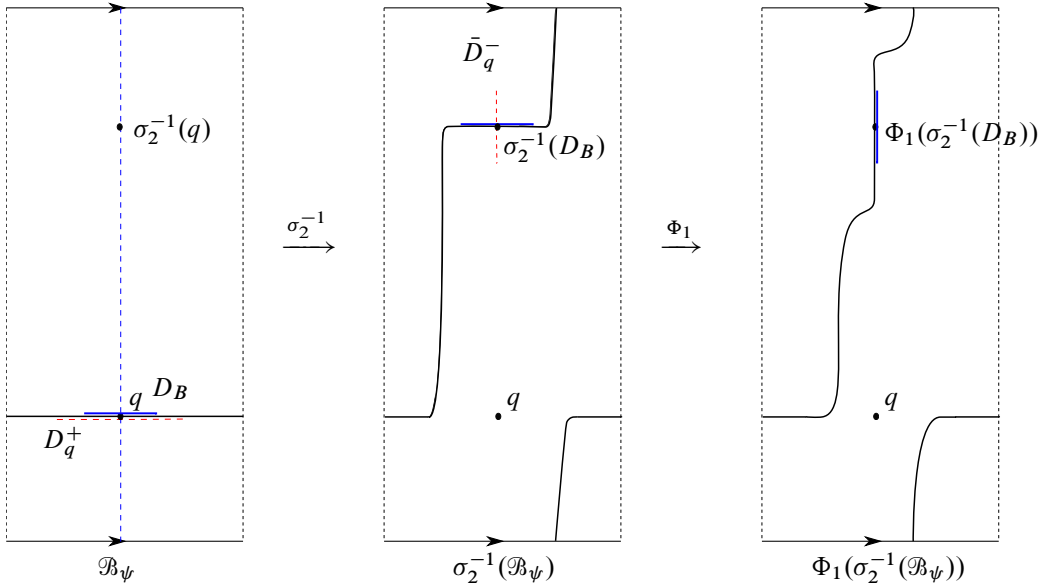
Figure 17: The blue curves represent $\widetilde{D}_B \cap \phi_{\beta_2}(W_y)$ in the left picture, $\sigma_2^{-1}(\widetilde{D}_B) \cap \phi_{\beta_2}(W_y)$ in the middle picture, and $\Phi_1(\sigma_2^{-1}(\widetilde{D}_B)) \cap \phi_{\beta_2}(W_y)$ in the right picture.

We choose a neighborhood $U \subset \beta_2$ of $\sigma_2^{-1}(q)$ and a Darboux chart $\phi_q : T^*U \xrightarrow{\sim} \mathbb{R}^4$ such that $\phi_q(\sigma_2^{-1}(q))$ is the origin. We remark that $T^*\beta_2$ denotes a neighborhood of β_2 in M, which is symplectomorphic to the cotangent bundle of β_2. Thus, for a subset U of β_2, one can assume that T^*U is a subset of M.

For convenience, let $\phi_q(x) = (x_1, x_2)$ where $x_i \in \mathbb{R}^2$. Then there is a Hamiltonian isotopy

$$(5\text{-}1) \qquad \Phi_t(x) = \begin{cases} (\phi_q^{-1} \circ H_{t\delta(c_1\|x_1\|+c_2\|x_2\|)} \circ \phi_q)(x) & \text{if } x \in T^*U, \\ x & \text{if } x \notin T^*U, \end{cases}$$

where c_i is a positive constant, $\|\cdot\|$ is the standard norm on $\mathbb{R}^2$, and H_t and δ are defined as follows: let $H_t : \mathbb{R}^4 \to \mathbb{R}^4$ be a Hamiltonian isotopy given by

$$H_t = \begin{pmatrix} \cos t & 0 & -\sin t & 0 \\ 0 & \cos t & 0 & -\sin t \\ \sin t & 0 & \cos t & 0 \\ 0 & \sin t & 0 & \cos t \end{pmatrix},$$

and let $\delta : [0, \infty) \to \mathbb{R}$ be a smooth decreasing function such that $\delta(x) = \frac{\pi}{2}$ for all $x < 1$ and $\delta(x) = 0$ for all $x > 2$.

Figure 17 represents the case of dim $M = 2$. We note that the rectangles in Figure 17 represent a support of σ_2^{-1}. By choosing proper constants c_i, we obtain a small disk $D_B \subset S_4$ such that

$$(\Phi_1 \circ \sigma_2^{-1})(D(\partial D_B)) \subset D(\partial S_6').$$
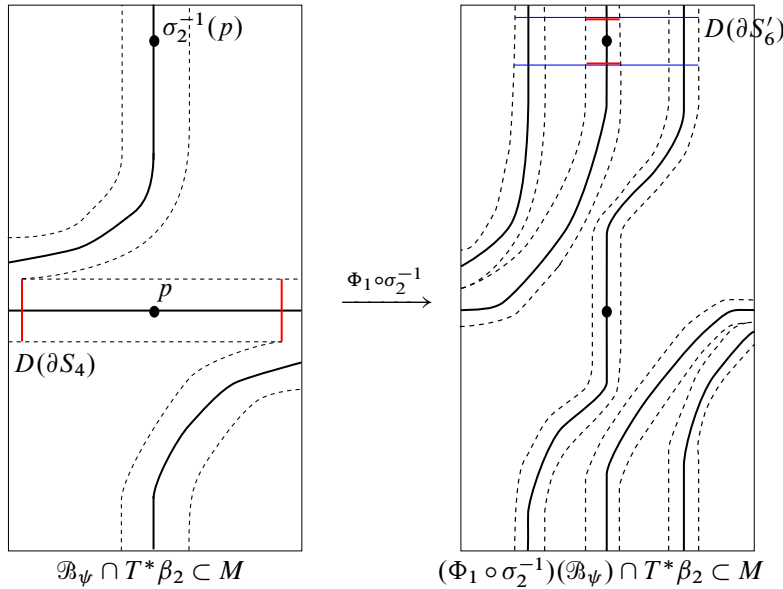
Figure 18: Left: the whole rectangle is a neighborhood of β_2 in M; thick black curves are parts of $\mathcal{B}_\psi$ and the dashed black curves are $N(\mathcal{B}_\psi)$; the red thick curves represent $D(\partial S_4)$. Right: thick black curves are $(\Phi_1 \circ \sigma_2^{-1})(\mathcal{B}_\psi)$, dashed black curves are $(\Phi_1 \circ \sigma_2^{-1})(N(\mathcal{B}_\psi))$, blue curves are $D(\partial S_6')$, and thick red curves represent the part of $D(\partial S_6')$ where $D(\partial S_4)$ contributes.

On a small neighborhood of D_B, σ_2^{-1} agrees with the antipodal map of $\phi_{\beta_2}(T^*\beta_2) \simeq T^*S^2$, as we mentioned in Remark 5.3. Then we obtain a map

$$f_1 \colon S^1 \times (\mathbb{D}^2)^\circ \simeq \pi^{-1}(\partial D_B) \xrightarrow{\Phi_1 \circ \sigma_2^{-1}} D(\partial S_6') \simeq S^1 \times \mathbb{D}^2, \quad (\theta, x, y) \mapsto (\theta + \pi, -r_1 x, -r_1 y).$$

The first and the last identifications are the natural identifications mentioned in Remark 4.11. The reason we consider the natural identification is for notational convenience, ie in order to write f_1 as a map on $(\theta, x, y) \in S^1 \times \mathbb{D}^2$. Then, the strands of b_6' in $N(B)$ is given by $f_1(b_4)$.

Figure 18 is a picture summarizing the whole process obtaining strands of b_6' in the first group, or equivalently, the picture explains how b_4 contributes on the construction of b_6', in the case dim $M = 2n = 2$.

In order to study the construction of strands of b_6' in $N(R)$ and $N(G)$, one should consider

$$\widetilde{D}(\partial S_4) := \bigcup_{p \in \text{Locus}(\mathcal{B}') \cap \partial S_4} F_p.$$

It is easy to check that $\widetilde{D}(\partial S_4)$ is a $\mathbb{D}^n$–bundle over ∂S_4 and $D(\partial S_4) \subset \widetilde{D}(\partial S_4)$.

Together with $\widetilde{D}(\partial S_4)$, we observe how b_6 contributes on the construction of b_6'. First, one can observe that b_6 and $L \cap (\widetilde{D}(\partial S_4) \setminus D(\partial S_4))$ are isotopic to each other. The isotopy connecting them is along the fibers on some regular disks such that the union of regular disks has ∂S_4 and ∂S_6 as their boundaries.
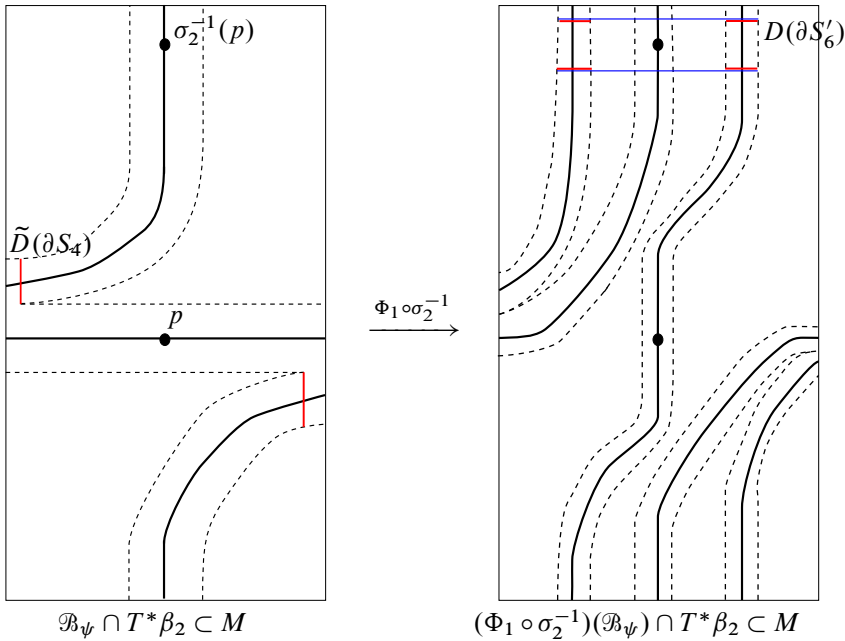
Figure 19: Left: the whole rectangle is a neighborhood of β_2 in M, thick black curves are parts of $\mathcal{B}_\psi$ and the dashed black curves are $N(\mathcal{B}_\psi)$, and the red thick curves represent $\widetilde{D}(\partial S_4)$. Right: thick black curves are $(\Phi_1 \circ \sigma_2^{-1})(\mathcal{B}_\psi)$, dashed black curves are $(\Phi_1 \circ \sigma_2^{-1})(N(\mathcal{B}_\psi))$, blue curves are $D(\partial S_6')$, and thick red curves represent the part of $D(\partial S_6')$ where $D(\partial S_4)$ contributes.

More precisely, the union of regular disks (resp. fibers on them) is homeomorphic to $S^{n-1} \times [0, 1]$ (resp. a disk bundle over $S^{n-1} \times [0, 1]$). The boundary of $S^{n-1} \times [0, 1]$ corresponds to ∂S_4 and ∂S_6.

Similarly, one can observe that $L \cap (\widetilde{D}(\partial S_4) \setminus D(\partial S_4))$ and b_6 are isotopic to each other. The isotopy connecting them is the intersection of L and the fibers on the regular disks.

Second, one can describe the contribution of $L \cap (\widetilde{D}(\partial S_4) \setminus D(\partial S_4))$ on the contribution of b_6'. The contributions are given as two functions as the contribution of b_4 is described by the function f_1. For the case of $n = 2$ and under the identification defined in Remark 5.4, the two functions denoted by f_2 and f_3 are

$$f_2 \colon S^1 \times \mathbb{D}^2 \to S^1 \times \mathbb{D}^2, \quad (\theta, x, y) \mapsto (\theta, r_0 \cos \theta + r_2 x, r_0 \sin \theta + r_2 y),$$

and

$$f_3 \colon S^1 \times \mathbb{D}^2 \to S^1 \times \mathbb{D}^2,$$

$$(\theta, x, y) \mapsto (\theta, -r_0 \cos \theta + r_2(x \cos 2\theta - y \sin 2\theta), -r_0 \sin \theta + r_2(x \sin 2\theta + y \cos 2\theta)),$$

Similar to Figure 18, Figure 19 summarizes the whole process obtaining strands of b_6' in the second and third groups, or equivalently, the picture explains how $\widetilde{D}(\partial S_4)$ contributes on the construction of b_6', for the case of dim $M = 2n = 2$.

Remark 5.5   (1) The constant r_1 is determined by the choice of an identification $\phi_{\beta_2} : T^*S^2 \xrightarrow{\sim} T^*\beta_2$, the fixed Dehn twist τ in Remark 2.2, and so on. However, r_1 has to be smaller than 1. This is because $\mathrm{Im}(f_1)$, $\mathrm{Im}(f_2)$ and $\mathrm{Im}(f_3)$ are mutually disjoint, since they correspond to $N(B)$, $N(R)$ and $N(G)$, respectively. Moreover, r_0 and r_2 are also positive numbers smaller than 1.

(2) Note that r_0 and r_2 are positive constants which are determined by specific choices. However, r_0 and r_2 have to satisfy $r_1 + r_2 < r_0$, since $\mathrm{Im}(f_1)$, $\mathrm{Im}(f_2)$ and $\mathrm{Im}(f_3)$ are mutually disjoint.

(3) To obtain f_1, we used a Hamiltonian isotopy Φ_t. Similarly, to obtain f_2 and f_3, we need a Hamiltonian isotopy.

The situation for b_4' is analogous. We obtain three maps g_1, g_2 and g_3 in the same way. At the end, b_4' is represented by $g_1(b_6) \sqcup g_2(b_6) \sqcup g_3(b_6)$. This proves Lemma 5.1 for the case of σ_2^{-1}.

Note that maps f_i and g_j are given by specific maps acting on $S^1 \times \mathbb{D}^2$, but we would like to consider them as maps on $\widetilde{\mathrm{Br}}_{\partial S_k}$ for some k. We summarize the effect of σ_2^{-1} as the matrix

$$\Sigma_{2,\mathscr{B}_\psi} = \begin{pmatrix} \mathrm{id} & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{id} & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathrm{id} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & g_1 + g_2 + g_3 \\ 0 & 0 & 0 & 0 & \mathrm{id} & 0 \\ 0 & 0 & 0 & f_1 & 0 & f_2 + f_3 \end{pmatrix}.$$

Thus,

$$\begin{pmatrix} b_1' \\ b_2' \\ b_3' \\ b_4' \\ b_5' \\ b_6' \end{pmatrix} = \Sigma_{2,\mathscr{B}_\psi} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ g_1(b_6) \sqcup g_2(b_6) \sqcup g_3(b_6) \\ b_5 \\ f_1(b_4) \sqcup f_2(b_6) \sqcup f_3(b_6) \end{pmatrix}.$$

Remark 5.6 In surface theory, we can do linear algebra on weights, but in a higher-dimensional case, we cannot do linear algebra with the matrix $\Sigma_{2,\mathscr{B}_\psi}$ because there is no module structure on $\widetilde{\mathrm{Br}}_{\partial S_i}$. In other words, the matrix $\Sigma_{2,\mathscr{B}_\psi}$ and sums of functions, for example $g_1 + g_2 + g_3$, are for notational convenience. Thus, the title of Section 5.1 is an abuse of terminologies.

5.3 Effect of τ_0

The situation for τ_0 is similar to that for σ_2^{-1}. For example, by observing how τ_0 acts on $D(\partial S_1)$, we obtain

$$h_1 : S^1 \times \mathbb{D}^2 \to S^1 \times \mathbb{D}^2,$$

explaining the contribution of b_1 on the construction of b_3''. Then, h_1 is given by a translation on S^1 and a scaling on $\mathbb{D}^2$, as f_1 is. Similarly, we obtain h_2 and h_3, which explain the contributions of b_3 on the construction of b_3''. The maps h_2 and h_3 are of the same type as f_2 and f_3, respectively, ie

$$h_2(\theta, x, y) = (\theta \text{ or } \theta + \pi, \pm r_1 \cos \theta + r_2 x, \pm r_1 \sin \theta + r_2 y),$$

$$h_3(\theta, x, y) = \big(\theta \text{ or } \theta + \pi, \pm r_1 \cos \theta + r_2(x \cos 2\theta - y \sin 2\theta), \pm r_1 \sin \theta + r_2(x \sin 2\theta + y \cos 2\theta)\big),$$

where r_1 and r_2 are constants.

We say that a map is of *scaling type* if a map is of the same type as f_1, in other words, if the map is given by a translation on S^1 and a scaling on $\mathbb{D}^2$. This is because the formula defining the map is given by a scaling on fibers. The maps of scaling type explain how the braids along the singular disk centered at p or antipodes of p, $b(L, S_p(\mathcal{B}_\psi))$ or $b(L, S_p^\pm)$, contribute on the braid along the singular points centered at the same points, $b\big(\delta(L), S_p(F_\delta(\mathcal{B}_\psi))\big)$ or $b(\delta(L), S_p^\pm)$, when one applies a Dehn twist δ.

We say that a map is of *the first (resp. second) singular type* if a map is of the same type as f_2 (resp. f_3). This is because they are related to a creation of new singular component. The maps of the first and second singular types explain how the braid $b(L, S_p^+)$ contributes on the construction of the braid $b\big(\delta(L), S_p(F_\delta(\mathcal{B}_\psi))\big)$.

To summarize, if b_i contributes the construction of b_j' and if the center of a singular disk corresponding to b_i is either the same point or the antipodal point of the center of the singular disk corresponding to b_j', maps of these three types explain the contribution of b_i on the construction of b_j'. Note that the center of a singular disk is defined in Remark 3.21.

The maps of these three types explain the effects of σ_2^{-1} on $\mathcal{B}$. However, to explain the effects of τ_0 on $\mathcal{B}_\psi$, we need maps of one more type. The reason is given in Figure 20, roughly. We note that the rectangles in Figure 20 are the support of τ_0 in M where M is given in Figure 15, ie a neighborhood of $\alpha \subset M$.

More precise reasoning is as follows. We note that α has two plumbing points, unlike β_i which has only one plumbing point. Thus, when we apply τ_0, b_i can contribute to b_j'' even if the centers of singular disks corresponding to b_i and b_j'' are neither the same nor antipodes of each other. For example, $L \cap \pi^{-1}(\pi(N_p))$ is stretched by τ_0. The stretched part $\tau_0\big(L \cap \pi^{-1}(\pi(N_p))\big)$ has intersection with $D(\partial S_4'')$ and $D(\partial S_5'')$ as one can see in Figure 20. Thus, b_4'' has some strands corresponding to $\tau_0\big(L \cap \pi^{-1}(\pi(N_p))\big) \cap D(\partial S_4)$. These strands are the contribution of b_3 on the construction of b_4''. Similarly, b_3 contributes to the construction of b_5'', and b_6 contributes to the constructions of b_1'' and b_2''.

To describe the contribution of b_3 on b_4'', without loss of generality, we assume that there is no singular value for

$$\tau_0\big(L \cap \pi^{-1}(\pi(N_p))\big) \cap D(S_4) \xrightarrow{\pi} S_4,$$

by Remark 3.21. Thus, $\tau_0\big(L \cap \pi^{-1}(\pi(N_p))\big) \cap D(S_4)$ is a union of disjoint Lagrangian disks on $D(S_4)$. We note that $D(S_4)$ is a disk bundle over $(S_4')^\circ$ which is an open disk. Thus, on the boundary $D(\partial S_4')$,
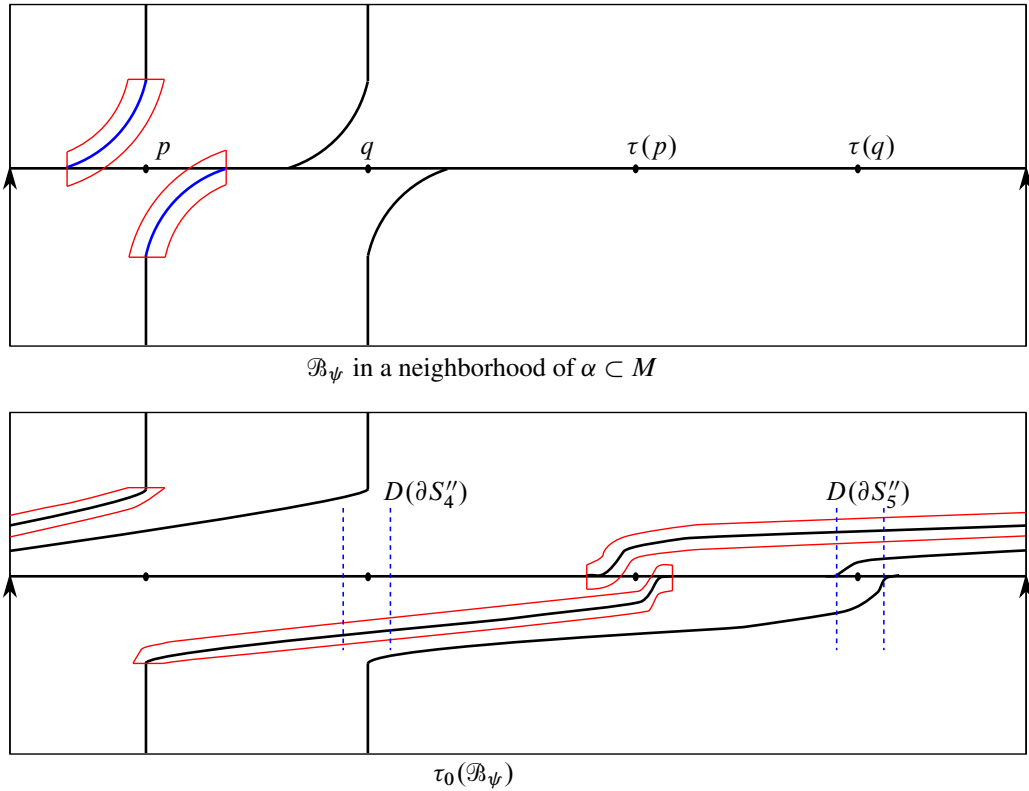
Figure 20: Top: the thick black and blue curves are $\mathcal{B}_\psi$ in a neighborhood of $\alpha \subset M$; in particular, the blue curves are N_p and the red parts are a fibered neighborhood of N_p, ie $\pi^{-1}(\pi(N_p))$. Bottom: the thick curves are $\tau_0(\mathcal{B}_\psi)$, the red parts are extended neighborhood of N_p by applying τ_0, and blue dashed lines are $D(\partial S''4)$ and $D(\partial S_5'')$.

b_3 contributes to b_4'' by adding strands near $\tau_0(N_p) \cap D(\partial S_4)$ which are not braided to each other. The number of the added strands is the same as the number of strands of b_3. In the same way, b_3 contributes to the construction of b_5''.

Remark 5.7 In the above argument, we said that the added strands are not braided to each other. To be more rigorous, we should specify the meaning of "not braided". We remark that $D(\partial S_4'')$ is identified with $S^{n-1} \times \mathbb{D}^n$ by the specific identification given in Remark 5.4. The added strands are not braided in $S^{n-1} \times \mathbb{D}^n$ after the identification.

As we did before, we would like to describe the added strands as an image of a function defined on $D(\partial S) = S^{n-1} \times \mathbb{D}^n$. In Section 5.3, we consider the case of $\dim(M) = 4$ as we did in Section 5.2 under the identifications given in Remark 5.4.

Let h_t be the function defined on $S^1 \times \mathbb{D}^2$. As we explained in Section 5.2, we expect that $h_t(b_3)$ can explain the contribution of b_3. However, for this case, $h_t(b_3)$ cannot do that. This is because the important

factor is the number of strands of b_3, not that b_3 is braided. Thus, we define a trivial braid b_3^t such that b_3^t and b_3 have the same number of strands as

$$b_3^t := \varphi_3^{-1}(\{(\theta, x_0, y_0) \in S^1 \times \mathbb{D}^2 \mid (0, x_0, y_0) \in \varphi_3(b_3)\}) \subset D(\partial S_3).$$

Then, one obtains

$$h_t \colon S^1 \times \mathbb{D}^2 \xrightarrow{\varphi_1} \pi^{-1}(\partial S_1) \xrightarrow{\Phi_1 \circ \tau_0} \pi^{-1}(\partial S_4) \overset{\varphi_4'}{\simeq} S^1 \times \mathbb{D}^2, \quad (\theta, x, y) \mapsto (\theta, r_0 x + c_1, r_0 y + c_2),$$

where r_0 is a positive constant number less than 1 and Φ_1 is a Hamiltonian isotopy. We note that in Section 5.2, we needed a Hamiltonian isotopy. In a similar way, we can construct a Hamiltonian isotopy Φ_1. Then $h_t(\bar{b}_1^\circ)$ represents the added strands in b_4', which correspond to $\tau_0\big(L \cap \pi^{-1}(\pi(N_p))\big)$.

Similarly, if b_i contributes the construction of b_j'' and if the center of a singular disk corresponding to b_i is neither the same point nor the antipodal point of the center of the singular disk corresponding to b_j'', then the contribution of b_i on b_j'' can be described by a map like h_t. A map is of *trivial type* if a map is of the same type with h_t, because a map of trivial type adds strands which are not braided with each other.

Then, we can describe the effect of τ_0 on $\mathcal{B}_\psi$ as a matrix

$$T_{0, \mathcal{B}_\psi} = \begin{pmatrix} 0 & i & 0 & 0 & 0 & h_t \\ h_1 & 0 & h_2 + h_3 & 0 & 0 & i_t \\ 0 & 0 & \mathrm{id} & 0 & 0 & 0 \\ 0 & 0 & h_t & 0 & i & 0 \\ 0 & 0 & i_t & h_1 & 0 & h_2 + h_3 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{id} \end{pmatrix}.$$

Among the entries, h_1, i and id are of scaling type, h_2 and h_3 are of the first and second singular types, and h_t and i_t are of trivial type.

Remark 5.8 A ψ of generalized Penner type is a product of Dehn twists. In the general case, when we apply ψ, each Dehn twist is followed by a Hamiltonian isotopy as σ_2^{-1} is followed by Φ_t in step two. Let $\psi_H = (\Phi_{1,1} \circ \delta_1) \circ \cdots \circ (\Phi_{l,1} \circ \delta_l)$, where $\psi = \delta_1 \circ \cdots \circ \delta_l$, δ_i is a Dehn twist, and $\Phi_{i,t}$ is a Hamiltonian isotopy which follows δ_i.

After applying the Hamiltonian isotopy, the effect of a Dehn twist τ_i (resp. σ_j^{-1}) on $\mathcal{B} \in \mathbb{B}$ is described by a matrix $T_{i, \mathcal{B}}$ (resp. $\Sigma_{j, \mathcal{B}}$), whose entries are sums of maps of four types.

6 Proof of Theorem 1.5

In Sections 4 and 5, we generalized the notion of weights and linear algebra on weights. In this section, we prove our main theorem, ie Theorem 1.5, by using those generalizations.

6.1   Limit of a sequence of braids

By Lemma 5.1, one obtains braid sequences $\{b(\psi^m(L), S_i)\}_{m \in \mathbb{N}}$, where L is carried by $\mathcal{B}_\psi$, and S_i is a singular disk of $\mathcal{B}_\psi^*$. In the present subsection, we construct a limit of $\{b(\psi^m(L), S_i)\}_{m \in \mathbb{N}}$ as $m \to \infty$.

We argue with the above example, ie

$$M = P(\alpha, \beta_1, \beta_2), \quad \psi = \tau_0 \circ \sigma_1^{-1} \circ \sigma_2^{-1}, \quad \dim M = 4$$

For convenience, let

$$\mathcal{B} := \mathcal{B}_\psi, \quad \mathcal{B}' := F_{\sigma_2^{-1}}(\mathcal{B}), \quad \mathcal{B}'' := F_{\sigma_1^{-1}}(\mathcal{B}'),$$

and let S_i, S_i' and S_i'' denote singular disks of $\mathcal{B}$, $\mathcal{B}'$ and $\mathcal{B}''$. Using notation from Sections 5.2 and 5.3, we have matrices $T_{0,\mathcal{B}''}$, $\Sigma_{1,\mathcal{B}'}$ and $\Sigma_{2,\mathcal{B}}$. Then we obtain $\Psi = T_{0,\mathcal{B}''} \cdot \Sigma_{1,\mathcal{B}'} \cdot \Sigma_{2,\mathcal{B}}$ by defining a multiplication of maps as the composition of them. Note that a product of two arbitrary matrices is not defined since a composition of two arbitrary functions is not defined. For example, an input of $\Sigma_{2,\mathcal{B}}$ and an output of $T_{0,\mathcal{B}''}$ are tuples of braids on singular disks of $\mathcal{B}^*$. Thus, $\Sigma_{2,\mathcal{B}} \cdot T_{0,\mathcal{B}''}$ is defined. However, $T_{0,\mathcal{B}''} \cdot \Sigma_{2,\mathcal{B}}$ is not defined since an input of $T_{0,\mathcal{B}''}$ is a tuple of braids on singular disks of $\mathcal{B}^*$, but an output of $\Sigma_{2,\mathcal{B}}$ is a tuple of braids on singular disks of $\mathcal{B}'^*$.

Let $b_{i,m} = b(\psi^m(L), S_i)$. Then

$$\begin{pmatrix} b_{1,m} \\ b_{2,m} \\ b_{3,m} \\ b_{4,m} \\ b_{5,m} \\ b_{6,m} \end{pmatrix} := \Psi^m \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix}.$$

Thus, in order to keep track of braid sequences $\{b_{i,m}\}_{m \in \mathbb{N}}$, it is enough to keep track of Ψ^m.

Every entry of Ψ^m is a sum of compositions of $3m$ maps. The image of a composition of $3m$ maps is a solid torus. By Remark 5.5, the radius of each solid torus appearing in Ψ^m decreases exponentially and converges to zero as $m \to \infty$.

In order to be more precise, we consider ψ_H which is defined in Remark 5.8. One observes

$$b_{i,m} \subset \psi_H^m(N(\mathcal{B}_\psi)) \cap D(\partial S_i)$$

for all $m \in \mathbb{N}$ and $i = 1, \ldots, 6$. Let

$$B_{i,m} := \psi_H^m(N(\mathcal{B}_\psi)) \cap D(\partial S_i).$$

Then $B_{i,m}$ is the disjoint union of solid tori. Each solid torus in $B_{i,m}$ is the image of a composition of $3m$ maps, appearing in Ψ^m. Conversely, for each composition of $3m$ maps appearing in Ψ^m, the image is a solid torus contained in $B_{i,m}$. The radii of solid tori in $B_{i,m}$ are decreasing exponentially and are converging to zero as $m \to \infty$.

Since $B_{i+1,m} \subset B_{i,m}$ for all $m \in \mathbb{N}$, there is a limit

$$B_{i,\infty} := \lim_{m \to \infty} B_{i,m} = \bigcap_{m \in \mathbb{N}} B_{i,m}.$$

Thus, $B_{i,\infty}$ is the union of infinite strands as a subset of $D(\partial S_i)$ and

$$\lim_{m \to \infty} b_{i,m} = B_{i,\infty}$$

as a sequence of closed sets in $D(\partial S_i)$.

Remark 6.1   (1)   We have constructed a sequence of specific representatives $\{b_{i,m}\}_{m \in \mathbb{N}}$ such that

$$\lim_{m \to \infty} b_{i,m} = B_{i,\infty}.$$

For the purposes of extending the lamination to the singular and regular disks in Sections 6.2 and 6.3, we assume that the limit $B_{i,\infty}$ is a specific closed subset in $D(\partial S_i)$.

(2)   Each strand of $B_{i,\infty}$ corresponds to an infinite sequence $\{f_m\}_{m \in \mathbb{N}}$ such that $f_1 \circ \cdots \circ f_{3m}$ appears in Φ^m for all $m \in \mathbb{N}$.

6.2   Lagrangian lamination on a singular disk

Let ψ be of generalized Penner type and let L be a Lagrangian submanifold which is carried by $\mathcal{B}_\psi$. In the previous sections, on each singular disk S_i, we gave an inductive description of a sequence $\{b_{i,m} := b(\psi^m(L), S_i)\}_{m \in \mathbb{N}}$. There is a limit $B_{i,\infty}$ of the sequence, which is independent of L. In this present subsection, we will construct a Lagrangian lamination $\mathcal{L}_i \subset \pi^{-1}(S_i)$ from $B_{i,\infty}$.

Lemma 6.2   *Let ψ be of generalized Penner type. For each singular disk S_i of $\mathcal{B}_\psi$, there is a Lagrangian lamination $\mathcal{L}_i \subset D(S_i)$, such that if L is a Lagrangian submanifold of M which is carried by $\mathcal{B}_\psi$, then for every $m \in \mathbb{N}$, there is a Lagrangian submanifold L_m which is Hamiltonian isotopic to $\psi^m(L)$ and $L_m \cap D(S_i)$ converges to $\mathcal{L}_i$ as a sequence of closed subsets.*

Proof   Let ψ be of generalized Penner type, ie $\psi = \delta_1 \circ \cdots \circ \delta_l$, where δ_k is a Dehn twist τ_i or σ_j^{-1}. We will use similar notation as the previous subsections; for example, S_i denotes a singular disk of $\mathcal{B}_\psi$, Ψ denotes a matrix corresponding to ψ, $\varphi_i : D(\partial S_i) \xrightarrow{\sim} S^{n-1} \times \mathbb{D}^n$ denotes the identification induced from the fixed coordinate chart on S_i, and so on.

We will assume that L_m in Lemma 6.2 is $\psi_H^m(L)$ where ψ_H is defined in Remark 5.8. Then $\mathcal{L}_i$ is the limit of $\psi_H^m(L) \cap D(S_i)$ as $m \to \infty$. Thus, $\mathcal{L}_i \cap D(\partial S_i)$ is the limit of $\psi_H^m(L) \cap D(\partial S_i)$, ie $\mathcal{L}_i \cap D(\partial S_i) = D_{i,\infty}$. We will construct a Lagrangian lamination $\mathcal{L}_i$ when $B_{i,\infty}$ is given. Then we will prove that Lemma 6.2 holds with the constructed $\mathcal{L}_i$.

Construction of $\mathcal{L}_i$ As we mentioned in Remark 6.1, each strand of $B_{i,\infty}$ is identified with an infinite sequence $\{f_m\}_{m \in \mathbb{N}}$ such that $f_1 \circ \cdots \circ f_{lk}$ appears in Ψ^k for all $k \in \mathbb{N}$. For each strand $\{f_m\}_{m \in \mathbb{N}}$ of $B_{i,\infty}$, we will construct a Lagrangian submanifold of $D(S_i)$ whose boundary agrees with the strand $\{f_m\}_{m \in \mathbb{N}}$ in the construction part.

First, for a given strand $\{f_m\}_{m \in \mathbb{N}}$, let f_1 be of trivial type. Then the strand is identified with a sphere

$$\{(\theta, x_1, \ldots, x_n) \mid \theta \in S^{n-1}\} \subset S^{n-1} \times \mathbb{D}^n \overset{\varphi_i}{\simeq} D(\partial S_i),$$

where x_i is a constant. A subsequence $\{f_m\}_{m \geq 2}$ determines constants x_i. Let

$$D := \{(p, x_1, \ldots, x_n) \mid p \in S_i\} \subset \mathbb{D}^n \times \mathbb{D}^n \overset{\varphi_i}{\simeq} D(S_i).$$

Then $\varphi_i(D)$ is a Lagrangian disk in $D(S_i)$, whose boundary agrees with the strands $\{f_m\}_{m \in \mathbb{N}}$.

Second, let f_1 be not of trivial type, but there exists $m \in \mathbb{N}$ such that f_m is of trivial type. Let $k > 1$ be the smallest number such that f_k is of trivial type appearing in $\{f_m\}_{m \in \mathbb{N}}$. Then $\tilde{\psi} = \delta_{k_0} \circ \cdots \circ \delta_l \circ \delta_1 \circ \cdots \circ \delta_{k_0-1}$, where $k_0 \cong k \pmod{l}$, is of generalized Penner type satisfying the following: $\mathcal{B}_{\tilde{\psi}}$ has a singular disk $\tilde{S}_j$ such that $\tilde{B}_{j,\infty}$, the limit of the braid sequence corresponding to $\tilde{\psi}$ and $\tilde{S}_j$, has a strand identified with $\{f_m\}_{m \geq k}$. Thus, there is a Lagrangian disk in $D(\tilde{S}_j)$ whose boundary agrees with $\{f_m\}_{m \geq k}$. Let D denote the Lagrangian disk in $D(\tilde{S}_j)$. Then there is a connected component of

$$\left((\Phi_{1,1} \circ \delta_1) \circ \cdots \circ (\Phi_{k_0,1} \circ \delta_k) \right)(D) \cap D(S_i)$$

whose boundary is $\{f_m\}_{m \in \mathbb{N}}$, where $\Phi_{i,t}$ is a Hamiltonian isotopy mentioned in [Remark 5.8](#).

To summarize, if there is at least one map of trivial type in $\{f_m\}_{m \in \mathbb{N}}$, then we have a Lagrangian submanifold in $D(S_i)$, whose boundary agrees with $\{f_m\}_{m \in \mathbb{N}}$. Let $\mathcal{L}_{i,\infty}$ be the union of those Lagrangian submanifolds.

Finally, suppose that f_m is not of trivial type for any $m \in \mathbb{N}$. Then, for all $k \in \mathbb{N}$, we will construct a sequence $\{f_m^k\}_{m \in \mathbb{N}}$ for each $k \in \mathbb{N}$, satisfying

(1) $\{f_m^k\}_{m \in \mathbb{N}}$ is a strand of $B_{i,\infty}$;

(2) if $m \leq kl$, then $f_m^k = f_m$;

(3) there exists a constant $N_k \in \mathbb{N}$ such that $f_{kl+N_k}^k$ is of trivial type.

To prove the existence of these sequences $\{f_m^k\}_{m \in \mathbb{N}}$ for all $k \in \mathbb{N}$, we use the fact that the limits $B_{i,\infty}$ depend only on ψ and are independent of L. Let k be a fixed positive integer. Then $f_1 \circ \cdots \circ f_{kl}$ explains an impact of $b_{i,0} = b(L, S_i)$ on $b_{j,k} = b(\psi^k(L), S_j)$ for some i and j.

Let consider $\tilde{b}_{i,m} = b\left(\psi^m(\psi^N(L)), S_i \right) = b_{i,m+N}$ for a sufficiently large integer N. Then $\tilde{b}_{i,0}$ is given by a union of images of $g_1 \circ \cdots \circ g_{Nl}$ which appears in the ith row of Ψ^N. If we assume that there is at least one compact case having two or more plumbing points, then for a sufficiently large N, there exists a sequence of functions $g_1, \ldots, g_{Nl}$ such that $g_1 \circ \cdots \circ g_{Nl}$ appears in the ith row of Ψ^N and g_t is of trivial type for some $t \in [1, Nl]$. The reason is as follows: First, when we apply a Dehn twist along the compact core with two or more plumbing points, a function g_t of trivial type appears. The function g_t appears in a specific row. By applying ψ sufficiently many times, ie N times, one can guarantee that g_t appears in i^{th} row. This is because every Dehn twist along each compact core appears in ψ.

In $\tilde{b}_{i,kl} = b_{i,kl+Nl}$, there is a strand satisfying the last two conditions, and thus it guarantees the existence of $\{f_m^k\}_{m \in \mathbb{N}}$ for all $k \in \mathbb{N}$, assuming at least one compact core has two or more plumbing points. We note that the assumption excludes only one case, the plumbing of one positive and one negative sphere plumbed at only one point. The excluded case can be easily handled directly. For more detail, see Remark 6.3

Without loss of generality, there is a strand $\{f_m^k\}_{k \in \mathbb{N}}$ of $B_{i,\infty}$ for each $k \in \mathbb{N}$. These strands converge to $\{f_m\}_{m \in \mathbb{N}}$ as $k \to \infty$. Moreover, by definition of $\mathscr{L}_{i,\infty}$, the boundary of $\mathscr{L}_{i,\infty}$ contains strands $\{f_m^k\}_{m \in \mathbb{N}}$ for all $k \in \mathbb{N}$. Thus, the strand $\{f_m\}_{m \in \mathbb{N}}$ is contained in the boundary of $\mathscr{L}_i := \overline{\mathscr{L}_{i,\infty}}$, ie the closure of $\mathscr{L}_{i,\infty}$.

Remark 6.3 If there is no sphere with two or more plumbing points, then every sphere is plumbed at only one point. Thus, there is exactly one positive sphere and one negative sphere plumbed at one point. In this case, we can construct a Lagrangian lamination $\mathscr{L}$ on M by spinning. This is because only two spheres are plumbed, thus there is a plenty of symmetry, which comes from the symmetry of spheres. Then, $\mathscr{L}_i := \mathscr{L} \cap D(S_i)$ is a Lagrangian lamination which we want to construct in Lemma 6.2.

Convergence to $\mathscr{L}_i$ Let $L_m := \psi_H^m(L)$. We defined ψ_H in the fourth step of the proof of Lemma 5.1. We will prove that $L_m \cap D(S_i)$ converges to $\mathscr{L}_i$.

First, we will show that

$$(6\text{-}1) \qquad \lim_{m \to \infty} (L_m \cap D(S_i)) = \lim_{m \to \infty} \left(\psi_H^m(N(\mathscr{B}_\psi)) \cap D(S_i) \right).$$

Since $\psi_H(N(\mathscr{B}_\psi)) \subset N(\mathscr{B}_\psi)$,

$$\psi_H^{m+1}(N(\mathscr{B}_\psi)) \cap D(S_i) \subset \psi_H^m(N(\mathscr{B}_\psi)) \cap D(S_i) \quad \text{for all } m \in \mathbb{N}.$$

Thus, we have the limit

$$\lim_{m \to \infty} \left(\psi_H^m(N(\mathscr{B}_\psi)) \cap D(S_i) \right) = \bigcap_m \left(\psi_H^m(N(\mathscr{B}_\psi)) \cap D(S_i) \right).$$

If we equip M with a Riemannian metric g, then $d_H\left(\psi_H^m(\mathscr{B}_\psi), \psi_H^m(N(\mathscr{B}_\psi)) \right)$, where d_H is the Hausdorff metric induced by g, converges to zero as $m \to \infty$ for the same reason that $B_{i,m} := \psi_H^m(N(\mathscr{B}_\psi)) \cap D(\partial S_i)$ converges to an infinite braid $B_{i,\infty}$ in the last part of Section 6.1.

Since for a large integer N_0, L_{N_0} intersects $D(S_j)$ for any singular disk S_j, and $L_{m+N_0} \cap D(S_j)$ intersects every connected component of $\psi_H^m(N(\mathscr{B}_\psi)) \cap D(S_i)$. Thus,

$$0 \leq \lim_{m \to \infty} d_H(L_{m+N_0} \cap D(S_i), \psi_H^m(N(\mathscr{B}_\psi)) \cap D(S_i))$$

$$\leq \lim_{m \to \infty} [d_H(L_{m+N_0} \cap D(S_i), \psi_H^m(\mathscr{B}_\psi) \cap D(S_i)) + d_H(\psi_H^m(\mathscr{B}_\psi) \cap D(S_i), \psi^m(N(\mathscr{B}_\psi)) \cap D(S_i))]$$

$$\leq \lim_{m \to \infty} 2 d_H(\psi_H^m(\mathscr{B}_\psi) \cap D(S_i), \psi_H^m(N(\mathscr{B}_\psi)) \cap D(S_i))$$

$$= 0.$$

This proves (6-1). Let $\mathbb{L}_i$ be the limit in (6-1).

Second, we show that $\mathbb{L}_i$ is $\mathscr{L}_i$. By the construction of $\mathscr{L}_i$, we know that

$$\mathscr{L}_i \subset \psi_H^m(N(\mathscr{B}_\psi)) \cap D(S_i) \quad \text{for every } m \in \mathbb{N}.$$

It implies that $\mathscr{L}_i \subset \mathbb{L}_i$. Moreover,

$$\mathscr{L}_i \cap \pi^{-1}(\partial S_i) = \mathbb{L}_i = B_{i,\infty} \cap D(S_i).$$

Because every connected component of $\mathbb{L}_i$ has a boundary on ∂S_i, this shows $\mathscr{L}_i = \mathbb{L}_i$. □

6.3 Lagrangian lamination on a regular disk

In the previous subsection, we constructed Lagrangian laminations on singular disks, when boundary data for singular disks were given. In the present subsection, first, we will define boundary data for a regular disk. Second, we will construct Lagrangian laminations on regular disks from the given data. Finally, we will prove Theorem 1.5 as a corollary of Lemmas 6.2 and 6.5.

Before defining the boundary data, we remark that, $\overline{\pi^{-1}(R_i^\circ)}$ is symplectomorphic to $DT^*\mathbb{D}^n$, where $\mathbb{D}^n$ is a disk, by Remark 4.5. Similar to Definition 4.10, let $D(R_j)$ (resp. $D(\partial R_j)$) denote the $\mathbb{D}^n$–bundle $\overline{\pi^{-1}(R_i^\circ)}$ (resp. $\partial\overline{\pi^{-1}(R_i^\circ)}$) over R_j (resp. ∂R_j).

We define a data $c_{j,m}$ on the boundary of a regular disk R_j for $\psi^m(L)$, by setting

$$c_{j,m} := L_m \cap D(\partial R_j).$$

We defined $L_m := \psi_H^m(L)$ in the proof of Lemma 6.2. Note that $c_{j,m}$ is a closed subset, not a class of a closed subset.

To obtain a limit of $c_{j,m}$, we consider

$$C_{j,m} := \psi_H^m(N(\mathscr{B}_\psi)) \cap D(\partial R_j),$$

as we did in Section 6.1. Since $\psi_H^m(N(\mathscr{B}_\psi)) \subset N(\mathscr{B}_\psi)$, $C_{j,m+1} \subset C_{j,m}$. Moreover, $C_{j,m}$ is the union of solid tori in $D(\partial R_j)$ when $n = 2$, or the union of $S^{n-1} \times \mathbb{D}^n$ for general n. If a symplectic manifold M is equipped with a Riemannian metric g, we can measure the radii of solid tori in $C_{j,m}$. The radii decrease exponentially and converge to zero as $m \to \infty$, for the same reason that radii of solid tori comprising $B_{i,m}$ decrease exponentially and converge to zero as $m \to \infty$ in Section 6.1. The limit of $c_{j,m}$ is given by

$$C_{j,\infty} = \lim_{m\to\infty} C_{j,m} = \bigcap_m C_{j,m}.$$

The next step is to smooth R_j. A regular disk R_j has corners. We will replace R_j with a smooth disk R_j'. This is because, at the end, a Lagrangian lamination will be given as graphs of closed sections. By smoothing R_j, it will be easier to handle closed sections.

To smooth R_j, we subtract a tubular neighborhood $N(\partial R_j) \subset R_j$ from R_j. Let $R_j' := R_j \setminus N(\partial R_j)$. Then R_j' is a smooth disk. We replace R_j with R_j'. To finish smoothing, we need to obtain boundary data for R_j' from $c_{j,m}$.

Each connected component of $c_{j,m}$ can be identified with a section of a bundle $D(\partial R_j)$ over ∂R_j. We can extend this section to a closed section of a bundle $\pi^{-1}(N(\partial R_j))$ over $N(\partial R_j)$ by computations. Then the graph of the extended section is a Lagrangian submanifold of $\pi^{-1}(N(\partial R_j))$. The boundary of the Lagrangian submanifold on $\partial R'_j$ makes up the boundary data for R'_j.

From now, we assume that a regular disk R_j is a smoothed disk. Lemma 6.4 claims that for a given data $c_{j,m}$ on a smoothed regular disk R_j, we can construct a Lagrangian submanifold $N_{j,m} \subset D(R_i)$ such that $\partial N_{j,m} = c_{j,m} \cap D(R_i)$.

Lemma 6.4 *Let Q be a closed subset of $\partial T^* \mathbb{D}^n$ such that there exists a disjoint union L of Lagrangian disks in $T^* \mathbb{D}^n$, which are transversal to fibers, such that $L \cap \partial T^* \mathbb{D}^n = Q$. Then we can construct a Lagrangian submanifold L uniquely up to Hamiltonian isotopy through Lagrangians transverse to the fibers.*

Proof To prove Lemma 6.4, we consider a identification $\varphi \colon \partial T^* \mathbb{D}^n \xrightarrow{\sim} S^{n-1} \times \mathbb{D}^n$ which is defined as follows. If there is a global coordinate charts of the zero section $\mathbb{D}^n$ of $T^* \mathbb{D}^n$, then it induces an identification between $\mathbb{D}^n \times \mathbb{D}^n$ and $T^* \mathbb{D}^n$. By restricting the identification on $\partial T^* \mathbb{D}^n$, one obtains $\varphi \colon \partial T^* \mathbb{D}^n \xrightarrow{\sim} S^{n-1} \times \mathbb{D}^n$. With the fixed identification φ, $\varphi(Q) = \varphi(\partial L)$ is isotopic to a union of spheres

$$\{S^{n-1} \times p_1, \ldots, S^{n-1} \times p_m \mid p_i \in \mathbb{D}^n, m \text{ is the number of component of } L\}.$$

This is because $\varphi(L)$ is a union of Lagrangian disks in $\mathbb{D}^n \times \mathbb{D}^n \overset{\varphi}{\simeq} T^* \mathbb{D}^n$.

The proof of Lemma 6.4 consists of two parts: the construction of L and the uniqueness of L.

Construction We start the proof with the simplest case, ie Q consists of only one strand.

By fixing coordinate charts on $\mathbb{D}^n$, we can write down Q as a section of a disk bundle $\partial T^* \mathbb{D}^n$ over $\partial \mathbb{D}^n$, ie

$$Q := \{f_1(x_1, \ldots, x_n) dx_1 + \cdots + f_n(x_1, \ldots, x_n) dx_n \mid x_1^2 + \cdots + x_n^2 = 1\}.$$

Then, the simplest case is proved by determining a function $\phi \colon \mathbb{D}^n \to \mathbb{R}$ such that $d\phi = f_1 dx_1 + \cdots + f_n dx_n$ on $\partial \mathbb{D}^n$. The graph of $d\phi$ is a Lagrangian submanifold which we would like to find. Note that there are infinitely many ϕ satisfying the conditions, but the Hamiltonian isotopy class of the graph of $d\phi$ is unique through Lagrangians transverse to the fibers.

If Q has two or more connected components l_i, then we can write l_i as a section over $\partial \mathbb{D}^n$. For each i, we need to determine functions $\phi_i \colon \mathbb{D}^n \to \mathbb{R}$ such that $d\phi_i$ agrees with l_i on $\partial \mathbb{D}^n$. Moreover, to avoid self-intersection, they should not be equal, ie $d\phi_i \neq d\phi_j$ for all $i \neq j$. Then, the union of graphs of $d\phi_i$ on $T^* \mathbb{D}^n$ is a Lagrangian submanifold L which we want to construct.

We discuss with the simplest nontrivial case, ie Q has two connected components l_0 and l_1, and the dimension $2n = 4$. Without loss of generality, we assume that l_0 is the zero section. Furthermore, we can assume that $\phi_0 \equiv 0$. We only need to determine ϕ_1 such that $d\phi_1$ does not vanish everywhere.
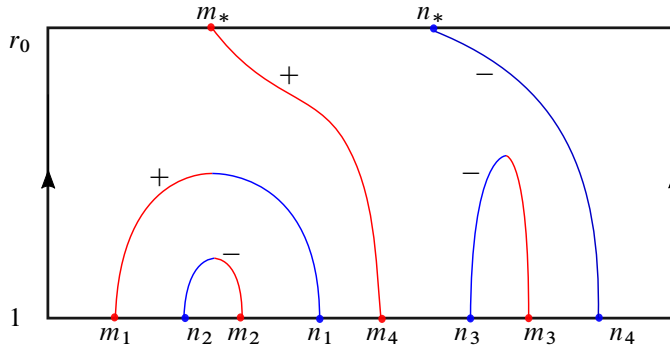
Figure 21: Example of a collection $\mathscr{C}$ on $[r_0, 1] \times S^1$.

We assume that there exists ϕ_1 satisfying the conditions. Then we will collect combinatorial data from ϕ_1, and we will construct a function $\tilde{\phi}_1$ satisfying conditions given by the combinatorial data. Through this, we will see what combinatorial data we need. We will end the construction by explaining how to obtain the combinatorial data from the given Q.

For convenience, we will use the polar coordinates instead of (x, y)–coordinates on $\mathbb{D}^2$. Let r_0 be a small positive number. We restrict the function ϕ_1 on $[r_0, 1] \times S^1$. On $\{1\} \times S^1 = \partial \mathbb{D}^2$, $d\phi_1$ agrees with l_1. On $\{r_0\} \times S^1$, $d\phi_1$ is approximately a constant section

$$a\,dx + b\,dy = a(\cos\theta\,dr - r_0 \sin\theta\,d\theta) + b(\sin\theta\,dr + r_0 \cos\theta\,d\theta),$$

where $d\phi_1(0, 0) = a\,dx + b\,dy$ and (x, y) are the standard coordinate charts of $\mathbb{D}^2$.

We remark that on $\{r_0\} \times S^1$, the pair of graphs of $d\phi_i|_{\{r_0\} \times S^1}$ represents the trivial braid under the identification induced from the (x, y)–coordinates. Similarly, on $[r_0, 1] \times S^1$, the pair $(d\phi_0 \equiv 0, d\phi_1)$ implies an isotopy between two representatives of the trivial braid.

For every $r_* \in [r_0, 1]$, we can find all local maxima and minima of a function

$$\theta \mapsto \phi_1(r_*, \theta).$$

We mark (r_*, θ_*) as a red (resp. blue) point if the above function has a local maxima (resp. minima) at θ_*. If $r_* = 1$, there are same number of red/blue marked points on $\{1\} \times S^1$, and there is only one red/blue marked point on $\{r_0\} \times S^1$. On $[r_0, 1] \times S^1$, we have a collection $\mathscr{C}$ of curves shaded red and blue. If a curve in $\mathscr{C}$ is not a circle, then the curve has two end points on the boundary of $[r_0, 1] \times S^1$. There are exactly two curves connecting both boundary components of $[r_0, 1] \times S^1$, and those two curves have end points of the same color.

If we write $d\phi_1 = f\,d\theta + g\,dr$, then f is zero on curves in $\mathscr{C}$. Since $d\phi_1$ does not vanish, g cannot be zero on the curves. Thus, we can assign the sign of g for each curve. Figure 21 is an example of a collection $\mathscr{C}$.

Conversely, if we have a collection $\mathscr{C}$ of curves such that each curve is shaded red and blue and is equipped with a sign, then we can draw a graph of $\tilde{\phi}_1$ roughly. This is because the collection $\mathscr{C}$ determines the sign of horizontal directional derivative of $\tilde{\phi}_1$, ie $d\tilde{\phi}_1(\partial_\theta)$ on every point of $[r_0, 1] \times S^1$, and vertical directional derivative of $\tilde{\phi}_1$, ie $d\tilde{\phi}_1(\partial_r)$ on the curves. From these, one obtains a (rough) graph of $\tilde{\phi}_1$. Thus, in order to determine a function ϕ_1, it is enough to determine a collection $\mathscr{C}$ of curves in $[r_0, 1] \times S^1$ from the given Q.

From now on, we will construct a collection $\mathscr{C}$ from the given Q. For the given Q, we assume that a connected component l_0 of Q is the zero section, without loss of generality. For the other connected component l_1, one has $f_1, g_1 \colon S^1 \to \mathbb{R}$ such that l_1 is the graph of $f_1 \, d\theta + g_1 \, dr$ on $\{1\} \times S^1 = \partial \mathbb{D}^2$. We know that Q represents the trivial braid with respect to the standard (x, y)–coordinates of $\mathbb{D}^2$. Thus, there is an isotopy $\Gamma \colon [r_0, 1] \times S^1 \to \mathbb{D}^2$ such that

$$\Gamma(1, \theta) = (f(\theta), g(\theta)), \quad \Gamma(r_0, \theta) = (Ar_0 \cos \theta, A \sin \theta),$$

$$\Gamma(t, \theta) \neq (0, 0) \quad \text{for all } (t, \theta) \in [r_0, 1] \times S^1,$$

where A is a constant.

For every $r \in [r_0, 1]$, let $\gamma_r(\theta) = \Gamma(r, \theta)$. Then, γ_r is a closed curve in $\mathbb{D}^2$, for all r. Moreover, Γ is a path connecting γ_1 and γ_{r_0} in the loop space of $(\mathbb{D}^2)^\circ$ without touching the origin.

We mark (r, θ) on $[r_0, 1] \times S^1$ as a red (resp. blue) point if $\gamma_r(\theta)$ intersects dr–axis from right to left (resp. from left to right). These marked points comprise curves in $[r_0, 1] \times S^1$, and we have a collection $\mathscr{C}$ of curves, shaded red and blue, in $[r_0, 1] \times S^1$. We know that γ_1 has intersection points. The number of intersection points is an even number. When r decreases, there is a series of creations/removals of intersection points, which are given by finger moves along the dr–axis. Each finger move does not touch the origin. Thus, for a curve in $\mathscr{C}$, every intersection point composing the curve lies on either the positive dr–axis or the negative dr–axis. Then, we can assign a sign for each curve in $\mathscr{C}$.

Figure 22 is an example of Γ, corresponding to the case described by Figure 21.

The upper left of Figure 22 is γ_1 and the upper right is γ_{r_0}. Through the first arrow, we observe a finger move removing two intersection points. Those two intersection points correspond to m_2, a local maxima shaded red, and n_2, a local minima shaded blue. Thus, we obtain a curve connecting m_2 and n_2 in Figure 21. Moreover, the intersection points lie in the negative part of the dr–axis. Thus, we assign a negative sign to the curve. Similarly, we observe there are finger moves removing intersection points. We obtain curves connecting m_i and n_i for $i = 1, 2, 3$ in Figure 21. After the finger moves, there are only two intersection points corresponding to m_* and n_*, and we obtain curves connecting m_4 (resp. n_4) and m_* (resp. n_*).

We have constructed a collection $\mathscr{C}$ of curves on $[r_0, 1] \times S^1$ from an isotopy Γ. Thus, we can obtain a function $\phi_1 \colon [r_0, 1] \times S^1 \to \mathbb{R}$. In order to complete the proof, we need to extend ϕ_1 into a small disk with radius r_0. To extend ϕ_1, we assume that

$$\phi_1(x, y) = Ar \sin \theta = Ay$$

on the small disk.
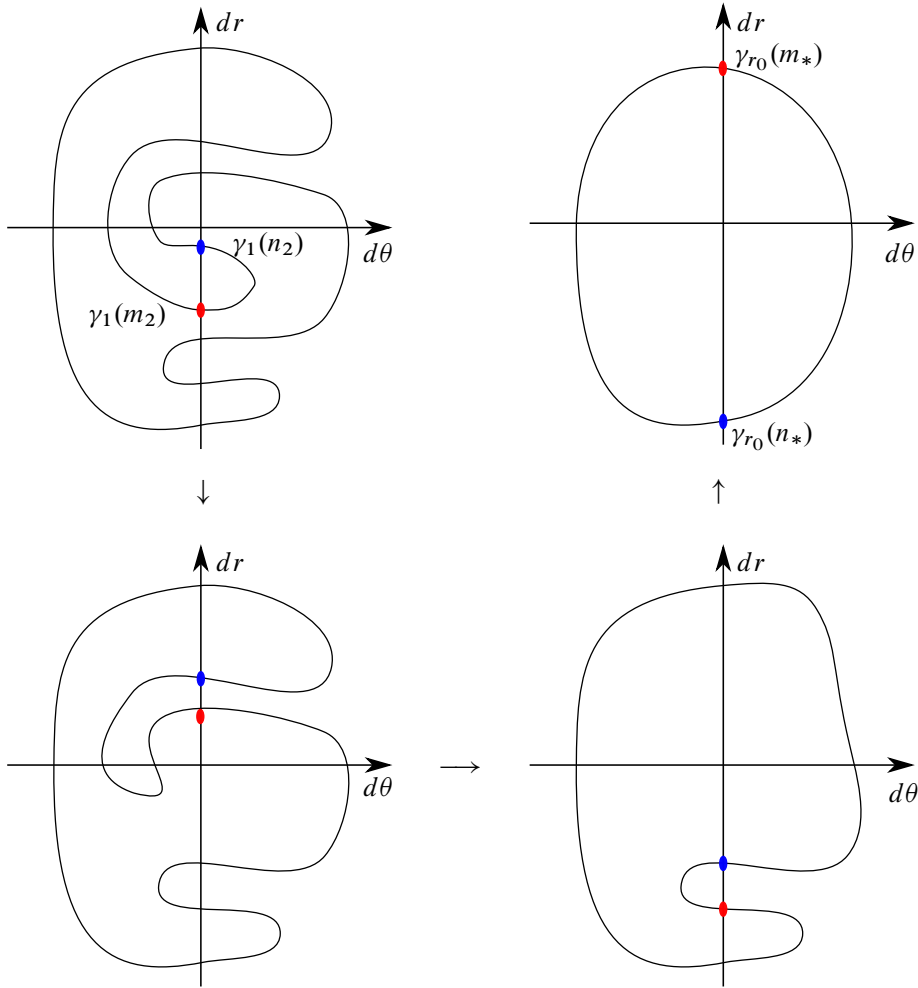
Figure 22: Creation of a collection $\mathscr{C}$.

The situation for the general case is analogous. If Q has more connected components l_i for $i = 0, \ldots, k$, then we have to determine $\phi_i : \mathbb{D}^2 \to \mathbb{R}$ such that $d\phi_i = l_i$ on $\partial \mathbb{D}^2$, and $d\phi_i \neq d\phi_j$ for all $i \neq j$. We fix an isotopy Γ, and obtain a collection $\mathscr{C}$ of curves on $[r_0, 1] \times S^1$ from Γ. Each curve in $\mathscr{C}$ encodes restrictions on $d\phi_i - d\phi_j$ for some i and j. More precisely, $(\phi_i - \phi_j)$ has a local maxima (resp. minima) in the horizontal direction, only at a point of a curve shaded red (resp. blue), and $(d\phi_i - d\phi_j)(\partial_r)$ has the sign assigned on the curve. For the case of general dimension $2n$, we obtain combinatorial data from Q, ie a collection of curves on $[r_0, 1] \times S^{n-1}$ assigned a sign, and construct functions on $\mathbb{D}^n$ from the combinatorial data.

Uniqueness Recall that the construction consists of three steps. First, we choose an isotopy Γ connecting Q and the trivial representative of the trivial braid. Then, we obtained a collection $\mathscr{C}$ of curves from Γ, such that each curve encodes restrictions on $d\phi_i - d\phi_j$. The last step is to construct a set of functions $\{\phi_i : \mathbb{D}^n \to \mathbb{R}\}$.

The construction depends on choices in the first and last steps. More precisely, for the first step, the choice of isotopy Γ is not unique. If we choose an isotopy Γ, then there is a unique collection $\mathscr{C}$. However, a set $\{\phi_i\}$ of functions, which is constructed from the collection $\mathscr{C}$, is not unique. We will show that the Hamiltonian isotopy class of L, through Lagrangians transverse to the fibers, is independent of those choices.

First, we discuss the choice in the third step. Let us assume that we have a collection $\mathscr{C}$ of curves in $[r_0, 1] \times S^{n-1}$ and two sets of functions $\{\phi_i\}_i$ and $\{\zeta_i\}_i$ satisfying the restrictions encoded by $\mathscr{C}$. Then, by setting $\eta_{i,t} := (1-t)\phi_i + t\zeta_i$, we obtain a family of sets of functions such that every member of the family satisfies the restrictions encoded by $\mathscr{C}$.

Let L_t be the Lagrangian submanifold corresponding to $\{\eta_{i,t}\}$ for a fixed t. Then L_t is a Lagrangian isotopy connecting L_0, corresponding to $\{\phi_i\}$, and L_1, corresponding to $\{\zeta_i\}$. Since L_t is a disjoint union of Lagrangian disks in $T^*\mathbb{D}^n$, L_0 and L_1 are Hamiltonian isotopic. Thus, the Hamiltonian class of L through Lagrangians transverse to the fibers is independent of the choice of functions for the third step of the construction.

Before discussing the choice of the first step, note that a continuous change on a collection $\mathscr{C}$ does not make a change on the Hamiltonian isotopy class. More precisely, let $\mathscr{C}_0 = \{\gamma_1, \dots, \gamma_N\}$ be a collection of curves and let $\{\phi_i\}$ be a set of functions corresponding to $\mathscr{C}_0$. If $\{\gamma_{k,t}\}$ is a continuous family of curves with respect to t such that $\gamma_{k,0} = \gamma_k$ for all k, then we can obtain a continuous family $\{\phi_{1,t}, \dots, \phi_{N,t}\}$ such that $\phi_{i,0} = \phi_i$ and $\{\phi_{1,t}, \dots, \phi_{N,t}\}$ corresponds to $\mathscr{C}_t := \{\gamma_{1,t}, \dots, \gamma_{N,t}\}$. Then, it is easy to check that the Hamiltonian isotopy class of the union of graphs of $d\phi_{i,t}$ in $T^*\mathbb{D}^n$, through Lagrangians transverse to the fibers, is independent of t.

Finally, we will discuss the choice of Γ. Let Γ_0 and Γ_1 be two isotopies obtained from the given Q in the first step. Then we can understand Γ_0 and Γ_1 as paths on the loop space of the configuration space of $(\mathbb{D}^n)^\circ$. Since the loop space is simply connected, there is a continuous family $\{\Gamma_t\}_{t \in [0,1]}$ connecting γ_0 and γ_1.

Let $\mathscr{C}_t$ be the collection of curves obtained from Γ_t and let $\{\phi_i\}$ be a set of functions constructed from $\mathscr{C}_0$. There is $\{\phi_{i,t}\}$ corresponding to $\mathscr{C}_t$ such that $\phi_{i,0} = \phi_i$. Then, if L_t is the union of graphs of $d\phi_{i,t}$, then the Hamiltonian class of L_t is independent of t. This shows the uniqueness of L, up to Hamiltonian isotopy, through Lagrangians transverse to the fibers. $\square$

For a smoothed regular disk R_j, there is a sequence of data $c_{j,m}$ for each $m \in \mathbb{N}$. Then, we can construct a sequence of Lagrangian submanifolds $N_{j,m} \subset D(R_j)$ such that $N_{j,m} \cap \partial D(R_j) = c_{j,m}$. The following lemma, Lemma 6.5, claims that we can construct $N_{j,m}$ wisely, so that $N_{j,m}$ converges to a Lagrangian lamination $\mathcal{N}_j$ as m goes to ∞.

Lemma 6.5 *It is possible to construct $N_{j,m} \subset D(R_j)$ so that the sequence $N_{j,m}$ converges to a Lagrangian lamination $\mathcal{N}_j \subset D(R_j)$ as $m \to \infty$.*

Proof Let the boundary condition $c_{j,m}$ be the set $\{l_{1,m}, \ldots, l_{N_m,m}\}$, where $l_{i,m}$ is a connected component of $c_{j,m}$, or equivalently, $l_{i,m}$ is a strand of the braid represented by $c_{j,m}$. Note that $C_{j,m}$ is a disjoint union of solid tori in $D(\partial R_j)$, which is defined at the beginning of the present subsection. Then we can divide $c_{j,m}$ into a partition such that $l_{i,m}$ and $l_{j,m}$ are in the same subset if and only if $l_{i,m}$ and $l_{j,m}$ are in the same solid torus (or $S^{n-1} \times \mathbb{D}^n$, for a higher-dimensional case) in $C_{j,m}$. After that, we randomly choose a connected component $l_{s,m}$ from each subset of the partition.

By Lemma 6.4, there is $\phi_{s,m} \colon R_j \to \mathbb{R}$ such that $d\phi_{s,m} = l_{s,m}$ on ∂R_j. Then $\Gamma(d\phi_{s,m})$ is a Lagrangian disk in $D(R_j)$, where $\Gamma(d\phi_{s,m})$ is the graph of $d\phi_{s,m}$. We can choose a neighborhood $N(\Gamma(\phi_{s,m}))$ of $\Gamma(d\phi_{s,m})$ in $D(R_j)$, such that $N(\Gamma(d\phi_{s,m})) \simeq T^*\mathbb{D}^n$ and $N(\Gamma(d\phi_{s,m})) \cap D(\partial R_j)$ is the torus in $C_{j,m}$ containing $l_{s,m}$. Moreover, we can assume that

$$d_H(N(\phi_{s,m}), \Gamma(d\phi_{s,m})) < 2r^m,$$

where d_H is the Hausdorff metric induced by a fixed Riemannian metric and $r < 1$ is a small positive number.

We apply Lemma 6.4 to $\{l_{t,m+1} \in c_{j,m+1} \mid l_{t,m+1} \subset N(\Gamma(d\phi_{s,m}))\}$ in $N(\Gamma(d\phi_{s,m})) \simeq T^*\mathbb{D}^n$. Then we can construct $\phi_{t,m+1} \colon R_j \to \mathbb{R}$ such that $d\phi_{i,m+1} = l_{t,m+1}$ on ∂R_j and $\Gamma(d\phi_{s,m})$ is contained in $N(\phi_{s,m+1})$. We repeat this procedure inductively on $m \in \mathbb{N}$.

Let l be a strand of $C_{j,\infty}$. Then there is a sequence $l_{i_m,m} \in c_{j,m}$ such that $l_{i_m,m}$ converges to l. If we construct $\phi_{i,m}$ by repeating the above procedure, we know that

$$d_H(\Gamma(d\phi_{i_m,m}), \Gamma(d\phi_{i_n,n})) < 4r^{\max(m,n)}.$$

Thus, $d\phi_{i_m,m}$ converges. Moreover, by assuming that $\phi_{i,m}(p) = 0$ for every i and m, where p is a center of R_j, $\phi_{i_m,m}$ converges to a function ϕ. Then $\Gamma(d\phi)$ is a Lagrangian disk in $D(R_j)$ whose boundary is l, the stand of $C_{j,\infty}$. The union of $\Gamma(d\phi)$ is the Lagrangian lamination $\mathcal{N}_j$ which $N_{j,m}$ converges to. □

Proof of Theorem 1.5 By Lemma 6.2, there is a Lagrangian lamination $\mathcal{L}_i$ in $D(S_i)$, and by Lemma 6.5, there is a Lagrangian lamination $\mathcal{N}_j$ in $D(R_j)$. Moreover, every Lagrangian lamination agrees with each other along boundaries. Thus, we can glue them. Then we obtain a Lagrangian lamination $\mathcal{L}$ in M. □

6.4 A generalization

In the previous sections, we assumed that ψ is of generalized Penner type. In the present subsection, we discuss a symplectic automorphism $\psi \colon (M, \omega) \to (M, \omega)$, not necessarily to be of generalized Penner type, with some assumptions. In other words, we prove the following theorem.

Theorem 6.6 *Let $\psi \colon M \xrightarrow{\sim} M$ be a symplectic automorphism and let $\mathcal{B}_\psi$ be a Lagrangian branched submanifold such that $\psi(\mathcal{B}_\psi)$ is carried by $\mathcal{B}_\psi$. If the associated branched manifold $\mathcal{B}_\psi$ admits a*

decomposition into singular and regular disks, then there is a Lagrangian lamination $\mathcal{L}$ such that if L is a Lagrangian submanifold of M which is carried by $\mathcal{B}_\psi$ and if L is compatible with the decomposition of $\mathcal{B}_\psi^*$, then there is a Lagrangian submanifold L_m for all $m \in \mathbb{N}$, which is Hamiltonian isotopic to $\psi^m(L)$ and converges to $\mathcal{L}$ as closed sets as $m \to \infty$.

First, we assume that there is a Lagrangian branched submanifold $\mathcal{B}_\psi$ such that $\psi(\mathcal{B}_\psi)$ is (weakly) carried by $\mathcal{B}_\psi$. Then if a Lagrangian submanifold L is (weakly) carried by $\mathcal{B}_\psi$, then $\psi(L)$ is carried by $\mathcal{B}_\psi$. This is because the proof of Lemma 3.19 carries over with no change.

As mentioned in Section 4.2, we assume that $\mathcal{B}_\psi^*$ admits a decomposition into a union of finite number of singular disks $S_i \simeq \mathbb{D}^n$ and regular disks $R_j \simeq \mathbb{D}^n$.

Proof of Theorem 6.6 First, we define data on the boundary of each singular and regular disk, in the same way we did for the case of ψ of generalized Penner type. Then, on a regular disk R_j, the proofs of Lemma 6.4 and Lemma 6.5 carry over with no change. Thus, we can construct a Lagrangian lamination on $D(R_j)$.

On a singular disk S_i, we define the boundary data in the same way. In other words, the boundary data is defined by the isotopy class of $\psi^m(L) \cap D(\partial S_i)$. We also can obtain a matrix Ψ, which explains how the sequences of braids are constructed inductively. However, the rest of the proof of Lemma 6.2 does not carry over. This is because in the proof of Lemma 6.2, functions of trivial type have a key role. To use the same proof, we need to show that there are enough functions of trivial type. However, the assumptions cannot imply the existence of enough functions of trivial type.

For a singular disk S_i, let $\{f_m\}_{m \in \mathbb{N}}$ be a strand of the limit braid on S_i. We note that each strand can be identified to an infinite sequence of functions. We forget specific functions f_m, but remember their types. Then, we obtain a sequence of types. The sequence of types determines the "shape" of strand, for example, how many times the strand is rotated.

We can construct a symplectomorphism ϕ which is of generalized Penner type such that $\mathcal{B}_\phi$ has a singular disk S such that the limit braid assigned on S has a strand of the same shape. In Section 4.3, we constructed a Lagrangian submanifold $L_0 \subset D(S)$ such that ∂L_0 is the strand. Since $D(S) \simeq D(S_i)$, we assume that L_0 is a Lagrangian submanifold in $D(S_i)$. By scaling and translating L_0 inside $D(S_i)$, we obtain a Lagrangian submanifold whose boundary agrees with the strand.

The rest of the proof is the same as the proof of Theorem 1.5.   □

7 Application to Lagrangian Floer homology

One natural question following the construction of stable/unstable Lagrangian lamination is: how can we understand those constructed Lagrangian laminations in terms of Fukaya category? The purpose of Section 7 is to introduce one possible view-point of answering the question. More precisely, we expect

that a symplectic automorphism of Penner type will induce a *pseudo-Anosov autoequivalence* in terms of Fan, Filip, Haiden, Katzarkov and Liu [4].

Remark 7.1 There are two different definitions of pseudo-Anosov autoequivalence. One is defined by Dimitrov, Haiden, Katzarkov and Kontsevich in [3] and the other is defined in [4].

Roughly, we expect that, for a given ϕ of Penner type, by counting intersection numbers of a Lagrangian submanifold L and the stable/unstable Lagrangian laminations, we can define a *mass function* for ϕ. Then, ϕ will induce a pseudo-Anosov autoequivalence with respect to that mass function.

We do not prove the above claim in the current paper. However, we prove Theorem 7.3 which relates the intersection numbers with Lagrangian Floer theory.

In Section 7.1, we state Theorem 7.3. In Section 7.2, we will give a proof of Theorem 7.3. Moreover, we will prove Lemmas 7.7 and 7.8, in order to weaken the difficulties of applying Theorem 7.3 together with Example 7.9.

Remark 7.2   (1)   In order to do Lagrangian Floer theory, we should choose a suitable almost complex structure J. We will discuss our choice of almost complex structure in Section 7.1; see Remark 7.6.

(2)   If M is a surface, ie a 2–dimensional symplectic manifold, then $\widetilde{L}_i = L_i$, and Theorem 7.3 is claiming that the rank of Lagrangian Floer homology of L_1 and L_2 is the same to the intersection number of L_1 and L_2. This is already proven in [3, Lemma 2.18].

7.1 Setting

First, we state Theorem 7.3. Then, we will define the terms in Theorem 7.3.

Theorem 7.3 *Let M be a plumbing space of Penner type, and let $\eta\colon M \xrightarrow{\ \sim\ } M$ be the involution associated to M. Assume that a transversal pair $L_1, L_2 \subset M$ of Lagrangian submanifolds satisfies*

(1)   *$\eta(L_i) = L_i$ for $i = 0, 1$;*

(2)   *if $\widetilde{L}_i = L_i \cap M_i$, then $\widetilde{L}_i$ is a Lagrangian submanifold of $\widetilde{M}$ such that $\widetilde{L}_0$ and $\widetilde{L}_1$ are not isotopic to each other;*

(3)   *$L_0 \cap L_1 = \widetilde{L}_0 \cap \widetilde{L}_1$;*

(4)   *L_0 and L_1 are not isotopic to each other.*

Then

(7-1) $$\dim HF^0(L_1, L_2) + \dim HF^1(L_1, L_2) = i(\widetilde{L}_1, \widetilde{L}_2),$$

where $HF^k(L_1, L_2)$ denotes $\mathbb{Z}/2$–graded Lagrangian Floer homology over the Novikov ring of characteristic 2 and $i(\widetilde{L}_1, \widetilde{L}_2)$ denotes the geometric intersection number of $\widetilde{L}_1$ and $\widetilde{L}_2$ in the fixed surface $\widetilde{M}$.

In Section 7, we assume that our symplectic manifold M is a plumbing space

$$M = P(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l)$$

of Penner type defined as follows.

Definition 7.4 A plumbing space $M = P(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l)$ is of *Penner type* if α_i and β_j satisfy

(1) $\alpha_1, \ldots, \alpha_m$ and $\beta_1, \ldots, \beta_l$ are n–dimensional spheres,

(2) $\alpha_i \cap \alpha_j = \varnothing$ and $\beta_i \cap \beta_j = \varnothing$ for all $i \neq j$.

Note that $P(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l)$ is defined in Section 2.1.

From now on, we will define an involution $\eta \colon M \xrightarrow{\sim} M$, which is associated to M.

Involution η_0 on T^*S^n First, we will define an involution η_0 on T^*S^n. Let

$$S^n = \{x \in \mathbb{R}^{n+1} \mid |x| = 1\},$$
$$T^*S^n = \{(x, y) \in S^n \times \mathbb{R}^{n+1} \mid x \in S^n, \langle x, y \rangle = 0\}.$$

Then we define $\eta_0 \colon T^*S^n \xrightarrow{\sim} T^*S^n$ by

$$\eta_0(x_1, \ldots, x_{n+1}, y_1, \ldots, y_{n+1}) = (x_1, x_2, -x_3, \ldots, -x_{n+1}, y_1, y_2, -y_3, \ldots, -y_{n+1}).$$

Let

$$W_0 = \{(\cos\theta, \sin\theta, 0, \ldots, 0) \mid \theta \in [0, 2\pi]\} \subset S^n,$$
$$T^*S = \{(\cos\theta, \sin\theta, 0, \ldots, 0, -r\sin\theta, r\cos\theta, 0, \ldots, 0) \mid \theta \in [0, 2\pi], r \in \mathbb{R}\} \subset T^*S^n.$$

Then it is easy to check that T^*W_0 is the set of fixed points of η_0, ie $\eta_0^{\text{fixed}} = T^*W_0$.

Involution η associated to M First, we will construct an involution η_{α_i} and η_{β_j} on $T^*\alpha_i$ and $T^*\beta_j$ for every i and j. Note that $T^*\alpha_i, T^*\beta_j \subset M$.

For each α_i, we will choose a great circle $W_{\alpha_i} \subset \alpha_i$ such that W_{α_i} contains every plumbing point of α_i. Then there is a symplectic isomorphism $\phi_{\alpha_i} \colon T^*S^n \xrightarrow{\sim} T^*\alpha_i$ such that $\phi_{\alpha_i}(S^n) = \alpha_i$ and $\phi_{\alpha_i}(W_0) = W_{\alpha_i}$. One obtains an involution $\eta_{\alpha_i} \colon T^*\alpha_i \xrightarrow{\sim} T^*\alpha_i$ by setting

$$\eta_{\alpha_i} := \phi_{\alpha_i} \circ \eta_0 \circ (\phi_{\alpha_i})^{-1}.$$

Similarly, one obtains an involution $\eta_{\beta_j} \colon T^*\beta_j \xrightarrow{\sim} T^*\beta_j$.

Without loss of generality, one can assume that $\eta_{\alpha_i}(x) = \eta_{\beta_j}(x)$ for every $x \in T^*\alpha_i \cap T^*\beta_j$. Finally, the involution $\eta \colon M \xrightarrow{\sim} M$ is defined by

$$\eta(x) := \begin{cases} \eta_{\alpha_i}(x) & \text{if } x \in T^*\alpha_i, \\ \eta_{\beta_j}(x) & \text{if } x \in T^*\beta_j. \end{cases}$$

Let $\widetilde{M}$ be the set of fixed points of η, ie $\widetilde{M} = \{x \in M \mid \eta(x) = x\}$. It is easy to check that $\widetilde{M}$ is a 2–dimensional symplectic submanifold of M. Moreover, $\widetilde{M}$ is symplectomorphic to a plumbing space $P(S_{\alpha_1}, \ldots, S_{\alpha_m}, S_{\beta_1}, \ldots, S_{\beta_l})$ of Penner type. Note that S_{α_i} and S_{β_j} are embedded circles in α_i and β_j.

Definition 7.5 (1) The above η is called *the involution associated to M*.

(2) The above $\tilde{M}$ is called *the fixed surface of M*.

Remark 7.6 It is easy to check that our setting is a special case of Seidel and Smith [13]. More precisely, [13] considers Lagrangian Floer cohomology on a symplectic manifold carrying a symplectic involution. Under various topological hypothesis, the authors proved a localization theorem, and the theorem implies a Smith-type inequality which is closely related to (7-1).

As a basic setup of Lagrangian Floer homology, [13] contains some analytic background; for example, the choice of almost complex structures. We follow their settings in order to do Lagrangian Floer homology. We refer the reader to [13, Section 3].

7.2 Proof of Theorem 7.3

Let M be a plumbing space of Penner type, η the associated involution of M, and L_0 and L_1 a transversal pair of Lagrangian submanifolds such that

(1) $\eta(L_i) = L_i$;

(2) $\tilde{L}_i = L_i \cap \tilde{M}$ is a Lagrangian submanifold of $\tilde{M}$;

(3) $L_0 \cap L_1 = \tilde{L}_0 \cap \tilde{L}_1$;

(4) L_0 and L_1 are not isotopic to each other.

We will compute $\mathbb{Z}/2$–graded Lagrangian Floer homology $HF^*(L_0, L_1)$ over the Novikov field Λ of characteristic 2. To do this, we will prove that chain complexes $CF^*(L_0, L_1)$ and $CF^*(\tilde{L}_0, \tilde{L}_1)$ have the same generators and the same differential maps.

First, it is easy to show that $CF^*(L_0, L_1)$ and $CF^*(\tilde{L}_0, \tilde{L}_1)$ have the same generators since L_0 and L_1 satisfy that $L_0 \cap L_1 = \tilde{L}_0 \cap \tilde{L}_1$. Thus, $CF^*(L_0, L_1) = CF^*(\tilde{L}_0, \tilde{L}_1)$ as vector spaces.

Second, let ∂ (resp. $\tilde{\partial}$) denote the differential map on $CF^*(L_0, L_1)$ (resp. $CF^*(\tilde{L}_0, \tilde{L}_1)$). Then

$$\partial(p) = \sum_{\substack{q \in L_0 \cap L_1 \\ [u]:\mathrm{ind}([u])=1}} (\#\mathcal{M}(p, q; [u], J)) T^{\omega([u])} q,$$

where J is an almost complex structure on M, u is a holomorphic strip connecting p and q, and $\mathcal{M}(p, q; [u], J)$ is the moduli space of holomorphic strips. We skip the foundational details of the definition of ∂.

One can easily check that $\eta \circ u$ is also a holomorphic strip connecting p and q. Assume that for a holomorphic strip u, the image of u is not contained in $\tilde{M}$. Then u and $\eta \circ u$ will be canceled together in $\partial(p)$, since the Novikov field Λ is of characteristic 2. Thus, in order to define the differential map ∂, it is enough to count holomorphic strips u such that the image of u is contained in $\tilde{M}$.

On the other hand, in order to define $\tilde{\partial}\colon CF^*(\tilde{L}_0, \tilde{L}_1) \to CF^*(\tilde{L}_0, \tilde{L}_1)$, one needs to count the holomorphic strips on $\tilde{M}$. Thus, $\partial(p) = \tilde{\partial}(p)$ for all $p \in L_0 \cap L_1 = \tilde{L}_0 \cap \tilde{L}_1$.

Under the assumptions, $HF^*(L_0, L_1) = HF^*(\widetilde{L}_0, \widetilde{L}_1)$. Note that the former is defined on M^{2n}, but the latter is defined on a surface $\widetilde{M}$. Thus, it is enough to check that

$$\dim HF^0(\widetilde{L}_1, \widetilde{L}_2) + \dim HF^1(\widetilde{L}_1, \widetilde{L}_2) = i(\widetilde{L}_1, \widetilde{L}_2).$$

By Remark 7.2, [3, Lemma 2.18] completes the proof. □

7.3 Example 7.9

In the present subsection, we will prove Lemmas 7.7 and 7.8 in order to slightly weaken the difficulty of applying Theorem 7.3. Then, we will give Example 7.9.

The difficulty of applying Theorem 7.3 is that there are too many conditions which L_0 and L_1 should satisfy. Lemmas 7.8 and 7.7 will give us plenty of Lagrangians satisfying the conditions after Hamiltonian isotopies.

Before giving the statement of Lemmas 7.7 and 7.8, we will establish notation. Since

$$M = P(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l)$$

is a plumbing space of Penner type, we can construct a set $\mathbb{B}$ of Lagrangian branched submanifolds of M as we did in Section 3.4. Every Lagrangian branched submanifold $\mathscr{B} \in \mathbb{B}$ is a union of (parts of) α_i and β_j and Lagrangian connected sums α_i and β_j. However, there are two possible Lagrangian connected sums of α_i and β_j at each plumbing point $p \in \alpha_i \cap \beta_j$. They are $\alpha_i \#_p \beta_j$ and $\beta_j \#_p \alpha_i$. By assuming that α_i is a positive sphere and β_j is a negative sphere, one considers the Lagrangian connected sum $\beta_j \#_p \alpha_i$, not $\alpha_i \#_p \beta_j$. Similarly, by assuming that α_i is negative and β_j is positive, one can construct another set $\mathbb{B}^{\mathrm{op}}$ of Lagrangian branched submanifolds.

Lemma 7.7 *Let* $\mathscr{B}_1, \mathscr{B}_2 \in \mathbb{B} \cup \mathbb{B}^{\mathrm{op}}$. *Then there is a Hamiltonian isotopy* $\Phi_t : M \to M$ *such that*

(1) $\Phi_t \circ \eta = \eta \circ \Phi_t$,

(2) $\mathscr{B}_0 \pitchfork \Phi_1(\mathscr{B}_1)$,

(3) *for every* $q \in \mathscr{B}_0 \cap \Phi_1(\mathscr{B}_1)$, q *is not a plumbing point or the antipodal point of a plumbing point.*

Proof Since $\mathscr{B}_1$ is a union of (parts of) compact cores and their Lagrangian connected sums, we will construct Hamiltonian isotopies perturbing each compact core α_i and β_j. Then, one obtains a perturbation of $\mathscr{B}_1$ as a union of (parts of) perturbations of α_i, β_j and Lagrangian connected sums of perturbed α_i and β_j.

First, we choose a smooth function $f_i : \alpha_i \to \mathbb{R}$ with isolated critical points such that

(1) for every plumbing point $p \in \alpha_i$, $f_i(p) = f_i(-p) = 0$, where $-p$ is the antipodal point of p on α_i;

(2) every critical point q of f_i lies on S_{α_i} and $q \neq p, -p$ for any plumbing point $p \in \alpha_i$;

(3) $|df_i(x)| < \epsilon$ for all $x \in \alpha_i$ and for a sufficiently small fixed positive number ϵ;

(4) $f_i \circ \eta_{\alpha_i} = f_i$, where η_{α_i} is the involution on $T^*\alpha_i$ defined in Section 7.1.

We remark that

$$T^*\alpha_i \overset{\phi_{\alpha_i}}{\simeq} T^*S^n = \{(x, y) \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \mid |x| = 1, \langle x, y \rangle = 0\},$$

where $\phi_{\alpha_i} : T^*S^n \xrightarrow{\sim} T^*\alpha_i$ is the identification which we used in Section 7.1. Also, we remark that in (3), $|df_i(x)|$ is given by the standard metric on $\mathbb{R}^{2n+2}$.

Then, we can extend f_i to $\tilde{f}_i : T^*\alpha_i \to \mathbb{R}$ as follows. Let $\delta : [0, \infty) \to \mathbb{R}$ be a smooth decreasing function such that

$$\delta([0, \epsilon]) = 1, \delta([2\epsilon, \infty)) = 0.$$

We set

$$\tilde{f}_i : T^*\alpha_i \to \mathbb{R}, \tilde{f}_i(x, y) = \delta(|y|) f_i(x).$$

We get $\tilde{g}_j : T^*\beta_j \to \mathbb{R}$ in the same way.

These Hamiltonian functions $\tilde{f}_i$ and $\tilde{g}_j$ induce Hamiltonian isotopies on $T^*\alpha_i$ and $T^*\beta_j$. Moreover, these Hamiltonian isotopies could be extended on the plumbing space M since the Hamiltonian isotopies have compact supports on $T^*\alpha_i$ and $T^*\beta_j$.

Let $\Phi_{\alpha_i, t} : M \xrightarrow{\sim} M$ be the (extended) Hamiltonian isotopy associated to $\tilde{f}_i$. It is easy to check that

$$\Phi_{\alpha_i, t} \circ \eta = \eta \circ \Phi_{\alpha_i, t},$$
$$\Phi_{\alpha_i, t}(\alpha_k) = \alpha_k \quad \text{if } k \neq i,$$
$$\Phi_{\alpha_i, t}(\beta_j) = \beta_j \quad \text{for all } j,$$
$$\Phi_{\alpha_i, 1}(\alpha_i) = \Gamma(df_i),$$

where $\Gamma(df_i)$ is the graph of df_i in $T^*\alpha_i \subset M$. Similarly, one can obtain a Hamiltonian isotopy $\Phi_{\beta_j, t} : M \xrightarrow{\sim} M$ for each β_j.

Let

$$\Phi_t = \prod_{\beta_j} \Phi_{\beta_j, t} \circ \prod_{\alpha_i} \Phi_{\alpha_i, t}.$$

It is easy to check that Φ_t satisfies the first condition of Lemma 7.7. Moreover, one can assume that $\Phi_1(\mathcal{B}_1)$ is constructed from $\Phi_1(\alpha_i)$ and $\Phi_1(\beta_j)$. Thus, it is easy to prove that $\mathcal{B}_0$ and $\Phi_1(\mathcal{B}_1)$ satisfy the second and the last conditions of Lemma 7.7. $\qquad \square$

We will now explain how Lemma 7.7 weakens a difficulty of applying Theorem 7.3. The difficulty we will consider is the last condition of Theorem 7.3, ie $L_0 \cap L_1 = \tilde{L}_0 \cap \tilde{L}_1$. The other conditions can be weakened by a similar way.

Assume that L_0 (resp. L_1) is a Lagrangian submanifold which is carried by $\mathcal{B}_0$ (resp. $\mathcal{B}_1$) in $\mathbb{B} \cup \mathbb{B}^{\text{op}}$. Note that $\Phi_1(L_1)$ is carried by $\Phi_1(\mathcal{B}_1)$, where Φ_1 is the Hamiltonian isotopy constructed in Lemma 7.7.

We will count the numbers of intersections $L_0 \cap \Phi_1(L_1)$ and $\widetilde{L}_0 \cap \Phi_1(\widetilde{L}_1)$. If these numbers are the same, then $L_0 \cap \Phi_1(L_1) = \widetilde{L}_0 \cap \Phi_1(\widetilde{L}_1)$.

First, we remark that $\widetilde{L}_0$ (resp. $\Phi_1(\widetilde{L}_1)$) is a curve carried by a train track $\mathcal{B}_0 \cap \widetilde{M}$ (resp. $\Phi_1(\mathcal{B}_1) \cap \widetilde{M}$). Then, $\widetilde{L}_0$ (resp. $\Phi_1(\widetilde{L}_1)$) has weights on the train track $\mathcal{B}_0 \cap \widetilde{M}$ (resp. $\Phi_1(\mathcal{B}_1) \cap \widetilde{M}$). Moreover, the number of $\widetilde{L}_0 \cap \Phi_1(\widetilde{L}_1)$ is

$$\sum_{x \in \mathcal{B}_0 \cap \Phi_1(\mathcal{B}_1)} (\text{the weight of } \widetilde{L}_0 \text{ at } x) \cdot (\text{the weight of } \Phi_1(\widetilde{L}_1) \text{ at } x).$$

To count the number of $L_0 \cap \Phi_1(L_1)$, we can assume that $L_0 \cap \Phi_1(L_1)$ is contained in a small neighborhood of $\mathcal{B}_0 \cap \Phi_1(\mathcal{B}_1)$. Since L_0 is carried by $\mathcal{B}_0$, not strongly carried by, L_0 can have singular points. However, the singular points are "close" to one of plumbing points or the antipodes of plumbing points. Since the intersection points of $\mathcal{B}_0$ and $\Phi_1(\mathcal{B}_1)$ are not plumbing points or their antipodes, every $x \in L_0 \cap \Phi_1(L_1)$ is a regular point of L_0 (resp. $\Phi_1(L_1)$). It means that the number $|L_0 \cap \Phi_1(L_1)|$ is also given by

$$\sum_{x \in \mathcal{B}_0 \cap \Phi_1(\mathcal{B}_1)} (\text{the weight of } \widetilde{L}_0 \text{ at } x) \cdot (\text{the weight of } \Phi_1(\widetilde{L}_1) \text{ at } x).$$

Thus, $|L_0 \cap \Phi_1(L_1)| = |\widetilde{L}_0 \cap \Phi_1(\widetilde{L}_1)|$.

Lemma 7.8 *Let L_0 and L_1 be carried by $\mathcal{B}_0, \mathcal{B}_1 \in \mathbb{B} \cup \mathbb{B}^{\mathrm{op}}$. Then there is a Hamiltonian isotopy Φ_t such that*

$$L_0 \cap \Phi_1(L_1) = \widetilde{L}_0 \cap \Phi_1(\widetilde{L}_1).$$

Thus, if L_0 and L_1 are carried by $\mathcal{B}_0, \mathcal{B}_1 \in \mathbb{B} \cup \mathbb{B}^{\mathrm{op}}$, and if L_0 and L_1 satisfy conditions (1), (2) and (4) of Theorem 7.3, then one can apply Theorem 7.3 for L_0 and $\Phi_1(L_1)$.

Example 7.9 Let ψ_0 and ψ_1 be symplectomorphisms of Penner type, ie ψ_0 and ψ_1 are products of positive (resp. negative) powers of τ_i and negative (resp. positive) powers of σ_j, where τ_i and σ_j are Dehn twists along α_i and β_j respectively. Assume that L_0 (resp. L_1) is a Lagrangian submanifold of M, which is generated from one of compact cores by applying ψ_0 (resp. ψ_1), ie

$$L_0 = \psi_0(\alpha_k) \text{ or } \psi_0(\beta_j), \quad L_1 = \psi_1(\alpha_k) \text{ or } \psi_1(\beta_j).$$

Then $\eta(L_i) = L_i$ since

$$\eta(\alpha_i) = \alpha_i, \quad \eta(\beta_j) = \beta_j, \quad \eta \circ \tau_i = \tau_i \circ \eta, \quad \eta \circ \sigma_j = \sigma_j \circ \eta \quad \text{for all } i, j.$$

Moreover, $\widetilde{L}_i = \psi_i(\widetilde{\alpha}_k)$ or $\psi_i(\widetilde{\beta}_j)$. Thus, $\widetilde{L}_i$ is a Lagrangian submanifold of $\widetilde{M}$. Finally, L_i is carried by $\mathcal{B}_{\psi_i}$.

Thus, if L_0 and L_1 are not isotopic to each other, then one can apply Theorem 7.3.

References

[1] **D Auroux**, *A beginner's introduction to Fukaya categories*, from "Contact and symplectic topology" (F Bourgeois, V Colin, A Stipsicz, editors), Bolyai Soc. Math. Stud. 26, Bolyai, Budapest (2014) 85–136 MR Zbl

[2] **A J Casson**, **S A Bleiler**, *Automorphisms of surfaces after Nielsen and Thurston*, Lond. Math. Soc. Student Texts 9, Cambridge Univ. Press (1988) MR Zbl

[3] **G Dimitrov**, **F Haiden**, **L Katzarkov**, **M Kontsevich**, *Dynamical systems and categories*, from "The influence of Solomon Lefschetz in geometry and topology" (L Katzarkov, E Lupercio, F J Turrubiates, editors), Contemp. Math. 621, Amer. Math. Soc., Providence, RI (2014) 133–170 MR Zbl

[4] **Y-W Fan**, **S Filip**, **F Haiden**, **L Katzarkov**, **Y Liu**, *On pseudo-Anosov autoequivalences*, Adv. Math. 384 (2021) art. id. 107732 MR Zbl

[5] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Math. Ser. 49, Princeton Univ. Press (2012) MR Zbl

[6] **W Floyd**, **U Oertel**, *Incompressible surfaces via branched surfaces*, Topology 23 (1984) 117–125 MR Zbl

[7] **J Maher**, *Random walks on the mapping class group*, Duke Math. J. 156 (2011) 429–468 MR Zbl

[8] **D Nadler**, *Arboreal singularities*, Geom. Topol. 21 (2017) 1231–1274 MR Zbl

[9] **U Oertel**, *Incompressible branched surfaces*, Invent. Math. 76 (1984) 385–410 MR Zbl

[10] **R C Penner**, *A construction of pseudo-Anosov homeomorphisms*, Trans. Amer. Math. Soc. 310 (1988) 179–197 MR Zbl

[11] **R C Penner**, **J L Harer**, *Combinatorics of train tracks*, Ann. of Math. Stud. 125, Princeton Univ. Press (1992) MR Zbl

[12] **P Seidel**, *Lagrangian two-spheres can be symplectically knotted*, J. Differential Geom. 52 (1999) 145–171 MR Zbl

[13] **P Seidel**, **I Smith**, *Localization for involutions in Floer cohomology*, Geom. Funct. Anal. 20 (2010) 1464–1501 MR Zbl

[14] **W P Thurston**, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. 19 (1988) 417–431 MR Zbl

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

[16] **A Weinstein**, *Symplectic manifolds and their Lagrangian submanifolds*, Adv. Math. 6 (1971) 329–346 MR Zbl

[17] **R F Williams**, *Expanding attractors*, Inst. Hautes Études Sci. Publ. Math. 43 (1974) 169–203 MR Zbl

[18] **W Wu**, *Exact Lagrangians in A_n–surface singularities*, Math. Ann. 359 (2014) 153–168 MR Zbl

Center for Geometry and Physics, Institute for Basic Science
Pohang, South Korea

sangjinlee@ibs.re.kr

https://sites.google.com/view/sangjinlee/home

msp

A strong Haken theorem

MARTIN SCHARLEMANN

Suppose $M = A \cup_T B$ is a Heegaard split compact orientable 3–manifold and $S \subset M$ is a reducing sphere for M. Haken (1968) showed that there is then also a reducing sphere S^* for the Heegaard splitting. Casson and Gordon (1987) extended the result to ∂–reducing disks in M and noted that in both cases S^* is obtained from S by a sequence of operations called 1–surgeries. Here we show that in fact one may take $S^* = S$.

57K35

It is a foundational theorem of Haken [4] that any Heegaard splitting $M = A \cup_T B$ of a closed orientable reducible 3–manifold M is reducible; that is, there is an essential sphere in the manifold that intersects T in a single circle. Casson and Gordon [1, Lemma 1.1] refined and generalized the theorem, showing that it applies also to essential disks, when M has boundary. More specifically, if S is a disjoint union of essential disks and 2–spheres in M then there is a similar family S^*, obtained from S by ambient 1–surgery and isotopy, such that each component of S^* intersects T in a single circle. In particular, if M is irreducible, so S consists entirely of disks, S^* is isotopic to S.

There is of course a more natural statement, in which S does not have to be replaced by S^*. I became interested in whether the natural statement is true because it would be the first step in a program to characterize generators of the Goeritz group of S^3; see Freedman and the author [3; 8]. Inquiring of experts, I learned that this more natural statement had been pursued by some, but not successfully. Here we present such a proof. A reader who would like to get the main idea in a short amount of time could start with the example in Section 11. Recently, Hensel and Schultens [6] have proposed an alternative proof that applies when M is closed and S consists entirely of spheres.

Here is an outline of the paper: Sections 1 and 2 are mostly a review of what is known; particularly the use of verticality in classical compression bodies, those which have no spheres in their boundary. We wish to allow sphere components in the boundary, and Section 3 explains how to recover the classical results in this context. Section 4 shows how to use these results to inductively reduce the proof of the main theorem to the case when S is connected. The proof when S is connected (the core of the proof) then occupies Sections 6 through 10.

1 Introduction and review

All manifolds considered will be orientable and, unless otherwise described, also compact. For M a 3–manifold, a closed surface $T \subset M$ is a *Heegaard surface* in M if the closed complementary components A and B are each compression bodies, defined below. This structure is called a *Heegaard splitting* and is typically written $M = A \cup_T B$. See, for example, [7] for an overview of the general theory of Heegaard surfaces. Among the foundational theorems of the subject is the following [1].

Suppose T is a Heegaard surface in a Heegaard split 3–manifold $M = A \cup_T B$ and D is a ∂ reducing disk for M, with $\partial D \subset \partial_- B \subset \partial M$.

Theorem 1.1 (Haken, Casson–Gordon) *There is a ∂–reducing disk E for M such that*

- $\partial E = \partial D$,
- *E intersects T in a single essential circle (ie E ∂–reduces T).*

Note that D and E are isotopic if M is irreducible; but if M is reducible then there is no claim that D and E are isotopic.

There is a similar foundational theorem, by Haken alone [4], that if M is reducible, there is a reducing sphere for M that intersects T in a single circle (ie it is a reducing sphere for T). But Haken made no claim that the reducing sphere for T is isotopic to a given reducing sphere for M.

The intention of this paper is to fill this gap in our understanding. We begin by retreating to a more general setting. For our purposes, a *compression body* C is a connected 3–manifold obtained from a (typically disconnected) closed surface $\partial_- C$ by attaching 1–handles to one end of a collar of $\partial_- C$. The closed connected surface $\partial C - \partial_- C$ is denoted $\partial_+ C$. This differs from what may be the standard notion in that we allow $\partial_- C$ to contain spheres, so C may be reducible. Put another way, we take the standard notion, but then allow the compression body to be punctured finitely many times. In particular, the compact 3–manifolds whose Heegaard splittings we study may have spheres as boundary components.

Suppose then that $M = A \cup_T B$ is a Heegaard splitting, with A and B compression bodies as above. A disk/sphere set $(S, \partial S) \subset (M, \partial M)$ is a properly embedded surface in M such that each component of S is either a disk or a sphere. A sphere in M is called *inessential* if it either bounds a ball or is parallel to a boundary component of M; a disk is inessential if it is parallel to a disk in ∂M. S may contain such inessential components, but these are easily dismissed, as we will see.

Definition 1.2 The Heegaard splitting T is *aligned* with S (or vice versa) if each component of S intersects T in at most one circle.

For example, a reducing sphere or ∂–reducing disk for T, typically defined as a sphere or disk that intersects T in a single essential circle, are each important examples of an aligned disk/sphere. This new

terminology is introduced in part because, in the mathematical context of this paper, the word "reduce" is used in multiple ways that can be confusing. More importantly, once we generalize compression bodies as above, so that some boundary components may be spheres, there are essential spheres and disks in M that may miss T entirely and others that may intersect T only in curves that are inessential in T. We need to take these disks and spheres into account.

Theorem 1.3 *Suppose that* $(S, \partial S) \subset (M, \partial M)$ *is a disk/sphere set in* M. *Then there is an isotopy of* T *such that afterwards* T *is aligned with* S.

Moreover, such an isotopy can be found so that, after the alignment, the annular components $S \cap A$, *if any, form a vertical family of spanning annuli in the compression body* A, *and similarly for* $S \cap B$.

The terminology "vertical family of spanning annuli" is defined in Section 2.

Note that a disk/sphere set S may contain inessential disks or spheres, or essential disks whose boundaries are inessential in ∂M. Each of these are examples in which the disk or sphere could lie entirely in one of the compression bodies and so be disjoint from T. In the classical setting, Theorem 1.3 has this immediate corollary:

Corollary 1.4 (strong Haken) *Suppose* ∂M *contains no sphere components. Suppose* $S \subset M$ (*resp.* $(S, \partial S) \subset (M, \partial M)$) *is a reducing sphere* (*resp.* ∂–*reducing disk*) *in* M. *Then* S *is isotopic to a reducing sphere* (*resp.* ∂–*reducing disk*) *for* T.

The assumption in Corollary 1.4 that there are no sphere components in ∂M puts us in the classical setting, where any reducing sphere S for M must intersect T.

2 Verticality in aspherical compression bodies

We first briefly review some classic facts and terminology for an aspherical compression body C, by which we mean that $\partial_- C$ contains no sphere components. Later, sphere components will add a small but interesting amount of complexity to this standard theory. See [7] for a fuller account of the classical theory. Unstated in that account (and others) is the following elementary observation, which further supports the use of the term "aspherical":

Proposition 2.1 *An aspherical compression body* C *is irreducible.*

Proof Let Δ be the cocores of the 1–handles used in the construction of C from the collar $\partial_- C \times I$. If C contained a reducing sphere S, that is a sphere that does not bound a ball, a standard innermost disk argument on $S \cap \Delta$ would show that there is a reducing sphere in the collar $\partial_- C \times I$. But since C is assumed to be aspherical, $\partial_- C$ contains no spheres, and it is classical that a collar of a closed orientable surface that is not a sphere is irreducible. (For example, its universal cover is a collar of R^2; the interior of this collar is R^3; and R^3 is known to be irreducible by the Schoenflies theorem [10].) □
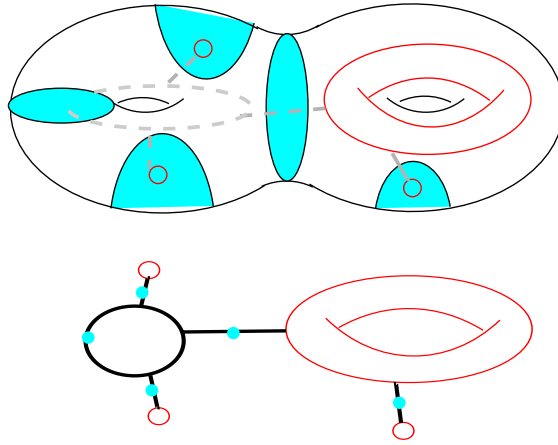
Figure 1: 2–handles and dual spine in a compression body.

Definition 2.2 A properly embedded family $(\Delta, \partial\Delta) \subset (C, \partial C)$ of disks is a *complete collection of meridian disks* for C if $C - \eta(\Delta)$ consists of a collar of $\partial_- C$ and some 3–balls.

That there is such a family of disks follows from the definition of a compression body: take Δ to be the cocores of the 1–handles used in the construction. Given two complete collections Δ and Δ' of meridian disks in an aspherical compression body, it is possible to make them disjoint by a sequence of 2–handle slides, viewing the disks as cocores of 2–handles. (The slides are often more easily viewed dually, as slides of 1–handles.) The argument in brief is this: If Δ and Δ' are two complete collections of meridians, an innermost disk argument (which relies on asphericity) can be used to remove all circles of intersection. A disk cut off from Δ' by an outermost arc γ of $\Delta' \cap \Delta$ in Δ' determines a way of sliding the 2–handle in Δ containing γ over some other members of Δ to eliminate γ without creating more intersection arcs. Continue until all arcs are gone. (A bit more detail is contained in Phase 2 of the proof of Proposition 3.4.)

Visually, one can think of the cores of the balls and 1–handles as a properly embedded graph in C, with some valence 1 vertices on $\partial_- C$, so that the union Σ of the graph and $\partial_- C$ has C as its regular neighborhood. Σ is called a *spine* of the compression body. As already noted, a spine for C is far from unique, but one can move from any spine to any other spine by sliding ends of edges in the graph over other edges, or over components of $\partial_- C$, dual to the 2–handle slides described above. (See [9] or [7].) For most arguments it is sufficient and also simplifying to disregard any valence-one vertex that is not on $\partial_- C$ and the "canceling" edge to which it is attached (but these do briefly appear in the proof of Corollary 5.5); to disregard all valence-two vertices by amalgamating the incident edges into a single edge; and, via a slight perturbation, to require all vertices not on $\partial_- C$ to be of valence three. We can, by edge slides, ensure that only a single edge of the spine is incident to each component of $\partial_- C$; this choice of spine is also sometimes useful.

The spine can be defined as above even when $\partial_- C$ contains spheres. Figure 1 shows a schematic picture of a (nonaspherical) compression body, viewed first with its (aqua) two-handle structure and then its dual

1–handle (spinal) structure. $\partial_- C$ is the union of a torus and 3 spheres; the genus two $\partial_+ C$ appears in the spinal diagram only as an imagined boundary of a regular neighborhood of the spine.

Definition 2.3 A properly embedded arc α in a compression body C is *spanning* if one end of α lies on each of $\partial_- C$ and $\partial_+ C$. Similarly, a properly embedded annulus in C is *spanning* if one end lies in each of $\partial_- C$ and $\partial_+ C$. (Hence, each spanning arc in a spanning annulus is also spanning in the compression body.)

A disjoint collection of spanning arcs α in a compression body is a *vertical family of arcs* if there is a complete collection Δ of meridian disks for C such that

- $\alpha \cap \Delta = \varnothing$ and
- for N, the components of $C - \Delta$ that are a collar of $\partial_- C$, there is a homeomorphism

$$h \colon \partial_- C \times (I, \{0\}) \to (N, \partial_- C)$$

  such that $h(\mathfrak{p} \times I) = \alpha$, where $\mathfrak{p}$ is a collection of points in $\partial_- C$.

A word of caution: We will show in [Proposition 2.8](#) that any two vertical arcs with endpoints on the same component $F \subset \partial_- C$ are properly isotopic in C. This is obvious if the two constitute a vertical family. If they are each vertical, but not as a vertical family, proof is required because the collection of meridian disks referred to in [Definition 2.3](#) may differ for the two arcs.

There is a relatively simple but quite useful way of characterizing a vertical family of arcs. To that end, let α be a family of spanning arcs in C and $\hat{p} = \alpha \cap \partial_- C$ be their endpoints in $\partial_- C$. An embedded family c of simple closed curves in $\partial_- C$ is a *circle family associated to α* if $\hat{p} \subset c$.

Lemma 2.4 *Suppose α is a family of spanning arcs in an aspherical compression body C.*

- *Suppose α is vertical and c is an associated circle family. Then there is a family $\mathscr{A}$ of disjoint spanning annuli in C such that $\mathscr{A}$ contains α and $\mathscr{A} \cap \partial_- C = c$.*
- *Suppose, on the other hand, there is a collection $\mathscr{A}$ of disjoint spanning annuli in C that contains α. Suppose further that in the family of circles $\mathscr{A} \cap \partial_- C$ associated to α, each circle is essential in $\partial_- C$. Then α is a vertical family.*

Proof One direction is clear: suppose α is a vertical family and $h \colon \partial_- C \times (I, \{0\}) \to (N, \partial_- C)$ is the homeomorphism from [Definition 2.3](#); then $h(c \times I)$ is the required family of spanning annuli (after the technical adjustment, from general position, of moving the circles $h(c \times \{1\})$ off the disks in $h(\partial_- C \times \{1\})$ coming from the family Δ of meridian disks for C).

For the second claim, let Δ be any complete collection of meridians for C and consider the collection of curves $\Delta \cap \mathscr{A}$. If $\Delta \cap \mathscr{A} = \varnothing$ then $\mathscr{A}$ is a family of incompressible spanning annuli in the collar
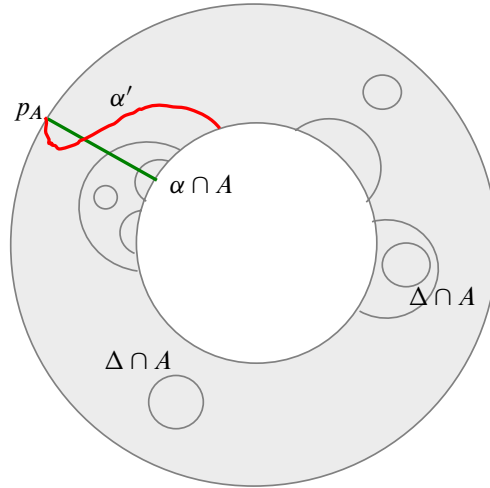
Figure 2: The spanning arc α' avoids $\Delta \cap A$.

$\partial_- C \times I$ and, by standard arguments, any family of incompressible spanning annuli in a collar is vertical. Furthermore, any family of spanning arcs in a vertical annulus can visibly be isotoped rel one end of the annulus to be a family of vertical arcs. So we are left with the case $\Delta \cap \mathcal{A} \neq \varnothing$.

Suppose $\Delta \cap \mathcal{A}$ contains a simple closed curve, necessarily inessential in Δ. If that curve were essential in a component $A \in \mathcal{A}$, then the end $A \cap \partial_- C \subset c$ would be nullhomotopic in C. Since the hypothesis is that each such circle is essential in $\partial_- C$, this would contradict the injectivity of $\pi_1(\partial_- C) \to \pi_1(C)$.

We conclude that each component of $\Delta \cap \mathcal{A}$ is either an inessential circle in $\mathcal{A}$ or an arc in $\mathcal{A}$ with both ends on $\partial_+ C$, since $\partial\Delta \subset \partial_+ C$. Such arcs are inessential in $\mathcal{A}$.

Consider what this means in a component $A \in \mathcal{A}$; let $c_A = A \cap \partial_- C \in c$ be the end of A in $\partial_- C$. It is easy to find spanning arcs α' in A with ends at the points $p_A = \hat{p} \cap c_A$, chosen so that α' avoids all components of $\Delta \cap A$. See Figure 2. But, as spanning arcs, $\alpha \cap A$ and α' are isotopic in A rel c_A (or, if one prefers, one can picture this as an isotopy near A that moves the curves $\Delta \cap A$ off of $\alpha \cap A$). After such an isotopy in each annulus, Δ and α are disjoint. Now apply classic innermost disk, outermost arc arguments to alter Δ until it becomes a complete collection of meridians disjoint from $\mathcal{A}$, the case we have already considered. More details of this classic argument appear in Phase 2 of the proof of Proposition 3.4. $\qquad \square$

Lemma 2.4 suggests the following definition.

Definition 2.5 Suppose $\mathcal{A}$ is a family of disjoint spanning annuli in C and α is a collection of disjoint spanning arcs in $\mathcal{A}$, with at least one arc of α in each annulus of $\mathcal{A}$. $\mathcal{A}$ is a *vertical family of annuli* if and only if α is a vertical family of arcs.

Note that for $\mathcal{A}$ to be vertical we do not require that $\mathcal{A}$ be incompressible in C. This adds some complexity to our later arguments, particularly the proof of Proposition 3.8.

Proposition 2.6 *Suppose $\mathscr{A}$ is a vertical family of annuli in an aspherical compression body C. Then there is a complete collection of meridian disks for C that is disjoint from $\mathscr{A}$.*

Proof Let $\alpha \subset \mathscr{A}$ be a vertical family of spanning arcs as given in Definition 2.5. Since α is a vertical family of arcs, there is a complete collection Δ of meridian disks for C that is disjoint from α, so Δ intersects $\mathscr{A}$ only in inessential circles, and arcs with both ends incident to the end of ∂A at $\partial_+ C$. As noted in the proof of Lemma 2.4, a standard innermost disk, outermost arc argument can be used to alter Δ to be disjoint from $\mathscr{A}$. $\qquad\square$

Corollary 2.7 *Suppose $(\mathscr{D}, \partial\mathscr{D}) \subset (C, \partial_+ C)$ is an embedded family of disks that is disjoint from an embedded family of vertical annuli $\mathscr{A}$ in an aspherical compression body C. Then there is a complete collection of meridian disks for C that is disjoint from $\mathscr{A} \cup \mathscr{D}$.*

Proof Proposition 2.6 shows that there is a complete collection disjoint from $\mathscr{A}$. But the same proof (which exploits asphericity through its use of Lemma 2.4) works here, if we also augment the curves $\Delta \cap \mathscr{A}$ with the circles $\Delta \cap \mathscr{D}$. $\qquad\square$

Proposition 2.8 *Suppose F is a component of $\partial_- C$ and α and β are vertical arcs in C with endpoints $p, q \in F$. Then α and β are properly isotopic in C.*

Notice that the proposition does not claim that α and β are parallel, so in particular they do not necessarily constitute a vertical family. Indeed the isotopy from α to β that we will describe may involve crossings between α and β.

Proof Since C is aspherical, genus$(F) \geq 1$ and there are simple closed curves $c_\alpha, c_\beta \subset F$ such that

- $p \in c_\alpha$ and $q \in c_\beta$,
- c_α and c_β intersect in a single point.

Since α and β are each vertical, it follows from Lemma 2.4 that there are spanning annuli A_α and A_β in C that contain α and β, respectively, and whose ends on F are c_α and c_β, respectively. Since c_α and c_β intersect in a single point, this means that among the curves in $A_\alpha \cap A_\beta$ there is a single arc γ that spans each annulus, and no other arcs are incident to F. The annulus A_α then provides a proper isotopy from the spanning arc α to γ and the annulus A_β provides a proper isotopy from γ to β. Hence, α and β are properly isotopic in C. See Figure 3. $\qquad\square$

We now embark on a technical lemma that uses these ideas, which we will need later. Begin with a closed connected surface F that is not a sphere, and say that circles α and β *essentially intersect* if they are not isotopic to disjoint circles and have been isotoped so that $|\alpha \cap \beta|$ is minimized. Suppose $\hat{a} \subset F$ is an embedded family of simple closed curves, not necessarily essential, and p_1 and p_2 are a pair of points disjoint from $\hat{a}$. (We only will need the case of two points; the argument below extends to any finite number, with some loss of clarity in statement and proof.)
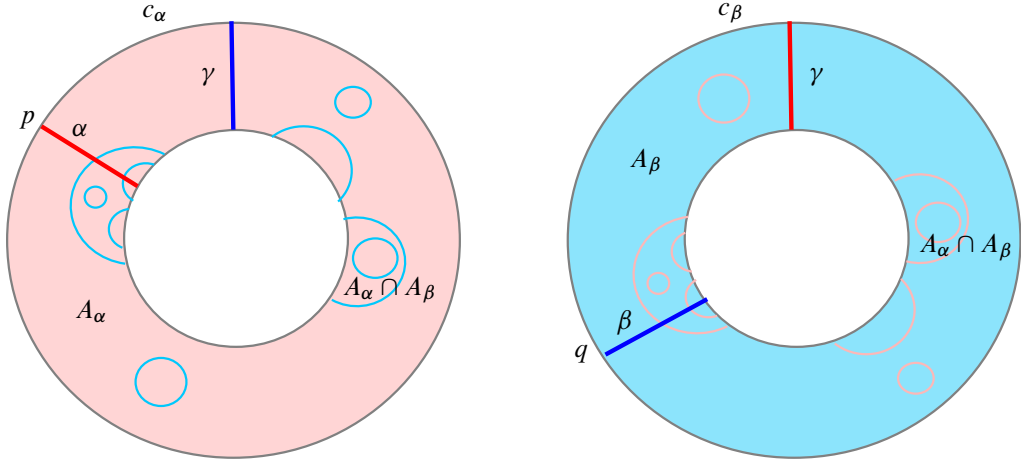
Figure 3: Arcs α and β both properly isotopic to γ.

Let $b' \subset F$ be a nonseparating simple closed curve in F that is not parallel to any $a \in \hat{a}$. For example, if all curves in $\hat{a}$ are separating, b' could be any nonseparating curve; if some curve $a \in \hat{a}$ is nonseparating, take b' to be a circle that intersects a once. Isotope b' in F so that it contains p_1 and p_2, and intersects $\hat{a}$ transversally if at all; call the result $b \subset F$. (Note that, following these requirements, $\hat{a}$ may not intersect b essentially, for example if an innermost disk in F cut off by an inessential $a \in \hat{a}$ contains p_i.) If b intersects $\hat{a}$, let q_i be points in $b \cap \hat{a}$ such that the subintervals $\sigma_i \subset b$ between p_i and q_i have interiors disjoint from $\hat{a}$ and are also disjoint from each other. Informally, we could say that q_i is the closest point in $\hat{a}$ to p_i along b, and σ_i is the path in b between p_i and q_i.

Since b is nonseparating there is a simple closed curve $x \subset F$ that intersects b exactly twice, with the same orientation (so the intersection is essential). Isotope x along b until the two points of intersection are exactly q_1 and q_2. See Figure 4.
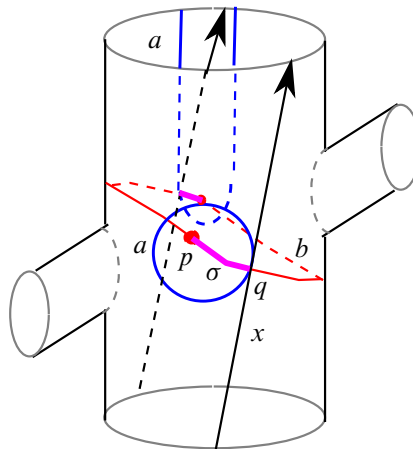
Figure 4: Preamble to Lemma 2.9.
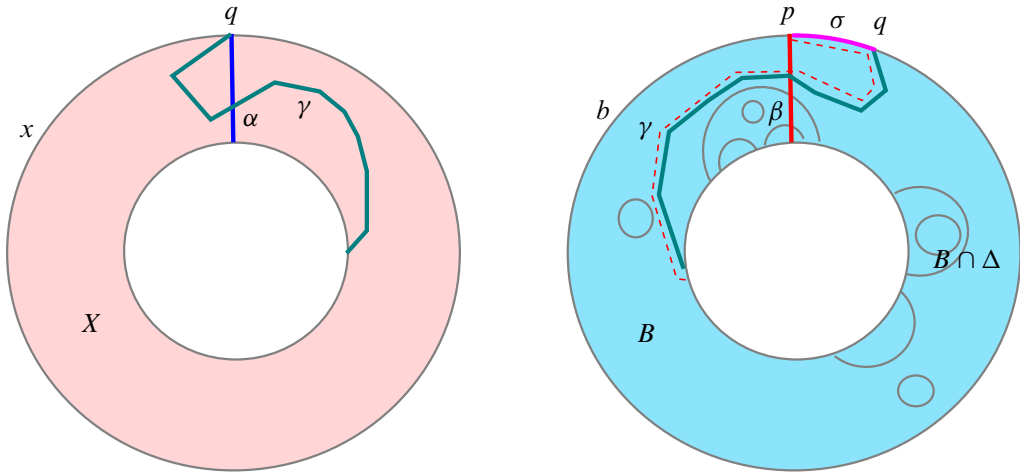
Figure 5: Concluding the proof of Lemma 2.9.

Lemma 2.9 *Let $(\mathcal{D}, \partial\mathcal{D}) \subset (C, \partial_+C)$ and $\mathcal{A} \subset C$ be as in Corollary 2.7. Suppose $\hat{\beta} = \{\beta_i\}$ for $i = 1, 2$ is a vertical family of arcs in C whose endpoints $p_i \in \partial_-C$ are disjoint from the family of circles $\hat{a} = \mathcal{A} \cap \partial_-C$ in ∂_-C. Then $\hat{\beta}$ can be properly isotoped rel $\{p_i\}$ so that it is disjoint from $\mathcal{A} \cup \mathcal{D}$.*

Proof We suppose that both components of $\hat{\beta}$ are incident to the same component F of ∂_-C. The proof is essentially the same (indeed easier) if they are incident to different components of ∂_-C. Let Δ be a complete family of meridian disks as given in Corollary 2.7, so $\mathcal{A}$ lies entirely in a collar of ∂_-C. Per Lemma 2.4, let $B \subset C$ be a spanning annulus that contains the vertical pair $\hat{\beta}$ and has the curve b (from the preamble to this lemma) as its end $B \cap F$ on F.

Suppose first that b is disjoint from $\hat{a}$ and consider $B \cap (\Delta \cup \mathcal{D} \cup \mathcal{A})$. If there were a circle c of intersection that is essential in B, then it could not be in $\Delta \cup \mathcal{D}$, since b does not compress in C. The circle c could not be essential in $\mathcal{A}$, since b was chosen so that it is not isotopic to any element of $\hat{a}$, and it can't be inessential there either again since b does not compress in C. We deduce that there can be no essential circle of intersection, so any circles in $B \cap (\Delta \cup \mathcal{D} \cup \mathcal{A})$ are inessential in B. Also, any arc of intersection must have both ends on ∂_+C since b is disjoint from $\hat{a}$. It follows that the spanning arcs $\hat{\beta}$ of B can be properly isotoped in B to arcs that avoid $\Delta \cup \mathcal{D} \cup \mathcal{A}$. So, note, they are in the collar of ∂_-C as well as being disjoint from $\mathcal{A} \cup \mathcal{D}$ as required.

Now suppose that b is not disjoint from $\hat{a}$ and let the points q_i, the subarcs σ_i of b and the simple closed curve $x \subset F$ be as described in the preamble to this lemma. By construction, each q_i is in the end of an annulus $A_i \subset \mathcal{A}$; let $\alpha_i \subset A_i$ be a spanning arc of A_i with an end on q_i. Since $\mathcal{A}$ is a vertical family of annuli, α_1 and α_2 are a vertical pair of spanning arcs. Per Lemma 2.4, there is a spanning annulus X that contains the α_i and has the curve x as its end $X \cap F$ on F. Since x essentially intersects b in these two points, $B \cap X$ contains exactly two spanning arcs γ_i, for $i = 1, 2$, each with one endpoint on the respective q_i.

In B the spanning arcs β_i can be properly isotoped rel p_i so that they are each very near the concatenation of σ_i and γ_i; in X the arcs γ_i can be properly isotoped rel q_i to α_i. See Figure 5. (One could also think of this as giving an ambient isotopy of the annulus B so that afterwards $\gamma_i = \alpha_i$.) The combination of these isotopies then leaves β_i parallel to the arc $\sigma_i \cup \alpha_i$. A slight push-off away from A_i leaves β_i disjoint from $\mathscr{A} \cup \mathscr{D}$, as required.   □

3 Verticality in compression bodies

We no longer will assume that compression bodies are aspherical. That is, $\partial_- C$ may contain spheres. We will denote by $\widehat{C}$ the aspherical compression body obtained by attaching a 3–ball to each such sphere.

Figure 1 shows a particularly useful type of meridian disk to consider when $\partial_- C$ contains spheres.

Definition 3.1 A complete collection Δ of meridian disks in a compression body C is a *snug collection* if, for each sphere $F \subset \partial_- C$, the associated collar of F in $C - \Delta$ is incident to exactly one disk $D_F \in \Delta$.

The use of the word "snug" is motivated by a simple construction. Suppose Δ is a snug collection of meridian disks for C and $F \subset \partial_- C$ is a sphere. Then the associated disk $D_F \subset \Delta$ is completely determined by a spanning arc α_F in the collar of F in $C - \Delta$, and vice versa: the arc α_F is uniquely determined by D_F, by the light-bulb trick, and once α_F is given, D_F is recovered simply by taking a regular neighborhood of $\alpha_F \cup F$; this regular neighborhood is a collar of F, and the end of the collar away from F itself is the boundary union of a disk in $\partial_+ C$ and a copy of D_F. With that description, we picture D_F as sitting "snugly" around $\alpha_F \cup F$. See Figure 6.

Following immediately from Definition 3.1 is:

Lemma 3.2 *Suppose C is a compression body and $\widehat{\Delta}$ is a collection of meridian disks for C that is a complete collection for the aspherical compression body $\widehat{C}$. Then $\widehat{\Delta}$ is contained in a snug collection for C.*

Proof For each sphere component F of $\partial_- C$, let α_F be a properly embedded arc in $C - \widehat{\Delta}$ from F to $\partial_- C$ and construct a corresponding meridian disk D_F as just described. Then the union of $\widehat{\Delta}$ with all these new meridian disks is a snug collection for C.   □
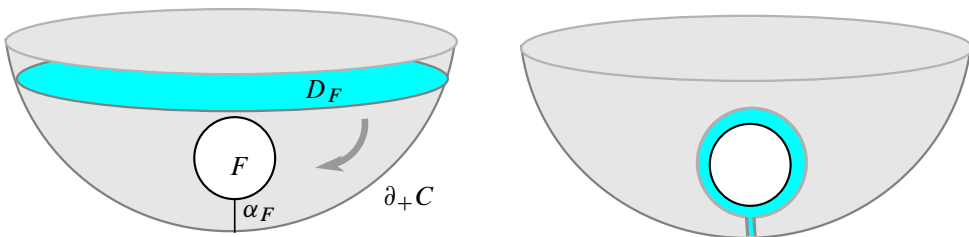
Figure 6: D_F snuggles down around $\alpha_F \cup F$.
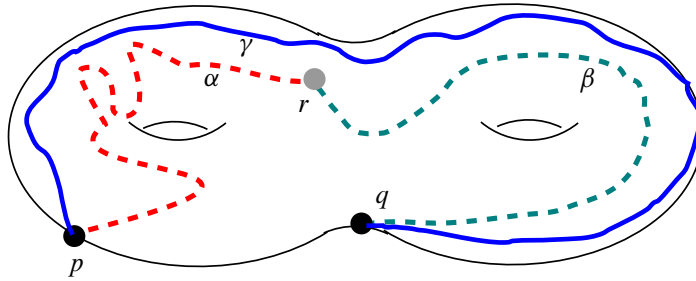
Figure 7

Following Definition 2.2 we noted that for an aspherical compression body, two complete collections of meridian disks can be handle-slid and isotoped to be disjoint. As a useful warm-up we will show that this is also true for snug collections, in case $\partial_- C$ contains spheres. This is the key lemma:

Lemma 3.3 *Suppose C is a compression-body with $p, q \in \partial_+ C$ and $r \in \text{int}(C)$. Suppose α and β are arcs from p and q, respectively, to r in C. Then there is a proper isotopy of β to α in C, fixing r.*

Proof Let Σ be a spine for the compression-body C. By general position, we may take Σ to be disjoint from the path $\alpha \cup \beta$. Since $\pi_1(\partial_+ C) \to \pi_1(C)$ is surjective there is a path γ in ∂C such that the closed curve $\alpha \cup \beta \cup \gamma$ is nullhomotopic in C. See Figure 7. Slide the end of β at q along γ to p so that β becomes an arc β' (parallel to the concatenation of γ and β) also from p to r, one that is homotopic to α rel endpoints. A sophisticated version of the light-bulb trick [5, Proposition 4] then shows that α and β' are isotopic rel endpoints. (Early versions of this paper appealed to the far more complex [2, Theorem 0] to provide such an isotopy.) $\square$

Proposition 3.4 *Suppose Δ and Δ' are snug collections of meridian disks for C. Then Δ can be made disjoint from Δ' by a sequence of handle slides and proper isotopies.*

Proof Let $\mathscr{F} = \{F_i\}$ for $1 \le i \le n$ be the collection of spherical boundary components of C. Since Δ (resp. Δ') is snug, to each F_i there corresponds a properly embedded arc α_i (resp. α_i') in C from F_i to $\partial_+ C$ and this arc determines the meridian disk in $D_i \subset \Delta$ (resp. $D_i' \subset \Delta'$) associated to F_i as described after Definition 3.1. The proof in the aspherical case (as outlined following Definition 2.2; see also [7]) was achieved by isotopies and slides reducing $|\Delta \cap \Delta'|$. In the general case the proof proceeds in two phases.

Phase 1 We will properly isotope the arcs $\{\alpha_i\}$ to $\{\alpha_i'\}$ for $1 \le i \le n$. The associated ambient isotopy of Δ in C may increase $|\Delta \cap \Delta'|$ but in this first phase we don't care. Once each $\alpha_i = \alpha_i'$, each snug disk D_i can be made parallel to D_i' by construction.

Pick a sphere component F_i with associated arcs α_i and α_i'. Isotope the end of α_i on F_i to the end r of α_i' at F_i. Temporarily attach a ball B to F_i and apply Lemma 3.3 to the arcs α and α', after which α and α' coincide. By general position, we can assume the isotopy misses the center b of B and by the

light-bulb trick that it never passes through the radius of B between b and r. Now use radial projection from b to push the isotopy entirely out of B and thus back into C.

Having established how to do the isotopy for a single α_i, observe that we can perform such an isotopy simultaneously on all α_i for $1 \le i \le n$. Indeed, anytime the isotopy of α_i is to cross α_j with $i \ne j$ we can avoid the crossing by pushing it along α_j, over the sphere F_j, and then back along α_j; in short, use the light-bulb trick.

Phase 2 We eliminate $\Delta \cap \Delta'$ by reducing $|\Delta \cap \Delta'|$, as in the aspherical case. After Phase 1, the disks $\{D_i\}$ for $1 \le i \le n$ are parallel to the disks $\{D_i'\}$ for $1 \le i \le n$; until the end of this phase we take them to coincide and also to be fixed, neither isotoped nor slid. Denote the complement in Δ (resp. Δ') of this collection of disks $\{D_i\}$ by $\hat{\Delta}$ (resp. $\hat{\Delta}'$), since they constitute a complete collection of meridians in $\hat{C}$. Moreover, the component of $C - \{D_i\}$ containing $\hat{\Delta}$ and $\hat{\Delta}'$ is homeomorphic to $\hat{C}$, so that is how we will designate that component.

Motivated by that last observation, we now complete the proof by isotoping and sliding $\hat{\Delta}$, much as in the aspherical case, to reduce $|\hat{\Delta} \cap \hat{\Delta}'|$. Suppose first there are circles of intersection and let $E' \subset \hat{\Delta}'$ be a disk with interior disjoint from $\hat{\Delta}$ cut off by an innermost such circle of intersection in $\hat{\Delta}'$. Then $\partial E'$ also bounds a disk $E \subset \hat{\Delta}$ (which may further intersect $\hat{\Delta}'$). Although C is no longer aspherical, the sphere $E \cup E'$ lies entirely in $\hat{C}$, which is aspherical, so $E \cup E'$ bounds a ball in $\hat{C}$, through which we can isotope E past E', reducing $|\hat{\Delta} \cap \hat{\Delta}'|$ by at least one.

Once all the circles of intersection are eliminated as described, we consider arcs in $\hat{\Delta} \cap \hat{\Delta}'$. An outermost such arc in $\hat{\Delta}'$ cuts off a disk E' from $\hat{\Delta}'$ that is disjoint from $\hat{\Delta}$; the same arc cuts off a disk E from $\hat{\Delta}$ (which may further intersect $\hat{\Delta}'$). The properly embedded disk $E \cup E' \subset \hat{C}$ has boundary on $\partial_+ \hat{C}$ and its interior is disjoint from Δ. The latter fact means that its boundary lies on one end of the collar $\hat{C} - \eta(\Delta)$ of a nonspherical component F of $\partial_- C$. But in a collar of F any properly embedded disk is ∂–parallel. Use the disk in the end of the collar (the other end from F itself) to which $E \cup E'$ is parallel to slide E past E' (possibly sliding it over other disks in Δ, including those in $\{D_i\}$), thereby reducing $|\hat{\Delta} \cap \hat{\Delta}'|$ by at least one.

Once $\hat{\Delta}$ and $\hat{\Delta}'$ are disjoint, slightly push the disks $\{D_i\}$ off the presently coinciding disks $\{D_i'\}$ so that Δ and Δ' are disjoint. $\qquad\square$

Energized by these observations we will now show that all the results of Section 2 remain true (in an appropriate form) in compression bodies that are not aspherical; that is, even when there are sphere components of $\partial_- C$. Here are the analogous results, with edits on statement in boldface, and proofs annotated as appropriate:

Lemma 3.5 (cf Lemma 2.4) *Suppose $\hat{\alpha}$ is a family of spanning arcs in compression body C.*

- *Suppose $\hat{\alpha}$ is vertical and c is an associated circle family. Then there is a family $\mathcal{A}$ of disjoint spanning annuli in C such that $\mathcal{A}$ contains $\hat{\alpha}$ and $\mathcal{A} \cap \partial_- C = c$.*

- *Suppose*, *on the other hand*, *there is a collection* $\mathscr{A}$ *of disjoint spanning annuli in* C *that contains* $\hat{\alpha}$. *Suppose further that*

  - **at most one arc in** $\hat{\alpha}$ **is incident to each sphere component of** ∂_-C, *and*
  - *in the family of circles* $\mathscr{A} \cap \partial_-C$ *associated to* $\hat{\alpha}$, *each circle* **lying in a nonspherical component of** ∂_-C *is essential.*

  Then α *is a vertical family.*

Proof　The proof of the first statement is unchanged.

For the second, observe that by [Lemma 2.4](#) there is a collection $\hat{\Delta}$ of meridian disks in $\hat{C}$ such that $\hat{\Delta}$ is disjoint from each arc $\alpha \in \hat{\alpha}$ that is incident to a nonspherical component of ∂_-C. By general position, $\hat{\Delta}$ can be taken to be disjoint from the balls $C - \hat{C}$ and so lie in C.

Now consider an arc $\alpha' \in \hat{\alpha}$ that is incident to a sphere F in ∂_-C. It may be that $\hat{\Delta}$ intersects α'. In this case, push a neighborhood of each point of intersection along α' and then over F. Note that this last operation is not an isotopy of $\hat{\Delta}$ in C, since it pops across F, but that's unimportant — afterwards the (new) $\hat{\Delta}$ is completely disjoint from α'. Repeat the operation for every component of $\hat{\alpha}$ that is incident to a sphere in ∂_-C, so that $\hat{\Delta}$ is disjoint from all of $\hat{\alpha}$. Now apply the proof of [Lemma 3.2](#), expanding $\hat{\Delta}$ by adding a snug meridian disk for each sphere in ∂_-C, using the corresponding arc in $\hat{\alpha}$ to define the snug meridian disk for spheres that are incident to $\hat{\alpha}$.　　　　□

Proposition 3.6　(cf [Corollary 2.7](#))　*Suppose* $(\mathscr{D}, \partial\mathscr{D}) \subset (C, \partial_+C)$ *is an embedded family of disks that is disjoint from an embedded family of vertical annuli* $\mathscr{A}$ *in* C. *Then there is a complete collection of meridian disks for* C *that is disjoint from* $\mathscr{A} \cup \mathscr{D}$.

Proof　Let $\alpha \subset \mathscr{A}$ be a vertical family of spanning arcs as given in [Definition 2.5](#). This means there is a complete collection Δ of meridian disks for C that is disjoint from α, so Δ intersects $\mathscr{A}$ only in inessential circles, and in arcs with both ends incident to the end of $\mathscr{A}$ at ∂_+C.

Let C' be the compression body obtained by attaching a ball to each sphere component of ∂_-C that is *not incident to* $\mathscr{A}$. Because Δ is a complete collection in C, it is also a complete collection in C', since attaching a ball to a collar of a sphere just creates a ball. Consider the curves $\Delta \cap (\mathscr{A} \cup \mathscr{D})$, and proceed as usual, much as in Phase 2 of the proof of [Proposition 3.4](#):

If there are circles of intersection, an innermost one in Δ cuts off a disk $E \subset \Delta$ and a disk $E' \subset (\mathscr{A} \cup \mathscr{D})$ which together form a sphere whose interior is disjoint from $\mathscr{A}$ and so bounds a ball in C'. In C', E' can be isotoped across E, reducing $|\Delta \cap (\mathscr{A} \cup \mathscr{D})|$. On the other hand, if there are no circles of intersection, then an arc of intersection γ outermost in $\mathscr{A} \cup \mathscr{D}$ cuts off a disk $E' \subset (\mathscr{A} \cup \mathscr{D})$ and a disk $E \subset \Delta$ which together form a properly embedded disk E'' in $C' - \Delta$ whose boundary lies on ∂_+C. Since E'' lies in $C' - \Delta$, it lies in a collar of ∂_-C' and so is parallel to a disk in the other end of the collar. (If the relevant component of ∂_-C' is a sphere, we may have to reset E to be the other half of the disk in Δ in which γ lies to accomplish this.) The disk allows us to slide E past E' and so reduce $|\Delta \cap (\mathscr{A} \cup \mathscr{D})|$.

The upshot is that eventually, with slides and isotopies, Δ can be made disjoint from $\Delta \cap (\mathscr{A} \cup \mathscr{D})$ *in* C'. The isotopies themselves can't be done in C, since sphere boundary components *disjoint from* $\mathscr{A}$ may get in the way, but the result of the isotopy shows how to alter Δ (not necessarily by isotopy) to a family of disks Δ' disjoint from $\mathscr{A} \cup \mathscr{D}$ that is complete in C'. Now apply the argument of Lemma 3.2, adding a snug disk to Δ' for each sphere component of $\partial_- C$ that was not incident to $\mathscr{A}$ and so bounded a ball in C'. These additional snug disks, when added to Δ', create a complete collection of meridian disks for C that is disjoint from $\mathscr{A} \cup \mathscr{D}$, as required. $\square$

Proposition 3.7 (cf Proposition 2.8) *Suppose α and β are vertical arcs in C with endpoints p and q in a component $F \subset \partial_- C$. Then α and β are properly isotopic in C.*

Proof If F is not a sphere, apply the argument of Proposition 2.8. If F is a sphere, apply Lemma 3.3. $\square$

Proposition 3.8 (cf Lemma 2.9) *Suppose $(\mathscr{D}, \partial \mathscr{D}) \subset (C, \partial_+ C)$ is an embedded family of disks that is disjoint from an embedded family of vertical annuli $\mathscr{A}$ in C. Suppose $\hat{\beta} = \{\beta_i\}$ for $i = 1, 2$ is a vertical family of arcs in C whose endpoints $p_i \in \partial_- C$ are disjoint from the family of circles $\hat{a} = \mathscr{A} \cap \partial_- C$ in $\partial_- C$. Then β can be properly isotoped rel $\{p_i\}$ so that it is disjoint from $\mathscr{A} \cup \mathscr{D}$.*

Proof The proof, like the statement, is essentially identical to that of Lemma 2.9, with this alteration when $F \subset \partial C_-$ is a sphere: Use Lemma 3.3 to isotope the vertical (hence parallel) pair $\hat{\beta}$ rel p_i until the arcs are parallel to the vertical family of spanning arcs of $\mathscr{A}$ that are incident to F. In particular, we can then take $\hat{\beta}$ to lie in the same collar $F \times I$ as $\mathscr{A}$ does, and to be parallel to $\mathscr{A}$ in that collar. It is then a simple matter, as in the proof of Lemma 2.9, to isotope each arc in $\hat{\beta}$ rel p_i very near to the concatenation of arcs σ_i disjoint from $\mathscr{A}$ and arcs α_i in $\mathscr{A}$ and, once so positioned, to push $\hat{\beta}$ off of $\mathscr{A} \cup \mathscr{D}$. $\square$

Let us now return to the world and language of Heegaard splittings with a lemma on verticality, closely related to ∂–reduction of Heegaard splittings.

Suppose $M = A \cup_T B$ is a Heegaard splitting of a compact orientable 3–manifold M and $(E, \partial E) \subset (M, \partial_- B)$ is a properly embedded disk, intersecting T in a single circle, so that the annulus $E \cap B$ is vertical in B and the disk $E \cap A$ is essential in A. Since $E \cap B$ is vertical, there is a complete collection of meridian disks Δ in the compression body B such that a component N of $B - \Delta$ is a collar of $\partial_- B$ in which $E \cap B$ is a vertical annulus. Parametrize E as a unit disk with center $b \in E \cap A$ and $E \cap B$ the set of points in E with radius $\frac{1}{2} \le r \le 1$. Let ρ be a vertical radius of E, with ρ_A the half in the disk $E \cap A$ and ρ_B the half in the annulus $E \cap B$.

Let $E \times [-1, 1]$ be a collar of the disk E in M and consider the manifold $M_0 = M - (E \times (-\epsilon, \epsilon))$, the complement of a thinner collar of E. It has a natural Heegaard splitting, obtained by moving the solid cylinders $(E \cap A) \times (-1, -\epsilon]$ and $(E \cap A) \times [\epsilon, -1)$ from A to B. Classically, this operation (when E is essential) is called ∂–reducing T along E [7, Definition 3.5]. We denote this splitting by $M_0 = A_0 \cup_{T_0} B_0$,
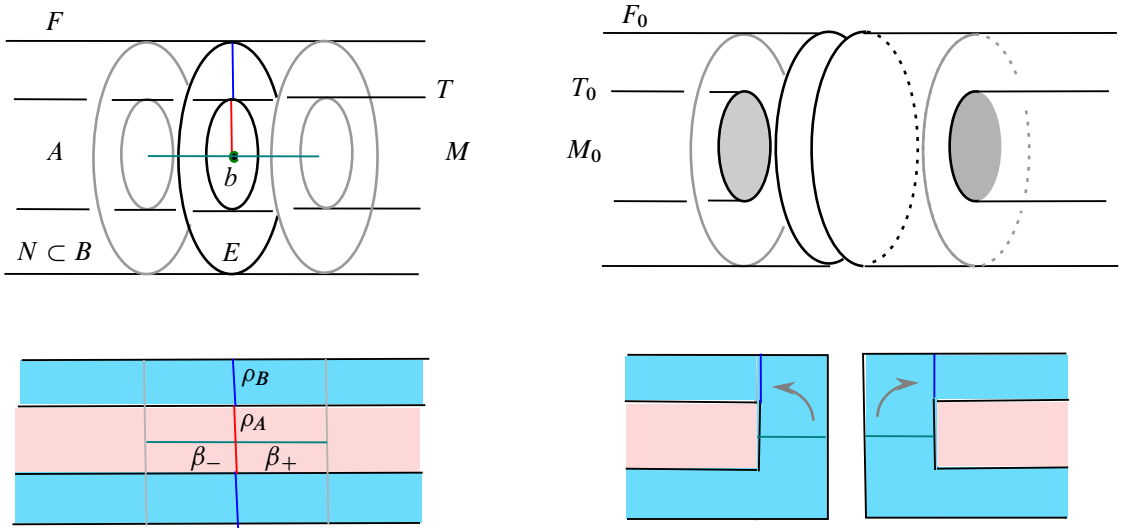
Figure 8

recognizing that if E is separating, it describes a Heegaard splitting of each component. Denote the spanning arcs $b \times [-1, -\epsilon]$ and $b \times [\epsilon, 1]$ in B_0 by β_- and β_+, respectively. See the top two panes of Figure 8, with a schematic rendering below.

Lemma 3.9 *The spanning arcs $\beta_\pm$ are a vertical family of arcs in B_0.*

Proof The complete collection of meridian disks Δ for B is disjoint from the annulus $E \cap B$, so remains in B_0. Viewed in the collar component $N \cong (F \times I)$ in the complement of Δ to which $E \cap B$ belongs, the operation described cuts the component $F \subset \partial_- B$ by $\partial E \subset F$, then caps off the boundary circles by disks to get a new surface F_0 and extends the collar structure to $F \times I$. The rectangles $\rho \times [\epsilon, 1]$ and $\rho \times [-1, -\epsilon]$ provide isotopies in M_0 from $\beta_\pm$ to the vertical arcs $\rho_B \times \{\pm 1\}$, illustrating that $\beta_\pm$ is a vertical family. $\square$

4 Reducing Theorem 1.3 to the case S is connected

To begin the proof of Theorem 1.3 note that (unsurprisingly) we may as well assume each component of S is essential; that is no sphere in S bounds a ball and no sphere or disk in S is ∂–parallel. This can be accomplished simply by isotoping all inessential components well away from T. So henceforth we will assume all components of S are essential, including perhaps disks whose boundaries are inessential in ∂M but which are not ∂–parallel in M.

Assign a simple notion of complexity (g, s) to the pair (M, T), with g the genus of T and s the number of spherical boundary components of M. We will induct on this pair, noting that there is nothing to prove if $g = 0$ and $s \leq 2$.

Suppose then that we are given a disk/sphere set $(S, \partial S) \subset (M, \partial M)$ in which all components are essential. We begin with:

Assumption 4.1 (inductive assumption) Theorem 1.3 is true for Heegaard splittings of manifolds that have lower complexity than that of (M, T).

With this inductive assumption we have:

Proposition 4.2 *It suffices to prove Theorem 1.3 for a single component S_0 of S.*

Proof Let $M = A \cup_T B$ be a Heegaard splitting, $S \subset M$ be a disk/sphere set, in which each component is essential in M, and let S_0 be a component of S that is aligned with T. The goal is to isotope the other components of S so that they are also aligned, using the inductive Assumption 4.1.

Case 1 S_0 is a sphere and $S_0 \cap T = \varnothing$ or an inessential curve in T.

If S_0 is disjoint from T, say $S_0 \subset B$, then it cuts off from M a punctured ball. This follows from Proposition 2.1, which shows that S_0 bounds a ball in the aspherical compression body $\hat{B}$ and so a punctured 3–ball in B itself. Any component of $S - S_0$ lying in the punctured 3–ball is automatically aligned, since it is disjoint from T. Removing the punctured 3–ball from B leaves a compression body B_0 with still at least one spherical boundary component, namely S_0. The Heegaard split $M_0 = A \cup_T B_0$ is unchanged, except there are fewer boundary spheres in B_0 than in B because S_0 is essential. Now align all remaining components of $S - S_0$ using the inductive assumption, completing the construction.

Suppose next that S_0 intersects T in a single circle that bounds a disk D_T in T, and S_0 can't be isotoped off of T. Then S_0 again bounds a punctured ball in M with $m \geq 1$ spheres of ∂M lying in A and $n \geq 1$ spheres of ∂M lying in B. S_0 itself is cut by T into hemispheres $D_A = S_0 \cap A$ and $D_B = S_0 \cap B$. A useful picture can be obtained by regarding D_A (say) as the cocore of a thin 1–handle in A connecting a copy A_+ of A with m fewer punctures to a boundary component $T_- = D_T \cup D_A$ of an m–punctured ball in A. In this picture, S_0 and T_- are parallel in $\hat{B}$; the interior of the collar between them has n punctures in B itself. See Figure 9.

Let β be the core of the 1–handle, divided by S_0 into a subarc β_+ incident to $T_+ = \partial A_+$ and β_- incident to the sphere T_-. Now cut M along S_0, dividing it into two pieces. One is a copy $M_+ = A_+ \cup_{T_+} B_+$ of M, but with m fewer punctures in A_+ and $n - 1$ fewer in B_+ (a copy of S_0 is now a spherical boundary component of B_+). The other is an $m + n + 1$ punctured 3–sphere M_-, Heegaard split by the sphere T_-. (Neither of the spanning arcs β_+ nor β_- play a role in these splittings yet.)

Now apply the inductive assumption to align T_+ and T_- with the disk/sphere set $S - S_0$ (not shown in Figure 9). Afterwards, reattach M_+ to M_- along the copies of S_0 in each. The result is again M, and S is aligned with the two parts T_- and T_+ in T. But to recover T itself, while ensuring that S remains aligned, we need to ensure that β can be properly isotoped rel S_0 so that it is disjoint from $S - S_0$. Such
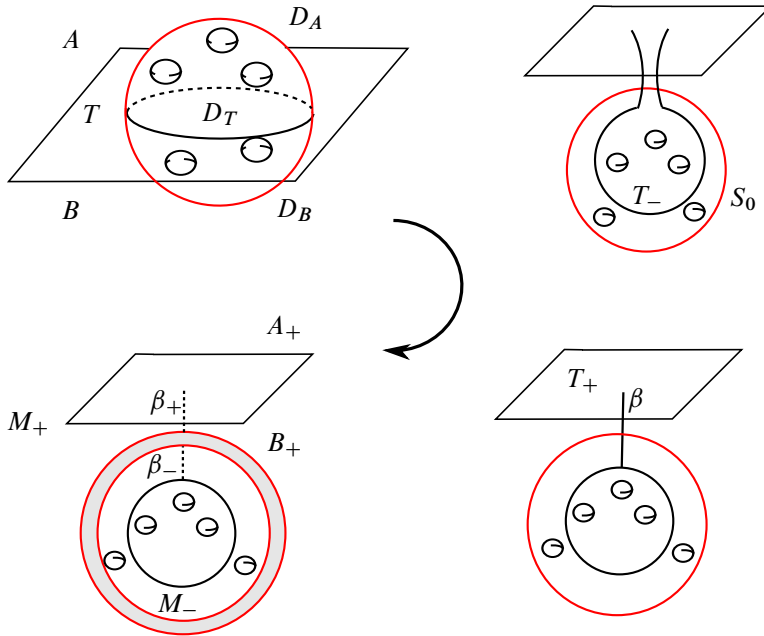
Figure 9: Clockwise through the inductive step in Case 1.

a proper isotopy of β will determine an isotopy of T by viewing β as the core of a tube (the remaining part of T) connecting T_+ to T_-. But once $S - S_0$ is aligned, the proper isotopy of β can be found by first applying Proposition 3.8 to β_+ and the family $S \cap B_0$ of disks and annuli in the compression body B_+ and then proceeding similarly with the arc β_- in M_-.

Case 2 S_0 is a sphere that intersects T in an essential curve.

As in Case 1, S_0 is cut by T into hemispheres $D_A = S_0 \cap A$ and $D_B = S_0 \cap B$ and we can consider D_A (say) as the cocore of a thin 1–handle in A. Continuing as in Case 1, denote the arc core of the 1–handle by β; S_0 again divides the arc β into two arcs which we label $\beta_\pm$.

If S_0 separates, then it divides M into two manifolds, say $M_\pm$ containing, respectively, $\beta_\pm$. Apply the same argument in each that was applied in Case 1 to the manifold M_+.

If S_0 is a nonseparating sphere then we can regard $S - S_0$ as a disk sphere set in the manifold $M_0 = M - \eta(S_0)$. Since S_0 is two-sided, two copies $S_\pm$ of S_0 appear as spheres in ∂M_0. Choose the labeling such that each arc $\beta_\pm$ has one end in the corresponding $S_\pm$. M_0 has lower complexity (the genus is lower) so the inductive assumption applies, and the spheres in $S - S_0$ can be aligned with T_0. Apply Proposition 3.8 to the arcs $\beta_\pm$ and then reconstruct (M, T), now with T aligned with S, as in Case 1.

Case 3 S_0 is a separating disk.

Suppose, with no loss of generality, that $\partial S_0 \subset \partial_- B$, so S_0 intersects A in a separating disk D_A and B in a separating vertical spanning annulus. As in the previous cases, let $M_\pm$ be the manifolds obtained

from M by cutting along S_0, β the core of the 1–handle in A whose cocore is D_A, and $\beta_\pm$ its two subarcs in $M_\pm$, respectively.

The compression body $A - \eta(D_A)$ consists of two compression bodies, $A_\pm$ in $M_\pm$, respectively. As described in the preamble to Lemma 3.9, the complement $B_\pm$ of $A_\pm$ in $M_\pm$ is a compression body, in which $\beta_\pm$ is a vertical spanning arc. So the surfaces $T_\pm$ obtained from T by compressing along D_A are Heegaard splitting surfaces for $M_\pm$, and the pairs $(M_\pm, T_\pm)$ have lower complexity than (M, T).

Now apply the inductive hypothesis: Isotope each of $T_\pm$ in $M_\pm$ so that they align with the components of $S - S_0$ lying in $M_\pm$. As in Case 1, apply Proposition 3.8 to each of $\beta_\pm$ and then reattach M_+ to M_- along disks in $\partial M_\pm$ centered on the points $\beta_\pm \cap \partial M_\pm$ and simultaneously reattach β_+ to β_- at those points. The result is an arc isotopic to β which is disjoint from $S - S_0$. Moreover, the original Heegaard surface T can be recovered from $T_\pm$ by tubing them together along β and, since β is now disjoint from $S - S_0$, all of T is aligned with S.

Case 4 S_0 is a nonseparating disk.

Near S_0 the argument is the same as in Case 3. Now, however, the manifold M_0 obtained by cutting along S_0 is connected. The construction of its Heegaard splitting $M_0 = A_0 \cup_{T_0} B_0$ and vertical spanning arcs $\beta_\pm$ proceeds as in Case 3, and, since genus$(T_0) = $ genus$(T) - 1$, we can again apply the inductive hypothesis to align $S - S_0$ with T_0.

If ∂S_0 separates the component F of $\partial_- B \subset \partial M$ in which it lies, say into surfaces $F_\pm$, the argument concludes just as in Case 3. If ∂S_0 is nonseparating in F, then we encounter the technical point that Proposition 3.8 requires that β be a vertical family of arcs. But this follows from Lemma 3.9.   □

5   Breaking symmetry: stem swaps

Applications of Lemma 3.3 extend beyond Propositions 2.8 and 3.8. But the arguments will require *breaking symmetry*: given a Heegaard splitting $M = A \cup_T B$ of a compact orientable 3–manifold M and Σ a spine for B, we can, and typically will, regard B as a thin regular neighborhood of Σ, with T as the boundary of that thin regular neighborhood. This allows general position to be invoked as if B were a graph embedded in M. Edge slides of Σ can be viewed as isotopies of T in M and therefore typically are of little consequence. We have encountered this idea in the previous section: the boundary of a tubular neighborhood of an arc β there represented an annulus in T; a proper isotopy of β was there interpreted as an isotopy of T. We can then regard A as the closure of $M - \eta(\Sigma)$; a properly embedded arc in A then appears as an arc whose interior lies in $M - \Sigma$ and whose endpoints may be incident to Σ. We describe such an arc as a properly embedded arc in A whose endpoints lie on Σ. This point of view is crucial to what follows; without it many of the statements might appear to be nonsense.

Let R be a sphere component of $\partial_- B$. Let Σ be a spine for B for which a single edge σ is incident to R.
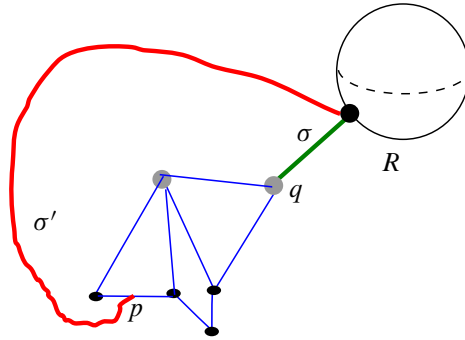
Figure 10: A stem swap for the case $p, q \notin \partial_- B \subset \Sigma$.

Definition 5.1 The complex $\sigma \cup R$ is called a *flower*, with σ the *stem* and R the *blossom*. The point $\sigma \cap R$ is the *base* of the blossom, and the other end of σ is the *base* of both the stem and the flower.

Now suppose σ' is a properly embedded arc in A from the base of the blossom R to a point p in $\Sigma - \sigma$. See Figure 10 for an example when p and q lie on edges of the spine.

Proposition 5.2 (stem swapping) *The complex Σ' obtained from Σ by replacing the arc σ with the arc σ' is, up to isotopy, also a spine for B. That is, T is isotopic in M to the boundary of a regular neighborhood of Σ'.*

Proof Given the spine Σ as described, there is a natural alternative Heegaard splitting for M in which R is regarded as lying in $\partial_- A$ instead of $\partial_- B$. It is obtained by deleting the flower $\sigma \cup R$ from Σ, leaving R as an additional component of $\partial_- A$. Call the resulting spine Σ_- and let A_+ be the complementary compression body (so $M = A_+ \cup_{T'} \eta(\Sigma_-)$). Apply the argument of Lemma 3.3 to A_+, with $\beta = \sigma$, $\alpha = \sigma'$ and $r = R$. (See Phase 1 of the proof of Proposition 3.4 for how we can regard the sphere R as the point r.) Let γ be the path in $\partial_+ A_+ = \partial(\eta(\Sigma_-))$ given by Lemma 3.3. Note that in Figure 10 some edges in the spine Σ_- are shown, but we do not claim that the path γ from Lemma 3.3 is a subgraph of Σ_-. Rather, the path is on the boundary of a *regular neighborhood* of Σ_- and does not necessarily project to an embedded path in Σ_- itself. Note further that after the stem swap the edge in Σ that contains p in its interior (if p is on an edge and not on $\partial_- B$) becomes two edges in Σ' and, dually, when q is not on $\partial_- B \subset \Sigma$, it is natural to concatenate the two edges of Σ that are incident to q into a single edge of Σ'.

Returning to the original splitting, sliding an end of σ along γ does not change the fact that Σ is a spine for B and, viewing T as the boundary of a regular neighborhood of Σ, the slide defines an isotopy of T in M. After the slide, according to Lemma 3.3, σ and σ' have the same endpoints at R and p; then σ can be isotoped to σ' rel its endpoints, completing the proof. (Note that passing σ through σ', as must be allowed to invoke Lemma 3.3, has no significance in this context.) $\qquad\square$

Definition 5.3 The operation of Proposition 5.2 in which we replace the stem σ with σ' is called a *stem swap*. If the base of the stem σ' is the same as that of σ, it is called a *local stem swap*.

Definition 5.4 Suppose $M = A \cup_T B$, and Σ is a spine for B. A sphere R_e that intersects Σ in a single point in the interior of an edge e is an *edge-reducing sphere* for Σ and the associated edge e is called a *reducing edge* in Σ.

There is a broader context in which we will consider stem swaps: Let $\mathfrak{R}$ be an embedded collection of edge-reducing spheres for Σ, chosen so that no edge of Σ intersects more than one sphere in $\mathfrak{R}$. (The latter condition, that each edge of Σ intersect at most one sphere in $\mathfrak{R}$, is discussed at the beginning of Section 8.) Let $M_\mathfrak{R}$ be a component of $M - \mathfrak{R}$ and $\mathfrak{R}_0 \subset \mathfrak{R}$ be the collection incident to $M_\mathfrak{R}$. (Note that a nonseparating sphere in $\mathfrak{R}$ may be incident to $M_\mathfrak{R}$ on both its sides. We will be working with each side independently, so this makes very little difference in the argument.)

For a sphere $R_e \in \mathfrak{R}_0$, and $e \in \Sigma$ the corresponding edge, the segment (or segments) $e \cap M_\mathfrak{R}$ can each be regarded as a stem in $M_\mathfrak{R}$, with blossom (one side of) R_e. A stem swap on this flower can be defined for an arc $\sigma' \subset M_\mathfrak{R}$ with interior disjoint from Σ that runs from the point $e \cap R_e$ to a point in $\Sigma \cap M_\mathfrak{R}$. Such a swap can be viewed in M as a way of replacing e with another reducing edge e' for R_e that differs from e inside of $M_\mathfrak{R}$, leaving the other segment (if any) of e inside $M_\mathfrak{R}$ alone.

Corollary 5.5 *If σ and σ' both lie in $M_\mathfrak{R}$, then the isotopy of T described in Proposition 5.2 can be assumed to take place entirely in $M_\mathfrak{R}$.*

Proof The manifold $M_\mathfrak{R}$ has a natural Heegaard splitting $M_\mathfrak{R} = A_\mathfrak{R} \cup_{T_0} B_\mathfrak{R}$ induced by that of M, in which each boundary sphere $R \in \mathfrak{R}_0$ is assigned to $\partial_- B_\mathfrak{R}$. We describe this construction:

Recall the setting: Σ is a spine for B and B itself is *a thin regular neighborhood* of Σ. Thus an edge-reducing sphere $R \in \mathfrak{R}$ intersects B in a tiny disk, centered at the point $R \cap \Sigma$. This disk is a meridian of the tubular neighborhood of the reducing edge that contains the point $R \cap \Sigma$. The rest of R, all but this tiny disk, is a disk lying in A. So R is a reducing sphere for the Heegaard splitting of M.

In the classical theory of Heegaard splittings — see eg [7] — such a reducing sphere naturally induces a Heegaard splitting for the manifold $\overline{M}$ obtained by reducing M along R; that is, $\overline{M}$ is obtained by removing an open collar $\eta(R)$ of the sphere R and *attaching* 3–*balls* to the two copies $R_\pm$ of R at the ends of the collar. The classical argument then gives a natural Heegaard splitting on each component of $\overline{M}$: replace the annulus $T \cap \eta(R)$ by equatorial disks in the two balls attached to $R_\pm$. Translated to our setting, the original spine Σ thereby induces a natural spine on each component of $\overline{M}$: the reducing edge is broken in two when $\eta(R)$ is removed, and at each side of the break, a valence-one vertex is attached, corresponding to the attached ball.

For understanding $M_\mathfrak{R}$, we don't care about $\overline{M}$ and the unconventional (because of the valence one vertex) spine just described. We care about the manifold $M - \eta(R)$, in which there are two new sphere boundary components created, but no balls are attached. But the classical construction suggests how to construct a natural Heegaard splitting for the manifold $M - \eta(R)$ and a natural spine for it: simply regard both spheres $R_\pm$ as new components of $\partial_- B$ and attach them at the breaks in the reducing edge where, above,
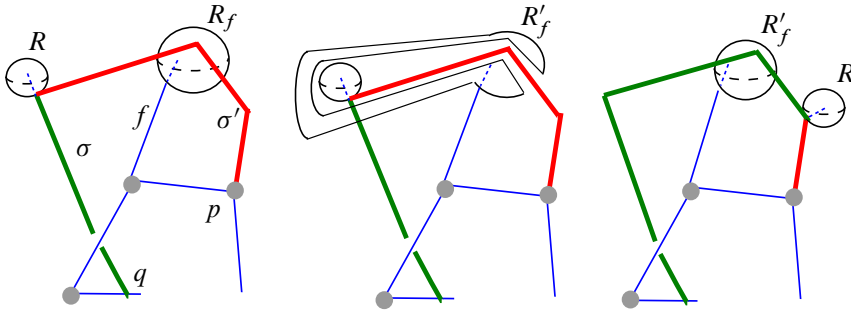
Figure 11: Blossoms R_f and R'_f.

we had added a valence 1 vertex. This Heegaard splitting for $M - \eta(R)$ is topologically equivalent to taking the classical construction of the splitting on $\overline{M}$ and removing two balls from the compression body $\overline{B}$.

When applied to all spheres in $\mathfrak{R}$ simultaneously, the result of this construction is a natural Heegaard splitting on each component of $M - \eta(\mathfrak{R})$. On $M_{\mathfrak{R}}$ it gives the splitting $A_{\mathfrak{R}} \cup_{T_0} B_{\mathfrak{R}}$ which was promised above, and also a natural spine $\Sigma_{\mathfrak{R}}$ for $B_{\mathfrak{R}}$. The required isotopy then follows, by applying Proposition 5.2 to the Heegaard splitting $M_{\mathfrak{R}} = A_{\mathfrak{R}} \cup_{T_0} B_{\mathfrak{R}}$, with $B_{\mathfrak{R}}$ a thin regular neighborhood of the spine $\Sigma_{\mathfrak{R}}$. $\quad\square$

Suppose, in a stem swap, that σ' intersects an edge-reducing sphere R_f, with associated edge $f \neq \sigma$. See the first panel of Figure 11. (Note that f is an edge in Σ but if $p \in f$ then f becomes two edges in Σ'.) Although R_f is no longer an edge-reducing sphere for Σ', there is a natural way to construct a corresponding edge-reducing sphere R'_f for Σ', one that intersects f in the same point, but now intersects σ instead of σ'. At the closest point in which σ' intersects R_f, tube a tiny neighborhood in R_f of the intersection point to its end at R and then around R. Repeat until the resulting sphere is disjoint from σ', as shown in the second panel of Figure 11. One way to visualize the process is to imagine ambiently isotoping R'_f, in a neighborhood of σ', to the position of R_f, as shown in the third panel of Figure 11. The effect of the ambient isotopy is as if R is a bead sitting on the embedded arc $\sigma \cup \sigma'$ and the ambient isotopy moves the bead along this arc and through R_f. We will call R'_f the *swap-mate* of R_f (and vice versa).

Here is an application.

Suppose R_0 is a reducing sphere for a reducing edge $e_0 \in \Sigma$ and $\sigma \subset e_0$ is one of the two segments into which R_0 divides e_0. Let $\sigma' \subset A - R_0$ be an arc whose ends are the same as those of σ but is otherwise disjoint from σ. Let e'_0 be the arc obtained from e by replacing σ with σ'. Let $\eta(R_0)$ be the interior of a collar neighborhood of R_0 on the side away from σ.

Viewing $\sigma \cup R_0$ as a flower in the manifold $M - \eta(R_0)$, and the substitution of σ' for σ as a local stem swap, it follows from the proof of Proposition 5.2 that the 1–complex Σ' obtained from Σ by replacing e_0 with e'_0 is also a spine for B. That is, T is isotopic in M to the boundary of a regular neighborhood of Σ'. Moreover, e'_0 remains a reducing edge in Σ' with edge-reducing sphere R_0.
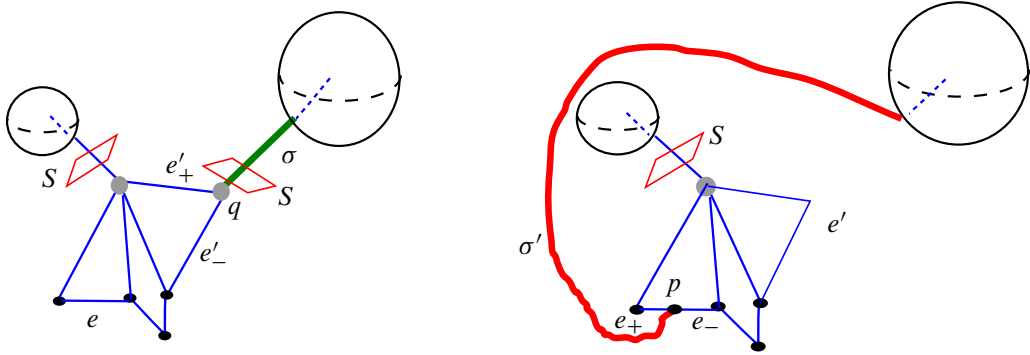
Figure 12: Spines Σ and Σ'.

With this as context, we have:

Lemma 5.6 *Suppose $\mathcal{E}$ is a collection of edges in Σ, with $e_0 \in \mathcal{E}$, and let $\mathcal{E}_r \subset \mathcal{E}$ be the set of reducing edges for Σ that lie in $\mathcal{E}$. Similarly, suppose $\mathcal{E}'$ is a collection of edges in Σ' containing the edge e_0' constructed above, and $\mathcal{E}_r' \subset \mathcal{E}'$ is the set of reducing edges for Σ' that lie in $\mathcal{E}'$. If $\mathcal{E}' - e_0' \subset \mathcal{E} - e_0$ then $\mathcal{E}_r' - e_0' \subset \mathcal{E}_r - e_0$.*

Proof Let f be an edge in $\mathcal{E}_r'$ other than e_0', and R_f' be a corresponding edge-reducing sphere for Σ'. Then R_f' is disjoint from e_0', so, although it may intersect e_0, any intersection points lie in $\sigma \subset e_0$. The swap-mate R_f of R_f' then may intersect σ' but by construction it will not intersect σ. Hence, R_f is disjoint from e_0 (as well as all edges of Σ other than f). Hence, R_f is an edge-reducing sphere for Σ and $f \in \mathcal{E}_r$. $\square$

Consider as usual a Heegaard splitting $M = A \cup_T B$, where B is viewed as a thin regular neighborhood of a spine Σ. Suppose $\mathcal{E}$ is a collection of edges in Σ and $\mathcal{E}_r \subset \mathcal{E}$ is the set of reducing edges for Σ that lie in $\mathcal{E}$. (For example, $\mathcal{E}$ might be the set of edges that intersects a specific essential sphere S in M, as in the discussion that will follow Corollary 5.5. This motivates the appearance of the red parallelograms in Figure 12.) Suppose $\mathfrak{R}$ is an embedded collection of edge-reducing spheres for Σ, one associated to each edge in $\mathcal{E}_r$. Let $M_{\mathfrak{R}}$ be a component of $M - \mathfrak{R}$ and consider a sphere $R_0 \in \mathfrak{R}_0 \subset \partial M_{\mathfrak{R}}$. Then, as just described before Lemma 5.6, a segment of the associated reducing edge e_0 that lies in $M_{\mathfrak{R}}$ can be regarded in $M_{\mathfrak{R}}$ as a stem σ with blossom R_0. (The rest of e_0 is shown as a dotted extension in Figure 12.) Let σ' be another arc properly embedded in $M_{\mathfrak{R}}$ which has the same ends as σ but is otherwise disjoint from Σ, and let Σ' be the spine for B constructed as above for the local stem swap of σ to σ'. Notice that because $\mathrm{int}(M_{\mathfrak{R}})$ is disjoint from the spheres $\mathfrak{R}$, $\mathrm{int}(\sigma') \subset \mathrm{int}(M_{\mathfrak{R}})$ is also disjoint from $\mathfrak{R}$.

Proposition 5.7 *Suppose $\mathcal{E}'$ is a subcollection of the edges $\mathcal{E} - e_0$, together possibly with the edge e_0', and denote by $\mathcal{E}_r' \subset \mathcal{E}'$ the set of reducing edges for Σ' in $\mathcal{E}'$. There is a collection of edge-reducing spheres $\mathfrak{R}'$ for Σ', one associated to each edge in $\mathcal{E}_r'$, such that $\mathfrak{R}' \subset \mathfrak{R}$.*

Proof From Lemma 5.6 we know that $\mathscr{E}'_r - e'_0 \subset \mathscr{E}_r - e_0$. Since σ' is in $M_{\mathfrak{R}}$, it is disjoint from $\mathfrak{R}$, so for each edge f in $\mathscr{E}'_r - e'_0$ we can just use the corresponding edge reducing sphere for f in Σ. In the same vein, since R_0 is disjoint from σ', R_0 is an edge-reducing sphere for e'_0 in Σ'. $\quad\square$

There is an analogous result for more general stem swaps, but it is more difficult to formulate and prove. To that end, suppose $\sigma' \subset M_{\mathfrak{R}}$ has one end at the base of R_0 and the other at a point $p \in \Sigma$. Here p is not a vertex of Σ, nor a point in $\mathfrak{R}$, and $\text{int}(\sigma')$ is disjoint from Σ. If p lies on an edge of Σ, the edge is not one that is also incident to the base point q of σ.

Consider the stem swap as described in Proposition 5.2. After the stem swap, one difference between the two spines Σ and Σ' (other than the obvious switch from σ to σ') is that if p lies on an edge $e \subset \Sigma$ then e becomes two edges $e_{\pm}$ in Σ' and if the base point q of σ lies on an edge $e' \subset \Sigma'$ then e' began as two edges $e'_{\pm}$ in Σ. See Figure 12.

Definition 5.8 A collection of edges $\mathscr{E}'$ in Σ' is *consistent with the swap* of σ to σ' (or *swap-consistent*) if, when p and/or q lie on edges as just described, $\mathscr{E}'$ has these properties:

- $\mathscr{E}' - \{e_{\pm}, e', \sigma'\} \subset \mathscr{E}$.
- If either $e_{\pm}$ is in $\mathscr{E}'$ then $e \in \mathscr{E}$.
- If both $e'_{\pm} \notin \mathscr{E}'$ then $e' \notin \mathscr{E}'$. Or, equivalently, if $e' \in \mathscr{E}'$ then at least one of $e'_{\pm} \in \mathscr{E}$.
- Suppose e is a reducing edge in $\mathscr{E}$ with R_e the corresponding edge-reducing sphere in $\mathfrak{R}$. Then the segment e_+ or e_- not incident to R_e is not in $\mathscr{E}'$. There must be such a segment since by hypothesis $p \notin \mathfrak{R}$.

(In the case that p and/or q lie on $\partial_- B \subset \Sigma$, so the edges e and/or e' are not defined, statements about these edges are deleted.)

Lemma 5.9 *Suppose $\mathscr{E}'$ is consistent with the swap described above. Then there is collection of edge-reducing spheres $\mathfrak{R}'$ for Σ', one associated to each reducing edge in $\mathscr{E}'$, such that $\mathfrak{R}' \subset \mathfrak{R}$.*

Proof Consider any reducing edge $f \in \mathscr{E}'$. If $f = \sigma'$ use R_0 for the corresponding sphere in $\mathfrak{R}'$. In any other case, since f is a reducing edge for an edge in Σ', a corresponding edge-reducing sphere R'_f is automatically disjoint from $\text{int}(\sigma')$ since R'_f only intersects Σ' in a single point. Its swap-mate R_f is then an edge-reducing sphere for Σ, because it is disjoint from $\text{int}(\sigma)$. We do not know that $R_f \in \mathfrak{R}$ and in fact it can't be if $\text{int}(\sigma')$ intersects R_f, since σ' was chosen, following Proposition 5.7, to be in $M_{\mathfrak{R}}$. With this in mind, consider the possibilities:

If $f \notin \{e_{\pm}, e', \sigma'\}$ then $f \in \mathscr{E}$, since $\mathscr{E}'$ is consistent with the swap. Then R_f is an edge-reducing sphere for f in Σ, so f is a reducing edge in $\mathscr{E}$. As originally defined prior to Proposition 5.7, $\mathscr{E}_r$ is the set of reducing edges in $\mathscr{E}$, so $f \in \mathscr{E}_r$. Since $\mathfrak{R}$ contains an edge-reducing sphere for each edge in $\mathscr{E}_r$,

$\mathfrak{R}$ contains an edge-reducing sphere for f. By construction this sphere is disjoint from both $\mathrm{int}(\sigma)$ and $\mathrm{int}(\sigma')$, the latter by choice of σ'. Include this as the sphere in $\mathfrak{R}'$ that corresponds to f.

As noted at the start, if $f = \sigma'$, use R_0.

If $f = e'$ then one of $e'_{\pm}$, say e'_+, is in $\mathscr{E}$, since $\mathscr{E}'$ is consistent with the swap. R'_f may as well be taken to pass through $e'_+ \subset e'$. Then R_f is an edge-reducing sphere for Σ that passes through e'_+. Hence, e'_+ is a reducing edge in $\mathscr{E}$. The edge-reducing sphere in $\mathfrak{R}$ corresponding to e'_+ is again disjoint from both $\mathrm{int}(\sigma)$ and $\mathrm{int}(\sigma')$. Include this as the sphere in $\mathfrak{R}'$ that corresponds to f.

If f is one of the edges $e_{\pm}$, say e_+, then $e \in \mathscr{E}$, since $\mathscr{E}'$ is consistent with the swap. As before, the sphere R_f shows that e is a reducing edge for Σ and so has a corresponding edge-reducing sphere R in $\mathfrak{R}$. Include it in $\mathfrak{R}'$ to correspond to $f = e_+$. The last condition in Definition 5.8 ensures that $e_- \notin \mathscr{E}'$, so no corresponding edge-reducing sphere is included in $\mathfrak{R}'$. In simple terms, R appears only once in $\mathfrak{R}'$. The condition also ensures that f is the subedge of e in Σ' that is incident to R. □

6 When $\partial S \subset \partial_- B \subset \partial M$: early considerations

We will begin the proof of Theorem 1.3 in the case that S is connected. In conjunction with Proposition 4.2, this will complete the proof of Theorem 1.3.

6.1 Preliminary remarks

What will be most important for our purposes is not that S is connected, but that S is entirely disjoint either from all of $\partial_- A$ or all of $\partial_- B$, as is naturally the case when S is connected. So we henceforth assume with no loss of generality that $\partial S \subset \partial_- B$. Following that assumption, the compression bodies A and B play very different roles in the proof. We will be studying spines of B and will take for A the complement in M of a regular neighborhood $\eta(\Sigma)$ of such a spine Σ. In particular, each sphere component R of $\partial_- B$ is part of Σ. As noted in the discussion of spines following Definition 2.2, we can choose Σ so that each sphere component R is incident to exactly one edge of Σ; in that case we are in a position to apply the key idea of stem swapping to alter Σ, as in Proposition 5.2.

In contrast, the sphere components of $\partial_- A$ play almost no role in the proof, other than requiring a small change in language. Since in Theorem 1.3 the isotopy class of S remains fixed (indeed, that is the point of the theorem), we must be careful not to pass any part of S through a sphere component of $\partial_- A$, but the constructions we make use of will avoid this. For example, underlying a stem swap in Σ is the slide and isotopy of an edge of Σ. (See Proposition 5.2.) But these can be made to avoid sphere components of $\partial_- A$, essentially by general position. More explicitly, let $\widehat{M}$ be the 3–manifold obtained from M by attaching a ball to each sphere component of $\partial_- A$. A slide or isotopy of an edge of Σ can avoid the centers of these balls by general position, and then be radially moved outside the entire balls and back into A.

A more subtle problem arises when, for example, we want to use a classical innermost disk (or outermost arc) argument to move a surface F in A so that it is disjoint from S. In the classical setting we find a circle c in $F \cap S$ that bounds a disk $E_S \subset S - F$ and a disk $E_F \subset F$ and argue that one can isotope E_F past E_S, reducing the number of intersections, via a ball whose boundary is the sphere $E_F \cup E_S$. But the existence of such a ball requires A to be irreducible, an assumption that fails when $\partial_- A$ contains spheres. It will turn out that this fraught situation can always be avoided here by *redefining* F to be the surface obtained by a simple disk-exchange, replacing $E_F \subset F$ with a push-off of $E_S \subset S$.

A useful way to visualize and describe this process of redefining F is to imagine, both in the argument and in the figures, a host of bubbles floating around in A, corresponding to sphere components of $\partial_- A$. These bubbles cannot pass through S (or Σ), but typically each bubble can pass "through" other surfaces we construct, in the sense that, when needed, the constructed surface F can be redefined to pass on the other side of the bubble. As shorthand for this process (which we have already seen in Phase 2 of the proof of Proposition 3.4) we will describe the process as a *porous isotopy* of F (equivalent to an actual isotopy in $\widehat{M}$), since the bubbles appear to pass through F.

6.2 The argument begins

Let Σ denote a spine of B and, as usual, take B to be a thin regular neighborhood of Σ.

Let $(\Delta, \partial\Delta) \subset (A, T)$ be a collection of meridian disks for A that constitute a complete collection of meridian disks for $\widehat{A}$, the compression body obtained from A by capping off all spherical boundary components by balls. Let $B_+ = B \cup \eta(\Delta)$; since Δ is complete for $\widehat{A}$, the complement of B_+ is the union of punctured balls and a punctured collar of $\partial_- A \subset \partial M$. The deformation retraction of B to Σ will carry Δ to disks in $M - \Sigma$; continue to denote these by Δ.

Suppose an edge e of Σ is disjoint from Δ. A point on e corresponds to a meridian of B whose boundary lies on ∂B_+. If it is inessential in ∂B_+ then it bounds a disk in A, so such a meridian can be completed to a sphere intersecting e in a single point. In other words, e is a reducing edge of Σ.

The other possibility is that the boundary of the meridian disk for e is essential on ∂B_+, so it, together with an essential curve in $\partial_- A$, bounds an essential spanning annulus $a_e \subset A$. Together, the meridian disk of e and the annulus a_e comprise a boundary reducing disk for M, in fact one that also ∂–reduces the splitting surface T. (In particular, the disk is aligned with T.) We will eliminate from consideration this possibility by a straightforward trick, which we now describe.

Lemma 6.1 *There is a collection $\mathscr{C} \subset \partial_- A$ of disjoint essential simple closed curves with the property that $\mathscr{C}$ intersects any essential simple closed curve in $\partial_- A$ that bounds a disk in M.*

Proof Suppose A_0 is a genus $g \geq 1$ component of $\partial_- A$. By standard duality arguments, the collection $K \subset A_0$ of simple closed curves that compress in M can generate at most a g–dimensional subspace of $H_1(A_0, \mathbb{R}) \cong \mathbb{R}^{2g}$. More specifically, one can find a nonseparating collection $c_1, \ldots, c_g$ of disjoint simple closed curves in A_0 such that $\mathscr{C}_- = \bigcup_{i=1}^{g} c_i$ generates a complementary g–dimensional subspace

of $H_1(A_0, \mathbb{R})$, and therefore essentially intersects any *nonseparating* curve in K. It is easy to add to $\mathscr{C}_-$ a further disjoint collection of $2g - 3$ simple closed curves, each nonseparating, so that the result $\mathscr{C}_0 \subset A_0$ has complement a collection of $2g - 2$ pairs of pants. Any curve in A_0 that is disjoint from $\mathscr{C}_0$ is parallel to a curve in $\mathscr{C}_0$ and so must be nonseparating. Since it is disjoint from $\mathscr{C}_- \subset \mathscr{C}_0$ it cannot be in K.

Do the same in each component of $\partial_- A$; the result is the required collection C.   □

Following Lemma 6.1 add to the collection of disks Δ the disjoint collection of annuli

$$\mathscr{C} \times I \subset \partial_- A \times I \subset M - B_+,$$

and continue to call the complete collection of meridional disks and these spanning annuli Δ. Then a meridian of an edge e of Σ that is disjoint from the (newly augmented) Δ cannot be part of a ∂–reducing disk for T and so must be part of a reducing sphere. Since the collection S of reducing spheres and ∂–reducing disks we are considering have no contact with $\partial_- A$, arcs of $S \cap \Delta$ are nowhere incident to $\partial_- A$. Additionally, no circle in $S \cap \Delta$ can be essential in an annulus in $\mathscr{C} \times I$, since no circle in $\mathscr{C}$ bounds a disk in M. Hence, the annuli which we have added to Δ intersect S much as a disk would: each circle of intersection bounds a disk in the annulus and each arc of intersection cuts off a disk from the same end of the annulus. As a result, the arguments cited below, usually applied to disk components of Δ, apply also to the newly added annuli components $\mathscr{C} \times I$.

7   Reducing edges and S

Lemma 7.1   *Suppose a spine Σ for B and a collection Δ of meridians and annuli, as just described, have been chosen to minimize the pair $(|\Sigma \cap S|, |\partial \Delta \cap S|)$ (lexicographically ordered, with Σ, S and Δ all in general position). Then Σ intersects $\mathrm{int}(S)$ only in reducing edges.*

Notes:

- We do not care about the number of circles in $\Delta \cap S$.
- If S is a disk and intersects Σ transversally only in $\partial S \subset \partial_- B$, then S is aligned with $T = \partial(\eta(\Sigma))$ and intersects B in a vertical annulus, completing the proof of Theorem 1.3 in this case. In addition, S is a ∂–reducing disk for T if ∂S is essential in $\partial_- B$.
- If S is a sphere and intersects Σ transversally only in a single point, then S is aligned with T, completing the proof of Theorem 1.3 in this case. Moreover, if the circle $S \cap T$ is essential in T, S is a reducing sphere for T.

Proof   Recall from a standard proof of Haken's theorem — see eg [7; 9, Proposition 2.2] — that $(\Sigma \cup \Delta) \cap S$ (ignoring circles of intersection) can be viewed as a graph Γ in S in which points of $\Sigma \cap S$ are the vertices and $\Delta \cap S$ are the edges. As discussed in [9] in the preamble to Proposition 2.2 there, this is accomplished

by extending the disks and annuli Δ via a retraction $B \to \Sigma$ so that it becomes a collection of disks and annuli whose embedded interior is disjoint from Σ and whose singular boundary lies on Σ. When S is a disk we will, with slight abuse of notation, also regard ∂S as a vertex in the graph, since it lies in $\partial_- B \subset \Sigma$. (This can be made sensible by imagining capping off ∂S by an imaginary disk outside of M.)

Borrowing further from the preamble to [9, Proposition 2.2], an edge in Γ is a loop if both ends lie on the same vertex, called the base vertex for the loop. A loop is *inessential* if it bounds a disk in S whose interior is disjoint from Σ, otherwise it is *essential*. A vertex in Γ is *isolated* if it is incident to no edge in Γ.

It is shown in [9] that if Σ and Δ are chosen to minimize the pair $(|\Sigma \cap S|, |\partial \Delta \cap S|)$ then

- there are no inessential loops,
- any innermost loop in the graph Γ bounds a disk in S that contains only isolated vertices, and
- if there are no loops in Γ then every vertex is isolated.

It follows that either S is disjoint from Σ (so it is aligned and we are done) or there is at least one isolated vertex. An isolated vertex represents a point p in an edge e of Σ which is incident to no element of Δ. The point p defines a meridional disk D_B of $B = \eta(\Sigma)$, and the fact that the curve $\partial D_B \subset \partial_+ A$ is disjoint from Δ ensures that ∂D_B is parallel to a curve in $\partial_- A$ that is inessential. Thus ∂D_B also bounds a disk D_A in A. Then $D_A \cup D_B$ is a reducing sphere, so e is a reducing edge in Σ. This establishes the original Haken theorem and, if there are no loops at all, also Lemma 7.1. That there are no loops is what we now show.

Consider an innermost loop, consisting of a vertex $p \in \Sigma \cap S$ and an edge lying in a component D of Δ. Together, they define a circle c in S that bounds a disk $E \subset S$ whose interior, by the argument of [9, Proposition 2.2], contains only isolated vertices and so intersects Σ only in reducing edges. Remembering that we are taking $A = M - \eta(\Sigma)$, the 3–manifold $A_- = A - \eta(D)$ can be viewed as $M - \eta(D \cup \Sigma)$, so c is parallel in E to a circle c' in ∂A_- bounding a subdisk E_- of E. E_- is the complement in E of the collar in E between c and c'. Since E_- intersects Σ only in reducing edges, it follows immediately that c' is nullhomotopic in A_- and then by Dehn's lemma that it bounds an embedded disk E' entirely in A_-.

By standard innermost disk arguments we can find an E' such that its interior is disjoint from Δ. Now split D in two by compressing the loop to the vertex along E' and replace D in Δ by these two pieces, creating a new complete (for $\hat{A}$) collection of disks and annuli Δ', with $|\partial \Delta' \cap S| \leq |\partial \Delta \cap S| - 2$. Since we have introduced no new vertices, this contradicts our assumption that $(|\Sigma \cap S|, |\partial \Delta \cap S|)$ is minimal. $\square$

Note that the new Δ' may intersect S in many more circles than Δ did, but we don't care.

8 Edge-reducing spheres for Σ

Recall from Section 5 that, given a reducing edge e in Σ, an associated edge-reducing sphere R_e is a sphere in M that passes once through e. Any other edge-reducing sphere R'_e passing once through e is

porously isotopic to R_e in M (ie isotopic in $\widehat{M}$) via edge-reducing spheres. Indeed, the segment of e between the points of intersection with Σ provides an isotopy from the meridian disk $R_e \cap B$ to $R'_e \cap B$; this can be extended to a porous isotopy of $R_e \cap A$ to $R'_e \cap A$ since $\widehat{A}$ is irreducible. So R_e is well-defined up to porous isotopy.

Let Σ be a spine for B in general position with respect to the disk/sphere S, and suppose $\mathscr{E}$ is a collection of edges in Σ. Let $\mathfrak{R}$ be a corresponding embedded collection of edge-reducing spheres transverse to S, one for each reducing edge in $\mathscr{E}$. Let $|\mathfrak{R} \cap S|$ denote the number of components of intersection.

Definition 8.1 The *weight* $w(\mathfrak{R})$ of $\mathfrak{R}$ is $|\mathfrak{R} \cap S|$. Porously isotope $\mathfrak{R}$ via edge-reducing spheres so that its weight is minimized, and call the result $\mathfrak{R}(\mathscr{E})$. Then the *weight* $w(\mathscr{E})$ of $\mathscr{E}$ is $w(\mathfrak{R}(\mathscr{E}))$.

Consider the stem swap as defined in Proposition 5.2 and Corollary 5.5 and suppose $\mathscr{E}'$ is a collection of edges in Σ that is swap-consistent with $\mathscr{E}$.

Lemma 8.2 *There is a collection $\mathfrak{R}'$ of edge-reducing spheres for Σ', one for each reducing edge in $\mathscr{E}'$ such that $w(\mathfrak{R}') \leq w(\mathfrak{R})$.*

Proof This is immediate from Lemma 5.9. □

Corollary 8.3 *Suppose in Lemma 8.2 that $\mathfrak{R}$ is $\mathfrak{R}(\mathscr{E})$. Then $w(\mathscr{E}') \leq w(\mathscr{E})$.*

Proof Let $\mathfrak{R}'$ be the collection of spheres given in Lemma 8.2. By definition $w(\mathscr{E}') \leq w(\mathfrak{R}')$ so, by Lemma 8.2,
$$w(\mathscr{E}') \leq w(\mathfrak{R}') \leq w(\mathfrak{R}) = w(\mathfrak{R}(\mathscr{E})) = w(\mathscr{E}). \qquad \square$$

Here is a motivating example: For Σ a spine of B in general position with respect to S, let $\mathscr{E}$ be the set of edges that intersect S, with the set of edge-reducing spheres $\mathfrak{R} = \mathfrak{R}(\mathscr{E})$ corresponding to the reducing edges of $\mathscr{E}$. As usual, let $M_{\mathfrak{R}}$ be a component of $M - \mathfrak{R}$ and $\mathfrak{R}_0$ be the collection of spheres in $\partial M_{\mathfrak{R}}$ that comes from $\mathfrak{R}$. Suppose R_0 is a sphere in $\mathfrak{R}_0$ with stem σ, and suppose σ' is an arc in $M_{\mathfrak{R}}$ from the base of R_0 to a point p in an edge e of Σ, very near an end vertex of e, so that the subinterval of e between p and the end vertex does not intersect S.

Perform an edge swap and choose $\mathscr{E}'$ to be the set of edges in Σ' that intersect S.

Proposition 8.4 $\mathscr{E}'$ *is swap-consistent with* $\mathscr{E}$.

Proof All but the last property of Definition 5.8 is immediate, because S will intersect an edge if and only if it intersects some subedge. The last property of Definition 5.8 follows from our construction: since σ' lies in a component $M_{\mathfrak{R}}$ of $M - \mathfrak{R}$, the point p lies between the sphere in $\mathfrak{R}$ corresponding to e and an end vertex v of e, and the segment of e between p and v is disjoint from S by construction and therefore not in $\mathscr{E}'$. □

Define the weight $w(\Sigma)$ of Σ to be $w(\mathcal{E})$, and similarly $w(\Sigma') = w(\mathcal{E}')$.

Corollary 8.5 *Given a stem swap as described in Propositions 5.2 or 5.7 for $\mathfrak{R}(\mathcal{E})$, $w(\Sigma') \leq w(\Sigma)$.*

Proof This follows immediately from Proposition 8.4 and Corollary 8.3. □

We will need a modest variant of Corollary 8.5 that is similar in proof but a bit more complicated. As before, let $\mathcal{E}$ be the set of edges in a spine Σ that intersect S, with the set of edge-reducing spheres $\mathfrak{R} = \mathfrak{R}(\mathcal{E})$ corresponding to the reducing edges of $\mathcal{E}$. Suppose $e_0 \in \mathcal{E}$ with corresponding edge-reducing sphere $R_0 \in \mathfrak{R}$. Then, by definition,

$$w(\Sigma) = w(\mathcal{E}) = w(\mathfrak{R}) = w(\mathfrak{R} - R_0) + w(R_0) = w(\mathfrak{R} - R_0) + |R_0 \cap S|.$$

Let $\mathfrak{R}_- = \mathfrak{R} - R_0$, $\mathcal{E}_- = \mathcal{E} - e_0$ and $M_{\mathfrak{R}_-}$ be the component of $M - \mathfrak{R}_-$ that contains R_0. Perform an edge swap *in $M_{\mathfrak{R}_-}$* as in the motivating example: replace the stem σ of a sphere $\mathfrak{a}$ in $\mathfrak{R}_-$ with σ', an arc in $M_{\mathfrak{R}_-}$ from the base of $\mathfrak{a}$ to a point p in an edge e of Σ, very near an end vertex of e, so that the subinterval of e between p and the end vertex does not intersect S. Notice that, in this set-up, R_0 is essentially invisible: the new stem σ' is allowed to pass through R_0. The swap-mate R_0' of R_0 is an edge-reducing sphere for e_0 in Σ' that is disjoint from $\mathfrak{R}_- = \mathfrak{R} - R_0$

As in the motivating example, let $\mathcal{E}'$ be the set of edges in Σ' that intersects S and further define $\mathcal{E}_-' = \mathcal{E}' - e_0$.

Proposition 8.6 $\qquad\qquad w(\Sigma') \leq w(\Sigma) - |R_0 \cap S| + |R_0' \cap S|.$

Proof As in the motivating example, $\mathcal{E}_-'$ is consistent with the swap, so by Lemma 5.9 there is a collection $\mathfrak{R}_-' \subset \mathfrak{R}_- = \mathfrak{R} - R_0$ of edge-reducing spheres associated to the edge-reducing spheres of $\mathcal{E}_-'$. Then $\mathfrak{R}_-' \cup R_0'$ is a collection of edge-reducing spheres for $\mathcal{E}'$. Thus,

$$w(\Sigma') = w(\mathcal{E}') \leq w(\mathfrak{R}_-') + w(R_0') \leq w(\mathfrak{R}_-) + w(R_0') = w(\mathfrak{R}) - w(R_0) + w(R_0')$$
$$= w(\Sigma) - w(R_0) + w(R_0'). \qquad \square$$

9   Minimizing $w(\mathfrak{R}) = |\mathfrak{R} \cap S|$

Following Lemma 7.1, consider all spines that intersect S only in reducing edges, and define $\mathcal{E}$ for each such spine to be as in the motivating example from Section 8: the collection of edges that intersect S. Let Σ be a spine for which $w(\Sigma) = w(\mathcal{E})$ is minimized and let $\mathfrak{R}(\Sigma)$ denote the corresponding collection of edge-reducing spheres for Σ. In other words, among all such spines and collections of edge-reducing spheres, choose that which minimizes the number $|\mathfrak{R} \cap S|$ of (circle) components of intersection.

Proposition 9.1 $\mathfrak{R}(\Sigma)$ *is disjoint from S.*

Note that for this proposition we don't care about how often the reducing edges of the spine Σ intersects S. We revert to the notation $\mathfrak{R}$ for $\mathfrak{R}(\Sigma)$.

Proof Suppose, contrary to the conclusion, $\mathfrak{R} \cap S \neq \varnothing$. Among the components of $\mathfrak{R} \cap S$, pick c to be one that is innermost in S. Let $E \subset S$ be the disk that c bounds in S and let $M_{\mathfrak{R}}$ be the component of $M - \mathfrak{R}$ in which E lies. Let $R_0 \in \mathfrak{R}$ be the edge-reducing sphere on which c lies, $e_0 \subset \Sigma$ the corresponding edge, p be the base $e_0 \cap R_0$ of R_0, and $D \subset R_0$ be the disk c bounds in $R_0 - p$. Finally, as in Proposition 8.6 let $\mathfrak{R}_- = \mathfrak{R} - R_0$ and $M_{\mathfrak{R}-} \supset M_{\mathfrak{R}}$ be the component of $M - \mathfrak{R}_-$ that contains R_0.

Claim 1 *After local stem swaps as in Proposition 5.7 we can take e_0 to be disjoint from E.*

Let $v_\pm$ be the vertices at the ends of e_0, with $e_\pm$ the incident components of $e_0 - p$. In a bicollar neighborhood of R_0, denote the side of R_0 incident to $e_\pm$ by, respectively, $R_\pm$, with the convention that a neighborhood of ∂E is incident to R_+. It is straightforward to find a point $p' \in R_0$ and arcs $e'_\pm$ in $M_{\mathfrak{R}} - E$, each with one end at the respective vertex $v_\pm$ and other end incident to p' via the respective side $R_\pm$.

It is not quite correct that replacing each of $e_\pm$ with $e'_\pm$ is a local stem swap, since the arcs are incident to R_0 at different points. But this can be easily fixed: Let γ be an arc from p' to p in R_0 and $\gamma_\pm$ be slight push-offs into $R_\pm$. Then replacing each $e_\pm$ with, respectively, $e'_\pm \cup \gamma_\pm$ is a local stem swap. Attach the two arcs at $p \in R_0$ to get a new reducing edge e'_0 for R_0, and then use the arc γ to isotope e'_0 back to the reducing edge $e'_+ \cup e'_-$, which is disjoint from E, as required. See Figure 13. Revert to e_0, p, etc as notation for $e'_+ \cup e'_-$, now disjoint from E.

Claim 2 *After local stem swaps we can assume that each stem that intersects E, intersects it always with the same orientation.*

Figure 14 shows how to use a local stem swap to cancel adjacent intersections with opposite orientations, proving the claim.

Notice that if E is nonseparating in $M_{\mathfrak{R}}$ we could do a local stem swap so that each stem intersects E algebraically zero times. Following Claim 2, this implies that we could make all stems disjoint from E. Once E intersects no stems, replace the subdisk D of R_0 that does not contain p with a copy of E. The result R'_0 is still an edge-reducing sphere for e_0, but the circle c (and perhaps more circles) of intersection with S has been removed. That is,

$$w(R'_0) = |R'_0 \cap S| \leq |R_0 \cap S| - 1 = w(R_0) - 1.$$

Hence, $w(\Sigma') < w(\mathfrak{R}) = w(\Sigma)$, contradicting our hypothesis that $w(\Sigma)$ is minimal.

So we henceforth proceed under the assumption that E is separating, but hoping for the same conclusion: that we can arrange for all stems to be disjoint from E, so that R'_0 as defined above leads to the same contradiction. Since E is separating, a stem that always passes through E with the same orientation can pass through at most once. So we henceforth assume that each stem that intersects E intersects it exactly once.
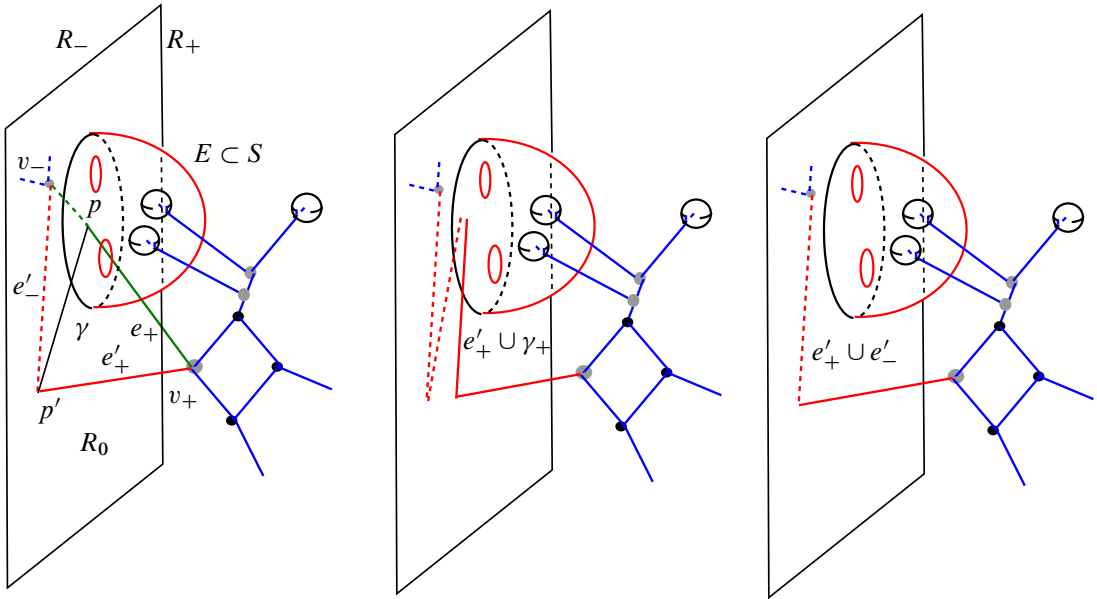
Figure 13: Making e_0 disjoint from E by local stem swaps.

In a bicollar neighborhood of the disk E, let E_+ be the side of E on which v_+ lies, and E_- be the other side of E. Consider a stem σ of a boundary sphere $\mathfrak{a}$ of $M_{\mathfrak{R}-}$. If σ intersects E, the subsegment of $\sigma - E$ that is incident to the blossom $\mathfrak{a}$ passes through one of $E_\pm$. Let $\hat{\sigma}_\pm$ be the collection of those stems intersecting E for which this subsegment passes through, respectively, $E_\pm$. If $\sigma \in \hat{\sigma}_+$, it is straightforward to find an alternative stem σ' from $\mathfrak{a}$ to a point very near v_+ so that σ' misses E. A stem swap to σ' is as in Proposition 5.2, and so by Corollary 8.5 does not increase weight. Hence, we have proven:

Claim 3 *After stem swaps, we may assume that each stem that intersects E is in $\hat{\sigma}_-$.*

Following Claim 3, we move to swap those stems in $\hat{\sigma}_-$ for ones that are disjoint from E. Let σ be the stem of a boundary sphere $\mathfrak{a}$ of $M_{\mathfrak{R}-}$, and assume that $\sigma \in \hat{\sigma}_-$. Then it is straightforward to find an alternative stem σ' for $\mathfrak{a}$ that is disjoint from E and ends in a point very near v_-, for example by concatenating an arc in E_- with an arc in R_- and an arc parallel to e_-. See Figure 15. A problem is, that such an arc intersects the disk $D \subset R_0$, so, after such a swap, R_0 is no longer an edge-reducing sphere
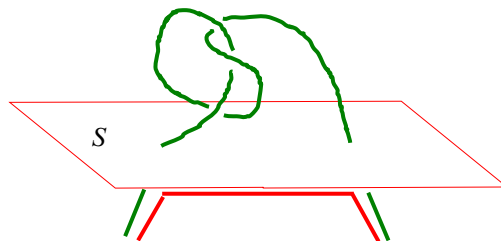
Figure 14: A local stem swap.
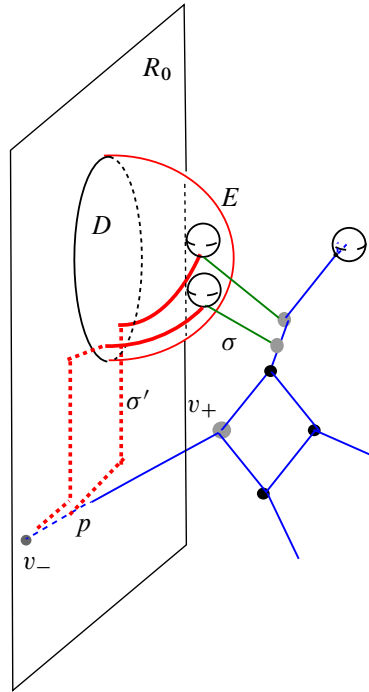
Figure 15

for the new spine. However, if such swaps are performed simultaneously on all stems in $\hat{\sigma}_-$, we have seen that the swap-mate of R_0 is an edge-reducing sphere for the new spine Σ', as required. But observe in Figure 15 that the swap-mate is exactly R_0'! So we can now appeal to Proposition 8.6:

$$w(\Sigma') \leq w(\Sigma) - |R_0 \cap S| + |R_0' \cap S| \leq w(\Sigma) - 1.$$

The contradiction proves Proposition 9.1. □

10 Conclusion

Proposition 10.1 *Suppose Σ intersects S only in reducing edges, and the associated set $\mathfrak{R}$ of edge-reducing spheres is disjoint from S. Then T can be isotoped (via edge slides of Σ) so that S is aligned with T.*

Proof We will proceed by stem swaps, chosen so that they do not affect the hypothesis that $\mathfrak{R} \cap S = \varnothing$. Let $M_{\mathfrak{R}}$ be the component of $M - \mathfrak{R}$ that contains S, and $\mathfrak{R}_0 \subset \partial M_{\mathfrak{R}}$ the collection of sphere components that come from $\mathfrak{R}$. In $M_{\mathfrak{R}}$ each $\mathfrak{a} \in \mathfrak{R}_0$ is the blossom of a flower whose stem typically intersects S. (A nonseparating sphere in $\mathfrak{R}$ may appear twice in $\mathfrak{R}_0$, with one or both stems intersecting S.) Denote by $\hat{\sigma}$ the collection of all stems of $\mathfrak{R}_0$ that intersect S. The proof will be by induction on $|\hat{\sigma} \cap S|$. If $|\hat{\sigma} \cap S| = 0$ then either S is a sphere disjoint from Σ and therefore aligned, or S is a disk. In the latter case our convention of which compression body to call B has $\partial S \subset \partial_- B \subset \Sigma$, so $T \cap S$ is a single circle
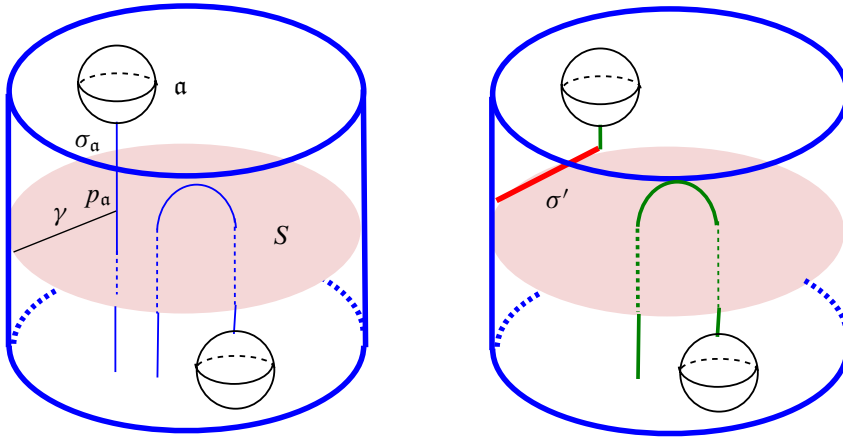
Figure 16: Swap lowering $|\hat{\sigma} \cap S|$, for S a disk.

parallel to ∂S in S. Again this means that S is aligned. Suppose then that $|\hat{\sigma} \cap S| > 0$ and inductively assume that the proposition is known to be true for lower values of $|\hat{\sigma} \cap S|$. Consider the possibilities:

Case 1 S is a disk.

Since $|\hat{\sigma} \cap S| > 0$ there is a blossom $\mathfrak{a} \in \mathfrak{R}_0$ with stem $\sigma \in \hat{\sigma}$. Let $\sigma_\mathfrak{a} \subset \sigma$ be the segment of $\sigma - S$ whose interior is disjoint from S and whose endpoints are the blossom $\mathfrak{a}$ and a point $p_\mathfrak{a}$ in S. Let γ be an arc in S that runs from $p_\mathfrak{a}$ to ∂S that avoids all other points of $\hat{\sigma} \cap S$. Push the arc $\gamma \cup \sigma_\mathfrak{a}$ off of S in the direction of $\sigma_\mathfrak{a}$ so that it becomes a stem σ' for $\mathfrak{a}$. Do a stem swap from σ to σ', and let Σ' be the result. See Figure 16. Since σ' is disjoint from S, σ is thereby removed from $\hat{\sigma}$, lowering $|\hat{\sigma} \cap S|$ by at least one. The stem swap does not affect other reducing edges or their edge-reducing spheres, so the latter remain disjoint from S. By Proposition 5.2 Σ' is still a spine of B, so T is isotopic in M to a regular neighborhood of Σ'. The inductive hypothesis implies that then T can be isotoped so that S is aligned with T, as required.

Case 2 S is a sphere.

Although S could be nonseparating in M, it cannot be nonseparating in $M_{\mathfrak{R}}$. Here is the argument: Suppose $S \subset M_{\mathfrak{R}}$ is nonseparating. If $\hat{\sigma}$ were disjoint from S then S would have no intersections with the Heegaard surface T at all and so $S \subset A$. But in a compression body such as A, all spheres separate, a contradiction. We will inductively reach the same contradiction by showing that if $\hat{\sigma}$ does intersect S there is a local stem swap that lowers $|\hat{\sigma} \cap S|$: Since S is nonseparating there is a circle c in $M_{\mathfrak{R}} - \Sigma$ that intersects S in a single point p. Let γ be a path in S from p to a point in $\sigma \cap S$, where $\sigma \in \hat{\sigma}$ and γ is chosen so that its interior is disjoint from $\hat{\sigma}$. Band sum σ to γ along a band perpendicular to S, with γ as its core. The result is an edge σ' that is obtained from σ by a local stem swap and intersects S in one fewer point than σ does, as required. See Figure 17.

So S is separating in $M_{\mathfrak{R}}$. This implies that no stem can intersect S more than once algebraically and so, following local stem swaps as in Claim 2 of Proposition 9.1 (see Figure 14), no more than once geometrically. If no stem intersects S at all, then $S \subset A$ and so S is aligned, finishing the proof.
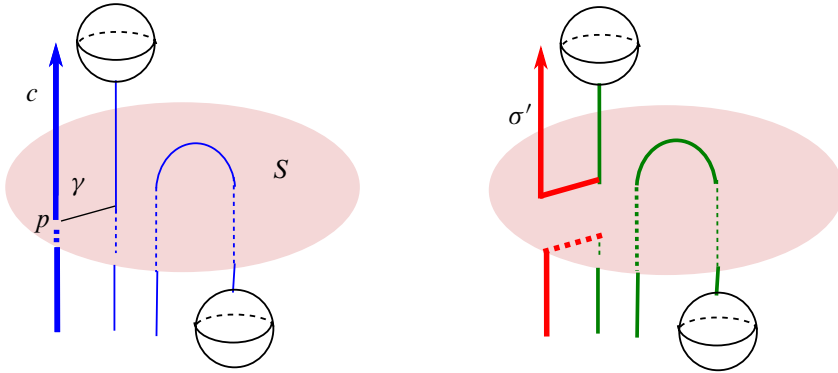
Figure 17: Swap lowering $|\hat{\sigma} \cap S|$, for S a nonseparating sphere.

Suppose, on the other hand, there is at least one stem $\sigma_!$ that intersects S exactly once. Repeat the argument of Case 1 for all stems other than $\sigma_!$, using the point $p_! = \sigma_i \cap S$ in place of ∂S in the argument. The result is that, after a sequence of stem swaps, all stems other than $\sigma_!$ are disjoint from S. This means that $S \cap \Sigma$ consists of the single point $p_!$. In other words, T intersects S in a single circle, and so S is aligned. □

The sequence of Proposition 4.2, Lemma 7.1, and Propositions 9.1 and 10.1 establishes Theorem 1.3. □

11 The Zupan example

Some time ago, Alex Zupan proposed a simple example for which the strong Haken theorem seemed unlikely (personal communication, 2019). The initial setting is of a Heegaard split 3–manifold $M = A \cup_T B$ that is the connected sum of compact manifolds M_1, M_2 and M_3, as shown in Figure 18. The blue indicates the spine Σ of B, say and, following our convention throughout the proof, B is to be thought of as a thin regular neighborhood of Σ. The spine is not shown inside of the punctured summands M_1 and M_2 because those parts are irrelevant to the argument; psychologically it's best to think of these as spherical boundary components of M lying in $\partial_- B$, so M_1 and M_2 are balls.
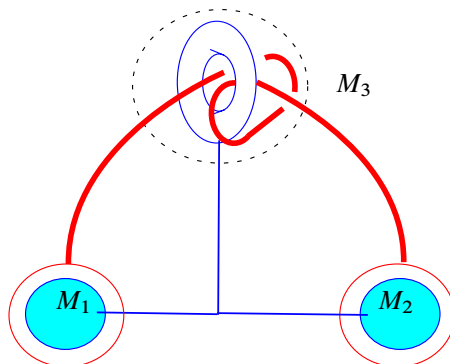
Figure 18: The initial setting.
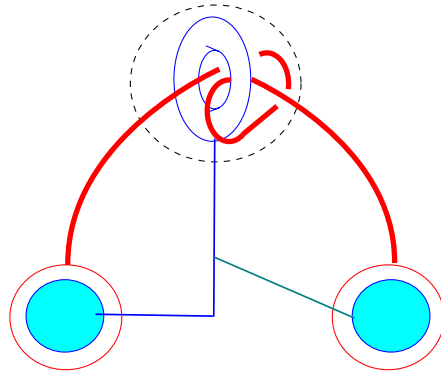
Figure 19: One blue edge now teal.

In the figure, M_3 is a solid torus and what we see is the punctured M_3, lying in M as a summand. We will continue the argument for this special case, in which M_3 is a solid torus and M_1 and M_2 are balls, but the argument works in general. An important role is played by the complement A of Σ outside M_1 and M_2. This is a solid torus: indeed, the region in the figure between the torus and the cyan balls is a twice punctured solid torus; A is obtained by removing both a collar of the torus boundary component and the blue arcs, all part of Σ. Removing the collar does not change the topology, but removing the blue arcs changes the twice-punctured solid torus into an unpunctured solid torus A.

Zupan proposed the following sort of reducing sphere S for M: the tube sum of the reducing spheres for M_1 and M_2 along a tube in M_3 which can be arbitrarily complicated. The outside of the tube is shown in red in Figure 18. The reducing sphere S is not aligned with T because it intersects Σ in two points, one near each of M_1 and M_2. The goal is then to isotope T through M so that it will be aligned with S. This is done by modifying Σ by what is ultimately a stem swap, and we will describe how the stem swap is obtained by an edge-slide of Σ. The edge-slide induces an isotopy of T in M because T is the boundary of a regular neighborhood of Σ. Note that in such an edge slide, passing one of the blue arcs through the red tube is perfectly legitimate.
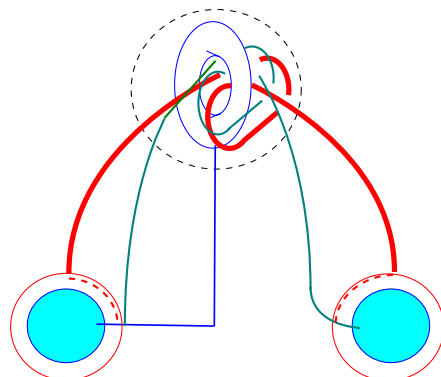
Figure 20: Teal edge now homotopic to red tube.
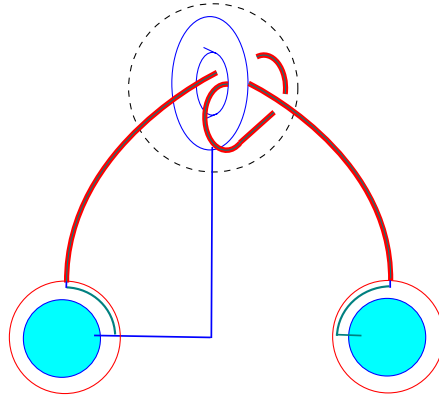
Figure 21: Teal edge isotoped into red tube.

Figure 19 is the same, but we have distinguished part of Σ (the rightmost edge) by turning it teal and beginning to slide it on the rest of the spine.

Now we invoke the viewpoint and notation of Proposition 5.2: There is a related Heegaard splitting of M available to us, in which the sphere boundary component at M_2 is not viewed as part of $\partial_- B$ but as part of $\partial_- A$, and the teal arc is also added to A. This changes A into a punctured solid torus A_+ and the spine of its complement into Σ_-, obtained by deleting from Σ both the teal edge and the sphere boundary component at M_2.

And so we apply Lemma 3.3, with A_+ playing the role of compression-body C; the boundary sphere at M_2 playing the role of the point r; the other end of the teal arc playing the role of q; the teal arc playing the role of β; and the union of the core of the red tube and the two dotted arcs in Figure 20 playing the role of α. Specifically, as the proof of Lemma 3.3 describes, because $\pi_1(\partial A_+) \to \pi_1(A_+)$ is surjective, and the slides take place in ∂A_+, one can slide the end of the teal arc around on the rest of Σ_- (technically on the boundary of a thin regular neighborhood of Σ_-) until it is *homotopic* rel endpoints to the path that is the union of the core of the tube of S and the two dotted red arcs shown in Figure 20. Hass and Thompson [5, Proposition 4] then shows that α and β are isotopic rel endpoints.

The result of the isotopy is shown in Figure 21; the teal edge now goes right through the tube, never intersecting S. Thus S now intersects Σ in only a single point, near the boundary sphere at M_1. In other words, S is aligned with T.

References

[1] **A J Casson**, **C M Gordon**, *Reducing Heegaard splittings*, Topology Appl. 27 (1987) 275–283 MR Zbl

[2] **M Freedman**, **M Scharlemann**, *Dehn's lemma for immersed loops*, Math. Res. Lett. 25 (2018) 1827–1836 MR Zbl

[3] **M Freedman**, **M Scharlemann**, *Uniqueness in Haken's theorem*, Michigan Math. J. 74 (2024) 119–142 MR

[4]   **W Haken**, *Some results on surfaces in* 3*–manifolds*, from "Studies in modern topology" (P J Hilton, editor), Stud. Math. 5, Math. Assoc. Amer., Englewood Cliffs, NJ (1968) 39–98 MR Zbl

[5]   **J Hass**, **A Thompson**, *Neon bulbs and the unknotting of arcs in manifolds*, J. Knot Theory Ramifications 6 (1997) 235–242 MR Zbl

[6]   **S Hensel**, **J Schultens**, *Strong Haken via sphere complexes*, preprint (2021) arXiv 2102.09831

[7]   **M Scharlemann**, *Heegaard splittings of compact* 3*–manifolds*, from "Handbook of geometric topology" (R J Daverman, R B Sher, editors), North-Holland, Amsterdam (2002) 921–953 MR Zbl

[8]   **M Scharlemann**, *Generating the Goeritz group of* S^3, preprint (2020) arXiv 2011.10613

[9]   **M Scharlemann**, **A Thompson**, *Thin position and Heegaard splittings of the* 3*–sphere*, J. Differential Geom. 39 (1994) 343–357 MR Zbl

[10] **A Schoenflies**, *Beiträge zur Theorie der Punktmengen, III*, Math. Ann. 62 (1906) 286–328 MR Zbl

Department of Mathematics, University of California, Santa Barbara
Santa Barbara, CA, United States

mgscharl@math.ucsb.edu

Right-angled Artin subgroups of right-angled Coxeter and Artin groups

PALLAVI DANI

IVAN LEVCOVITZ

We determine when certain natural classes of subgroups of right-angled Coxeter groups (RACGs) and right-angled Artin groups (RAAGs) are themselves RAAGs. We characterize finite-index *visual RAAG* subgroups of 2–dimensional RACGs. As an application, we show that any 2–dimensional, one-ended RACG with planar defining graph is quasi-isometric to a RAAG if and only if it is commensurable to a RAAG. Additionally, we give new examples of RACGs with nonplanar defining graphs which are commensurable to RAAGs.

Finally, we give a new proof of a result of Dyer: every subgroup generated by conjugates of RAAG generators is itself a RAAG.

1 Introduction

Let Γ be a finite simplicial graph with vertex set $V(\Gamma)$ and edge set $E(\Gamma)$. The *right-angled Artin group* (RAAG for short) associated to Γ is the group A_Γ given by the presentation

$$A_\Gamma = \langle V(\Gamma) \mid st = ts \text{ for all } (s,t) \in E(\Gamma)\rangle.$$

This article is concerned with the following question. Given a finite set S of elements in a group, when is the group generated by S isomorphic to a RAAG in the "obvious" way (ie with S as the "standard" RAAG generating set)? To make this precise, we define the notion of *RAAG system*.

Definition 1.1 (RAAG system) Let G be any group with generating set S. Let Δ be the graph whose vertex set is in bijection with S and which has an edge between distinct $s,t \in S \equiv V(\Delta)$ if and only if s and t commute. We call Δ the *commuting graph* associated to S. There is a canonical homomorphism $\phi\colon A_\Delta \to G$ extending the bijection $V(\Delta) \to S$. We say that (G,S) is a *RAAG system* if ϕ is an isomorphism. In particular, $(A_\Gamma, V(\Gamma))$ is a RAAG system for any RAAG A_Γ.

The *right-angled Coxeter group* (RACG for short) associated to the finite simplicial graph Γ is the group W_Γ given by the presentation

$$W_\Gamma = \langle V(\Gamma) \mid s^2 = 1 \text{ for all } s \in V(\Gamma), st = ts \text{ for all } (s,t) \in E(\Gamma)\rangle.$$

In this article we study subgroups G generated by particular natural subsets S of right-angled Coxeter and Artin groups, and we give characterizations for when (G, S) is a RAAG system or a finite-index RAAG system.

A theorem of Davis and Januszkiewicz [2000] states that every RAAG is commensurable to some RACG. This leads to the following question addressing the converse:

Question 1.2 Which RACGs are commensurable to RAAGs?

A RACG that is commensurable to a RAAG is, in particular, quasi-isometric to a RAAG. By considering different quasi-isometry invariants, one sees that the converse to the Davis–Januszkiewicz theorem above is far from being true. For instance, there are many RACGs that are one-ended hyperbolic (such as virtual hyperbolic surface groups), while no RAAG is both one-ended and hyperbolic. Furthermore, RAAGs have linear, quadratic or infinite divergence [Behrstock and Charney 2012], whereas the divergence of a RACG can be a polynomial of any degree [Dani and Thomas 2015]. Restricting to RACGs of at most quadratic divergence is still not enough to guarantee they are quasi-isometric to RAAGs. For instance, the Morse boundary of a RAAG with quadratic divergence is always totally disconnected [Charney and Sultan 2015; Cordes and Hume 2017], while the Morse boundary of a RACG of quadratic divergence can have nontrivial connected components [Behrstock 2019]. The above examples show that there are numerous families of RACGS which are not quasi-isometric and, hence, not commensurable to any RAAG. Within the subclass of one-ended RACGs with planar, triangle-free defining graphs, Nguyen and Tran [2019] characterize those quasi-isometric to RAAGs. Theorem B below answers Question 1.2 in this setting.

We note that every RACG (indeed, every Coxeter group) is virtually special, and therefore has a finite-index subgroup which is a subgroup of a RAAG [Haglund and Wise 2010]. However, this subgroup is not of finite index in the RAAG, which would be required for establishing commensurability.

One approach to proving that a RACG is commensurable to a RAAG is to look for finite-index subgroups that are isomorphic to RAAGs. We focus on a class of subgroups of RACGs, introduced by LaForge [2017] in his PhD thesis, that are logical candidates for being RAAGs. Given a RACG defined by a graph Γ and two nonadjacent vertices $s, t \in V(\Gamma)$, it follows that st is an infinite-order element of W_Γ. There is then a correspondence between edges of the complement graph Γ^c with such infinite-order elements of Γ. Given a subgraph Λ of Γ^c, let G be the subgroup generated by $E(\Lambda)$ (thought of as infinite-order elements of W_Γ). As G is generated by the edges of Λ, we may as well assume that Λ has no isolated vertices. A natural question is:

Question 1.3 When is $(G, E(\Lambda))$ a finite-index RAAG system?

If $(G, E(\Lambda))$ is indeed a RAAG system, then G is called a *visual RAAG subgroup* of W_Γ. LaForge obtained some necessary conditions for such subgroups to be visual RAAGs.

We say that W_Γ is 2–*dimensional* if Γ is triangle-free. Our first main theorem gives an exact characterization of the finite-index visual RAAG subgroups of 2–dimensional RACGs in terms of graph-theoretic conditions:

Theorem A *Let W_Γ be a 2–dimensional RACG. Let Λ be a subgraph of Γ^c with no isolated vertices, and let G be the subgroup generated by $E(\Lambda)$. Then the following are equivalent.*

(1) *$(G, E(\Lambda))$ is a RAAG system and G is finite index in W_Γ.*

(2) *$(G, E(\Lambda))$ is a RAAG system and G has index either two or four in W_Γ (and exactly four if W_Γ is not virtually free).*

(3) *Λ has at most two components and satisfies conditions $\mathscr{R}_1$–$\mathscr{R}_4$, $\mathscr{F}_1$ and $\mathscr{F}_2$.*

The conditions $\mathscr{R}_1$–$\mathscr{R}_4$, $\mathscr{F}_1$ and $\mathscr{F}_2$ in the above theorem are algorithmically checkable graph-theoretic conditions on Γ and Λ. See Section 3 for precise definitions of these conditions.

In Section 5 we provide several applications to concrete families of RACGs. In particular we prove:

Theorem B *Let W_Γ be a 2–dimensional, one-ended RACG with planar defining graph. Then W_Γ is quasi-isometric to a RAAG if and only if it contains an index 4 subgroup isomorphic to a RAAG.*

A complete description of which RACGs considered in Theorem B are quasi-isometric to RAAGs is given by Nguyen and Tran [2019, Theorem 1.2]. Theorem B shows these are actually commensurable to RAAGs.

We also give two families of RACGs defined by nonplanar graphs which contain finite-index RAAG subgroups (see Corollaries 5.1 and 5.2). These cannot be obtained by applying the Davis–Januszkiewicz constructions to the defining graphs of the RAAGs they are commensurable to. For the family in Corollary 5.1, we use work of Bestvina, Kleiner and Sageev on RAAGs [Bestvina et al. 2008], to conclude the RACGs are quasi-isometrically distinct. We believe that the methods from this article may be used to further study commensurability of RACGs.

The proof of Theorem A consists of two main parts. One part involves obtaining an understanding of when G is of finite index, leading to conditions $\mathscr{F}_1$ and $\mathscr{F}_2$. To obtain these, we use *completions of subgroups*, introduced in [Dani and Levcovitz 2021]. The other aspect consists of obtaining criteria to recognize when $(G, E(\Lambda))$ is a RAAG system. To do so, we prove the following theorem by careful analysis of disk diagrams:

Theorem C *Let W_Γ be a RACG. Let Λ be a subgraph of Γ^c with no isolated vertices and at most two components. Then the subgroup $(G, E(\Lambda)) < W_\Gamma$ is a RAAG system if and only if $\mathscr{R}_1$–$\mathscr{R}_4$ are satisfied.*

Conditions $\mathscr{R}_1$, $\mathscr{R}_2$, and a condition more or less equivalent to $\mathscr{R}_3$ were known to be necessary for $(G, E(\Lambda))$ to be a RAAG system by work of LaForge [2017]. We show in Example 3.13 that they are not sufficient. We introduce a fourth graph-theoretic condition $\mathscr{R}_4$ to obtain a complete characterization of all visual RAAG subgroups defined by subgraphs of Γ^c with at most two components. The bulk of the proof of Theorem C consists of showing that the conditions $\mathscr{R}_1$–$\mathscr{R}_4$ are sufficient.

Note that, unlike in Theorem A, there is no assumption on the dimension of the RACGs in Theorem C. On the other hand, there is an additional assumption in Theorem C, namely that the subgraph Λ of Γ can have at most two components.

When Λ contains more than two components, the situation becomes much more complex. We show that additional graph-theoretic conditions are necessary to generalize the Theorem C to this setting (see Lemmas 3.32 and 3.34). Remarkably, a consequence of these conditions is that if Γ is triangle-free and $(G, E(\Lambda))$ is a finite-index RAAG system, then Λ can have at most two components. This fact is crucial to the proof of Theorem A, which does not have any assumption on the number of components of Λ. Additionally, we are aware that even more conditions are necessary than those in this article, but we do not have a complete conjectural list of conditions that would be sufficient to characterize visual RAAGs.

We next turn our attention to RAAG subgroups of RAAGs. A classical theorem on Coxeter groups, proven independently by Deodhar [1989] and Dyer [1990], states that reflection subgroups of Coxeter groups (ie those generated by conjugates of generators) are themselves Coxeter groups. In fact, Dyer proves an analogous result for the class of groups defined by *reflection systems* (see [Dyer 1990] for the definition), which includes Coxeter groups as well as RAAGs. Specifically, he shows that subgroups generated by conjugates of standard generators are themselves in this class. As RAAGs are the only torsion-free groups in this class, one obtains the following result. Here, we define a *generalized RAAG reflection* to be an element of a RAAG A_Δ that is conjugate to a generator in $V(\Delta)$.

Theorem D [Dyer 1990] *Let $\mathcal{T}$ be a finite set of generalized RAAG reflections in the RAAG A_Γ. Then the subgroup $G < A_\Gamma$ generated by $\mathcal{T}$ is a RAAG.*

We thank Luis Paris for informing us that this result is contained in [Dyer 1990], and the explanation in the preceding paragraph. We include our proof of Theorem D, as our geometric approach is very different from that of Dyer, which is algebraic and uses cocycles. Our proof uses a characterization of RAAG systems in terms of the deletion condition, given by Basarab [2002]. We use disk diagrams to show that subgroups generated by generalized RAAG reflections satisfy the criteria in Basarab's characterization.

We note that, although G (from Theorem D) is a RAAG, $(G, \mathcal{T})$ is not necessarily a RAAG *system* and in general G is not isomorphic to the RAAG A_Δ where Δ is the commuting graph corresponding to $\mathcal{T}$. Kim and Koberda [2013] show that there exists a subgroup of G (generated by sufficiently high powers of the elements of $\mathcal{T}$) which is isomorphic to A_Δ.

Genevois, as well as an anonymous referee, pointed out to us that a proof of Theorem D may be possible using [Genevois 2017, Theorem 10.54] (see also [Genevois 2019, Theorem 3.24]).

Acknowledgements

that led to Corollary 4.9, and Hung Tran for encouraging us to look at the examples considered in Theorem B. Finally, we would like to thank Jason Behrstock, Anthony Genevois, Garret LaForge, Kim Ruane and the referees for helpful comments and conversations.

2 Background

2.1 Basic terminology and notation

Let G be a group with generating set S. We say that $w = s_1 \cdots s_n$, with $s_i \in (S \cup S^{-1})$ for $1 \leq i \leq n$, is a *word over S* or a *word in G*. If the words w and w' represent the same element of G, then we say that w' is an *expression* for w and write $w' \simeq w$. We say the word $w = s_1 \cdots s_n$ is *reduced* (or *reduced over S* for emphasis) if given $w' = t_1 \cdots t_m \simeq w$, it follows that $n \leq m$.

2.2 Right-angled Coxeter and Artin groups

Coxeter groups can be characterized as those groups which are generated by involutions and which satisfy the deletion condition; see Definition 2.1 below (for a proof of this fact, see [Davis 2015, Theorem 3.3.4]). By work of Basarab [2002], RAAGs can be characterized in a similar manner (see Theorem 2.2 below). This characterization will be utilized in Section 6.

Definition 2.1 (deletion condition) Let G be a group generated by S. We say that (G, S) satisfies the *deletion condition* if, given any word w over S, either w is reduced or $w = s_1 \cdots s_k$ and there exist $1 \leq i < j \leq k$ such that $s_1 \cdots \hat{s}_i \cdots \hat{s}_j \cdots s_k$ is an expression for w.

The result below directly follows from a result of Basarab.

Theorem 2.2 [Basarab 2002] *Let G be a group generated by S such that $S \cap S^{-1} = \varnothing$ and $1 \notin S$. Then (G, S) is a RAAG system if and only if*

(1) *every s in S has infinite order, and*

(2) *(G, S) satisfies the deletion condition.*

Proof If (G, S) is a RAAG system, then G is torsion-free [Charney 2007], so (1) holds. Furthermore, (G, S) satisfies (2) by [Basarab 2002, Corollary 1.4.2] (see also [Bahls 2005, page 31, Exercise 17] for a simpler proof in this setting). The converse also follows from a direct application of [Basarab 2002, Corollary 1.4.2]. □

We now define certain moves which can be performed on a word that produce another expression for it. These moves provide a solution to the word problem for RAAGs and RACGs (see Theorem 2.4 below).

Definition 2.3 (Tits moves) Let G be a group generated by S. Let $w = s_1 \cdots s_n$ be a word over S. If s_i and s_{i+1} commute for some $1 \leq i < n$, then the word $s_1 \cdots s_{i-1} s_{i+1} s_i s_{i+2} \cdots s_n$ is an expression for w obtained by a *swap operation* performed on w, which *swaps* s_i and s_{i+1}. If $s_i = s_{i+1}^{-1}$ for some $1 \leq i < n$, then $s_1 \cdots s_{i-1} s_{i+2} \cdots s_n$ is an expression for w is obtained by a *deletion operation* performed on w. A *Tits move* is either a swap operation or a deletion operation. We say a word is *Tits reduced* if no sequence of Tits moves can be performed on the word to obtain an expression with fewer generators.

Theorem 2.4 below shows that RAAGs and RACGs admit a nice solution to the word problem. This solution to the word problem for RACGs is a well-known result of Tits [1969], a version of which holds more generally for all Coxeter groups. The result below in the setting of RAAGs follows from a theorem of Basarab [2002, Theorem 1.4.1] which generalizes Tits' result (see also [Green 1990, Theorem 3.9]).

Theorem 2.4 [Tits 1969; Basarab 2002] *Let A_Γ be either a RAAG or a RACG. Then*:

(1) *If w_1 and w_2 are reduced words over $V(\Gamma)$ representing the same element of G, then w_2 can be obtained from w_1 by Tits swap moves.*

(2) *Given any word w over $V(\Gamma)$, a reduced expression for w can be obtained by applying Tits moves to w.*

We will often not refer directly to the above theorem, and we will instead simply say that a given RAAG or RACG *admits a Tits solution to the word problem*.

The next two lemmas are well known and will often be implicitly assumed.

Lemma 2.5 *Let A_Γ either be a RAAG or RACG. Then $s, t \in V(\Gamma)$ commute as elements of A_Γ if and only if (s, t) is an edge of Γ.*

Proof One direction of the claim follows from the definitions of a RAAG and a RACG. If A_Γ is a RACG, then the other direction follows from [Björner and Brenti 2005, Proposition 4.1.2].

Now suppose that A_Γ is a RAAG, and let $s, t \in V(\Gamma)$ be nonadjacent vertices. Suppose, for a contradiction, that $w = sts^{-1}t^{-1} \simeq 1$. Let D be a disk diagram with boundary w (see Section 2.3 for a reference for disk diagrams). This disk diagram contains exactly two intersecting hyperplanes: one labeled by s and one labeled by t. However, this is a contradiction as a pair of hyperplanes whose labels are nonadjacent vertices of Γ cannot intersect. □

Lemma 2.6 *Let W_Γ be a RACG, and let $s, t, q, r \in V(\Gamma)$ be such that s and t do not commute, and r and q do not commute. Then $(st)(qr) \simeq (qr)(st)$ if and only if*

(1) *there is a square in Γ formed by s, q, t, and r;*

(2) *$t = q$ and $s = r$; or*

(3) *$t = r$ and $s = q$.*

Proof Clearly each of (1), (2) and (3) implies that $(st)(qr) \simeq (qr)(st)$.

To prove the converse, suppose that $(st)(qr) \simeq (qr)(st)$. Suppose first that $t = q$, and consequently $stqr \simeq sr$. As s and t do not commute and q and r do not commute, this is only possible if $r = t$. Thus, (2) holds.

If $s = q$, as $qrts \simeq tsqr$, we apply the same argument to conclude that $t = r$, showing (3) holds. By similar arguments, if $s = r$ then $t = q$, and if $t = r$ then $s = q$. Thus, we may assume that s, t, q and r are all distinct vertices of Γ. In this case we again conclude by Tits' solution to the word problem, that if $stqr \simeq qrst$ then s, q, t and r form a square in Γ. □

2.3 Disk diagrams

We give a brief background on disk diagrams as they are used in our setting, and we refer the reader to [Sageev 1995; Wise 2021] for the general theory of disk diagrams over cube complexes. We then give some preliminary lemmas that are needed in later sections.

Let A_Δ be a RAAG, and let $w = s_1 \cdots s_n$, with $s_i \in V(\Delta)$, be a word equal in A_Δ to the identity, ie $w \simeq 1$. There exists a Van Kampen diagram D with boundary label w, and we call this planar 2–complex a *disk diagram in A_Δ with boundary label w*. We now describe some additional properties of D in our setting. The edges of D are oriented and labeled by generators in $V(\Delta)$. A *path in D* is a path γ in the 1–skeleton of D, traversing edges $e_1, \ldots, e_m$, and the label of γ is the word $a_1 \cdots a_m$ where, for each $1 \le i \le m$, a_i is the label of e_i if e_i is traversed along its orientation, and a_i^{-1} is the label of e_i if e_i is traversed opposite to its orientation. Every cell in D is a square that has a boundary path with label $aba^{-1}b^{-1}$ for some commuting generators a and b in $V(\Delta) \cup V(\Delta)^{-1}$.

There is a base vertex $p \in \partial D$ and an orientation on D, such that the smallest closed path δ which traverses the boundary of D in the clockwise orientation starting at p and traversing every edge outside the interior of D has label w. We call δ the *boundary path* of D. Note that if D contains an edge e not contained in a square, then necessarily δ traverses e exactly twice.

If W_Γ is a RACG and w is a word over $V(\Gamma)$ equal in W_Γ to the identity, then we define a disk diagram D in W_Γ with boundary w similarly. However, as each generator in $V(\Gamma)$ is an involution, we do not need to orient the edges of D.

Let D be a disk diagram and $q = [0, 1] \times [0, 1]$ be a square in D. The subset $\{\frac{1}{2}\} \times [0, 1] \subset q$ (similarly, $[0, 1] \times \{\frac{1}{2}\} \subset q$) is a *midcube*. The midpoint of an edge in D is also defined to be a *midcube*. A *hyperplane* in D is a minimal nonempty collection H of midcubes in D with the property that given any midcube $m \in H$ and a midcube m' in D such that $m \cap m'$ is contained in an edge of D, it follows that $m' \in H$. We say that H is dual to an edge e if the midpoint of e is in H.

Since opposite edges in every square in D have the same label, it follows that every edge intersecting a fixed hyperplane H has the same label. We call this the *label of the hyperplane*. Since adjacent sides
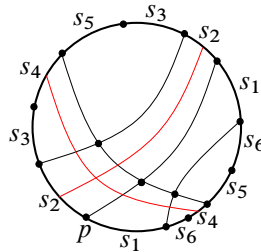
Figure 1: A disk diagram in a RACG with boundary the word $s_2s_3s_4s_5s_3s_2s_1s_6s_5s_4s_6s_1$ and base vertex p. Two hyperplanes are shown in red. As these hyperplanes intersect, it must be that s_4 commutes with s_2.

of a square have distinct labels which commute, it follows that no hyperplane self-intersects, and if two hyperplanes intersect, then their labels correspond to distinct, commuting generators. (See Figure 1 for an example of a disk diagram and some of its hyperplanes.)

Definition 2.7 (maps preserving boundary combinatorics) Let D and D' be disk diagrams, and let δ and δ' respectively be their boundary paths. Let $E = \{e_1, \ldots, e_m\}$ (resp. $E' = \{e'_1, \ldots, e'_n\}$) be the edges traversed by δ (resp. δ'). More precisely, e_i (resp. e'_i) is the ith edge traversed by δ (resp. δ') for each i. Observe that every hyperplane of D is dual to two edges $e_j, e_k \subset E$ for some $j \neq k$. (It could be that $e_j = e_k$, thought of as edges of D.) A similar statement holds for D'.

Let $F \subset E$ and $F' \subset E'$, and let $\psi: F \to F'$ be a bijection. We say that ψ *preserves boundary combinatorics* if for every pair of edges $e, f \in F$ which are dual to the same hyperplane of D, their images $\psi(e)$ and $\psi(f)$ are dual to the same hyperplane of D'.

Note that if Ψ preserves boundary combinatorics, then Ψ^{-1} does as well.

A pair of hyperplanes H and H' in a disk diagram D form a *bigon* if they intersect in at least two distinct points. The following lemma, first proven in [Sageev 1995, Theorem 4.3], guarantees that we can always choose a disk diagram without bigons. The boundary combinatorics statement below is guaranteed by the proof of this fact in [Wise 2021, Lemma 2.3, Corollary 2.4].

Lemma 2.8 [Sageev 1995; Wise 2021] *Given a disk diagram D with boundary label w, there exists a disk diagram D' also with boundary label w such that D' does not contain any bigons. Moreover, the natural bijection between the edges traversed by the boundary paths of D and D' induced by the label w preserves boundary combinatorics.*

Remark 2.9 In light of Lemma 2.8, for the rest of this paper we will always assume that any disk diagrams we consider do not have bigons.

Remark 2.10 Let α be a path with label $s_1 \cdots s_n$ in some disk diagram. The "edge of α with label s_i" is understood to be the ith edge α traverses (even though there may be several edges of α with the same label as this edge. A similar statement holds when we refer to subpaths of α.

Given a disk diagram with boundary label w, we will often want to produce a new disk diagram with boundary label w', where w' is obtained from w by a Tits move, and such that boundary combinatorics are preserved on appropriate subsets of the boundary paths. The following lemma exactly describes how we can perform these operations.

Lemma 2.11 *Let D be a disk diagram over the group W, where W is either a RACG or a RAAG. Suppose the boundary path of D traverses the edges $e_1, \ldots e_n$ and has label $w = s_1 \cdots s_n$.*

(1) *If s_r and s_{r+1} (taken modulo n) are distinct and commute for some $1 \le r \le n$, then there is a disk diagram D' whose boundary path traverses the edges $e'_1, \ldots, e'_n$ and has label $s_1 \cdots s_{i+1} s_i \cdots s_n$. Furthermore, the map ψ preserves boundary combinatorics, where ψ is defined by $\psi(e_r) = e'_{r+1}$, $\phi(e_{r+1}) = e'_r$, and $\psi(e_j) = e'_j$ for $j \ne r, r+1$.*

(2) *If $s_r = s_{r+1}^{-1}$ (taken modulo n) for some $1 \le r \le n$, then there is a disk diagram D' with boundary label $s_1 \cdots s_{r-1} s_{r+2} \cdots s_n$. Moreover, the natural map from edges traversed by the boundary path of D' to edges traversed by the boundary path of D preserves boundary combinatorics.*

(3) *Given any generator (or inverse of a generator) s and any r, with $1 \le r \le n$, it follows that there exists a disk diagram D' with boundary label $s_1 \cdots s_r (ss^{-1}) s_{r+1} \cdots s_n$. Moreover, the natural map from edges traversed by the boundary path of D to the edges traversed by the boundary path of D' preserves boundary combinatorics.*

Proof We first prove (1). Let q be a square whose edges are labeled consecutively by s_r, s_{r+1}, s_r^{-1} and s_{r+1}^{-1}. We form the disk diagram D' by identifying consecutive edges of q labeled by s_r and s_{r+1} to the edges of ∂D labeled by s_r and s_{r+1} (these edges must be distinct as $s_r \ne s_{r+1}$). The claim is readily checked.

We next prove (2). Let e and f be the edges of ∂D labeled respectively by s_r and s_{r+1}. Suppose first that e and f are distinct. In this case, form the disk diagram D' by identifying e and f, ie "fold" these edges together. On the other hand, if $e = f$, then as D has boundary label w, it must follow that e is a spur, ie an edge attached to D that is not contained in any square and which contains a vertex of valence 1. In this case we can remove the edge e from D to obtain D'. In either case, the claim is readily checked.

To show (3), form D' by inserting a spur edge with label s to the vertex traversed by the boundary path of D between s_r and s_{r+1}. □

3 Visual RAAG subgroups of right-angled Coxeter groups

In this and the next section we study visual RAAG subgroups of RACGs, as described in the introduction. We begin by describing some notation that will be used throughout these sections.

Let Γ be a graph, and let W_Γ be the corresponding RACG. Let Γ^c denote the complement of Γ, that is, the graph with the same vertex set as Γ, which has an edge between two (distinct) vertices if and only if

the corresponding vertices are not adjacent in Γ. Let Λ be a subgraph of Γ^c with no isolated vertices, ie one in which every vertex of Λ is contained in some edge.

We form a new graph $\Theta = \Theta(\Gamma, \Lambda)$ which we think of as a graph containing the edges of both Γ and Λ. More formally, $V(\Theta) = V(\Gamma)$ and $E(\Theta) = E(\Gamma) \cup E(\Lambda)$. Note that as $E(\Lambda) \subset \Gamma^c$, it follows that Θ is simplicial. We refer to edges of Θ that correspond to edges of Γ (resp. Λ) as Γ–edges (resp. Λ–edges).

A Λ–edge between vertices a and b corresponds to an inverse pair of infinite-order elements of W_Γ, namely ab and ba. By a slight abuse of terminology, we will use the term Λ–edge to refer to one of these elements and vice versa. We identify $E(\Lambda)$ with a subset of W_Γ by arbitrarily choosing one of the two infinite-order elements corresponding to each Λ–edge, and we define G^Θ to be the subgroup of W_Γ generated by $E(\Lambda)$. As we are dealing with subgroups generated by $E(\Lambda)$, there is no loss in generality in assuming that Λ has no isolated vertices. The goal of this section is to study when $(G^\Theta, E(\Lambda))$ is a RAAG system.

Let Δ be the commuting graph corresponding to $E(\Lambda)$ (as defined in the introduction), and let A_Δ be the corresponding RAAG. Recall that, by definition, $(G^\Theta, E(\Lambda))$ is a RAAG system if and only if the natural homomorphism $\phi \colon A_\Delta \to G^\Theta$ extending the bijection between $V(\Delta)$ and $E(\Lambda)$ is an isomorphism. As ϕ is always surjective, we would like to understand when ϕ is injective.

For the remainder of this section, we fix Γ, Λ, Θ, A_Δ, and ϕ as above. Furthermore, we will use the following terminology. The path γ in Θ *visiting vertices* $x_1, x_2, \ldots, x_n$ is defined to be the path which starts at x_1, passes through the remaining vertices in the order listed, and ends at x_n. We say that γ is simple if $x_i \neq x_j$ for $i \neq j$, and that γ is a loop if $x_1 = x_n$. Finally, γ is a cycle if it is a loop with $n \geq 3$, such that $x_i \neq x_j$ unless $\{i, j\} = \{1, n\}$. We call a path (resp. cycle) in Θ consisting only of Γ–edges a Γ–path (resp. Γ–cycle). We define Λ–paths and Λ–cycles similarly.

We begin by describing some graph-theoretic conditions on Θ which are consequences of either G^Θ being a RAAG or of $(G^\Theta, E(\Lambda))$ being a RAAG system.

Conditions $\mathcal{R}_1$ and $\mathcal{R}_2$, defined below, when combined, are equivalent to LaForge's star-cycle condition. LaForge [2017, Lemma 8.2.1] proves that $\mathcal{R}_1$ and $\mathcal{R}_2$ are necessary conditions for $(G^\Theta, E(\Lambda))$ to be a RAAG system. We include proofs here for completeness.

Definition 3.1 (condition $\mathcal{R}_1$) We say that Θ satisfies *condition $\mathcal{R}_1$* if it does not contain a Λ–cycle.

Lemma 3.2 [LaForge 2017] *If $(G^\Theta, E(\Lambda))$ is a RAAG system, then Θ satisfies $\mathcal{R}_1$.*

Proof Suppose Θ does not satisfy $\mathcal{R}_1$. Then it contains a Λ–cycle, say with vertices $a_1, \ldots, a_k$, where $k \geq 3$, such that for each $i \pmod{k}$, a_i is connected to a_{i+1} by a Λ–edge. Let g_i be the generator of A_Δ (or its inverse) corresponding to the (oriented) Λ–edge $a_i a_{i+1}$. As the a_i's are along a cycle, no Λ–edge is repeated, and we have that $g_i \neq g_j^{-1}$ for all $i \neq j$. This, together with the fact that RAAGs
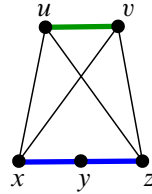
Figure 2

satisfy the deletion condition (see Theorem 2.2), implies that $g = g_1 g_2 \cdots g_k$ is a nontrivial element of A_Δ. Moreover, $\phi(g) = (a_1 a_2)(a_2 a_3) \cdots (a_k a_1) = 1$, so g is in the kernel of ϕ, and therefore ϕ is not injective. □

Definition 3.3 (condition $\mathscr{R}_2$) We say that Θ satisfies *condition $\mathscr{R}_2$* if each component of $\Lambda \subset \Theta$ (with the natural inclusion) is an induced subgraph of Θ.

Lemma 3.4 [LaForge 2017] *If G^Θ is a RAAG, then Θ satisfies $\mathscr{R}_2$.*

Proof Suppose Θ does not satisfy $\mathscr{R}_2$, and let u and v be a pair of vertices in a component of Λ, such that u and v are adjacent in Θ. It follows that u and v are connected by a Γ–edge, and therefore they commute. Since u and v are in the same component of Λ, there is a simple Λ–path from u to v whose vertices (in order) are $u = a_1, \ldots, a_k = v$. Note that $k \geq 3$, since Θ is a simplicial graph. For $1 \leq i \leq k - 1$, let g_i be the generator of A_Δ (or its inverse) corresponding to the Λ–edge $a_i a_{i+1}$, and let $g = g_1 g_2 \cdots g_{k-1}$. The element g is a nontrivial element of A_Δ, as RAAGs satisfy the deletion condition by Theorem 2.2.

We now have that $\phi(g)^2 = \big((a_1 a_2)(a_2 a_3) \cdots (a_{k-1} a_k)\big)^2 = (a_1 a_k)^2 = (uv)^2 = 1$, since u and v commute. This implies that G^Θ has torsion. Thus, G^Θ cannot be a RAAG as RAAGs are torsion-free [Charney 2007]. □

Our next condition, $\mathscr{R}_3$, is motivated by the following example.

Example 3.5 Let Θ be the graph in Figure 2, where the Γ edges are black and the Λ edges are colored. Since u and v each commute with x and z, the commutator $[uv, xz]$ represents the trivial element in W_Γ. Now observe that $[uv, xz] \simeq (uv)(xy)(yz)(vu)(zy)(yz)$, which is a product of Λ–edges, and therefore represents an element g of G^Θ. Now we can see that $(G^\Theta, E(\Lambda))$ is not a RAAG system: if it were, then it would be possible to show that g is trivial in G^Θ using only swap and deletion moves involving RAAG generators. However, since y does not commute with u and v, no such moves are possible (see Lemma 2.6). On the other hand, if there had been Γ edges, from y to both u and v, then there would be no contradiction.

A Λ–edge word similar to the one in the above example can be constructed whenever Γ has a square whose vertices alternate between two components of Λ. The example suggests that for such a Λ to define
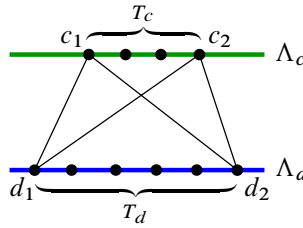
Figure 3: In the figure, the colored parts consist of Λ–edges, and the black parts consist of Γ–edges. The condition $\mathcal{R}_3$ says that if Θ contains a black square as shown, then every vertex of T_c is joined by a Γ–edge to every vertex of T_d.

a RAAG, the "intermediate" vertices in Λ between the endpoints of the square must all mutually commute. This is made precise in the definition of $\mathcal{R}_3$ (Definition 3.8) and Lemma 3.11 below. Before stating these, we introduce some terminology, which will be used throughout this section.

Definition 3.6 (2–component paths and cycles) We say the Γ–path γ in Θ is a 2–*component path* if γ visits vertices (in order) $c_1, d_1, c_2, d_2, \ldots, c_n, d_n$ for some $n \geq 1$ (where d_n could be omitted if $n > 1$) such that the c_i's all lie in a single component Λ_c of Λ, and the d_i's all lie in a single component $\Lambda_d \neq \Lambda_c$ of Λ. If it is important to emphasize the components visited by γ, we will call it a $\Lambda_c \Lambda_d$–path.

A 2–*component loop* is a 2–component path visiting $c_1, d_1, \ldots, c_n, d_n, c_{n+1}$ such that $c_1 = c_{n+1}$. A 2–*component cycle* is a 2–component loop which is a Γ–cycle. A 2–component cycle of length four will be called a 2–*component square*.

Definition 3.7 (Λ–convex hull) We define the Λ–*convex hull* of a set $X \subset V(\Theta)$ to be the convex hull of X in Λ.

Definition 3.8 (condition $\mathcal{R}_3$) We say that Θ satisfies *condition $\mathcal{R}_3$* if the following holds for every 2–component square in Θ. Consider a 2–component square in Θ visiting vertices c_1, d_1, c_2 and d_2, where $c_1, c_2 \in \Lambda_c$, $d_1, d_2 \in \Lambda_d$, and Λ_c and Λ_d are distinct components of Λ. Then the graph Γ contains the join of $V(T_c)$ and $V(T_d)$, where T_c and T_d are the Λ–convex hulls of $\{c_1, c_2\}$ and $\{d_1, d_2\}$ respectively. (See Figure 3.)

We will often need to utilize an expression for a word in W_Γ which is the product of Λ–edges. This construction is the content of the following definition.

Definition 3.9 (Λ–edge words) Suppose Θ satisfies condition $\mathcal{R}_1$, and let w be a word in W_Γ such that $w = (a_1 a_1')(a_2 a_2') \cdots (a_n a_n')$, where a_i and a_i' are in the same Λ–component of Θ for each $1 \leq i \leq n$. As Θ satisfies $\mathcal{R}_1$, there is a unique simple Λ–path from a_i to a_i'. Let $a_i = a_1^i, \ldots, a_{m_i}^i = a_i'$ be the vertices visited by this path. Form the word

$$w' = ((a_1^1 a_2^1)(a_2^1 a_3^1) \cdots (a_{(m_1-1)}^1 a_{m_1}^1)) \cdots ((a_1^n a_2^n)(a_2^n a_3^n) \cdots (a_{(m_n-1)}^n a_{m_n}^n)).$$

We call w' the Λ–*edge word* associated to w. Note that w' is well-defined, as long as Θ satisfies $\mathcal{R}_1$. In particular if $(G^\Theta, E(\Lambda))$ is a RAAG system, then w' is well-defined by Lemma 3.2. Also note that $w \simeq w'$ and w' is a product of Λ–edges.

Remark 3.10 Suppose that Θ satisfies $\mathcal{R}_1$ and that $a, a' \in \Theta$ are two vertices in the same Λ–component. Let $w' = (a_1 a_2)(a_2 a_3) \cdots (a_{n-1} a_n)$ be the Λ–edge word associated to $w = a a'$ (in particular $a = a_1, a_2, \ldots, a_n = a'$ is the unique simple Λ–path from a to a'). We remark that given a Λ–edge xy of Θ, there is at most one occurrence of one of xy or yx in w'. This fact will be relevant in the proofs of the next two lemmas.

Before diving into the next lemma, we briefly discuss some of the ideas used in its proof, and the proof of Lemma 3.16. In each case, we will have a word w over the RACG W_Γ representing the identity element. We then find a Λ–edge word w' associated to w as in Definition 3.9. The word w' has a natural decomposition into Λ–edges, $w' = (s_1 s_1') \cdots (s_n s_n')$. Moreover, there is a RAAG generator $g_i \in \Delta$ associated to each $s_i s_i' = \phi(g_i)$. By a slight abuse of notation, we also think of $w' = g_1 \cdots g_n$ as a word over the RAAG A_Δ. Doing so, we consider a disk diagram D *in the RAAG A_Δ* with boundary $g_1 \cdots g_n$. The edges of D are labeled by the g_i's. To simplify things, by another abuse of notation we also think of these edges as labeled by the Λ–edges $s_i s_i'$. We use the intersection patterns of hyperplanes in D to deduce commuting relations between the generators of the RAAG. Consequently, this gives us commuting relations between the Λ–edges and for generators in the RACG W_Γ.

Lemma 3.11 *If $(G^\Theta, E(\Lambda))$ is a RAAG system, then Θ satisfies $\mathcal{R}_3$.*

Remark 3.12 (comparison of Lemma 3.11 with Laforge's chain–chord condition) LaForge [2017, Lemma 8.2.3] introduced a necessary condition, called the chain–chord condition, which, if interpreted in the language of joins and 2–component cycles, is close to our condition $\mathcal{R}_3$. We note that there are errors in the statement and proof of [LaForge 2017, Lemma 8.2.3].

Proof of Lemma 3.11 Suppose there is a 2–component square γ in Θ visiting vertices c_1, d_1, c_2 and d_2 as in condition $\mathcal{R}_3$. Let Λ_c and Λ_d be the components of Λ respectively containing $\{c_1, c_2\}$ and $\{d_1, d_2\}$. Let T_c and T_d be the Λ convex hulls respectively of $\{c_1, c_2\}$ and $\{d_1, d_2\}$. By Lemma 3.2, there is a unique simple Λ–path from c_1 to c_2 (resp. d_1 to d_2) and this path is equal to T_c (resp. T_d).

Let w denote the commutator $[c_1 c_2, d_1 d_2]$. The existence of γ tells us that c_1 and c_2 both commute with d_1 and d_2, so w represents the identity in W_Γ.

Let w_1, w_2 and w' be the Λ–edge words associated to respectively $c_1 c_2$, $d_1 d_2$ and w. As ϕ is injective, w' represents the trivial element of A_Δ, and there is a disk diagram D over A_Δ with boundary label w'. We warn that the edges of D are labeled by Λ–edges, ie generators of A_Δ. We will analyze hyperplanes of this diagram.
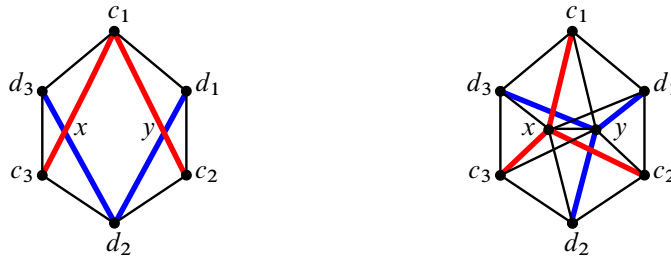
Figure 4: The graph on the left concerns Example 3.13 and the graph of the right concerns Example 3.14.

Let p_{w_1}, p_{w_2}, $p_{w_1^{-1}}$ and $p_{w_2^{-1}}$ be the paths in ∂D with labels w_1, w_2, w_1^{-1} and w_2^{-1} respectively. For $i \in \{1, 2\}$, the word w_i (thought of as a word over $V(\Delta) = E(\Lambda)$) does not contain any repeated letters (or their inverses) in $V(\Delta)$ by Remark 3.10. Consequently, a hyperplane is dual to at most one edge of p_{w_1} (resp. p_{w_2}, $p_{w_1^{-1}}$ and $p_{w_2^{-1}}$). Furthermore, w_1 and w_2 are words over $E(T_c)$ and $E(T_d)$ respectively. As Λ_c and Λ_d are distinct components of Λ, a hyperplane dual to an edge of p_{w_1} must be dual to an edge of $p_{w_1^{-1}}$ and vice versa. A similar statement holds for hyperplanes dual to p_{w_2} and $p_{w_2^{-1}}$.

It follows that every hyperplane dual to p_{w_1} intersects every hyperplane dual to p_{w_2}. Consequently, every Λ–edge in the word w_1 commutes with every Λ–edge in the word w_2. Since ϕ is a homomorphism, the Coxeter group elements corresponding to these Λ–edges must commute as well. By Lemma 2.6 each vertex of T_c commutes with each vertex of T_d. □

The next example shows that the conditions obtained so far are not sufficient for $(G^\Theta, E(\Lambda))$ to be a RAAG system.

Example 3.13 Let Γ be a hexagon, and let Λ be the graph with two components shown on the left side in Figure 4. It is clear that $\mathcal{R}_1, \mathcal{R}_2$, and $\mathcal{R}_3$ are satisfied. However, by considering the word

$$w = (c_1 c_2)(d_1 d_2)(c_2 c_3)(d_2 d_3)(c_3 c_1)(d_3 d_1)$$

we can see that $(G^\Theta, E(\Lambda))$ is not a RAAG system. Specifically, the commutation relations specified by Γ–edges show that $w \simeq 1$ in W_Γ. Moreover w can be expressed as a product of Λ–edges using the Λ–edge words corresponding to each parenthetical element. However, it is not possible to reduce this word to the empty word using just swap and deletion moves involving the Λ–edges, and as a result, $(G^\Theta, E(\Lambda))$ cannot be a RAAG system. A rigorous proof of this fact follows from Lemma 3.16 below.

Example 3.13 shows that at least one additional condition is needed in order to obtain a characterization of visual RAAGs, and suggests that this condition may be a generalization of $\mathcal{R}_3$ involving longer 2–component cycles instead of squares. It is tempting to conjecture that, given any $\Lambda_c \Lambda_d$–cycle with corresponding Λ–convex hulls T_c and T_d, the graph Γ contains the join of $V(T_c)$ and $V(T_d)$ (as is the case when the cycle has length four, by Lemma 3.11 above). However, the following example shows this is not necessarily true for longer cycles.
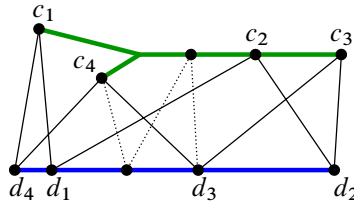
Figure 5: This figure illustrates condition $\mathcal{R}_4$. The green subgraph is T_c and the blue subgraph is T_d. The condition says that any edge in the 2–component cycle (shown in solid black edges) is part of a square of Γ with two vertices in T_c and two in T_d. This is illustrated for the edge from d_3 to c_4. The dotted lines are Γ–edges which are not necessarily in the 2–component cycle.

Example 3.14 In Figure 4, let Θ be the graph on the right where Γ–edges are black and Λ edges are colored. Observe that Λ has two components, colored red and blue. Consider the 2–component cycle visiting vertices $c_1, d_1, c_2, d_2, c_3, d_3$ and c_1. Then T_c is the entire red tree and T_d is the entire blue tree. However, Γ does not contain the join of $V(T_c)$ and $V(T_d)$. (For example, there is no edge in Γ connecting c_1 and d_2.) On the other hand, $(G^\Theta, E(\Lambda))$ is a RAAG system in this case. (See Corollary 5.1 for a proof.)

Despite the fact that $\mathcal{R}_3$ does not generalize to a necessary condition on longer cycles in the obvious way, the following weaker statement does turn out to be necessary to guarantee that $(G^\Theta, E(\Lambda))$ is a RAAG system and is missing from [LaForge 2017].

Definition 3.15 (condition $\mathcal{R}_4$) We say that Θ satisfies *condition $\mathcal{R}_4$* if the following holds. Let γ be any $\Lambda_c\Lambda_d$–cycle in Θ visiting vertices $c_1, d_1, c_2, d_2, \ldots, c_n, d_n, c_1$ for some $n \geq 2$. Let T_c and T_d be the Λ–convex hulls of $\{c_1, \ldots, c_n\}$ and $\{d_1, \ldots, d_n\}$ respectively. Then every edge of γ is contained in a 2–component square of Θ with two vertices in T_c and two vertices in T_d. (See Figure 5.)

The next lemma shows $\mathcal{R}_4$ is necessary for $(G^\Theta, E(\Lambda))$ to be a RAAG system.

Lemma 3.16 *If $(G^\Theta, E(\Lambda))$ is a RAAG system, then Θ satisfies $\mathcal{R}_4$.*

Proof Let γ be a $\Lambda_c\Lambda_d$–cycle visiting vertices $c_1, d_1, \ldots, c_n, d_n, c_1$, and let T_c and T_d be as in Definition 3.15. Let w be the word

$$(1) \qquad w = (c_1 c_2)(d_1 d_2)(c_2 c_3)(d_2 d_3) \cdots (d_{n-1} d_n)(c_n c_1)(d_n d_1).$$

Then $w \simeq 1$ in W_Γ. To see this, note that for each i, we know that c_i commutes with d_{i-1} and d_i (where i is taken mod n). Using this we can cancel the c_i for $i > 1$ in pairs to get

$$w \simeq c_1 d_1 d_2 d_2 d_3 \cdots d_{n-1} d_n c_1 d_n d_1 \simeq c_1 d_1 c_1 d_1 \simeq 1.$$

Let w' be the Λ–edge word associated to w. Let D be a disk diagram over A_Δ with boundary label w'. As in the proof of Lemma 3.11, edges of D are labeled by Λ–edges, which are thought of as generators of A_Δ.

Color the part of the boundary of D and the hyperplanes coming out of it green if they correspond to Λ–edges from Λ_c and blue if they correspond to Λ–edges from Λ_d. Now we see from the structure of w' that ∂D alternates between green and blue stretches, and a stretch of a given color corresponds to a simple path in the corresponding component of Λ. It follows from Remark 3.10 that a hyperplane of a given color must start and end in different stretches of that color.

Let $L = |E(T_c)|$ denote the number of Λ–edges in T_c. We will prove that condition $\mathscr{R}_4$ holds for γ by induction on (n, L). The conclusion of the lemma is obvious for γ corresponding to $(2, L)$ for any L, since the cycle itself is a square. This includes the base case, when $n = 2$ (ie γ is a square) and T_c is an edge. Now let $n > 2$, and assume the claim is true for all (n', L') such that either $n' < n$ or $n' = n$ and $L' < L$.

By Lemma 3.2, T_c and T_d are trees. Now suppose c_j is a leaf of T_c, and let xc_j be the Λ–edge incident to c_j in T_c. Since $c_i \neq c_j$ for all $i \neq j$ (by the definition of a 2–component cycle), we know that xc_j occurs exactly once in w' (as part of the subword of w' representing $c_{j-1}c_j$) and $c_j x$ occurs exactly once in w' (as part of the subword representing $c_j c_{j+1}$). It follows there is a unique hyperplane H labeled xc_j which is dual to both the path whose label is an expression for $c_j c_{j+1}$ and the path whose label is an expression for $c_{j-1} c_j$. Moreover, the subword w'' of w' between these two subwords is the product of Λ–edges which is an expression for $d_{j-1} d_j$. It follows that every hyperplane dual to the path in ∂D labeled w'' must intersect the hyperplane H. By Lemma 2.6, both x and c_j commute (in W_Γ) with each letter of $V(\Gamma)$ used in the word w''. In particular, d_{j-1} and d_j each commute with x.

Now there are two possibilities. Suppose first that $x = c_t$ for some $t \neq j$. Since $t \neq j$ and $n > 2$ (which implies that γ has more than four edges), it follows that either $c_t d_{j-1}$ or $c_t d_j$ is a diagonal of γ. We can use this diagonal to cut γ into two 2–component cycles γ_1 and γ_2 as follows. Assume $c_t d_j$ is a diagonal δ of γ (the other case is analogous), and let β_1 and β_2 be the two components of γ obtained by removing the vertices labeled c_t and d_j. Set $\gamma_1 = \beta_1 \cup \delta$ and $\gamma_2 = \beta_2 \cup \delta$. Note γ_1 and γ_2 each have strictly fewer vertices than γ. For $i = 1, 2$ let T_c^i and T_d^i be the components of the Λ–convex hull of γ_i contained respectively in Λ_c and Λ_d. By the induction hypothesis, we see that every edge in γ_i is part of a square in Γ with two vertices in $T_c^i \subset T_c$ and two in $T_d^i \subset T_d$. Since each edge of γ is either an edge of γ_1 or of γ_2, the claim follows for this case.

On the other hand, suppose that $x \neq c_i$ for any $1 \leq i \leq n$. Consider the new 2–component cycle γ' obtained from γ by replacing the edges $d_{j-1} c_j$ and $c_j d_j$ with $d_{j-1} x$ and $x d_j$. As $x \neq c_i$ for any $1 \leq i \leq n$, this does not violate the requirement that 2–component cycles do not repeat vertices. Let T_c' and T_d' be the components of the Λ–convex hull of γ' contained respectively in Λ_c and Λ_d. Since c_j is a leaf of T_c, it follows that $|E(T_c')| < |E(T_c)|$, and we also have that $|V(\gamma')| = |V(\gamma)| = n$. We now apply the induction hypothesis to conclude that each edge of γ' is part of a square of Γ with two vertices in $T_d' = T_d$ and two vertices in $T_c' \subset T_c$. This means that this property holds automatically for all edges of γ, except possibly $d_{j-1} c_j$ and $c_j d_j$. However, these edges are part of the square in T_c with vertices x, c_j, d_{j-1} and d_j. Thus, the claim follows for this case as well. $\qquad\square$

The following proposition summarizes Lemmas 3.2, 3.4, 3.11 and 3.16.

Proposition 3.17 *If $(G^\Theta, E(\Lambda))$ is a RAAG system, then Θ satisfies $\mathcal{R}_1$–$\mathcal{R}_4$.*

If Λ has at most two components, then it turns out that there are no additional obstructions to $(G^\Theta, E(\Lambda))$ being a RAAG system. More precisely:

Theorem 3.18 *Suppose Λ has at most two components. Then $(G^\Theta, E(\Lambda))$ is a RAAG system if and only if $\mathcal{R}_1$–$\mathcal{R}_4$ are satisfied.*

Proof outline Proposition 3.17 constitutes one direction of the theorem. The following strategy will be used to prove that $\mathcal{R}_1$–$\mathcal{R}_4$ imply that $(G^\Theta, E(\Lambda))$ is a RAAG system. We wish to show that the image of every nontrivial element of A_Δ under ϕ is nontrivial in W_Γ.

Towards a contradiction, we assume that there exists some nontrivial $g \in A_\Delta$ such that $\phi(g) = 1$. Then there is a disk diagram D whose boundary label is a word in Λ–edges which represents $\phi(g)$. We will put this word in a certain normal form which will be defined in terms of the configuration of hyperplanes in D.

To define the normal form, we first show that the set of all hyperplanes can be partitioned into subsets that we call "closed chains of hyperplanes" (see Definition 3.20 and Figure 6). Properties of hyperplanes and closed chains can be translated into information about the graph Θ and vice versa (see Observations 3.19, 3.23 and 3.24). Next, we prove in Lemma 3.25 that we can fix a particular closed chain $\mathcal{H}$ which is "maximally nested" in a certain sense. Specifically, $\mathcal{H}$ has a distinguished hyperplane H_0 such that every other closed chain either intersects H_0 or is separated from the rest of $\mathcal{H}$ by H_0 (see Figure 8).

Our normal form is defined in terms of the fixed closed chain $\mathcal{H}$. We first choose a basepoint p on ∂D which is the endpoint of an edge of ∂D dual to H_0. (This has the effect of possibly replacing our original element $g \in A_\Delta$ with a conjugate.) Let w be the label of ∂D read clockwise starting at p. We show in Claim 3.26 that w, D and $\mathcal{H}$ may be replaced by an equivalent word $\widetilde{w}$ and corresponding disk diagram $\widetilde{D}$ and maximally nested closed chain $\widetilde{\mathcal{H}}$, with the property that the Λ–edges in $\widetilde{w}$ coming from $\widetilde{\mathcal{H}}$ are "as far right as possible", ie it is not possible to swap one of these Λ–edges with a Λ–edge to its right by a commutation relation. We consider $\widetilde{w}$ to be a word in normal form representing $\phi(g)$.

Finally, to complete the proof of Theorem 3.18, we will show (by analyzing interactions between closed chains in $\widetilde{D}$) that if $\mathcal{R}_1$–$\mathcal{R}_4$ are satisfied, then the normal form is violated.

Before we embark on the proof, we need to develop some preliminaries on disk diagrams, and on transferring information from the disk diagram D to the graph Θ. In what follows, we assume that D is a disk diagram whose boundary is a word w in the RACG W_Γ. Unlike in the proofs of Lemmas 3.11 and 3.16, we are now working in W_Γ rather than A_Δ, so the edges and hyperplanes of D are labeled by generators of W_Γ rather than elements of A_Δ corresponding to Λ–edges. As the words w we consider are the images of elements of A_Δ under ϕ, they have a natural decomposition into Λ–edges.
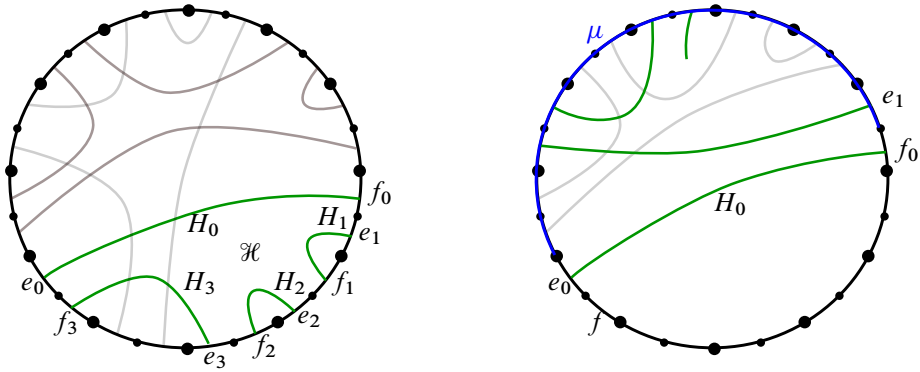
Figure 6: The figure on the left shows a disk diagram D such that the label of ∂D has a natural decomposition into Λ–edges, delineated by large black dots. The green hyperplanes form a closed chain of hyperplanes $\mathcal{H}$, as defined in Definition 3.20 (here we can take $\eta = \partial D$). Two other closed chains of hyperplanes are shown in gray. The figure on the right shows an impossible configuration pertaining to the proof of Lemma 3.21. The path μ from the lemma is colored blue.

We associate a color (red and green) to each component of Λ. Each hyperplane of D then inherits the color corresponding to the component of Λ in which its label lies. Thus, two edges of ∂D contained in the same Λ–edge are dual to hyperplanes of the same color.

Observation 3.19 *If Θ satisfies $\mathcal{R}_2$, then no two hyperplanes of the same color intersect. This is because if two hyperplanes intersect, then their labels are distinct and commute, and so are connected by a Γ–edge. Thus they cannot be in the same component of Λ, since each component of Λ is an induced subgraph of Θ, by $\mathcal{R}_2$.*

The hyperplanes of D can be partitioned into "closed chains of hyperplanes", as described in Definition 3.20 below. Although the proof of Theorem 3.18 only uses disk diagrams whose boundary labels are words in Λ–edges, the definition below applies to slightly more general disk diagrams, as this will be needed in Section 6.

Definition 3.20 (chains of hyperplanes) Let D be a disk diagram whose boundary ∂D contains a connected subpath η (possibly all of ∂D), such that the label of η is a word in Λ–edges. Let $H_0, \ldots, H_n$ be a sequence of distinct hyperplanes in D. Let e_i and f_i be the edges on ∂D that are dual to H_i. (See Figure 6 for an illustration when $n = 3$.) We say that $\{H_0, \ldots, H_n\}$ is a *chain* in D, if for all $0 \le i < n$, the edges f_i and e_{i+1} are contained in η and are dual to the same Λ–edge of η. Note that e_0 and f_n can be dual to edges not contained in η.

Additionally, if e_0 and f_n are contained in the same Λ–edge of η, we say that $\{H_0, \ldots, H_n\}$ is a *closed chain*. (Figure 6 shows three closed chains.)

Since the two hyperplanes dual to a Λ–edge have the same color, each chain also inherits a well-defined color.

Lemma 3.21 *If the label of ∂D is a word in Λ–edges, then every hyperplane of D is contained in a unique closed chain. Thus, there is a partition of the hyperplanes of a given color into closed chains.*

Proof Let H_0 be a hyperplane of D dual to edges e_0 and f_0 of ∂D. Assume H_0 is green. Let f and e_1 be edges of ∂D which pair with e_0 and f_0 respectively to form Λ–edges. We claim that f and e_1 are in the same component of $D \setminus H_0$. If not, there would be an odd number of edges in a part μ of ∂D between e_0 and f_0 (see the right side of Figure 6). Since the hyperplanes dual to the two edges of a Λ–edge have the same color, an odd number of these edges would be dual to green hyperplanes. This is a contradiction, since no green hyperplanes can cross H_0 by Observation 3.19, so there must be an even number of edges in μ dual to green hyperplanes.

The hyperplane H_1 dual to e_1 is green, and cannot cross H_0. Let f_1 be the other edge dual to H_1. If $f_1 = f$ we have a closed chain. Otherwise, there is an edge e_2 which pairs with f_1 to form a Λ–edge. By same argument as before, e_2 is in the same component of $D \setminus H_1$ as f_0 and there is a green hyperplane H_2 dual e_2 and another edge f_2, such that H_2 does not cross H_0 or H_1 (see the left side of Figure 6). Continuing this process we obtain a sequence of hyperplanes as in Definition 3.20. Since the number of possibilities for f_i reduces each time, eventually the process stops, with $f_n = f$ for some n, and $H_0, \ldots, H_n$ form a closed chain. $\qquad\square$

We say that a chain $\mathcal{H}$ intersects a hyperplane H if some $K \in \mathcal{H}$ intersects H. We say that chains $\mathcal{H}$ and $\mathcal{K}$ intersect if $\mathcal{K}$ intersects some $H \in \mathcal{H}$. We will need the following observation:

Observation 3.22 *If a hyperplane H intersects a closed chain $\mathcal{K}$, then it intersects $\mathcal{K}$ in exactly two distinct hyperplanes. To see this, note that given a hyperplane $K \in \mathcal{K}$, the hyperplanes in $\mathcal{K} \setminus \{K\}$ all lie in a single component of $D \setminus K$. It follows that if H intersects $\mathcal{K}$ more than twice, it must intersect some hyperplane of $\mathcal{K}$ twice. This contradicts the fact that D has no bigons (see Remark 2.9).*

The following two observations enable us to transfer information from the disk diagram D to the graph Θ.

Observation 3.23 (chains in D give Λ–paths in Θ) *Let $\mathcal{K} = \{K_0, \ldots, K_l\}$ be a chain in D, and for $0 \le i \le l$, let k_i be the label of K_i. Then by the definition of a chain, K_i and K_{i+1} are dual to the same Λ–edge in ∂D for each i, so there is an edge in Λ between k_i and k_{i+1}. It follows that $\mathcal{K}$ naturally defines a Λ–path in Θ visiting vertices $k_0, k_1, \ldots, k_l$. Moreover, if $\mathcal{K}$ is a closed chain, then the corresponding Λ–path is a loop. See Figure 7.*

Observation 3.24 (pairs of intersecting closed chains give 2–component loops in Θ) *Consider two closed chains which intersect, say a red chain $\mathcal{H}$ and a green chain $\mathcal{K}$. Let $H_1 \in \mathcal{H}$ and $K_1 \in \mathcal{K}$ be intersecting hyperplanes. By Observation 3.22, the hyperplane K_1 intersects $\mathcal{H}$ in a second hyperplane $H_2 \neq H_1$. Similarly, H_2 intersects $\mathcal{K}$ in a second hyperplane K_2. Proceeding in this way, we obtain a polygon with at least four sides, with sides alternating between red and green hyperplanes. See the left side of Figure 7.*
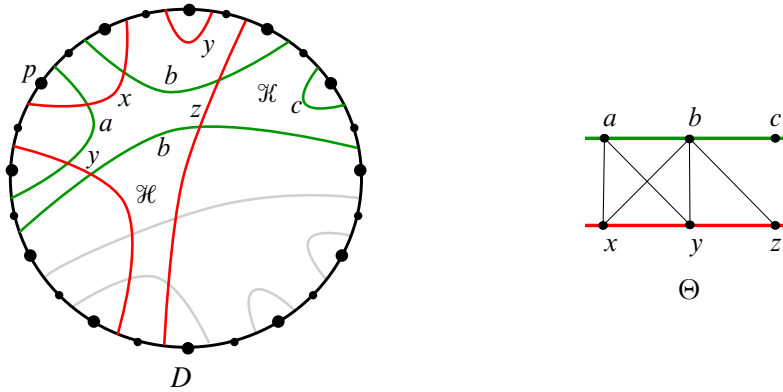
Figure 7: The figure illustrates Observations 3.23 and 3.24. On the left is a disk diagram D with red and green closed chains called $\mathcal{H}$ and $\mathcal{K}$ respectively. The graph on the right is a part of Θ. The labels of the hyperplanes in D correspond to vertices of Θ (in particular of Λ). Starting at the basepoint p and going around ∂D clockwise, the closed chain $\mathcal{H}$ defines a Λ–loop in Θ visiting vertices x, y, z, y and x, and the closed chain $\mathcal{K}$ defines a Λ–loop visiting vertices a, b, c, b and a. The polygon coming from the intersection of $\mathcal{H}$ and $\mathcal{K}$ defines a 2–component loop in Θ visiting vertices a, x, b, z, b, y and a. Observe that this 2–component loop is not a cycle.

Since an intersecting pair of hyperplanes corresponds to an edge of Γ, a 2–colored polygon of the type we just constructed defines a 2–component loop in Θ (where each edge of the 2–component loop comes from a corner of the 2–colored polygon). See Figure 7. We warn that the 2–component loop obtained from a 2–colored polygon in D may not be a 2–component cycle. (Note that a 2–component cycle is a 2–component loop in which all of the vertices are distinct, and there are at least two vertices in each component.)

In order to define a normal form for the word u from the proof outline, we will need to choose a closed chain in D with some special properties:

Lemma 3.25 *Let u and D be as in the proof outline. There exists a closed chain $\mathcal{H}$ of D, containing a distinguished hyperplane H_0, such that given any closed chain $\mathcal{K} \neq \mathcal{H}$, either*

(1) *$\mathcal{K}$ and $\mathcal{H} \setminus \{H_0\}$ lie in different components of $D \setminus H_0$, or*

(2) *$\mathcal{K}$ intersects H_0.*

Proof We iteratively construct a sequence of closed chains $\mathcal{H}^1, \mathcal{H}^2, \ldots$ with distinguished hyperplanes $H_0^1, H_0^2, \ldots$ such that for all $i > 1$,

(i) *$\mathcal{H}^i$ and $\mathcal{H}^{i-1} \setminus \{H_0^{i-1}\}$ lie in the same component of $D \setminus H_0^{i-1}$, and*

(ii) *H_0^{i-1} and $\mathcal{H}^i \setminus \{H_0^i\}$ lie in different components of $D \setminus H_0^i$.*

Let $\mathcal{H}^1$ and H_0^1 be arbitrary. Now for any j, if $\mathcal{H}^j$ and H_0^j do not satisfy the conclusion of the lemma, then there must exist another closed chain $\mathcal{H}^{j+1}$ which lies entirely in C_j, where C_j is the component
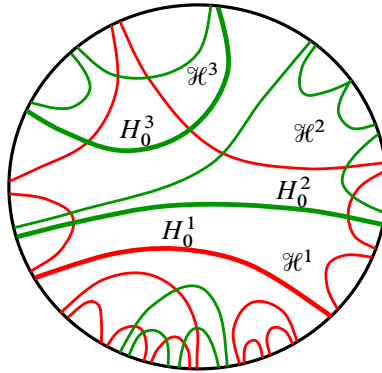
Figure 8: The figure illustrates the procedure for finding $\mathcal{H}$ and H_0 in Lemma 3.25. Each closed chain in the sequence is labeled in its interior. The hyperplanes H_0^i are shown in bold. In this example $\mathcal{H} = \mathcal{H}^3$ has the desired property.

of $D \setminus H_0^j$ containing $\mathcal{H} \setminus \{H_0^j\}$. (Figure 8 illustrates this for $j = 1, 2$.) There is a unique hyperplane in $\mathcal{H}^{j+1}$ satisfying condition (ii) above with $i = j + 1$, and we set this equal to H_0^{j+1}. Thus, we can produce a longer sequence of closed chains with properties (i) and (ii).

By construction, there is a nesting of components $C_1 \supset C_2 \supset C_3 \supset \cdots$, and it follows that $H_0^1, H_0^2, \ldots$ are distinct hyperplanes in D. As D has finitely many hyperplanes, this process can only be repeated finitely many times. Thus, $\mathcal{H}^j$ satisfies the claim for some j. $\qquad\square$

We are now ready to prove the theorem.

Proof of Theorem 3.18 As discussed, we need to show that if $\mathcal{R}_1$–$\mathcal{R}_4$ are satisfied, then the map $\phi: A_\Delta \to G^\Theta$ is injective. Let $g \in A_\Delta$ be a nontrivial element. Let $v = v_1 v_2 \cdots v_n$ be a reduced word over the set of the generators of A_Δ, which represents g. By the definition of A_Δ, we have that $\phi(v_i)$ is a Λ–edge of Θ, for $1 \le i \le n$. Then $u = \phi(v_1)\phi(v_2)\cdots\phi(v_n)$ is a concatenation of Λ–edges which represents $\phi(g)$. Towards a contradiction, we assume that u represents the identity element of W_Γ. Then there is a disk diagram D whose boundary label (read clockwise starting from some basepoint) is u. By Lemma 2.8 we may assume that D has no bigons.

An element has trivial image under ϕ if and if every element of its conjugacy class does. Thus, we may assume that g is of minimal length in its conjugacy class, where the length of an element is defined to be the minimal length of a word representing it.

We partition the hyperplanes of D into closed chains. (See Lemma 3.21.) By Lemma 3.25, we can choose a chain $\mathcal{H}$, with distinguished hyperplane H_0, such that given any other chain $\mathcal{K}$, either H_0 separates $\mathcal{H}$ from $\mathcal{H} \setminus \{H_0\}$, or $\mathcal{K}$ intersects H_0. Let $a_0, a_1, \ldots, a_s$ be the labels of the hyperplanes of $\mathcal{H}$, starting from H_0, and proceeding in order in the clockwise direction around ∂D. Then the Λ–edges $a_0 a_1, a_1 a_2, \ldots, a_{s-1} a_s, a_s a_0$ appear in ∂D in that order, possibly interspersed with some other Λ–edges.
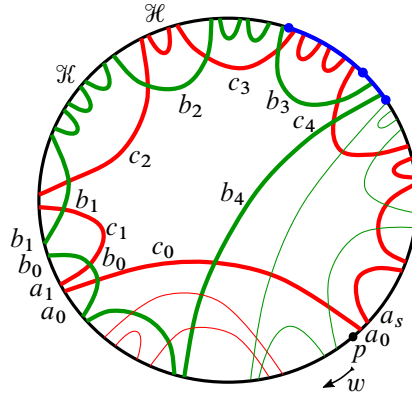
Figure 9: This example illustrates the proof of Theorem 3.18. The chain $\mathcal{H}$ satisfying Claim 3.26 is shown in thick red lines. In particular, no Λ–edge from $\mathcal{H}$ (except possibly the last one) commutes with the Λ–edge appearing after it in ∂D. The chain $\mathcal{K}$, which contributes the first Λ–edge not in $\mathcal{H}$ (after $a_0 a_1$), is shown in thick green lines. The polygon formed by the intersection of $\mathcal{H}$ and $\mathcal{K}$ induces a 2–component loop which visits (in this example) $c_0, b_0, c_1, b_1, c_2, b_2, c_3, b_3, c_4, b_4, c_0$. The blue subpaths of ∂D are the subpaths defined in Claim 3.28, with $i = 4$.

Let p denote the vertex on ∂D which is the endpoint of the Λ–edge from $\mathcal{H}$ labeled $a_s a_0$, read clockwise. (See Figure 9.) Let w be the word labeling ∂D clockwise, starting from p. Then w is a cyclic conjugate of u. Let x be the corresponding cyclic conjugate of v. Since v was chosen to be reduced, and since g (the element of A_Δ represented by v) is of minimal length in its conjugacy class by assumption, it follows that x is reduced.

We now show that we can modify D in such a way that the resultant boundary label is a word representing $w = \phi(x)$ which is in a certain normal form:

Claim 3.26   *There exists a disk diagram $\widetilde{D}$ such that the following hold.*

(1) *There is a closed chain $\widetilde{\mathcal{H}}$ in $\widetilde{D}$ which has a distinguished hyperplane $\widetilde{H}_0$ satisfying the criterion in Lemma 3.25. The labels of the hyperplanes of $\widetilde{\mathcal{H}}$ starting from $\widetilde{H}_0$ and proceeding clockwise, are $a_0, \ldots, a_s$ (ie they are the same labels as the labels of the hyperplanes in $\mathcal{H}$).*

(2) *Let $\tilde{p}$ be the endpoint of the Λ–edge $a_s a_0$ from $\widetilde{\mathcal{H}}$, and let $\widetilde{w}$ be the word labeling $\partial \widetilde{D}$ in the clockwise direction starting from $\tilde{p}$. Then $\widetilde{w} = \phi(\tilde{x})$, where $\tilde{x}$ is a reduced word in A_Δ obtained from x by Tits swap moves.*

(3) *The Λ–edges from $\widetilde{\mathcal{H}}$ appear as far right as possible in $\widetilde{w}$. More formally, the word $\widetilde{w}$ has no subword of the form $a_i a_{i+1} b b'$ such that $a_i a_{i+1}$ is one of the Λ–edges coming from $\widetilde{\mathcal{H}}$ with $0 \le i \le s$ (with indices mod s), $a_i a_{i+1} \ne b b'$, and $a_i a_{i+1}$ commutes with $b b'$.*

Proof   We construct $\widetilde{D}$ iteratively, starting with D. If $\mathcal{H}$, p, w and x are as defined above, then the first two conditions in the claim are satisfied. If (3) is not satisfied, then w has a subword $a_i a_{i+1} b b'$ as in (3). Since $a_s a_0$ is the last Λ–edge of w, we conclude that $a_i a_{i+1} \ne a_s a_0$.

Note that $a_i a_{i+1} \neq bb'$ (by condition (3)) and $a_i a_{i+1} \neq (bb')^{-1}$ (since x is reduced and $w = \phi(x)$). Then it follows from Lemma 2.6, that each of a_i and a_{i+1} commutes with each of b and b'. By applying Lemma 2.11(1) four times, we obtain a new disk diagram D' such that the label of $\partial D'$ is obtained from the label of ∂D by swapping the Λ–edges $a_i a_{i+1}$ and bb'. Moreover, the natural map ψ from the edges of ∂D to the edges of $\partial D'$ (defined in Lemma 2.11(1)) preserves boundary combinatorics. By applying Lemma 2.8 if necessary, we may assume that D' has no bigons, so hyperplanes in D' intersect at most once.

Since boundary combinatorics are preserved, ψ induces a bijection between the hyperplanes dual to ∂D and those dual to $\partial D'$. Since the transition from D to D' involves swapping a pair of Λ–edges, the label of $\partial D'$ is still a product of Λ–edges, and so the hyperplanes of D' can be partitioned into closed chains of hyperplanes. Moreover, ψ induces a bijection between the closed chains of hyperplanes in D and D'.

If $\mathcal{H}'$ and H_0' denote the images of $\mathcal{H}$ and H_0 respectively under ψ, it is clear that the labels of the hyperplanes of $\mathcal{H}'$, starting from H_0' and proceeding clockwise, are $a_0, \ldots, a_s$. We now prove that $\mathcal{H}'$ together with H_0' still satisfies the criterion in Lemma 3.25 required in (1).

Let $\mathcal{K}'$ be a closed chain in D', and let $\mathcal{K}$ be its preimage in D. Our choice of $\mathcal{H}$ implies that either H_0 separates $\mathcal{K}$ from $\mathcal{H} \setminus \{H_0\}$, or $\mathcal{K}$ intersects H_0. In the former case, H_0' still separates $\mathcal{K}'$ from $\mathcal{H}' \setminus \{H_0'\}$. This is because the swap performed does not involve any hyperplanes from chains which do not intersect H_0, since (as noted above) $a_i a_{i+1} \neq a_s a_0$.

On the other hand, suppose that $\mathcal{K}$ intersects H_0. By Observation 3.22, there are exactly two hyperplanes K_1 and K_2 in $\mathcal{K}$ which intersect H_0. If K_j, for $j = 1, 2$, is not dual to the Λ–edge labeled by bb', then the image of K_j intersects H_0'. Moreover, if $i \neq 0$, then it follows that the images of K_1 and K_2 in D' intersect the hyperplane H_0'. Thus, we only need to consider the case where the Λ–edge $a_0 a_1$ is swapped, and (up to relabeling) K_1 is dual to b and K_2 is dual to b'. In this case, K_1 and K_2 are dual to the same Λ–edge. It follows that no hyperplane in $\mathcal{K} \setminus \{K_1, K_2\}$ is contained in the same component of $D \setminus H_0$ as $\mathcal{H} \setminus \{H_0\}$. Thus, in D', no hyperplane of $\mathcal{K}'$ is contained in the same component of $D' \setminus H_0'$ as $\mathcal{H}' \setminus \{H_0'\}$. We have shown that $\mathcal{H}'$, with distinguished hyperplane H_0', satisfies the conclusion of Lemma 3.25.

Let p' be the vertex on $\partial D'$ which is the endpoint of the Λ–edge from $\mathcal{H}'$ labeled $a_s a_0$. Since the swap performed did not involve $a_s a_0$, the label w' of $\partial D'$, read clockwise from p', is obtained from w by swapping a single pair of Λ–edges, and its preimage in x' in A_Δ is obtained from x by swapping one pair of generators. This shows (2).

We have established that D', together with $\mathcal{H}'$, satisfies (1) and (2) of Claim 3.26. If (3) still fails, we may repeat the process above. Since each individual iteration involves moving one Λ–edge from the image of $\mathcal{H}$ to the right, this process eventually stops. After finitely many iterations, we arrive at a disk diagram $\widetilde{D}$ such that all three conditions hold. □

For the rest of the proof we assume, without loss of generality, that D, $\mathcal{H}$, p, w and x satisfy the conclusion of Claim 3.26.

We now analyze closed chains which intersect $\mathcal{H}$. First consider the case that there are no such chains. This includes the case when Λ has a single component. Since $\mathcal{H}$ is a closed chain, it defines a loop in Λ. (See Observation 3.23.) On the other hand, since no chains intersect $\mathcal{H}$, the union of the edges of ∂D dual to the hyperplanes of $\mathcal{H}$ is a continuous subpath (with label $(a_0 a_1)(a_1 a_2) \cdots (a_s a_0)$). Applying the following claim to this subpath, we conclude that the Λ–loop defined by $\mathcal{H}$ is a cycle. This contradicts $\mathcal{R}_1$. (The claim will be used again later in this proof.)

Claim 3.27 *Let v be a subpath of ∂D labeled by a product of Λ–edges. Suppose there exists a closed chain $\mathcal{X}$, such that each edge of v is dual to a hyperplane in $\mathcal{X}$. It follows that the label of v is $(x_1 x_2) \cdots (x_{n-1} x_n)$, where $x_1, x_2, \ldots, x_n$ are the labels of the hyperplanes of $\mathcal{X}$ dual to v, in order. Furthermore, the Λ–path through vertices $x_1, \ldots, x_n$ is simple.*

Proof The claim about the label of v is immediate. If the path through vertices $x_1, \ldots, x_n$ is not simple, then there is a Λ–loop through vertices $x_i, x_{i+1}, \ldots, x_{i+j} = x_i$ for some i and j. By $\mathcal{R}_1$, the image of this loop in Λ is a tree. Let x_r be a leaf of this tree, with $i < r < j$. It follows that $x_{r-1} = x_{r+1}$. Consequently, the label of v (and therefore of the word w) has a subword $(x_{r-1} x_r)(x_r x_{r-1})$. This is a contradiction, as it implies that the preimage x of w in A_Δ is not reduced. $\qquad \square$

Thus, we may assume that there is at least one chain intersecting $\mathcal{H}$. In particular, Λ has two components: say a red component Λ_a which contains the labels of $\mathcal{H}$, and a green component Λ_b. By Claim 3.26, each chain intersecting $\mathcal{H}$ intersects H_0. Let $\mathcal{K}$ be the "first" such chain, in the sense that the first Λ–edge from a chain other than $\mathcal{H}$ appearing in w to the right of $a_0 a_1$ is from $\mathcal{K}$. (See Figure 9.) By $\mathcal{R}_2$ and Observation 3.19, we conclude that $\mathcal{K}$ is green. Let $b_0, \ldots, b_{s'}$ be the labels of the hyperplanes of $\mathcal{K}$, where $b_0 b_1$ is the label of the first Λ–edge from $\mathcal{K}$ appearing in w to the right of $a_0 a_1$.

Now consider the 2–colored polygon in D whose sides alternate between hyperplanes in $\mathcal{H}$ and $\mathcal{K}$, as described in Observation 3.24. Let $c_0, d_0, \ldots, c_k, d_k$ be the labels of these sides, where $c_0 = a_0$, $d_0 = b_0$, and $c_0, \ldots, c_k$ (resp. $d_0, \ldots, d_k$) is a subsequence of $a_0, \ldots, a_s$ (resp. of $b_0, \ldots, b_{s'}$). (See Figure 9.)

The following technical claim about the hyperplanes dual to certain subpaths of ∂D associated to this 2–colored polygon will be needed in what follows:

Claim 3.28 *For $0 \leq i \leq k$, let e_i and f_i (resp. e_i' and f_i') be the edges dual to the hyperplane of $\mathcal{H}$ labeled c_i (resp. the hyperplane of $\mathcal{K}$ labeled d_i), where e_i (resp. e_i') appears before f_i (resp. f_i') reading clockwise from p.*

For $i > 0$, let η_i be the subpath of ∂D from (and including) f_{i-1} to (and including) e_i, and let μ_i be the subpath of ∂D from the endpoint of η_i to (and including) e_i'. (See Figure 10.) Then every edge of η_i (resp. μ_i) is dual to a hyperplane in $\mathcal{H}$ (resp. $\mathcal{K}$).
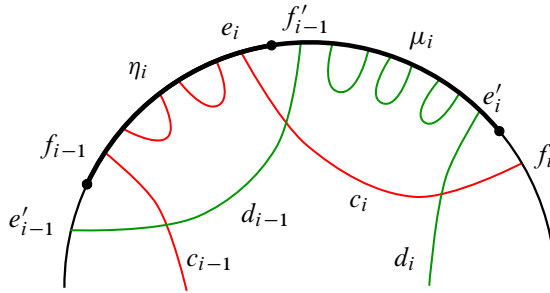
Figure 10: The paths η_i and μ_i from Claim 3.28 are shown in bold, delineated by dots. We remark that if $i = k$, then there could be additional hyperplanes not in $\mathcal{K}$ or $\mathcal{H}$ between the endpoint of μ_k and the start of the edge f_i.

Proof Suppose there is some hyperplane L dual to an edge e of η_i such that the closed chain $\mathcal{L}$ containing L is not equal to $\mathcal{H}$. From the definition of η_i, we conclude that e is on the same side of H_0 as $\mathcal{H} \setminus \{H_0\}$ in D, and by our choice of $\mathcal{H}$ (and Lemma 3.25), it follows that $\mathcal{L}$ intersects H_0. Therefore, L is green by Observation 3.19.

Let K denote the hyperplane of $\mathcal{K}$ labeled d_{i-1}. Then K separates e from $\mathcal{K} \setminus \{K\}$, so $L \notin \mathcal{K}$, ie $\mathcal{L} \neq \mathcal{K}$. If $d_{i-1} = b_0$, ie if K does intersect H_0, then our choice of $\mathcal{K}$ implies that K is the first hyperplane not in $\mathcal{H}$ dual to ∂D after the Λ–edge $a_0 a_1$, so such an L cannot exist. On the other hand, if K does not intersect H_0, then K separates e from H_0. So, in order to intersect H_0, the chain $\mathcal{L}$ must also intersect $\mathcal{K}$, which is a contradiction, since $\mathcal{L}$ and $\mathcal{K}$ are both green.

Now suppose $L \in \mathcal{L} \neq \mathcal{H}$ is dual to an edge e of μ_i. Since μ_i is only defined for $i > 0$, it is on the same side of H_0 as $\mathcal{H} \setminus \{H_0\}$, and consequently, the same holds for e. Therefore, we conclude as before that L is green.

Additionally, we conclude as before that the hyperplane $K \in \mathcal{K}$ labeled d_{i-1} does not intersect H_0. Now consider the subchain of $\mathcal{K}'$ of $\mathcal{K}$ consisting of the hyperplanes dual to all but the last edge e_i' of μ_i. Since K does not intersect H_0, it follows that e is separated from H_0 by some hyperplane in $\mathcal{K}'$. Thus, in order to intersect H_0, $\mathcal{L}$ must intersect $\mathcal{K}$, which is again a contradiction. $\square$

The 2–colored polygon obtained above gives a 2–component loop in Θ, as described in Observation 3.24. A priori this loop may not be a 2–component cycle, ie it is possible that $c_i = c_j$ or $d_i = d_j$ for some i and j. However, we now show that it contains a cycle. We will then be able to apply $\mathcal{R}_4$ to this cycle to make progress towards obtaining a contradiction to the normal form in Claim 3.26.

Claim 3.29 *Consider the 2–component loop in Θ visiting $c_0, d_0, \ldots, c_k, d_k, c_{k+1} = c_0$ defined above. There exist $0 \leq l \leq k - 1$ and $m \geq 2$, such that one of the two following subsequences of vertices (with indices taken mod $k + 1$) defines a 2–component cycle in Θ:*

(1) $c_l, d_l, c_{l+1}, \ldots, d_{l+m-1}, c_{l+m} = c_l$;

(2) $d_l, c_{l+1}, d_{l+1}, \ldots, c_{l+m}, d_{l+m} = d_l$.

Proof Observe that since $c_{k+1} = c_0$, the following set is nonempty:

$$\{j \mid c_i = c_{i+j} \text{ or } d_i = d_{i+j} \text{ for some } 0 \le i < k - 1 \text{ and } 1 \le j \le k + 1\}.$$

Let m denote its minimum value. We first show that $m \ge 2$, or equivalently that, for each $0 \le i \le k - 1$, both $c_i \ne c_{i+1}$ and $d_i \ne d_{i+1}$ are true. Suppose $c_i = c_{i+1}$ for some i. Consider the path η_{i+1} from Claim 3.28. It is labeled by Λ–edges, and every edge in it is dual to a hyperplane from $\mathcal{H}$. Then by Claim 3.27, it follows that η_{i+1} defines a simple Λ–path from the vertex c_i to the vertex c_{i+1}. However, this contradicts the assumption that $c_i = c_{i+1}$. This proves that for all $0 \le i \le k - 1$, we have $c_i \ne c_{i+1}$. The proof that $d_i \ne d_{i+1}$ is similar.

Now if l is such that $c_l = c_{l+m}$ (the case when $d_l = d_{l+m}$ is similar), then it readily follows from the minimality of m that the vertices $c_l, d_l, c_{l+1}, \ldots, c_{l+m-1}, d_{l+m-1}$ are distinct, and therefore define the desired cycle. $\square$

Continuing the proof of the theorem, we can now assume Θ has a 2–component cycle γ as in (1) from Claim 3.29. (The case in which Θ has a 2–component cycle as in (2) is similar.) Let T_c and T_d be the Λ–convex hulls of $\{c_l, \ldots, c_{l+m-1}\}$ and $\{d_l, \ldots, d_{l+m-1}\}$ respectively. Then T_c and T_d are trees by $\mathcal{R}_1$. Let c_j be a leaf of T_c with $c_j \ne c_0$. Then c_j labels a hyperplane $H_t \in \mathcal{H}$ for some $t \ne 0$, so $a_t = c_j$. Similarly, c_{j-1} labels a hyperplane H_{t-r} of $\mathcal{H}$, while d_{j-1} and d_j label hyperplanes $K_{t'}$ and $K_{t'+r'}$ respectively of $\mathcal{K}$, where $d_{j-1} = b_{t'}$ and $d_j = b_{t'+r'}$

Consider the paths η_j and μ_j defined in Claim 3.28. The last Λ–edge of η_j is $a_{t-1}a_t$. By Claim 3.28, the first edge of μ_j is dual to a hyperplane in $\mathcal{K}$. It follows that this must be $K_{t'}$, with label $b_{t'}$, for otherwise $K_{t'}$ would separate this edge from $\mathcal{K} \setminus K_{t'}$. It follows that the first Λ–edge of μ_j is $b_{t'}b_{t'+1}$, and that the word w has a subword $a_{t-1}a_t b_{t'}b_{t'+1}$.

To complete the proof, we will show that the presence of this subword violates the normal form established in Claim 3.26(3). Since the labels of $\mathcal{H}$ and $\mathcal{K}$ are from different components of Λ, it is immediate that $a_{t-1}a_t \ne b_{t'}b_{t'+1}$. We now show that $a_{t-1}a_t$ and $b_{t'}b_{t'+1}$ commute.

The 2–component cycle γ in Θ contains an edge with endpoints a_t and $b_{t'}$. Applying $\mathcal{R}_4$ to this edge, we conclude that there is a 2–component square visiting a_t, $b_{t'}$, a and b, where $a \in T_c$ and $b \in T_d$. Next, applying $\mathcal{R}_3$ to this 2–component square, we see that $b_{t'}$ commutes with the vertices of the Λ–convex hull of $\{a_t, a\}$. Claims 3.27 and 3.28 together imply that the path η_j induces a simple Λ–path visiting vertices $a_{t-r}, a_{t-r+1}, \ldots, a_t$. Consequently, the vertices along this path, and in particular a_{t-1}, are in T_c. Moreover, a_{t-1} is the unique vertex of T_c adjacent to a_t, since $a_t = c_j$ is a leaf of T_c. It follows that a_{t-1} is contained in the Λ–convex hull (which is the same as the T_c–convex hull) of $\{a_t, a\}$. Thus, a_{t-1} and $b_{t'}$ commute. The same reasoning, applied to the edge of γ with endpoints a_t and $b_{t+r'}$, implies that a_{t-1} and $b_{t+r'}$ commute.

Using the Γ–edges whose existence is implied by these two additional commutation relations, we obtain a 2–component square visiting a_t, $b_{t'}$, a_{t-1} and $b_{t'+r'}$. Applying $\mathcal{R}_3$ to this square, we conclude that a_t
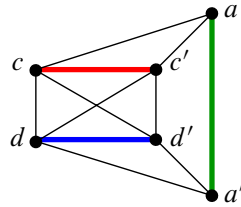
Figure 11: This figure illustrates condition $\mathscr{R}_5$. The red, blue and green segments are respectively T_c, T_d and T_a. Condition $\mathscr{R}_5$ states that any Λ–edge contained in the green segment must either commute with every Λ–edge in the red segment or must commute with every Λ–edge in the blue segment.

and a_{t-1} commute with each vertex in the Λ–convex hull of $\{b_{t'}, b_{t'+r'}\}$. By Claim 3.27, we see that the path μ_j from Claim 3.28 defines a simple Λ–path visiting $b_{t'}, b_{t'+1}, \ldots, b_{t'+r'}$. It follows that $b_{t'+1}$ is in the convex hull of $\{b_{t'}, b_{t'+r'}\}$, and consequently, a_t and a_{t-1} commute with $b_{t'+1}$.

Putting together the commutation relations established in the previous paragraphs, we conclude that $a_{t-1}a_t$ commutes with $b_{t'}b_{t'+1}$. This contradicts the fact that we have chosen D so that it satisfies (3) of Claim 3.26. □

3.1 Three or more Λ–components

In the case that Λ contains at most two components, Theorem 3.18 shows that $\mathscr{R}_1$–$\mathscr{R}_4$ are necessary and sufficient conditions that guarantee $(G^\Theta, E(\Lambda))$ is a RAAG system. In this subsection, we do not place any restriction on the number of components of Λ. We give an additional necessary condition $\mathscr{R}_5$ for $(G^\Theta, E(\Lambda))$ to be a RAAG system, and Example 3.31 shows this condition is independent of conditions $\mathscr{R}_1$–$\mathscr{R}_4$. The authors are aware that *even more* conditions are required in order to generalize Theorem 3.18 to this setting. These extra conditions are not included here, as they are complicated and the authors do not believe to yet possess the complete list of the necessary and sufficient conditions for this generalization.

We further show in this subsection that if Θ contains certain subgraphs and $(G^\Theta, E(\Lambda))$ is a RAAG system, then Γ must necessarily contain a triangle. These results are needed in the next section.

Definition 3.30 (condition $\mathscr{R}_5$) We say that Θ satisfies *condition $\mathscr{R}_5$* if the following holds. Let Λ_a, Λ_c and Λ_d be distinct components of Λ. Suppose we have vertices $a, a' \in \Lambda_a$, $c, c' \in \Lambda_c$ and $d, d' \in \Lambda_d$, such that Θ contains a 2–component square visiting c, d, c' and d'. Furthermore, suppose that c and c' are each adjacent to a in Γ and that d and d' are each adjacent to a' in Γ. (See Figure 11.) Let T_a, T_c and T_d be the Λ–convex hulls of $\{a, a'\}$, $\{c, c'\}$ and $\{d, d'\}$ respectively. Then given any Λ–edge xx' of T_a, the graph Γ contains either the join of $\{x, x'\}$ with $V(T_c)$ or the join of $\{x, x'\}$ with $V(T_d)$.

The following is a concrete example showing that when Λ has more than two components, the conditions $\mathscr{R}_1$–$\mathscr{R}_4$ are not sufficient to guarantee that $(G^\Theta, E(\Lambda))$ is a RAAG system.

Example 3.31 Let Γ be the graph whose vertex set is $\{a, a', c, c', d, d'\}$ and whose edge set is the set of black edges in Figure 11. Let $\Lambda \subset \Gamma^c$ consist of exactly three Λ–edges: aa', cc' and dd'. Then $\Theta = \Theta(\Gamma, \Lambda)$ satisfies conditions $\mathcal{R}_1$–$\mathcal{R}_4$ and does not satisfy condition $\mathcal{R}_5$. By Lemma 3.32 below, $(G^\Theta, E(\Lambda))$ is not a RAAG system.

We now show that condition $\mathcal{R}_5$ is necessary.

Lemma 3.32 If $(G^\Theta, E(\Lambda))$ is a RAAG system, then Θ satisfies condition $\mathcal{R}_5$.

Proof By Theorem 3.18, we may assume that Θ satisfies conditions $\mathcal{R}_1$–$\mathcal{R}_4$. Let $a, a' \in \Lambda_a$, $c, c' \in \Lambda_c$ and $d, d' \in \Lambda_d$ be as in Definition 3.30. Define the words $z_a = a'a$, $z_c = cc'$, $z_d = dd'$ and $z = [z_a z_c z_a^{-1}, z_d]$. By the commuting relations imposed in Definition 3.30, it follows that $z \simeq 1$ in W_Γ. Let w_a, w_c, w_d and w be the Λ–edge words corresponding respectively to z_a, z_c, z_d and z. Let D be a disk diagram over A_Δ with boundary label w.

Let γ_c, ζ_c, γ_d and ζ_d be the paths in ∂D labeled respectively by w_c, w_c^{-1}, w_d and w_d^{-1}. Note that no hyperplane is dual to two distinct edges of γ_c (resp. ζ_c, γ_d and ζ_d). This follows as z_c is a word in unique Λ–edges. Thus, every hyperplane dual to γ_c (resp. γ_d) is also dual to ζ_c (resp. γ_d).

Let α be a path in ∂D between γ_c and γ_d (which is labeled by w_a). Again, no hyperplane is dual to two distinct edges of α. Let xx' be a Λ–edge of T_a, and let H be the unique hyperplane dual to α with label xx'. Note that either H intersects every hyperplane dual to γ_c or H intersects every hyperplane dual to γ_d. Furthermore, every Λ–edge of T_c (resp. T_d) is the label of a hyperplane dual to γ_c (resp. γ_d). The claim now follows from Lemma 2.6, and the fact that intersecting hyperplanes correspond to commuting generators of A_Δ. $\qquad\square$

The following corollary shows that if Θ contains a configuration like that in the hypothesis of condition $\mathcal{R}_5$, then Γ must contain a triangle. This corollary is a warm-up to the more complicated Lemma 3.34.

Corollary 3.33 Suppose $(G^\Theta, E(\Lambda))$ is a RAAG system and Θ contains a set of vertices $\{a, a', b, b', c, c'\}$ satisfying the hypothesis of $\mathcal{R}_5$. Then Γ contains a triangle.

Proof Let $P = \{a, a', c, c', d, d'\}$ be a subset of vertices of Θ satisfying the hypothesis of $\mathcal{R}_5$. We call such a P a *configuration* in Θ. Keeping the same notation as in Definition 3.30, we call the number of vertices of T_a the *complexity* of P, and we prove the claim by induction on complexity. Note that $a = a'$ is possible in the hypothesis of $\mathcal{R}_5$, so the lowest possible complexity is $N = 1$. The corollary follows in this case, as Γ then contains a triangle spanned by the vertices $a = a'$, c and d.

Now let $N > 1$ and suppose the claim is true for all configurations P of smaller complexity. As $N > 1$, there is a vertex y such that ay is a Λ–edge of T_a. By Lemma 3.32, either y is adjacent in Γ to both c and c', or y is adjacent in Γ to both d and d'. In either case, we see that Θ contains a configuration of smaller complexity. $\qquad\square$
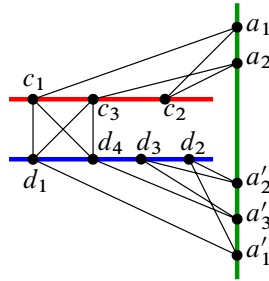
Figure 12: This figure illustrates the configuration described in Lemma 3.34 in the case $n = 3$ and $m = 4$. The black edges are edges of Γ. The red, green and blue parts consist of Λ–edges, and are all contained in Λ. The different colors indicate that they are in three distinct components of Λ.

The next lemma shows that if Θ contains certain subgraphs which generalize the configurations in the hypothesis of $\mathcal{R}_5$, then Γ must contain a triangle.

Lemma 3.34 *Let Λ_a, Λ_c and Λ_d be distinct components of Λ. Suppose Θ has a $\Lambda_a\Lambda_c$–path visiting $c_1, a_1, c_2, \ldots, a_{n-1}, c_n$, and a $\Lambda_a\Lambda_d$–path visiting $d_1, a'_1, d_2, \ldots, a'_{m-1}, d_m$, where $c_i \in \Lambda_c$, $d_i \in \Lambda_d$ and $a_i, a'_i \in \Lambda_a$ for all appropriate i. Further suppose that Θ contains a 2–component square visiting c_1, d_1, c_n and d_n. (See Figure 12.) If $(G^\Theta, E(\Lambda))$ is a RAAG system, then Γ has a triangle.*

Proof By Theorem 3.18 and Lemma 3.32, we may assume that Θ satisfies conditions $\mathcal{R}_1$–$\mathcal{R}_5$. Let $A = \{a_1, \ldots, a_{n-1}, a'_1, \ldots, a'_{m-1}\}$, $C = \{c_1, \ldots, c_n\}$ and $D = \{d_1, \ldots, d_m\}$ be vertices of Θ as in the statement of the lemma. We call such a triple (A, B, C) a configuration of Θ. Let T_a, T_c and T_d be the Λ–convex hulls of A, C and D respectively. We define the complexity of (A, C, D) to be the integer $N = |C| + |D| + |T_a|_E + |T_c|_E + |T_d|_E$, where $|X|_E$ denotes the number of edges in a graph X. The proof will be by induction on complexity of configurations.

By hypothesis, we have that $n, m \geq 2$ and $|T_c|_E, |T_d|_E \geq 1$. If $n = m = 2$ then Γ contains a triangle by Corollary 3.33. In particular, the base case follows.

We now fix a configuration $A = \{a_1, \ldots, a_{n-1}, a'_1, \ldots, a'_{m-1}\}$, $C = \{c_1, \ldots, c_n\}$ and $D = \{d_1, \ldots, d_m\}$ as above of complexity N, and we assume that the result holds for configurations of smaller complexity. By the previous paragraph, we may also assume (up to relabeling) that $n > 2$. We prove the lemma by showing that either Γ contains a triangle or Θ contains a configuration of smaller complexity.

Define α_{ac} and α_{ad} to respectively be the hypothesized $\Lambda_a\Lambda_c$–path and $\Lambda_a\Lambda_d$–path. We may assume that α_{ac} and α_{ad} are simple paths, for if not, we would be able to excise a loop to obtain a configuration of smaller complexity.

We claim that for all $1 < i < n$, we may assume that c_i does not lie on the simple path in Λ_c from c_1 to c_n. For suppose there exists such a vertex c_i. By $\mathcal{R}_3$, it follows that c_i commutes with both d_1 and d_m.

There then exists a $\Lambda_a \Lambda_c$ path visiting $c_1, a_1, \ldots, a_{i-1}, c_i$, and it follows that Θ contains a configuration of smaller complexity (obtained by replacing α_{ac} with this new path). Thus we may make this assumption without loss of generality. Furthermore, as $n \geq 3$, there exists an integer j such that c_j is a leaf vertex of T_c and such that $1 < j < m$. We fix such a vertex c_j.

Define the word z_c to be

$$z_c = (a_1' a_1)(c_1 c_2)(a_1 a_2)(c_2 c_3)(a_2 a_3) \cdots (a_{n-2} a_{n-1})(c_{n-1} c_n)(a_{n-1} a_1')$$

and define the word z_d, depending on the value of m, to be

$$z_d = \begin{cases} d_1 d_2 & \text{if } m = 2, \\ (d_1 d_2)(a_1' a_2')(d_2 d_3)(a_2' a_3') \cdots (a_{m-2}' a_{m-1}')(d_{m-1} d_m)(a_{m-1}' a_1') & \text{if } m > 2. \end{cases}$$

In W_Γ we have that $z_c \simeq a_1' c_1 c_n a_1'$ and $z_d \simeq a_1' d_1 d_m a_1'$. Let $z = [z_c, z_d]$. Note that as c_1 and c_n commute with d_1 and d_m in W_Γ,

$$z \simeq [a_1' c_1 c_n a_1', a_1' d_1 d_m a_1'] \simeq a_1' [c_1 c_n, d_1 d_m] a_1' \simeq 1.$$

Let w_c, w_d and w be the Λ–edge words associated to z_c, z_d and z respectively. Let D be a disk diagram over A_Δ with boundary label w. Let γ_c, ζ_c, γ_d and ζ_d be the subpaths of ∂D labeled respectively by w_c, w_c^{-1}, w_d and w_d^{-1}.

Let $y c_j$ be the Λ–edge of T_c incident to c_j. Since α_{ac} does not repeat vertices and since c_j is a leaf of T_c, it follows that w_c contains exactly two occurrences of the letter y contained in the subword labeled by $(y c_j)(a_{j-1} x_1)(x_1 x_2) \cdots (x_l a_j)(c_j y)$, where the x_i's are vertices in Λ_a. In particular, there are exactly four edges of ∂D (two on γ_c and two on ζ_c) labeled by either $y c_j$ or $c_j y$. Correspondingly, there are exactly two hyperplanes, H and H' in D labeled $y c_j$.

We claim that we may assume that H is dual to both γ_c and ζ_c, and the same is true for H'. For suppose otherwise, and suppose that H is dual to two edges of γ_c. (The case of H' is similar.) It follows that any hyperplane dual to the subpath of γ_c labeled by $(a_{j-1} x_1)(x_1 x_2) \cdots (x_l a_j)$ (which lies between the endpoints of H) must intersect H. Thus, in particular, $(a_{j-1} x_1)$ and $(x_l a_j)$ commute with $y c_j$, and applying Lemma 2.6, we conclude that y commutes with both a_{j-1} and a_j. We now show that we can replace α_{ac} with a new $\Lambda_a \Lambda_c$–path from c_1 to c_n such that $|T_a|_E$ is reduced, and thus Θ contains a smaller complexity configuration. If y is not equal to any c_k for any $1 \leq k \leq m$, then we obtain this path by simply replacing c_j with y in α_{ac}. On the other hand, if $y = c_k$ for some k, then we replace α_{ac} with the $\Lambda_a \Lambda_c$ path visiting $c_1, a_1, \ldots, c_k, a_j, c_{j+1}, a_{j+1}, \ldots, a_{n-1}, c_n$ if $k < j$ and perform a similar replacement if $k > j$. In either case, we have produced a configuration of smaller complexity. Thus, we now assume that each of H and H' is dual to both γ_c and ζ_c.

Let Q and Q' be the hyperplanes in D dual to the edges of γ_c labeled by $a_{j-1} x_1$ and $x_l a_j$ respectively. If both Q and Q' intersect $H \cup H'$, then we can conclude, as above, that y commutes with both a_{j-1} and a_j. We can then find a smaller complexity configuration as in the previous paragraph. Thus, we can

assume that either Q or Q' is dual to both γ_c and ζ_c. We assume that Q has this property (the case of Q' is similar).

We now examine hyperplanes dual to γ_d and ζ_d. If $m = 2$, then the unique hyperplane whose label contains d_1 is dual to both γ_d and ζ_d, and this hyperplane intersects both H and Q. Thus, d_1 commutes with both c_j and a_{j-1}. Since c_j commutes with a_{j-1}, it follows that Γ contains a triangle. On the other hand, if $m > 2$ by the same reasoning as before, we can assume there is a leaf vertex $d_{j'}$ of T_d and a hyperplane with label $y'd_{j'}$ that intersects both γ_d and ζ_d. This then implies that $d_{j'}$ commutes with both c_j and a_{j-1} and consequently, Γ contains a triangle.   □

4   Finite-index visual RAAGs

As in the previous section, given a simplicial graph Γ and a subgraph Λ of Γ^c with no isolated vertices, we set $\Theta = \Theta(\Gamma, \Lambda)$, and let G^Θ be the subgroup generated by $E(\Lambda)$. Our goal is to characterize graphs $\Lambda \subset \Gamma^c$ such that $(G^\Theta, E(\Lambda))$ is a RAAG system and G^Θ has finite index in W_Γ.

Suppose the graph Γ contains a vertex s which is Γ–adjacent to every other vertex of Γ. We say that s is a *cone vertex*. In this case, it easily follows that $W_{\Gamma \setminus s}$ has index 2 in W_Γ and that s cannot be contained in any Λ–edge.

We now recall a construction from [Dani and Levcovitz 2021] which will help us compute the index of G^Θ. The construction is general, but for simplicity, and as it is all that we use, we choose to only describe it in the context where Γ is triangle-free. We refer the reader to [Dani and Levcovitz 2021] for full details.

Let Γ be a triangle-free graph. We say a cell complex is Γ–labeled if every edge of the complex is labeled by a vertex of Γ. Let X be a Γ–labeled complex. Suppose two edges of X have the same label and a common endpoint. A *fold operation* produces a new complex from X by naturally identifying these two edges.

Suppose now that f_1 and f_2 are edges of X which share a common vertex u and whose labels $s_1, s_2 \in V(\Gamma)$ have an edge between them in Γ. Let c be a 2–cube with edges c_1, c_2, c_3 and c_4 such that $c_i \cap c_{i+1}$ is a vertex of c for each i mod 4. We label c_1 and c_3 by s_1, and c_2 and c_4 by s_2. A *square attachment operation* produces a new complex from X by attaching c to X by identifying c_1 to f_1 and c_2 to f_2. Note that, unlike in [Dani and Levcovitz 2021], we do not need to define cube attachments for higher-dimensional cubes, as we are in the case that Γ is triangle-free.

Finally, given a collection of 2–cubes in X with common boundary, we can produce a new complex from X by naturally identifying every 2–cube in this collection to a single 2–cube. In this case, we say a *cube identification operation* was performed to X.

We define a Γ–labeled complex Ω_0 associated to G^Θ as follows. First, we enumerate the Λ–edges as $s_1 t_1, \ldots, s_n t_n$, where s_i and t_i are the two endpoints of the ith Λ–edge. We set Ω_0 to be a bouquet of n circles, each of which is subdivided into two edges, such that the ith circle has label $s_i t_i$.

Next, we describe a series of complexes built iteratively from Ω_0. These are

$$\Omega_0 \to \Omega_1 \to \Omega_2 \to \cdots .$$

For each $i > 0$, the complex Ω_i is obtained by either a fold, square attachment or square identification operation performed to Ω_{i-1}. Furthermore, we assume that the order of operations is as follows: first all possible fold and square identifications are performed, then all possible square attachment operations are applied to the resulting complex, and these processes are alternated (see [Dani and Levcovitz 2021] for details).

Let Ω be the direct limit of such a sequence. We call Ω a *completion* of G^Θ. In [Dani and Levcovitz 2021] we show that properties of Ω reflect those of the subgroup G^Θ.

The index of G^Θ can be determined by properties of Ω. We say that a vertex u of a Γ–labeled complex has *full valence* if for any vertex $s \in \Gamma$, there is an edge incident to u with label s. Below we present a version of [Dani and Levcovitz 2021, Theorem 6.9] together with [Dani and Levcovitz 2021, Lemma 6.8] under the hypotheses which we will need:

Theorem 4.1 *Let Γ be a triangle-free graph with no cone vertex. A subgroup $G < W_\Gamma$ has finite index in W_Γ if and only if Ω is finite and every vertex of Ω has full valence. Furthermore, if G is indeed of finite index, then its index is exactly the number of vertices of Ω.*

We introduce two new properties below which will help us characterize when G^Θ has finite index in W_Γ.

Definition 4.2 (conditions $\mathscr{F}_1$ and $\mathscr{F}_2$) We say that $\Theta = \Theta(\Gamma, \Lambda)$ satisfies *condition $\mathscr{F}_1$* if given any $s \in V(\Theta)$ which is not a cone vertex of Γ, it follows that s is the endpoint of some Λ–edge. We say that Θ satisfies *condition $\mathscr{F}_2$* if given any distinct components Λ_s and Λ_t of Λ, and vertices s of Λ_s and t of Λ_t, there is a $\Lambda_s \Lambda_t$–path in Θ from s to t.

Remark 4.3 Suppose Γ is connected, Λ contains exactly two components and that $\Theta = \Theta(\Gamma, \Lambda)$ satisfies $\mathscr{R}_2$ and $\mathscr{F}_1$. Then Θ satisfies $\mathscr{F}_2$. For given any two vertices contained in different components of Λ, as Γ is connected, there is a Γ–path between them. Furthermore, this has to be a 2–component path as Θ satisfies $\mathscr{R}_2$, and the two Λ–components this path visits have to be the ones containing the chosen vertices (as there are only two Λ components). This remark will prove to be useful when verifying whether certain graphs satisfy $\mathscr{F}_2$.

Remark 4.4 Suppose $\Theta = \Theta(\Gamma, \Lambda)$ satisfies $\mathscr{F}_2$, and let Λ_1 and Λ_2 be distinct Λ–components. Then there exists an $\Lambda_1 \Lambda_2$–path between any two distinct vertices of Λ_1. To see this, let s and s' be distinct

vertices of Λ_1, and let t be vertex of Λ_2. By $\mathscr{F}_2$ there is a $\Lambda_1\Lambda_2$–path from s to t, and similarly there is a $\Lambda_1\Lambda_2$–path from t to s'. Combining these two paths gives a $\Lambda_1\Lambda_2$–path form s to s'.

Lemma 4.5 *Let Γ be a triangle-free graph with no cone vertex, and let Λ be a subgraph of Γ^c with no isolated vertices, such that $(G^\Theta, E(\Lambda))$ is a RAAG system. If Λ has at most $k \leq 2$ components and Θ satisfies $\mathscr{F}_1$ and $\mathscr{F}_2$, then G^Θ is of index 2^k in W_Γ.*

We remark that this proof readily generalizes to the case of arbitrary k. However, we only need the case $k \leq 2$.

Proof Let Ω_0 be the Γ–labeled complex defined above, and let Ω' be the complex obtained from Ω_0 by all possible fold operations.

Suppose first that Λ has one component. As Λ is connected, it is easily seen that Ω' consists of two vertices with an edge labeled by s between them for $s \in V(\Lambda)$. As Λ satisfies $\mathscr{R}_2$ by Proposition 3.17, no two vertices of Λ have an edge between them in Γ. Thus, no square attachments can be performed to Ω', and it follows that $\Omega = \Omega'$. Hence, Ω is finite and has exactly two vertices.

Note that by the description of $\Omega = \Omega'$ above, every vertex of Ω is adjacent to every edge of Ω. Also note that by condition $\mathscr{F}_1$, for every vertex $s \in \Gamma$ there is some edge in Ω labeled by s. From these two facts we deduce that every vertex of Ω has full valence. Thus, G^Θ has index 2 in W_Γ by Theorem 4.1.

Now suppose that Λ has two components Λ_1 and Λ_2. In this case, Ω' is readily seen to be a complex consisting of three vertices, u, v_1 and v_2, with an edge from u to v_i labeled s corresponding to each vertex s of Λ_i, for $i = 1, 2$. By condition $\mathscr{F}_1$, the vertex u has full valence. Furthermore, by $\mathscr{R}_2$, for each $i \in \{1, 2\}$, no two edges of Ω' that are each adjacent to both v_i and u have labels which are adjacent in Γ.

Let Ω'' be the complex obtained from Ω' by performing all possible square attachment operations to Ω', and let Ω''' be the complex obtained from Ω'' by all possible fold and square identification operations. In particular, $\Omega'' = \Omega_l$ and $\Omega''' = \Omega_k$ for some $0 \leq l \leq k$. Let s and s' be distinct vertices of Λ_1, and let t be any vertex of Λ_2. By condition $\mathscr{F}_2$, there is a $\Lambda_1\Lambda_2$–path whose vertices are $s, t_1, s_1, t_2, s_2, \ldots, t_m, s_m, t$ where $s_i \in \Lambda_1$ and $t_i \in \Lambda_2$ for all $1 \leq i \leq m$. Similarly, there is a $\Lambda_1\Lambda_2$–path whose vertices are $s', t'_1, s'_1, t'_2, s'_2, \ldots, t'_n, s'_n, t$ where $s'_i \in \Lambda_1$ and $t'_i \in \Lambda_2$ for all $1 \leq i \leq n$. Thus, Ω'' must contain length two paths, which do not intersect u, from v_1 to v_2 with each of the labels

$$t_1 s, t_1 s_1, t_2 s_1, t_2 s_2, \ldots, t_m s_{m-1}, t_m s_m, t s_m,$$

and similarly length two paths, which do not intersect u, from v_1 to v_2 with each of the labels

$$t'_1 s', t'_1 s'_1, t'_2 s'_1, t'_2 s'_2, \ldots, t'_n s'_{m-1}, t'_m s'_m, t' s_m.$$

It follows that the middle vertices of all these paths get folded to a single vertex v_3 in Ω'''. This analysis can be done for any $s, s' \in \Lambda_1$. Similar paths can also be produced for any $t, t' \in \Lambda_2$. It then follows that

Ω''' consists of exactly 4 vertices: u, v_1, v_2 and v_3. Furthermore, there is an edge with label s between v_1 and v_3 for each $s \in \Lambda_1$, and there is an edge with label t between v_2 and v_3 for each $t \in \Lambda_2$. Thus, every vertex of Ω''' can be seen to have full valence. Additionally, by condition $\mathscr{R}_2$, no additional square attachment operations can be performed to Ω'''. Hence, $\Omega = \Omega'''$. It follows that G^Θ has index exactly four in W_Γ. $\qquad \square$

The next lemma shows that $\mathscr{F}_1$ and $\mathscr{F}_2$ are necessary conditions for G^Θ to have finite index.

Lemma 4.6 *Let Γ be a triangle-free graph with no cone vertex, and let Λ be a subgraph of Γ^c with no isolated vertices, such that $(G^\Theta, E(\Lambda))$ is a RAAG system. If $G = G^\Theta$ is of finite index in W_Γ, then Θ satisfies $\mathscr{F}_1$ and $\mathscr{F}_2$.*

Proof We first check that condition $\mathscr{F}_1$ holds. Let Ω be a completion of $G := G^\Theta$ as described in the beginning of this section. Theorem 4.1 implies in particular that given any vertex $s \in \Gamma$ there is an edge of Ω with label s. This implies the vertex s is contained in some Λ–edge. Thus, $\mathscr{F}_1$ must hold.

We now check condition $\mathscr{F}_2$. Let $s \in \Lambda_s$ and $t \in \Lambda_t$ be as in the definition of condition $\mathscr{F}_2$ (Definition 4.2). If s commutes with t, then there is an edge in Γ between s and t, and we are done. So we may assume that s and t do not commute.

As G is of finite index, it follows that there exist $g_1, \ldots, g_n \in W_\Gamma$ such that $W_\Gamma = G g_1 \sqcup G g_2 \sqcup \cdots \sqcup G g_n$. Let $w_1, \ldots, w_n$ be reduced words representing $g_1, \ldots, g_n$, and let $K = \max\{|w_1|, \ldots, |w_n|\}$. Define the word $h = s_1 t_1 s_2 t_2 \cdots s_{K+4} t_{K+4}$ where $s_i = s$ and $t_i = t$ for all $1 \le i \le K + 4$. It readily follows from Tits' solution to the word problem (see Theorem 2.4) that h is reduced. Furthermore, we can write $h \simeq w w'$, where w and w' are words in W_Γ such that $w' = w_i$ for some $1 \le i \le n$ and w is a product of Λ–edges representing an element of G. We can form a disk diagram in W_Γ with boundary label $h w'^{-1} w^{-1}$. Let α_h, α_w and $\alpha_{w'}$ respectively be the corresponding paths along the boundary of D with labels respectively h, w and w^{-1}.

Note that as h is reduced, no hyperplane intersects α_h twice. Also note that any pair of hyperplanes emanating from α_h cannot intersect as s and t do not commute. As $|h| > |w'| + 4$, it follows that the hyperplanes H_{s_1}, H_{t_1}, H_{s_2} and H_{t_2}, dual respectively to the first four edges of α_h (namely those labeled by s_1, t_1, s_2 and t_2), must each intersect α_w. It must now be the case that there exists a chain of hyperplanes (see Definition 3.20) $H_{s_1} = H_0, H_1, \ldots, H_m = H_{s_2}$ and another chain of hyperplanes $H_{t_1} = H'_0, H'_1, \ldots, H'_n = H_{t_1}$. These two chains intersect, and by reasoning similar to that in Observation 3.24, it follows that there is a $\Lambda_s \Lambda_t$–path from s to t. $\qquad \square$

Lemma 4.7 *Let Γ be a triangle-free graph. Let Λ be a subgraph of Γ^c with no isolated vertices, such that $(G^\Theta, E(\Lambda))$ is a RAAG system and G^Θ has finite index in W_Γ. If Γ contains a cone vertex, then Λ contains exactly one component. If W_Γ is not virtually free, then Λ contains exactly two components. Otherwise, Λ contains at most two components.*

Proof Suppose first that Γ contains a cone vertex $s \in \Gamma$. We may assume that Γ does not consist of a single edge, as Λ would be empty in that case. As Γ is triangle-free in addition, there can be at most one cone vertex. Since Γ is triangle-free, it follows that $\Gamma' = \Gamma \setminus s$ is a graph with no edges and is therefore virtually free. Furthermore, every Λ–edge is contained in Γ', and G^Θ is a finite-index subgroup of $W_{\Gamma'}$. By Lemma 4.6, we conclude that $\Theta' = \Theta(\Gamma', \Lambda)$ satisfies condition $\mathscr{F}_2$. In particular, there is a Γ'–edge between any two Λ components. As Γ' does not have any edges, Λ has exactly one component and the claim follows in this case.

We now assume that Γ does not contain a cone vertex. Furthermore, by Lemma 4.6 we may assume that $\Theta = \Theta(\Gamma, \Lambda)$ satisfies $\mathscr{F}_1$ and $\mathscr{F}_2$, and that Θ satisfies $\mathscr{R}_1$–$\mathscr{R}_4$ by Proposition 3.17.

Suppose now that no two distinct Λ–edges commute. It follows that G^Θ is isomorphic to a free group, and since G^Θ is of finite index, W_Γ is virtually free. Suppose, for a contradiction, that Λ has three distinct components Λ_1, Λ_2 and Λ_3. Let s and t be distinct vertices of Λ_1. By Remark 4.4 there is a $\Lambda_1\Lambda_2$–path α_1 from s to t which we can assume does not repeat vertices. Similarly, there is a $\Lambda_1\Lambda_3$–path α_2 from s to t which does not repeat vertices. Observe that $s, t \in \alpha_1 \cap \alpha_2$. Starting at s and traveling along α_1, let x be the first vertex after s such that $x \in \alpha_1 \cap \alpha_2$. Then the subpath α_1' of α between s and x contains exactly two vertices of $\alpha_1 \cap \alpha_2$. Let α_2' be the subpath of α_2 between s and x. Note that $|\alpha_1'|, |\alpha_2'| \geq 2$, as every other vertex of α_1 is in Λ_2 and $\alpha_2 \cap \Lambda_2 = \varnothing$. It follows that $c = \alpha_1' \cup \alpha_2'$ is a cycle in Γ. Let c' be a subcycle of c which is an induced subgraph of Γ. If c' has three vertices, then this contradicts Γ being triangle-free. On the other hand, if c' has more than three vertices, then this contradicts W_Γ being virtually free. Thus, Λ can have at most two components and the claim follows in this case.

Suppose now there exist Λ–edges a_1a_2 and b_1b_2 which commute, with $a_1a_2 \neq (b_1b_2)^{\pm 1}$. These Λ–edges must be in different components of Λ by condition $\mathscr{R}_2$ and Lemma 2.6. In this case, W_Γ is not virtually free as it contains a subgroup isomorphic to $\mathbb{Z}^2$. Suppose, for a contradiction, that Λ contains at least three distinct Λ–edge components Λ_1, Λ_2 and Λ_3. Without loss of generality, we may assume that $a_1b_1 \in \Lambda_1$ and that $a_2b_2 \in \Lambda_2$. We will obtain a contradiction by showing that Γ must contain a triangle.

By Lemma 2.6, a_1, a_2, b_1 and b_2 form a square in Γ. By Remark 4.4, there is a $\Lambda_1\Lambda_3$–path from a_1 to a_2. Similarly, there is a $\Lambda_2\Lambda_3$–path from b_1 to b_2. Thus, Γ contains the configuration described in the statement of Lemma 3.34. That lemma then implies that Γ contains a triangle, a contradiction. $\square$

Theorem 4.8 *Let W_Γ be a 2–dimensional RACG. Let Λ be a subgraph of Γ^c with no isolated vertices, and let G^Θ be the subgroup of W_Γ generated by the Λ–edges. Then the following are equivalent:*

(1) *$(G^\Theta, E(\Lambda))$ is a RAAG system and G^Θ has finite index in W_Γ.*

(2) *$(G^\Theta, E(\Lambda))$ is a RAAG system and G^Θ has index either two or four in W_Γ (and exactly four if W_Γ is not virtually free).*

(3) *Λ has at most two components and Θ satisfies conditions $\mathscr{R}_1$–$\mathscr{R}_4$, $\mathscr{F}_1$ and $\mathscr{F}_2$.*

Proof Clearly (2) implies (1). To see the remaining implications, suppose first that Γ contains a cone vertex s. Then $\Gamma' = \Gamma \setminus s$ is a graph with no edges, and $W_{\Gamma'}$ is an index two subgroup of Γ. Suppose that (1) holds. By Lemma 4.7, Λ has exactly one component. By Theorem 3.18 and Lemma 4.6, $\Theta' = \Theta(\Gamma', \Lambda)$ satisfies conditions $\mathscr{R}_1$–$\mathscr{R}_4$, $\mathscr{F}_1$ and $\mathscr{F}_2$. Consequently, Θ satisfies these conditions as well. Thus (3) holds. By Lemma 4.5, we know that $(G^{\Theta'}, E(\Lambda))$ is a RAAG system of index 2 in $W_{\Gamma'}$, and thus $(G^\Theta, E(\Lambda))$ is a RAAG system of index four in W_Γ. Therefore (2) holds. Finally, if (3) holds then (1) holds by Theorem 3.18 and Lemma 4.5.

Now suppose that Γ does not have a cone vertex. If (1) holds, then by Lemma 4.7, Λ has exactly two components if W_Γ is not virtually free and at most two components otherwise. Thus (2) holds by Lemma 4.5. By Theorem 3.18 and Lemma 4.6, (3) holds. Finally if (3) holds, then (1) follows by Theorem 3.18 and Lemma 4.5. $\qquad\square$

Corollary 4.9 *Let W_Γ be a 2–dimensional RACG. Let Λ be a subgraph of Γ^c with no isolated vertices such that the subgroup $(G, E(\Lambda))$ is a finite-index RAAG system. Then either*:

(1) *The graph Γ does not contain any edges and $E(\Lambda)$ is a spanning tree in Γ^c. In particular, W_Γ is virtually free.*

(2) *The group W_Γ is not virtually free. Furthermore, the vertices of Γ can be 2–colored by red and blue (ie each edge of Γ connects a red vertex and a blue vertex) and G is isomorphic to the kernel of the homomorphism $\Psi \colon W_\Gamma \to \mathbb{Z}_2 \times \mathbb{Z}_2 = \langle r, b \mid r^2 = b^2 = 1 \rangle$ which maps red and blue generators of $V(\Gamma)$ to r and b respectively.*

Proof By Theorem 4.8, Λ has at most two components. Suppose first that Λ contains exactly one component. Again by Theorem 4.8, the graph Θ satisfies $\mathscr{R}_1$, $\mathscr{R}_2$, and $\mathscr{F}_1$. From these conditions, it follows that Γ cannot contain any edges and that $E(\Lambda)$ is a spanning tree in Γ^c. As Γ does not contain any edges, W_Γ is virtually free.

Suppose now that Λ has exactly two components. We color the vertices of one component red and the vertices of the other component blue. By $\mathscr{R}_2$, each edge of Γ connects a red vertex and a blue vertex, ie we have a 2–coloring of Γ. Furthermore, by the definition of Ψ, every Λ–edge (thought of as an element of G) is in the kernel of Ψ. As G is generated by such elements, it follows that $G < \ker(\Psi)$. By Theorem 4.8, G has index 4 in W_Γ. As $\ker(\Psi)$ has index 4 as well, it follows that G is isomorphic to $\ker(\Psi)$. $\qquad\square$

5 Applications

In this section we give concrete families of RACGs containing finite-index RAAG subgroups. These cannot be obtained by applying the Davis–Januszkiewicz constructions to the defining graphs of the RAAGs they are commensurable to.
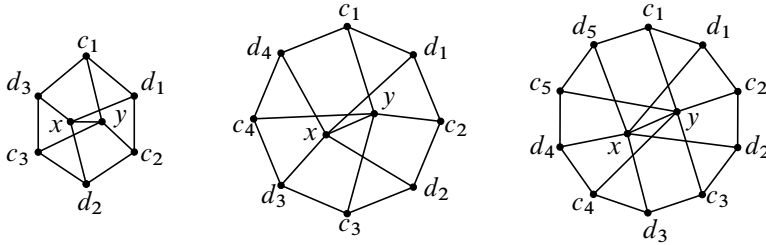
Figure 13: The figure illustrates the graphs Γ_n defined in Corollary 5.1 for $n = 3, 4, 5$.

5.1 Nonplanar RACGs commensurable to RAAGs

In this subsection, we construct two families of RACGs with nonplanar defining graphs containing finite-index RAAG subgroups. These will serve as a warm-up for Theorem 5.5.

We begin by constructing a family of quasi-isometrically distinct RACGs defined by the sequence of graphs Γ_n (shown in Figure 13) which are commensurable to RAAGs whose defining graphs are cycles.

Corollary 5.1 (to Theorem 4.8) *For $n \geq 3$, let Γ_n be the graph obtained by starting with a $2n$–gon whose vertices (in cycle order) are $c_1, d_1, c_2, d_2, \ldots, c_n, d_n$ and adding two vertices x and y, such that y is adjacent to c_i for each i, x is adjacent to d_i for each i, and x is adjacent to y (see Figure 13). Then*

(1) *the RACG W_{Γ_n} has a subgroup of index four that is isomorphic to (and hence is commensurable to) the RAAG $A_{C_{2n}}$, where C_{2n} is a cycle of length $2n$;*

(2) *W_{Γ_n} is not quasi-isometric to W_{Γ_m} for $m \neq n$.*

Proof Fix $n \geq 3$, and let Γ denote Γ_n. We define a graph $\Lambda \subset \Gamma^c$ as follows. Let Λ_x be the star graph consisting of the union of the edges of Γ^c from x to c_i for each i. Let Λ_y be the star graph consisting of the edges of Γ^c from y to d_i for each i. Let $\Lambda = \Lambda_x \cup \Lambda_y$. (See Figure 4 for an illustration of Λ in the case $n = 3$.)

We show below that $\Theta = \Theta(\Gamma, \Lambda)$ satisfies $\mathcal{R}_1$–$\mathcal{R}_4$, $\mathcal{F}_1$ and $\mathcal{F}_2$. Then it will follow from Theorem 4.8, that $(G^\Theta, E(\Lambda))$ is a RAAG system, and that G^Θ has index four in W_Γ. Moreover, it is easily checked that the commuting graph Δ associated to Λ (as defined in Section 2.2) is isomorphic to C_{2n}. Consequently, G^Θ is isomorphic to $A_{C_{2n}}$. Thus, this will show (1).

It is easy to verify $\mathcal{F}_1$, $\mathcal{R}_1$, and $\mathcal{R}_2$. Then by Remark 4.3, it follows that $\mathcal{F}_2$ holds as well. We now check $\mathcal{R}_3$. First note that there are exactly three squares in Γ containing the edge $c_1 d_1$, and each of these satisfies the property in $\mathcal{R}_3$. Now the fact that every square contains an edge of the $2n$–gon, together with the symmetry of the diagram, implies that $\mathcal{R}_3$ holds.

To check $\mathcal{R}_4$, let γ be a $\Lambda_x \Lambda_y$–cycle and let e be an edge of γ. By symmetry, we can assume that e is either $c_1 d_1$, $c_1 y$ or xy. Suppose first that $e = c_1 d_1$. Then γ contains either $d_n c_1$ or yc_1. In both cases,
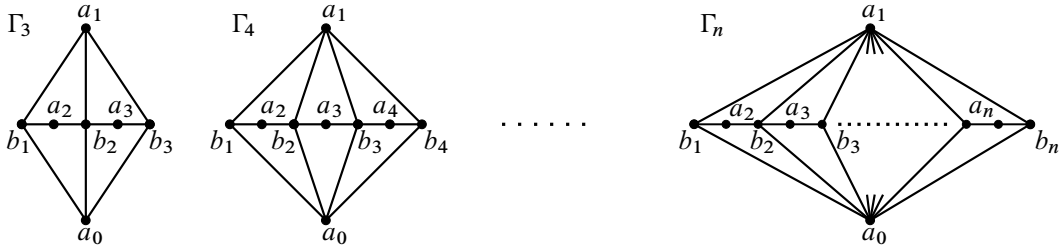
Figure 14: The figure defines the family of graphs Γ_n, for $n \geq 3$, used in Corollary 5.2.

the Λ–convex hull of the vertices of γ contains y. Similarly, γ contains either $d_1 x$ or $d_1 c_2$, and in both cases the Λ–convex hull of γ contains x. As $c_1 d_1$ is contained in the square $c_1 d_1 x y$, and x and y are in the appropriate convex hulls, it follows that $\mathcal{R}_4$ holds for the Γ–cycle γ and edge e.

Suppose now that $e = c_1 y$. It follows that γ contains either yx or yc_i for some $i > 1$. In each case, x is in the Λ–convex hull of γ. Furthermore, γ contains either $c_1 d_1$ or $c_1 d_n$. In the former case, the square $c_1 d_1 x y$ contains e and has vertices in the Λ–convex hull of the vertices in γ. In the latter case the same argument applies to the square $c_1 d_n x y$.

Finally, suppose that $e = xy$. By symmetry, we may assume that γ contains yc_1. Furthermore, yc_1 must be followed by either $c_1 d_2$ or $c_1 d_n$ in γ. Then, as in the previous paragraph, either the square $c_1 d_1 x y$ or the square $c_1 d_n x y$ contains e and has vertices in the Λ–convex hull of γ. Thus $\mathcal{R}_4$ is satisfied in all cases.

We have thus established that (1) holds, by showing that Θ satisfies $\mathcal{R}_1$–$\mathcal{R}_4$, $\mathcal{F}_1$ and $\mathcal{F}_2$. Consequently, for each n, we know that W_{Γ_n} is commensurable, and in particular quasi-isometric, to $A_{C_{2n}}$. Claim (2) then follows from [Bestvina et al. 2008].   □

Next, we give a family of RACGs whose defining graphs are not planar and are commensurable to RAAGs which are not atomic (as defined in [Bestvina et al. 2008]).

Corollary 5.2 *Given $n \geq 3$ and $k \geq 1$, let Δ_{nk} be the graph obtained by taking k copies of Γ_n (defined in Figure 14), and identifying them all along the subgraph induced by $V(\Gamma_n) \setminus \{a_0\}$. Thus Δ_{nk} has vertices $a_1, a_2, \ldots, a_n, b_1, \ldots, b_n$ and also $a_{01}, \ldots, a_{0k}$. (The left side of Figure 15 shows Δ_{42}.) Then $W_{\Delta_{nk}}$ contains an index four subgroup isomorphic to a RAAG.*

Proof Fix $n \geq 3, k \geq 1$ and let $\Delta = \Delta_{nk}$. We define Λ, a subgraph of Δ^c consisting of two components. The first component Λ_a is the union of the edges of Δ^c of the form $a_1 a_i$, where $2 \leq i \leq n$ and $a_1 a_{0j}$ for $1 \leq j \leq k$. The second component Λ_b is the path in Δ^c visiting $b_1, b_2, \ldots, b_n$. (See the right side of Figure 15 for an illustration of the case $n = 4$ and $k = 2$).

Let $\Theta = \Theta(\Delta, \Lambda)$ be as in the previous sections. We verify the properties $\mathcal{R}_1$–$\mathcal{R}_4$, $\mathcal{F}_1$ and $\mathcal{F}_2$. It will then follow from Theorem 4.8 that the subgroup generated by $E(\Lambda)$ is an index four visual RAAG subgroup.
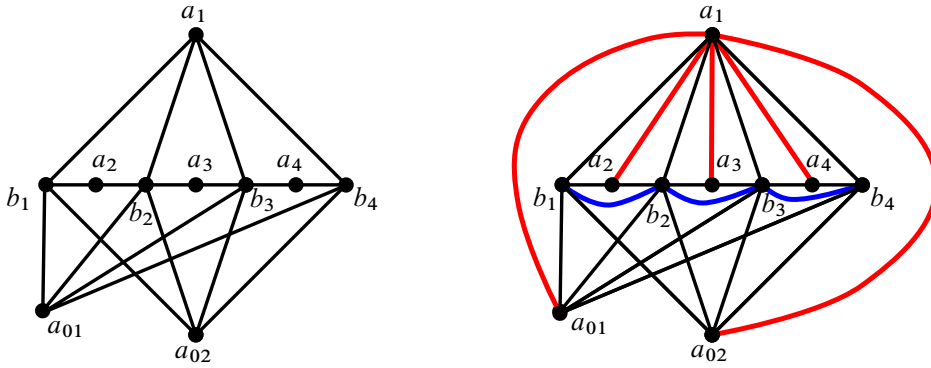
Figure 15: The figure shows Δ_{42} on the left, and the two components of Λ for the graph Δ_{42} on the right. The component Λ_a is shown in red and the component Λ_b is shown in blue.

The conditions $\mathscr{R}_1$, $\mathscr{R}_2$ and $\mathscr{F}_1$ are immediate. Condition $\mathscr{F}_2$ holds by Remark 4.3. We now check $\mathscr{R}_3$. Each square in Δ is of one of the following forms:

(1) $b_i a_{i+1} b_{i+1} a_1$ for $1 \le i \le n-1$;

(2) $b_i a_{i+1} b_{i+1} a_{0j}$ for $1 \le i \le n-1$ and $1 \le j \le k$;

(3) $b_i a_1 b_{i'} a_{0j}$ for $1 \le i < i' \le n$ and $1 \le j \le k$;

(4) $b_i a_{0j} b_{i'} a_{0j'}$ for $1 \le i < i' \le n$ and $1 \le j \le k$.

Condition $\mathscr{R}_3$ follows immediately for the first type of square, as the appropriate Λ–convex hulls do not contain any additional vertices not included in the vertex set of the square. For the second type of square, the convex hull in Λ_b does not contain any additional vertices, but the convex hull in Λ_a contains the additional vertex a_1, as this vertex lies on the Λ_a–path between a_{i+1} and a_{0j}. Since a_1 is adjacent to b_i and b_{i+1}, the condition $\mathscr{R}_3$ is verified for this type of 2–component square. For the third type, the Λ–convex hull of $\{a_1, a_{0j}\}$ does not contain any additional vertices of Λ, and the Λ–convex hull of $\{b_i, b_{i'}\}$ contains the additional vertices $b_{i+1}, \ldots, b_{i'-1}$. Since a_1 and a_{0j} are adjacent to each of these, $\mathscr{R}_3$ is verified for this type of 2–component square as well. Finally, for the last type, the Λ–convex hull of $\{a_{0j}, a_{0j'}\}$ contains the additional vertex a_1, and the Λ–convex hull of $\{b_i, b_{i'}\}$ contains the additional vertices $b_{i+1}, \ldots, b_{i'-1}$. Once again, it is easily verified that $a_0, a_{1j}, a_{1j'}$ are each adjacent to each of $b_i, \ldots, b_{i'}$. Thus $\mathscr{R}_3$ is verified.

Finally, we check $\mathscr{R}_4$. Let γ be a $\Lambda_a \Lambda_b$–cycle and let e be an edge of γ. First suppose e is of the form $a_i b_i$ for $2 \le i \le n$. In this case, γ necessarily passes through b_{i-1} and some a, where either $a = a_1$ or $a = a_{0j}$, for some $1 \le j \le k$. Thus the $\Lambda_a \Lambda_b$–square $b_{i-1} a_i b_i a$ satisfies the criterion in $\mathscr{R}_4$, since it contains e, and the two vertices a and b_{i-1} are contained in the Λ–convex hull of the vertices of γ. The case where e is of the form $a_i b_{i-1}$ for $2 \le i \le n$ is similar.

Now suppose e is of the form $a_1 b_i$ for some $1 \le i \le n$. Then γ necessarily passes through an edge of one of the following forms: $b_i a_{0j}$ for some $1 \le j \le k$, $b_i a_i$ or $b_i a_{i+1}$. In the first of these cases, γ

necessarily also passes through a vertex $b_{i'}$ for some $i' \neq i$, and $a_1 b_i a_{0j} b_{i'}$ is the desired square. If the edge is of the form $b_i a_i$ (resp. $b_i a_{i+1}$) then γ must also pass through b_{i-1} (resp. b_{i+1}), and the desired square is $a_1 b_i a_i b_{i-1}$ (resp. $a_1 b_i a_{i+1} b_{i+1}$). The case where e is of the form $a_{0j} b_i$ for some $1 \leq i \leq n$ and $1 \leq j \leq k$ is similar. This completes the verification of $\mathcal{R}_4$, and the corollary follows. $\qquad \square$

Remark 5.3 The RAAGs obtained in the above corollary do not have a tree for defining graph when $k \geq 2$ and $n \geq 3$. This is easy to check by computing the associated commuting graph.

5.2 2–Dimensional RACGs with planar defining graph

Nguyen and Tran [2019] characterized exactly which one-ended, 2–dimensional RACGs defined by planar nonjoin, CFS graphs are quasi-isometric to RAAGs. In this subsection, we use their work in conjunction with Theorem 4.8 to prove Theorem B from the introduction. Note that CFS is a graph-theoretic condition introduced in [Dani and Thomas 2015] to characterize RACGs with at most quadratic divergence. We omit the definition, as it is not needed here.

Remark 5.4 Any one-ended, 2–dimensional RACG that is quasi-isometric to a RAAG must have CFS defining graph. This follows as one-ended RAAGs have either linear or quadratic divergence [Behrstock and Charney 2012], and the defining graph of a 2–dimensional RACG with linear or quadratic divergence is CFS [Dani and Thomas 2015].

Recall that a graph Σ is a *suspension* if Σ decomposes as a join $\Sigma = \{a_1, a_2\} \star B$ where a_1 and a_2 are nonadjacent vertices. We also say that Σ is the *suspension* of the graph B. We use the notation $\Sigma_k(a, b)$ to denote the suspension graph $\{a_1, a_2\} \star \{b_1, \ldots, b_k\}$, and we say that a_1 and a_2 are the *suspension vertices*.

Let Γ be a graph which is connected, triangle-free, CFS and planar. Suppose that a planar embedding from Γ into the sphere $\mathbb{S}^2$ is fixed. Nguyen and Tran [2019] constructed a tree T (this is the *visual decomposition tree* of Section 3 of that paper) associated to Γ with the following properties. The vertices of T are in bijection with maximal suspension subgraphs of Γ. As Γ is triangle-free, every maximal suspension of Γ is of the form $\Sigma_k(a, b)$, where both $\{a_1, a_2\}$ and $\{b_1, \ldots, b_k\}$ are each sets of disjoint vertices of Γ, and $k \geq 3$ if T contains at least two vertices. Moreover, every vertex of Γ is contained in some suspension corresponding to a vertex of T. Two vertices of T corresponding to suspensions $\Sigma = \Sigma_k(a, b)$ and $\Sigma' = \Sigma_l(c, d)$ are connected by an edge if $\Sigma \cap \Sigma'$ is a 4–cycle C which separates $\mathbb{S}^2$ into two nontrivial components B_1 and B_2, such that $\Sigma_1 \setminus C \subset B_1$ and $\Sigma_2 \setminus C \subset B_2$. Moreover, it must follow (by the maximality of the suspensions) that $C = \{a_1, c_1, a_2, c_2\}$, ie C contains exactly the suspension vertices of Σ and Σ'.

If Γ (with the above assumptions) is a join, then it readily follows that Γ is quasi-isometric to a RAAG whose defining graph is a tree of diameter at most 2. Nguyen and Tran [2019, Theorem 1.2] showed that

if Γ is not a join, then W_Γ is quasi-isometric to a RAAG if and only if every vertex $v \in T$ has valence strictly less than k, where $\Sigma_k(a, b)$ is the maximal suspension in Γ corresponding to v. Moreover, they showed that such RAAGs have defining graph a tree of diameter at least 3. Below, we prove such RACGs are in fact commensurable to RAAGs.

Theorem 5.5 *Let W_Γ be a 2–dimensional, one-ended RACG with planar defining graph Γ. Then W_Γ is quasi-isometric to a RAAG if and only if it contains an index 4 subgroup isomorphic to a RAAG.*

Proof One direction of the theorem is obvious. Thus, we prove that if W_Γ satisfies these hypotheses and is quasi-isometric to a RAAG, then W_Γ contains an index 4 subgroup isomorphic to a RAAG. We do this by constructing a subgraph $\Lambda \subset \Gamma^c$ with two components and satisfying the hypotheses of Theorem 4.8.

Fix a planar embedding of Γ into the sphere $\mathbb{S}^2$. Note that by Remark 5.4 and the hypotheses of the theorem, it follows that Γ is triangle-free, CFS and planar. Thus, there exists a visual decomposition tree T associated to Γ as described above. Furthermore, as W_Γ is quasi-isometric to a RAAG, it follows from [Nguyen and Tran 2019, Theorem 1.2] that the valence of a vertex of T corresponding to the maximal suspension $\Sigma_k(a, b)$ is less than k.

Henceforth, to simplify notation, the word suspension will always refer to a maximal suspension, and will consequently correspond to a vertex of T. Given a suspension $\Sigma_k(a, b) = \{a_1, a_2\} \star B$ we say that a labeling $\{b_1, \ldots, b_k\}$ of the vertices of B is *cyclic* if the following holds. If C is a 4–cycle spanning the vertices $\{a_1, b_i, a_2, b_{i+1}\}$ for some $1 \le i \le k$ or spanning the vertices $\{a_1, b_1, a_2, b_k\}$, then every vertex of $\Sigma \setminus C$ is contained in a common component of $\mathbb{S}^2 \setminus C$. Observe that if E is a cycle corresponding to an edge of T incident to the vertex of T given by $\Sigma_k(a, b)$, then the planarity of Γ implies that E is one of the cycles C mentioned in the previous sentence.

Let N be the number of vertices of T. Let $T_1 \subset \cdots \subset T_N = T$ be a nested sequence of subtrees of T such that T_1 consists of a single vertex of T and T_i has exactly i vertices. Such choices are clearly possible. For each $1 \le i \le n$, let Γ_i be the subgraph of Γ spanned by every suspension that corresponds to a vertex of T_i. Note that $\Gamma_i \subset \Gamma_{i+1}$ for all $1 \le i < N$ and that $\Gamma_N = \Gamma$. We define a nested sequence of graphs $\Lambda_1 \subset \cdots \subset \Lambda_N$ such that for each $1 \le i \le N$, $\Lambda_i \subset \Gamma_i^c$ and the following hold:

(1) Let C be a 4–cycle corresponding to an edge of T that is incident to T_i. Then each pair of nonadjacent vertices in C is contained in a common edge of Λ_i.

(2) The graph Λ_i contains exactly two components, and $\Theta_i = \Theta(\Gamma_i, \Lambda_i)$ satisfies conditions $\mathcal{R}_1$–$\mathcal{R}_4$, $\mathcal{F}_1$ and $\mathcal{F}_2$.

The theorem clearly follows from this claim by using the graph $\Lambda = \Lambda_N \subset \Gamma^c$.

We first define Λ_1 corresponding to the vertex $T_1 = \{v\}$. Let $\Sigma = \Sigma_k(a, b) = \{a_1, a_2\} \star \{b_1, \ldots, b_k\}$ be the suspension corresponding to v, and assume that $\{b_1, \ldots, b_k\}$ is cyclic. As the valence of v in T is less than k, by possibly relabeling, we can assume that the 4–cycle $\{a_1, b_1, a_2, b_k\}$ does not correspond

to an edge of T. We define one component of Λ_1 to be the edge (a_1, a_2), and the other component of Λ_1 to consist of the edges $(b_1, b_2), (b_2, b_3), \ldots, (b_{k-1}, b_k)$. By the observation above and our choice of labeling, any 4–cycle C corresponding to an edge of T incident to v is of the form $\{a_1, b_i, a_2, b_{i+1}\}$ for some $1 \leq i \leq k-1$. Thus condition (1) follows. Condition (2) is readily verified.

Suppose now that we have defined the graph Λ_{n-1} corresponding to the tree T_{n-1} satisfying conditions (1) and (2). We now define Λ_n.

Let u be the unique vertex in $T_n \setminus T_{n-1}$, and let u' be the unique vertex of T_{n-1} that is adjacent to u. Let $\Sigma = \Sigma_k(a, b) = \{a_1, a_2\} \star \{b_1, \ldots, b_k\}$ and $\Sigma' = \Sigma_l(c, d) = \{c_1, c_2\} \star \{d_1, \ldots, d_l\}$ be the suspension graphs corresponding to u and u' respectively. Furthermore, suppose these labelings are cyclic. It follows that $E = \{a_1, c_1, a_2, c_2\}$ is the 4–cycle corresponding to the edge in T between u and u'. By possibly relabeling, we can assume that $c_1 = b_1$, $c_2 = b_k$, $a_1 = d_1$ and $a_2 = d_l$. As Λ_{n-1} satisfies (1) above, (a_1, a_2) and (c_1, c_2) are edges of Λ_{n-1}.

As the valence of u is less than k, there exist some $1 \leq j < k$ such that the 4–cycle $\{b_j, a_1, b_{j+1}, a_2\}$ does not correspond to an edge of T. We define $\Lambda_n \subset \Gamma_n^c$ to contain every edge of $\Lambda_{n-1} \subset \Gamma_{n-1}^c \subset \Gamma_n^c$ and additionally the edges

$$(b_1, b_2), (b_2, b_3), \ldots, (b_{j-1}, b_j), (b_{j+1}, b_{j+2}), \ldots, (b_{k-1}, b_k).$$

This corresponds to adding one or two line segments each to a distinct vertex of Λ_{n-1}. As Λ_{n-1} contains two components (by (2)) and does not contain any cycles (by $\mathscr{R}_1$), it follows that Λ_n contains two components and satisfies $\mathscr{R}_1$ as well. Furthermore, (1) and condition $\mathscr{F}_1$ (for Θ_n) follow from directly from our choices. Condition $\mathscr{F}_2$ then follows from [Remark 4.3](#).

We now check $\mathscr{R}_2$. Let $x, y \in \Lambda_n$ be vertices contained in the same component of Λ_n. If x and y are both contained in Λ_{n-1}, then the claim follows as Λ_{n-1} satisfies $\mathscr{R}_2$ and no new edges are added between vertices of Γ_{n-1} in forming Γ_n. If x and y are both contained in Σ, then by construction, they must lie in the same factor of the join Σ and there is no edge between them. The only case left to check is that x and y lie in different components of $\mathbb{S}^2 \setminus E$. However, in this case there is no edge between x and y as E separates x from y in the planar embedding.

We now check that $\mathscr{R}_3$ holds. Let C be a 2–component square in Λ_n. As E separates every vertex of $\Sigma \setminus E$ from every vertex in $(\Gamma_{n-1} \setminus E) \subset \Gamma_n$, it follows that either C lies in $\Gamma_{n-1} \subset \Gamma_n$ or C lies in Σ. In the first case the claim follows as Θ_{n-1} satisfies $\mathscr{R}_3$ (and noting that the convex hull of C in Λ_n lies in Λ_{n-1}). In the latter case, the claim is easily verified.

We now check $\mathscr{R}_4$. Let P be a 2–component cycle in Γ_n. If P lies entirely in Γ_{n-1} then every edge of P satisfies condition $\mathscr{R}_4$ as Θ_{n-1} satisfies $\mathscr{R}_4$. If P lies entirely in Σ, then $\mathscr{R}_4$ is easily verified. Thus, we may assume that P decomposes into two subpaths P_1 and P_2 such that $P_1 \subset \Gamma_{n-1}$ and $P_2 \subset \Sigma \setminus E$. As P does not repeat vertices, it follows that P_2 consists of just two edges (a_1, b_q) and (a_2, b_q) for some $2 \leq q \leq k$. As the valence of u' is less than l, there exists some $1 \leq q' < l$ and corresponding 4–cycle

$\{c_1, d_{q'}, c_2, d_{q'+1}\}$ such that every vertex of Γ_n is contained in a common component of $\mathbb{S}^2 \setminus C$. From this, we see that a_1 and a_2 are in different components of $\Gamma_{n-1} \setminus \{c_1, c_2\}$. Thus, P_1 must either contain c_1 or c_2. Suppose that P_1 contains c_1 (the other case is similar). The path P_1 does not contain both the edge (a_1, c_1) and the edge (a_2, c_1), for if it did, then P would either the equal 4–cycle $\{a_1, c_1, a_2, b_q\}$ or contain it as a subcycle. In the former case $P \subset \Sigma$, a case we have already ruled out, and in the latter case, P necessarily repeats a vertex (which is not allowed). We now define a cycle P' depending on which edges P_1 contains. We set $P' = (P_1 \setminus (a_1, c_1)) \cup (a_2, c_1)$ if $(a_1, c_1) \subset P_1$, $P' = (P_1 \setminus (a_2, c_1)) \cup (a_1, c_1)$ if $(a_2, c_1) \subset P_1$, and $P' = P_1 \cup (a_1, c_1) \cup (a_2, c_1)$ if P_1 does not contain either of (a_1, c_1) and (a_2, c_1). In each case, it follows that P' is a cycle in Γ_{n-1} containing every edge of P_1, except possibly (a_1, c_1) and (a_2, c_1). Additionally, every vertex of P' is a vertex of P, so the Λ–convex hull of P' is contained in the Λ–convex hull of P. From this and as Γ_{n-1} satisfies $\mathscr{R}_4$, it follows that every edge of P that is contained in P_1 satisfies $\mathscr{R}_4$ as well. Finally, every edge of $P \setminus P_1$ can been seen to satisfy $\mathscr{R}_4$ by using the 4–cycle $\{a_1, b_q, a_2, c_1\}$. $\qquad\qquad\square$

6 Generalized reflection subgroups of RAAGs

Let A_Γ be a RAAG. A *generalized RAAG reflection* is a conjugate of an element of $V(\Gamma)$, ie wsw^{-1} for some $s \in V(\Gamma) \cup V(\Gamma)^{-1}$ and w a word in A_Γ. Let $\mathcal{T}$ be a set of reduced generalized RAAG reflections. We say that $\mathcal{T}$ is *trimmed* if $\mathcal{T} \cap \mathcal{T}^{-1} = \varnothing$, and if given any two distinct generalized RAAG reflections wsw^{-1} and $w's'w'^{-1}$ in $\mathcal{T}$, no expression for w' has prefix ws^{-1} or prefix ws. The following lemma follows from a straightforward adaptation of the proof of [Dani and Levcovitz 2021, Lemma 10.1] to the setting of RAAGs.

Lemma 6.1 *Let $\mathcal{T}$ be a set of generalized RAAG reflections in the RAAG A_Γ, and let G be the subgroup generated by $\mathcal{T}$. Then G is generated by a trimmed set of generalized RAAG reflections which can be algorithmically obtained from $\mathcal{T}$.*

In this section, we give a new proof of a result of Dyer:

Theorem 6.2 [Dyer 1990] *Let $\mathcal{T}$ be a finite set of generalized RAAG reflections in A_Γ. Then the subgroup $G < A_\Gamma$ generated by $\mathcal{T}$ is a RAAG. Moreover, if $\mathcal{T}$ is trimmed then $(G, \mathcal{T})$ is a RAAG system.*

We will use the characterization of RAAGs in Theorem 2.2 to show that G is a RAAG. We first prove a series of lemmas about disk diagrams of a special type, namely, ones whose boundary labels are words over a trimmed set of generalized RAAG reflections.

The setup for these lemmas is as follows and will be fixed for the rest of this section. We fix a trimmed set $\mathcal{T}$ of reduced generalized RAAG reflections in A_Γ. Let $z = r_1 \cdots r_n$ be an expression for the identity element where $r_i = w_i s_i w_i^{-1} \in \mathcal{T}$ for each $1 \le i \le n$. Let D be a disk diagram whose boundary ∂D is labeled by z. For $1 \le i \le n$, let p_{r_i} be the subpath of ∂D which is labeled by r_i. Furthermore let p_{w_i} and

$p_{w_i^{-1}}$ denote the subpaths of ∂D labeled w_i and w_i^{-1} respectively, and let e_i denote the edge labeled s_i. Let H_i be the hyperplane dual to e_i, and let $\mathcal{H} = \{H_i\}_{i=1}^n$ be the collection of all such hyperplanes. Note that as r_i is a reduced word, no hyperplane is dual to two edges of p_{r_i} for any i.

In all of the following lemmas, arithmetic is taken modulo n.

Lemma 6.3 *For each $1 \leq i \leq n$, the hyperplane H_i does not intersect a hyperplane dual to p_{w_i} or a hyperplane dual to $p_{w_i^{-1}}$*

Proof Suppose H_i intersects a hyperplane K that is dual to an edge f of p_{w_i}. Without loss of generality, we may assume that f is the edge closest to e_i out of all possible choices for K. As no hyperplane is dual to two edges of p_{r_i}, it follows that every hyperplane dual to an edge of p_{w_i} which lies between e_i and f must intersect K. Thus, w_i has suffix the word $t_1 \cdots t_m$, where t_1 is the label of K and t_1 commutes with s_i, as well as with t_j for $2 \leq j \leq m$. This readily implies that r_i is not reduced, for in $r_i = w_i s_i w_i^{-1}$, an occurrence of the RAAG generator t_1 in w_i can be canceled with an occurrence of t_1^{-1} in w_i^{-1}. However, this is a contradiction as r_i is reduced by assumption. The argument for hyperplanes dual to $p_{w_i^{-1}}$ is analogous. □

Lemma 6.4 *For each $1 \leq i \leq n$, the hyperplane H_i is not dual to $p_{w_{i+1}}$, $p_{w_{i+1}^{-1}}$, $p_{w_{i-1}^{-1}}$ or $p_{w_{i-1}}$.*

Proof For a contradiction, suppose H_i is dual to an edge f of $p_{w_{i+1}}$. By [Lemma 6.3](), every hyperplane dual to an edge of $p_{w_i^{-1}}$ must also be dual to $p_{w_{i+1}}$. Write $s_i w_i^{-1} = t_1 \cdots t_m$ and $w_{i+1} = k_1 \cdots k_l$ where $t_j \in V(\Gamma)$ for $1 \leq j \leq m$ and $k_j \in V(\Gamma)$ for $1 \leq j \leq l$. The structure of the hyperplanes in D implies that w_{i+1} has an expression which begins with $t_m^{-1} \cdots t_1^{-1} = w_i s_i^{-1}$. This is a contradiction as $\mathcal{T}$ is trimmed. A similar argument shows that H_i is not dual to $p_{w_{i-1}^{-1}}$.

Suppose now that H_i is dual to $p_{w_{i-1}}$. By [Lemma 6.3](), it follows that H_{i-1} is dual to p_{w_i}. However, this is not possible by the same argument as above. Similarly, H_i cannot be dual to $p_{w_{i+1}^{-1}}$. □

The proof of the following lemma is similar to that of the previous one.

Lemma 6.5 *If $H_i = H_{i+1}$ for some $1 \leq i \leq n$ then $r_i \simeq r_{i+1}^{-1}$.* □

Lemma 6.6 *If H_i intersects H_{i+1}, then r_i and r_{i+1} commute. Furthermore, there is a disk diagram D' with boundary label $r_1 \cdots r_{i-1} r_{i+1} r_i r_{i+2} \cdots r_n$, such that the natural bijection, from e_i, e_{i+1} and the edges traversed by the subpath of the boundary path of D labeled by $r_{i+2} \cdots r_n r_1 \cdots r_{i-1}$ to the edges traversed by the corresponding subpaths of the boundary path of D' with the same labels, preserves boundary combinatorics.*

Proof Suppose H_i intersects H_{i+1}. By [Lemma 6.3](), every hyperplane dual to $p_{w_i^{-1}}$ is either dual to $p_{w_{i+1}}$ or intersects H_{i+1}. Similarly, every hyperplane dual to $p_{w_{i+1}}$ is either dual to $p_{w_i^{-1}}$ or intersects H_i. It then readily follows that w_i has a reduced expression ba_1 and w_{i+1} has a reduced

expression ba_2, where a_1, a_2 and b are words, such that the generators in the word $a_1 s_i$ are all distinct from and commute with the generators in the word $a_2 s_{i+1}$. Consequently, r_i commutes with r_{i+1}.

We now construct the disk diagram D'. By Tits' solution to the word problem, the expression ba_1 (resp. ba_2) can be obtained from w_i (resp. w_{i+1}) by sequentially permuting adjacent letters. Thus, by repeatedly applying Lemma 2.11(1), we obtain a disk diagram with boundary label

$$r_1 \cdots r_{i-1}(ba_1 s_i a_1^{-1} b^{-1})(ba_2 s_{i+1} a_2^{-1} b^{-1}) r_{i+2} \cdots r_n.$$

By repeatedly applying Lemma 2.11(2), we can "cancel" $b^{-1}b$ and obtain a disk diagram with boundary label

$$r_1 \cdots r_{i-1}(ba_1 s_i a_1^{-1})(a_2 s_{i+1} a_2^{-1} b^{-1}) r_{i+2} \cdots r_n.$$

Then, by repeatedly applying Lemma 2.11(1), we obtain a disk diagram with label

$$r_1 \cdots r_{i-1}(ba_2 s_{i+1} a_2^{-1})(a_1 s_i a_1^{-1} b^{-1}) r_{i+2} \cdots r_n.$$

By Lemma 2.11(3), we obtain a disk diagram with boundary label

$$r_1 \cdots r_{i-1}(ba_2 s_{i+1} a_2^{-1} b^{-1})(ba_1 s_i a_1^{-1} b^{-1}) r_{i+2} \cdots r_n.$$

Finally, by repeatedly applying Lemma 2.11(1), we obtain a disk diagram D' with boundary label

$$r_1 \cdots r_{i-1} r_{i+1} r_i r_{i+2} \cdots r_n.$$

Note that in each of these steps, the desired boundary combinatorics are preserved. $\square$

Lemma 6.7 *For every $1 \leq i \leq n$, there exists some $j \neq i$ such that $H_i = H_j$.*

Proof Suppose we have a disk diagram with boundary label $z = r_1 \cdots r_n$ such that, for some $1 \leq i \leq n$, the hyperplane H_i is dual to an edge f of ∂D where $f \neq e_j$ for all $1 \leq j \leq n$. We call any disk diagram which has such an H_i a *pathological diagram* with *pathology caused by H_i*. Given such a diagram, we define p to be a path along ∂D between e_i and f, which does not include e_i and f. We also let $\mathcal{H}'$ denote the set of H_j such that e_j is contained in p.

Given a pathological disk diagram D we may choose a hyperplane H_i causing the pathology together with a path p such that the set $\mathcal{H}'$ is minimal among all possible choices of H_i and p. After such a choice, we call $|\mathcal{H}'|$ the complexity of D. We will prove that pathological diagrams are not possible by induction on the complexity c of such a diagram. The base case, when $c = 0$, already follows from Lemma 6.4.

Now suppose we are given a pathological disk diagram D with pathology caused by H_i such that its complexity is $c = |\mathcal{H}'| > 0$, and suppose by induction there do not exist pathological disk diagrams of complexity smaller than c.

The edge $f \neq e_i$ of ∂D that is dual to H_i lies in a path $p_{r_{i'}}$ in ∂D labeled by $w_{i'} s_{i'} w_{i'}^{-1}$ for some $1 \leq i' \leq n$ where $i \neq i'$. Let Q denote the hyperplane $H_{i'}$. Note that Q may or may not be in $\mathcal{H}'$. We prove our claim by considering two cases:

Case 1 Every hyperplane in $\mathscr{H}'$ intersects H_i.

We first observe that $\mathscr{H}'$ is nonempty (since the complexity of D is positive) and does not consist of Q alone (by Lemma 6.4). Therefore, we may choose $K \in \mathscr{H}' \setminus Q$ such that no hyperplane in $\mathscr{H}' \setminus Q$ intersects H_i between $K \cap H_i$ and $H_i \cap e_i$. Let $1 \le l \le n$ be such that K is dual to $e_l \subset p_{r_l} \subset p$. Then for each j with $i < j < l$, the hyperplane H_j intersects $K = H_l$. Thus, by repeatedly applying Lemma 6.6, we can produce a new disk diagram with boundary label $r_1 \cdots r_l r_i \cdots r_{l-1} r_{l+1} \cdots r_n$. Furthermore, this new disk diagram is still pathological and has complexity smaller than D. However, this is not possible by our induction hypothesis.

Case 2 Some hyperplane $K \in \mathscr{H}'$ does not intersect H_i.

We can choose such a hyperplane K to be innermost, ie choose $K \in \mathscr{H}'$ such that K does not intersect H_i and such that any hyperplane of $\mathscr{H}'$ dual to the subpath of p between the edges dual to K intersects K. Since H_i and p were chosen to attain the complexity of D, it follows that K does not cause a pathology, and is dual to distinct edges e_l and $e_{l'}$ in p, where $1 \le l, l' \le n$. By relabeling the r_j's if necessary, we may assume that $l < l'$, and that the subpath of ∂D from e_l to $e_{l'}$ is contained in p. By repeatedly applying Lemma 6.6, we can produce a new pathological disk diagram D' with label $r_1 \cdots r_{l-1} r_{l+1} \cdots r_{l'-1} r_l r_{l'} \cdots r_n$ and where some hyperplane, which we still denote by K, is dual to both the edge labeled by e_l and the one labeled by $e_{l'}$. By Lemma 6.5, $r_l \simeq r_{l'}^{-1}$. Furthermore, by repeatedly applying Lemma 2.11(1) if necessary, we may assume that $r_l = r_{l'}^{-1}$ is the label of $\partial D'$.

We now produce a new disk diagram D'' by identifying the consecutive paths in $\partial D'$ labeled by r_l and $r_{l'}$, ie we fold these two paths together. If $K \ne Q$, then we have produced a new pathological disk diagram with complexity $c - 2$, contradicting the induction hypothesis. On the other hand, if $K = Q$, note that the image of H_i in D'' must intersect the path labeled by $r_i \cdots r_{l-1} r_{l+1} \cdots r_{l'-1}$ in $\partial D''$. Moreover we claim that it cannot be dual to an edge labeled by e_j for $i < j \le l' - 1$. Suppose it is dual to an edge labeled e_j. It follows that the hyperplane H_j in D is dual to an edge f' in p, such that $f' \ne e_k$ for any k, and such that the images of f and f' are identified in D''. This is a contradiction, as it implies that H_j causes a pathology of lower complexity than H_i. Thus, the image of H_i in D'' causes a pathology of complexity at most $c - 2$, which is again a contradiction.	$\square$

Proof of Theorem 6.2 As G can be generated by a trimmed set of generalized RAAG reflections (by Lemma 6.1), we assume without loss of generality that $\mathscr{T}$ is trimmed. We will show that $(G, \mathscr{T})$ is a RAAG system by applying Theorem 2.2. Note that $\mathscr{T} \cap \mathscr{T}^{-1} = \varnothing$ as $\mathscr{T}$ is trimmed. We check each condition of that theorem, by proving the corresponding two claims:

(i) Every $r \in \mathscr{T}$ has infinite order.

By definition, r is equal to a reduced word wsw^{-1} with $s \in V(\Gamma) \cup V(\Gamma)^{-1}$ and w a word in W_Γ. It follows that $ws^n w^{-1}$ is an expression for r^n. Moreover, as r is reduced, it readily follows from Theorem 2.4 that $ws^n w^{-1}$ is reduced as well. Hence, r has infinite order.

(ii) Given any word $w = a_1 \cdots a_m$, with $a_i \in \mathcal{T}$, either w is reduced over $\mathcal{T}$ or there is an expression for w of the form $a_1 \cdots \hat{a}_i \cdots \hat{a}_j \cdots a_m$.

Suppose $w = a_1 \cdots a_m$ is not reduced over $\mathcal{T}$. Let $w' = b_1 \cdots b_k$, with $b_i \in \mathcal{T}$ and $k < m$, be an expression for w which is reduced over $\mathcal{T}$. Form a disk diagram D with boundary label ww'^{-1}.

We relabel the generalized reflections in the word ww'^{-1} by setting $r_i = a_i$ for $1 \leq i \leq m$, and $r_{m+i} = b_{k-i+1}^{-1}$ (the ith generalized RAAG reflection in w'^{-1}) for $1 \leq i \leq k$. By Lemma 6.7, every $H \in \mathcal{H}$ is only dual to edges of ∂D labeled by s_i for some i, where $r_i = w_i s_i w_i^{-1}$. As $m > k$, there exists some hyperplane $H \in \mathcal{H}$ that is dual to two edges of the subpath p of ∂D labeled by w. Furthermore, we may choose an innermost such $H \in \mathcal{H}$, in the sense that every hyperplane in $\mathcal{H} \setminus H$ intersects p at most once.

Let e_l and $e_{l'}$ be the edges dual to H where $l < l' \leq m$. By repeatedly applying Lemma 6.6, we produce a disk diagram whose boundary label is

$$r_1 \cdots \hat{r}_l \cdots r_{l'-1} r_l r_{l'} \cdots r_n,$$

such that a hyperplane of $\mathcal{H}$ is still dual to the images of the edges e_l and $e_{l'}$ under the natural map between the boundaries of the disk diagrams. By Lemma 6.5, $r_l = r_{l'}^{-1}$. Thus, $r_1 \cdots \hat{r}_l \cdots \hat{r}_{l'} \cdots r_n$ is an expression for ww'^{-1}. Consequently, $r_1 \cdots \hat{r}_l \cdots \hat{r}_{l'} \cdots r_m = a_1 \cdots \hat{a}_l \cdots \hat{a}_{l'} \cdots a_m$ is an expression for w. □

References

[Bahls 2005]   **P Bahls**, *The isomorphism problem in Coxeter groups*, Imperial College Press, London (2005) MR Zbl

[Basarab 2002]   **Ş A Basarab**, *Partially commutative Artin–Coxeter groups and their arboreal structure*, J. Pure Appl. Algebra 176 (2002) 1–25 MR Zbl

[Behrstock 2019]   **J Behrstock**, *A counterexample to questions about boundaries, stability, and commensurability*, from "Beyond hyperbolicity" (M Hagen, R Webb, H Wilton, editors), Lond. Math. Soc. Lect. Note Ser. 454, Cambridge Univ. Press (2019) 151–159 MR Zbl

[Behrstock and Charney 2012]   **J Behrstock**, **R Charney**, *Divergence and quasimorphisms of right-angled Artin groups*, Math. Ann. 352 (2012) 339–356 MR Zbl

[Bestvina et al. 2008]   **M Bestvina**, **B Kleiner**, **M Sageev**, *The asymptotic geometry of right-angled Artin groups, I*, Geom. Topol. 12 (2008) 1653–1699 MR Zbl

[Björner and Brenti 2005]   **A Björner**, **F Brenti**, *Combinatorics of Coxeter groups*, Graduate Texts in Math. 231, Springer (2005) MR Zbl

[Charney 2007]   **R Charney**, *An introduction to right-angled Artin groups*, Geom. Dedicata 125 (2007) 141–158 MR Zbl

[Charney and Sultan 2015]   **R Charney**, **H Sultan**, *Contracting boundaries of* CAT(0) *spaces*, J. Topol. 8 (2015) 93–117 MR Zbl

[Cordes and Hume 2017]   **M Cordes**, **D Hume**, *Stability and the Morse boundary*, J. Lond. Math. Soc. 95 (2017) 963–988   MR Zbl

[Dani and Levcovitz 2021]   **P Dani**, **I Levcovitz**, *Subgroups of right-angled Coxeter groups via Stallings-like techniques*, J. Comb. Algebra 5 (2021) 237–295   MR Zbl

[Dani and Thomas 2015]   **P Dani**, **A Thomas**, *Divergence in right-angled Coxeter groups*, Trans. Amer. Math. Soc. 367 (2015) 3549–3577   MR Zbl

[Davis 2015]   **M W Davis**, *The geometry and topology of Coxeter groups*, from "Introduction to modern mathematics" (S-Y Cheng, L Ji, Y-S Poon, J Xiao, L Yang, S-T Yau, editors), Adv. Lect. Math. 33, International, Somerville, MA (2015) 129–142   MR Zbl

[Davis and Januszkiewicz 2000]   **M W Davis**, **T Januszkiewicz**, *Right-angled Artin groups are commensurable with right-angled Coxeter groups*, J. Pure Appl. Algebra 153 (2000) 229–235   MR Zbl

[Deodhar 1989]   **V V Deodhar**, *A note on subgroups generated by reflections in Coxeter groups*, Arch. Math. (Basel) 53 (1989) 543–546   MR Zbl

[Dyer 1990]   **M Dyer**, *Reflection subgroups of Coxeter systems*, J. Algebra 135 (1990) 57–73   MR Zbl

[Genevois 2017]   **A Genevois**, *Cubical-like geometry of quasi-median graphs and applications to geometric group theory*, PhD thesis, Aix-Marseille Université (2017)   arXiv 1712.01618

[Genevois 2019]   **A Genevois**, *Embeddings into Thompson's groups from quasi-median geometry*, Groups Geom. Dyn. 13 (2019) 1457–1510   MR Zbl

[Green 1990]   **E R Green**, *Graph products of groups*, PhD thesis, University of Leeds (1990)   Available at https://etheses.whiterose.ac.uk/236/

[Haglund and Wise 2010]   **F Haglund**, **D T Wise**, *Coxeter groups are virtually special*, Adv. Math. 224 (2010) 1890–1903   MR Zbl

[Kim and Koberda 2013]   **S-h Kim**, **T Koberda**, *Embedability between right-angled Artin groups*, Geom. Topol. 17 (2013) 493–530   MR Zbl

[LaForge 2017]   **G LaForge**, *Visible Artin subgroups of right-angled Coxeter groups*, PhD thesis, Tufts University (2017)   Available at https://www.proquest.com/docview/1986002761

[Nguyen and Tran 2019]   **H T Nguyen**, **H C Tran**, *On the coarse geometry of certain right-angled Coxeter groups*, Algebr. Geom. Topol. 19 (2019) 3075–3118   MR Zbl

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

[Tits 1969]   **J Tits**, *Le problème des mots dans les groupes de Coxeter*, from "Symposia mathematica, I", Academic, London (1969) 175–185   MR Zbl

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

Department of Mathematics, Louisiana State University
Baton Rouge, LA, United States
Department of Mathematics, Tufts University
Medford, MA, United States
pdani@math.lsu.edu,   ivan.levcovitz@gmail.com

Filling braided links with trisected surfaces

JEFFREY MEIER

We introduce the concept of a bridge trisection of a neatly embedded surface in a compact four-manifold, generalizing previous work with Alexander Zupan in the setting of closed surfaces in closed four-manifolds. Our main result states that any neatly embedded surface $\mathcal{F}$ in a compact four-manifold X can be isotoped to lie in bridge trisected position with respect to any trisection $\mathbb{T}$ of X. A bridge trisection of $\mathcal{F}$ induces a braiding of the link $\partial\mathcal{F}$ with respect to the open-book decomposition of ∂X induced by $\mathbb{T}$, and we show that the bridge trisection of $\mathcal{F}$ can be assumed to induce any such braiding.

We work in the general setting in which ∂X may be disconnected, and we describe how to encode bridge trisected surface diagrammatically using shadow diagrams. We use shadow diagrams to show how bridge trisected surfaces can be glued along portions of their boundary, and we explain how the data of the braiding of the boundary link can be recovered from a shadow diagram. Throughout, numerous examples and illustrations are given. We give a set of moves that we conjecture suffice to relate any two shadow diagrams corresponding to a given surface.

We devote extra attention to the setting of surfaces in B^4, where we give an independent proof of the existence of bridge trisections and develop a second diagrammatic approach using tri-plane diagrams. We characterize bridge trisections of ribbon surfaces in terms of their complexity parameters. The process of passing between bridge trisections and band presentations for surfaces in B^4 is addressed in detail and presented with many examples.

57K10, 57K40, 57K45

1 Introduction

The philosophy underlying the theory of trisections is that four-dimensional objects can be decomposed into three simple pieces whose intersections are well-enough controlled that all of the four-dimensional data can be encoded on the two-dimensional intersection of the three pieces, leading to new diagrammatic approaches to four-manifold topology. Trisections were first introduced for four-manifolds by Gay and Kirby in 2016 [10]. A few years later, the theory was adapted to the setting of closed surfaces in four-manifolds by the author and Zupan [27; 28]. The present article extends the theory to the general setting of neatly embedded surfaces in compact four-manifolds, yielding two diagrammatic approaches to the study of these objects: one that applies in general and one that applies when we restrict attention to surfaces in B^4.

1.1 Bridge trisections of surfaces in B^4

To introduce bridge trisections of surfaces in B^4, we must establish some terminology. First, let H be a three-ball $D^2 \times I$, equipped with a critical-point-free Morse function $D^2 \times I \to I$. Let $\mathcal{T} \subset H$ be a neatly embedded one-manifold such that the restriction of the Morse function to each component of $\mathcal{T}$ has either one critical point (a maximum) or none. If there are b components with one critical point and v with none, we call $(H, \mathcal{T})$ a (b, v)–*tangle*. Next, let Z be a four-ball $B^3 \times I$, equipped with a critical-point-free Morse function $B^3 \times I \to I$. Let $\mathcal{D} \subset Z$ be a collection of neatly embedded disks such that the restriction of the Morse function to each component of $\mathcal{D}$ has either one critical point (a minimum) or none. If there are c components with one critical point and v with none, we call $(Z, \mathcal{D})$ a (c, v)–*disk-tangle*. Finally, let $\mathbb{T}_0$ denote the *standard trisection* of B^4 — ie the decomposition $B^4 = Z_1 \cup Z_2 \cup Z_3$ in which, for each $i \in \mathbb{Z}_3$, the Z_i are four-balls, the pairwise intersections $H_i = Z_{i-1} \cap Z_i$ are three-balls, and the common intersection $\Sigma = Z_1 \cap Z_2 \cap Z_3$ is a disk.

A neatly embedded surface $\mathcal{F} \subset B^4$ is in $(b, \boldsymbol{c}, v)$–*bridge position* with respect to $\mathbb{T}_0$ if, for each $i \in \mathbb{Z}_3$,

(1) $\mathcal{F} \cap Z_i$ is a (c_i, v)–disk-tangle, where $\boldsymbol{c} = (c_1, c_2, c_3)$, and

(2) $\mathcal{F} \cap H_i$ is a (b, v)–tangle.

A definition very similar to this one was introduced independently in [2].

The trisection $\mathbb{T}_0$ induces the open-book decomposition of $S^3 = \partial B^4$ whose pages are the disks $S^3 \cap H_i$ and whose binding is $\partial \Sigma$. Let $\mathcal{L} = \partial \mathcal{F}$, and let $\beta_i = S^3 \cap \mathcal{D}_i$. Then $\mathcal{L} = \beta_1 \cup \beta_2 \cup \beta_3$ is braided about $\partial \Sigma$ with index v. Having outlined the requisite structures, we can state our existence result for bridge trisections of surfaces in the four-ball.

Theorem 3.17 *Let $\mathbb{T}_0$ be the standard trisection of B^4, and let $\mathcal{F} \subset B^4$ be a neatly embedded surface with $\mathcal{L} = \partial \mathcal{F}$. Fix an index v braiding $\hat{\beta}$ of $\mathcal{L}$. Suppose $\mathcal{F}$ has a handle decomposition with c_1 cups, n bands, and c_3 caps. Then, for some $b \in \mathbb{N}_0$, $\mathcal{F}$ can be isotoped to be in $(b, \boldsymbol{c}; v)$–bridge trisected position with respect to $\mathbb{T}_0$, such that $\partial \mathcal{F} = \hat{\beta}$, where $c_2 = b - n$.*

Explicit in the above statement is a connection between the complexity parameters of a bridge trisected surface and the numbers of each type of handle in a Morse decomposition of the surface. An immediate consequence of this correspondence is the fact that a ribbon surface admits a bridge trisection where $c_3 = 0$. It turns out that this observation can be strengthened to give the following characterization of ribbon surfaces in B^4. Again, $\boldsymbol{c} = (c_1, c_2, c_3)$, and we set $c = c_1 + c_2 + c_3$.

Theorem 3.21 *Let $\mathbb{T}_0$ be the standard trisection of B^4, and let $\mathcal{F} \subset B^4$ be a neatly embedded surface with $\mathcal{L} = \partial \mathcal{F}$. Let $\hat{\beta}$ be an index v braiding $\mathcal{L}$. Then the following are equivalent*:

(1) $\mathcal{F}$ *is ribbon.*

(2) $\mathcal{F}$ *admits a* $(b, \boldsymbol{c}; v)$–*bridge trisection filling* $\hat{\beta}$ *with* $c_i = 0$ *for some* i.

(3) $\mathcal{F}$ *admits a* $(b, 0; v + c)$–*bridge trisection filling a Markov perturbation* $\hat{\beta}^+$ *of* $\hat{\beta}$.

A bridge trisection turns out to be determined by its spine — ie the union $(H_1, \mathcal{T}_1) \cup (H_2, \mathcal{T}_2) \cup (H_3, \mathcal{T}_3)$, and each tangle $(H_i, \mathcal{T}_i)$ can be faithfully encoded by a planar diagram. It follows that any surface in B^4 can be encoded by a triple of planar diagrams whose pairwise unions are planar diagrams for split unions of geometric braids and unlinks. We call such triples *tri-plane diagrams*.

Corollary 4.2 *Every neatly embedded surface in B^4 can be described by a tri-plane diagram.*

In Section 4, we show how to read off the data of the braiding of $\mathcal{L}$ induced by a bridge trisection from a tri-plane for the bridge trisection, and we describe a collection of moves that suffice to relate any two tri-plane diagrams corresponding to a given bridge trisection. The reader concerned mainly with surfaces in B^4 can focus their attention on Sections 3 and 4, referring to the more general development of the preliminary material given in Section 2 when needed.

1.2 Bridge trisections of surfaces in compact four-manifolds

Having summarized the results of the paper that pertain to the setting of B^4, we now describe the more general setting in which X is a compact four-manifold with (possibly disconnected) boundary and $\mathcal{F} \subset X$ is a neatly embedded surface. To account for this added generality, we must expand the definitions given earlier for the basic building blocks of a bridge trisection. For ease of exposition, we will not record the complexity parameters, which are numerous in this setting; Section 2 contains compete details.

Let H be a compression body $(\Sigma \times I) \cup$ (3–dimensional 2–handles), where $\Sigma = \partial_+ H$ is connected and may have nonempty boundary, while $P = \partial_- H$ is allowed to be disconnected but cannot contain two-sphere components. We work relative to the induced Morse function. Let $\mathcal{T} \subset H$ be a neatly embedded one-manifold such that the restriction of the Morse function to each component of $\mathcal{T}$ has either one critical point (a maxima) or none. We call $(H, \mathcal{T})$ a *trivial tangle*. Let Z be a four-dimensional compression body $(P \times I \times I) \cup$ (4–dimensional 1–handles), where P is as above. We work relative to the obvious Morse function on Z. Let $\mathcal{D} \subset Z$ be a collection of neatly embedded disks such that the restriction of the Morse function to each component of $\mathcal{D}$ has either one critical point (a minima) or none. We call $(Z, \mathcal{D})$ a *trivial disk-tangle*.

Let X be a compact four-manifold, and let $\mathcal{F} \subset X$ be a neatly embedded surface. A *bridge trisection* of $(X, \mathcal{F})$ is a decomposition

$$(X, \mathcal{F}) = (Z_1, \mathcal{D}_1) \cup (Z_2, \mathcal{D}_2) \cup (Z_3, \mathcal{D}_3)$$

such that, for each $i \in \mathbb{Z}_3$,

(1) $(Z_i, \mathcal{D}_i)$ is a trivial disk-tangle, and

(2) $(H_i, \mathcal{T}_i) = (Z_{i-1}, \mathcal{D}_{i-1}) \cap (Z_i, \mathcal{D}_i)$ is a trivial tangle.

We let $(\Sigma, \boldsymbol{x}) = \partial_+(H_i, \mathcal{T}_i)$. The underlying trisection $X = Z_1 \cup Z_2 \cup Z_3$ induces an open-book decomposition on each component of $Y = \partial X$, and we find that the bridge trisection of $\mathcal{F}$ induces a braiding of $\mathcal{L} = \partial \mathcal{F}$ with respect to these open-book decompositions. Given this set-up, our general existence result can now be stated.

Theorem 8.1 *Let $\mathbb{T}$ be a trisection of a four-manifold X with $\partial X = Y$, and let (B, π) denote the open-book decomposition of Y induced by $\mathbb{T}$. Let $\mathcal{F}$ be a neatly embedded surface in X; let $\mathcal{L} = \partial \mathcal{F}$; and fix a braiding $\hat{\beta}$ of $\mathcal{L}$ about (B, π). Then $\mathcal{F}$ can be isotoped to be in bridge trisected position with respect to $\mathbb{T}$ such that $\partial \mathcal{F} = \hat{\beta}$. If $\mathcal{L}$ already coincides with the braiding β, then this isotopy can be assumed to restrict to the identity on Y.*

If H is not a three-ball, then $(H, \mathcal{T})$ cannot be encoded as a planar diagram, as before. However, H is determined by a collection of curves $\alpha \subset \Sigma \setminus \nu(\boldsymbol{x})$, and $\mathcal{T}$ is determined by a collection of arcs $\mathcal{T}^*$ and the points $\boldsymbol{x}$ in Σ, where the arcs of $\mathcal{T}^*$ connect pairs of points of $\boldsymbol{x}$. We call the data $(\Sigma, \alpha, \mathcal{T}^*, \boldsymbol{x})$, which determine the trivial tangle $(H, \mathcal{T})$, a *tangle shadow*. A triple of tangle shadows that satisfies certain pairwise-standardness conditions is called a *shadow diagram*. Because bridge trisections are determined by their spines, we obtain the following corollary.

Corollary 5.5 *Let X be a smooth, orientable, compact, connected four-manifold, and let $\mathcal{F}$ be a neatly embedded surface in X. Then $(X, \mathcal{F})$ can be described by a shadow diagram.*

A detailed development of shadow diagrams is given in Section 5, where it is described how to read off the data of the braiding of $\mathcal{L}$ induced by a bridge trisection from a shadow diagram corresponding to the bridge trisection. Moves relating shadow diagrams corresponding to a fixed bridge trisection are given. Section 6 discusses how to glue two bridge trisected surfaces so that the result is bridge trisected, as well as how these gluings can be carried out with shadow diagrams.

Section 7 gives some basic classification results, as well as a handful of examples to add to the many examples included throughout Sections 3–6. The proof of the main existence result, Theorem 8.1, is delayed until Section 8, though it requires only the content of Section 2 to be accessible. In Section 9, we discuss stabilization and perturbation operations that we conjecture are sufficient to relate any two bridge trisections of a fixed surface. A positive resolution of this conjecture would give complete diagrammatic calculi for studying surfaces via tri-plane diagrams and shadow diagrams.

Acknowledgements

2 Preliminaries

In this section, we give a detailed development of the ingredients required throughout the paper, establishing notation conventions as we go. This section should probably be considered as prerequisite for all the following sections, save for Sections 3 and 4, which pertain to the consideration of surfaces in the four-ball. The reader interested only in this setting may be able to skip ahead, referring back to this section only as needed.

2.1 Some conventions

Unless otherwise noted, all manifolds and maps between manifolds are assumed to be smooth, and manifolds are compact. The central objects of study here all have the form of a *manifold pair* (M, N), by which we mean that N is *neatly embedded* in M in the sense that $\partial N \subset \partial M$ and $N \pitchfork \partial M$ [15]. When N is compact (as it will always be here), N is properly embedded when it is neatly embedded and $\partial N \subset \partial M$ when N is properly embedded; the transversality condition on neat embeddings is not generally enjoyed by proper embeddings. Throughout, N will usually have codimension two in M. In any event, we let $\nu(N)$ denote the interior of a tubular neighborhood of N in M. If M is oriented, we let $\overline{(M, N)}$ denote the pair (M, N) with the opposite orientation and we call it the *mirror* of (M, N). We use the symbol $\sqcup$ to denote either the disjoint union or the split union, depending on the context. For example, writing $(M_1, N_1) \sqcup (M_2, N_2)$ indicates $M_1 \cap M_2 = \varnothing$. On the other hand, $(M, N_1 \sqcup N_2)$ indicates that N_1 and N_2 are *split* in M, by which we usually mean there are disjoint, codimension zero balls B_1 and B_2 in M (not necessarily neatly embedded) such that $N_i \subset \text{Int } B_i$ for each $i \in \{1, 2\}$.

2.2 Lensed cobordisms

Given compact manifold pairs (M_1, N_1) and (M_2, N_2) with $\partial(M_1, N_1) \cong \partial(M_2, N_2)$ nonempty, we normally think of a cobordism from (M_1, N_1) to (M_2, N_2) as a manifold pair (W, Z), where

$$\partial(W, Z) = ((M_1, N_1) \sqcup \overline{(M_2, N_2)}) \cup (\partial(M_1, N_1) \times I).$$

Thus, there is a cylindrical portion of the boundary. Consider the quotient space (W', Z') of (W, Z) obtained via the identification $(x, t) \sim (x, t')$ for all $x \in \partial M_1$ and $t, t' \in I$. The space (W', Z') is diffeomorphic to (W, Z), but

$$\partial(W', Z') = (M_1, N_1) \cup_{\partial(M_1, N_1)} \overline{(M_2, N_2)}.$$

We refer to (W', Z') as a *lensed cobordism*. An example of a lensed cobordism is the submanifold W' cobounded by two Seifert surfaces for a knot K in S^3 that are disjoint in their interior. If $W = M_1 \times I$, then we call W' a *product lensed cobordism*. An example of a product lensed cobordism is the submanifold W' cobounded by two pages of an open-book decomposition on an ambient manifold X. See Figure 1 for examples of lensed cobordisms between surfaces that contain 1–dimensional cobordisms as neat submanifolds.

We offer the following two important remarks regarding our use of lensed cobordisms.

Remark 2.1 Throughout this article, we will be interested in cobordisms between manifolds with boundary. For this reason, lensed cobordisms are naturally well-suited for our purposes. However, at times we will be discussing cobordisms between closed manifolds (eg null-cobordisms). In this case, lensed cobordisms do not make sense. We request that the reader remember to drop the adjective "lensed" upon consideration of such cases. For example, if (M, N) is any manifold pair with $N \subset \text{Int}(M)$ closed, then for the product lensed cobordism $(M, N) \times I$, we have that $M \times I$ is lensed, but $N \times I$ is not.

Remark 2.2 Lensed cobordisms do not admit Morse functions where (M_1, N_1) and (M_2, N_2) represent distinct level sets, since $(M_1, N_1) \cap (M_2, N_2) \neq \varnothing$. However, the manifold pair

$$(W'', Z'') = (W', Z') \setminus \nu(\partial(M_1, N_1))$$

does admit such a function and is trivially diffeomorphic to (W', Z'): We think of (W'', Z'') as being formed by "indenting" (W', Z') by removing $\nu(\partial(M_1, N_1))$. Note that there is a natural identification of (W'', Z'') with the original (ordinary) cobordism (W, Z). Since a generic Morse function on the cobordism W'' will not have critical points on its boundary, there is no loss of information here. We will have this modification in mind when we consider Morse functions on lensed cobordisms (W', Z'), which we will do throughout the paper. This subtlety illustrates that lensed cobordisms are unnatural in a Morse-theoretic approach to manifold theory, but we believe they are more natural in a trisection-theoretic approach.

2.3 Compression bodies

Given a surface Σ and a collection α of pairwise disjoint, simple closed curves on Σ, let Σ^α denote the surface obtained by surgering Σ along α. Let H denote the three-manifold obtained by attaching a collection $\mathfrak{h}_\alpha$ of three-dimensional 2–handles to $\Sigma \times [-1, 1]$ along $\alpha \times \{1\}$, before filling in any resulting sphere components with balls. As discussed in [Remark 2.1](#), in the case that Σ has nonempty boundary, we quotient out by the vertical portion of the boundary and view H as a lensed cobordism from $\partial_+ H = \Sigma$ to $\partial_- H = \Sigma^\alpha$. Considering H as an oriented manifold yields the decomposition

$$\partial H = \partial_+ H \cup_{\partial(\partial_+ H)} \overline{\partial_- H}.$$

The manifold H is called a (*lensed*) *compression body*. A collection $\mathscr{D}$ of disjoint, neatly embedded disks in a compression body H is called a *cut system* for H if $H \setminus \nu(\mathscr{D}) \cong (\partial_- H) \times I$ or $H \setminus \nu(\mathscr{D}) \cong B^3$, according with whether $\partial(\partial_+ H) = \partial(\partial_- H)$ is nonempty or empty, respectively. A collection of essential, simple closed curves on $\partial_+ H$ is called a *defining set of curves* for H if it is the boundary of a cut system for H.

In order to efficiently discuss compression bodies H for which $\partial_- H$ is disconnected, we will introduce the following terminology.

$(H_{2,0,1}, \mathcal{T}_{2,3})$ $(H_{2,1,1}, \mathcal{T}_{0,3})$ $(H_{0,(0,0),(1,1)}, \mathcal{T}_{3,(1,3)})$
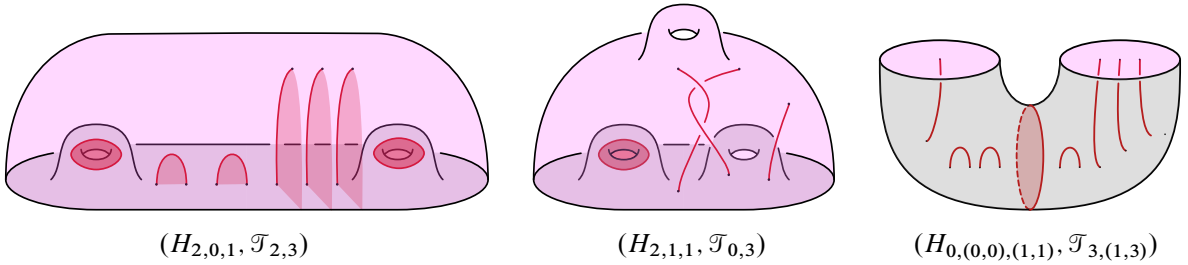
Figure 1: Three examples of trivial tangles inside lensed compression bodies.

Definition 2.3 Given $m \in \mathbb{N}_0$, an *ordered partition* of m is a sequence $\boldsymbol{m} = (m_1, \ldots, m_n)$ such that $m_j \in \mathbb{N}_0$ and $\sum m_j = m$. We say that such an ordered partition is of type (m, n). If $m_j > 0$ for all j, then the ordered partition is called *positive* and is said to be of type $(m, n)^+$. If $m_j = m'$ for all j, then the ordered partition is called *balanced*.

Let Σ_g denote the closed surface of genus g, and let $\Sigma_{g,f}$ denote the result of removing f disjoint, open disks from Σ_g. A surface Σ with $n > 1$ connected components is called *ordered* if there is an ordered partition $\boldsymbol{p} = (p_1, \ldots, p_n)$ of $p \in \mathbb{N}_0$ and a positive ordered partition $\boldsymbol{f} = (f_1, \ldots, f_n)$ of $f \in \mathbb{N}$ such that

$$\Sigma \cong \Sigma_{p_1, f_1} \sqcup \cdots \sqcup \Sigma_{p_n, f_n}.$$

We denote such an ordered surface by $\Sigma_{\boldsymbol{p}, \boldsymbol{f}}$, and we consider each $\Sigma_{p_j, f_j} \subset \Sigma_{\boldsymbol{p}, \boldsymbol{f}}$ to come equipped with an ordering of its f_j boundary components, when necessary. Note that we are requiring each component of the *disconnected* surface $\Sigma_{\boldsymbol{p}, \boldsymbol{f}}$ to have boundary.

Let $H_{g, \boldsymbol{p}, \boldsymbol{f}}$ denote the lensed compression body satisfying

(1) $\partial_+ H_{g, \boldsymbol{p}, \boldsymbol{f}} = \Sigma_{g, f}$, and

(2) $\partial_- H_{g, \boldsymbol{p}, \boldsymbol{f}} = \Sigma_{\boldsymbol{p}, \boldsymbol{f}}$.

If α is a defining set for such a compression body, then α consists of $(n-1)$ separating curves and $(g-p)$ nonseparating curves. See Figure 1 for three examples of lensed compression bodies, ignoring for now the submanifolds. Let H_{p_j, f_j} denote the product lensed cobordism from Σ_{p_j, f_j} to itself, and let

$$H_{\boldsymbol{p}, \boldsymbol{f}} = \bigsqcup_{j=1}^{\infty} H_{p_j, f_j}.$$

We refer to $H_{\boldsymbol{p}, \boldsymbol{f}}$ as a *spread*.

A lensed compression body H admits a Morse function $\Phi \colon H \to [-1, 3]$, which, as discussed in Remark 2.2, is defined on $H \setminus \nu(\partial(\partial_+ H))$, such that $\Phi(\partial_+ H) = -1$, $\Phi(\partial_- H) = 3$, and Φ has $(n-1) + (g-p)$ critical points, all of index two, and all lying in $\Phi^{-1}(2)$. We call such a Φ a *standard* Morse function for H. Every compression body admits a standard Morse function, even if it were built by capping off two-sphere components with 3–handles. These 3–handles can be assumed to cancel with 2–handles.

If 3–handles were required after 2–handles were attached to Σ along α, then some curves of α were unnecessary.

For a positive natural number I, we let $\boldsymbol{x}_I \subset \Sigma_{g,f}$ denote a fixed collection of I marked points.

2.4 Heegaard splittings and Heegaard-page splittings

Let M be an orientable three-manifold. A *Heegaard splitting* of M is a decomposition

$$M = H_1 \cup_\Sigma \overline{H}_2,$$

where $\Sigma \subset M$ is a neatly embedded surface $\Sigma_{g,f}$, and each H_i is a lensed compression body $H_{g,\boldsymbol{p},\boldsymbol{f}}$ with $\partial_+ H_i = \Sigma$. It follows that

$$\partial M = \overline{\partial_- H_1} \cup_{\partial \Sigma} \partial_- H_2.$$

We denote the Heegaard splitting by $(\Sigma; H_1, H_2)$, and we call it a $(g; \boldsymbol{p}, \boldsymbol{f})$–splitting, in reference to the relevant parameters. Note that our notion of Heegaard splitting restricts to the usual notion when M is closed, but is different from the usual notion when M has boundary. Our Heegaard splittings are a special type of sutured manifold decomposition. Since each of the H_i is determined by a defining set of curves α_i on Σ, the Heegaard splitting, including M itself, is determined by the triple $(\Sigma; \alpha_1, \alpha_2)$, which is called a *Heegaard diagram* for M.

Remark 2.4 We have defined Heegaard splittings so that the two compression bodies are homeomorphic, since this is the only case we will be interested in. Implicit in the set-up are matching orderings of the components of the $\partial_- H_i$ in the case that $|\partial_- H_i| > 1$. This will be important when we derive a Heegaard-page structure from a Heegaard splitting below. See also Remark 2.11.

A Heegaard splitting $(\Sigma; H_1, H_2)$ with $H_i \cong H_{g,\boldsymbol{p},\boldsymbol{f}}$ is called (m, n)–*standard* if there are cut systems $\mathscr{D}_i = \{D_i^l\}_{l=1}^{n-1+g-p}$ for the H_i such that

1. For $1 \le l \le n - 1$, we have $\partial D_1^l = \partial D_2^l$, and this curve is separating;
2. For $n \le l \le m + n - 1$, we have $\partial D_1^l = \partial D_2^l$, and this curve is nonseparating; and
3. For $m + n \le l, l' \le g - p$, we have $|\partial D_1^l \cap \partial D_2^{l'}|$ given by the Kronecker delta $\delta_{l,l'}$, and the curves ∂D_1^l and ∂D_2^l are nonseparating.

A Heegaard diagram $(\Sigma; \alpha_1, \alpha_2)$ is called (m, n)–*standard* if $\alpha_i = \partial \mathscr{D}_i$ for cut systems $\mathscr{D}_i$ satisfying these three properties. See Figure 2, left, for an example. In a sense, a standard Heegaard splitting is a "stabilized double". The following lemma makes this precise.

Lemma 2.5 Let $(\Sigma; H_1, H_2)$ be a (m, n)–standard Heegaard splitting with $H_i \cong H_{g,\boldsymbol{p},\boldsymbol{f}}$. Then

$$(\Sigma; H_1, H_2) = \left(\underset{j=1}{\overset{n}{\#}} ((\Sigma')^j; (H_1')^j, (H_2')^j) \right) \# (\Sigma''; H_1'', H_2''),$$

where $(H_1')^j \cong (H_2')^j \cong H_{p_j, f_j}$ for each $j = 1, \ldots, n$, and $(\Sigma''; H_1'', H_2'')$ is the standard genus $g - p$ Heegaard surface for $\#^m(S^1 \times S^2)$.

Proof   Consider the n regions of Σ cut out by the $n-1$ separating curves that bound in each compression body. After a sequence of handleslides, we can assume that all of the nonseparating curves of the α_i are contained in one of these regions. Once this is arranged, there is a separating curve δ in $\Sigma \setminus \nu(\alpha_1 \cup \alpha_2)$ that cuts off a subsurface Σ'' such that Σ'' has only one boundary component (the curve δ) and $g(\Sigma'') = g - p$. Since δ bounds in each of H_1 and H_2, we have that $(\Sigma; H_1, H_2) = (\Sigma'; H_1', H_2') \#_\delta (\Sigma''; H_1'', H_2'')$, such that the latter summand is the standard splitting of $\#^m(S^1 \times S^2)$, as claimed. The fact that the regions of Σ' cut out by the separating curves that bound in both handlebodies contain no other curves of the α_i means that these curves give the connected sum decomposition

$$(\Sigma'; H_1', H_2') = \Big(\overset{n}{\underset{j=1}{\#}} ((\Sigma')^j; (H_1')^j, (H_2')^j) \Big)$$

that is claimed. $\square$

Let H_1 and H_2 be two copies of $H_{g,\boldsymbol{p},\boldsymbol{f}}$, and let $h\colon \partial_+ H_1 \to \partial_+ H_2$ be a diffeomorphism. Let Y be the closed three-manifold obtained as the union of H_1 and H_2 along their boundaries such that $\partial_+ H_1$ and $\partial_+ H_2$ are identified via h and $\partial_- H_1$ and $\partial_- H_2$ are identified via the identity on $\partial_- H_{g,\boldsymbol{p},\boldsymbol{f}}$. The manifold Y is called a *Heegaard double* of $H_{g,\boldsymbol{p},\boldsymbol{f}}$ along h, and was introduced by Gompf, Scharlemann, and Thompson [13, Definition 4.4]. We say that a Heegaard double Y is (m, n)–*standard* if the Heegaard splitting $(\Sigma; H_1, H_2)$ is (m, n)–standard. Let $Y_{g,\boldsymbol{p},\boldsymbol{f}}$ denote the Heegaard double of a standard Heegaard splitting whose compression bodies are $H_{g,\boldsymbol{p},\boldsymbol{f}}$. The uniqueness of $Y_{g,\boldsymbol{p},\boldsymbol{f}}$ is justified by the following lemma, which is proved with slightly different terminology than that of [6, Corollary 14].

Lemma 2.6   *Let $M = H_1 \cup_\Sigma \overline{H}_2$ be a standard Heegaard splitting with $H_i \cong H_{g,\boldsymbol{p},\boldsymbol{f}}$. Then there is a unique (up to isotopy rel-∂) diffeomorphism $\mathrm{Id}_{(M,\Sigma)}\colon \partial_- H_1 \to \partial_- H_2$ such that the identification space $M/_{x \sim \mathrm{Id}_{(M,\Sigma)}(x)}$, where $x \in \partial_- H_1$, is diffeomorphic to the standard Heegaard double $Y_{g,\boldsymbol{p},\boldsymbol{f}}$.*

We now identify the total space of a standard Heegaard double. Let $\mathrm{Id}_{p_j, f_j}\colon \Sigma_{p_j, f_j} \to \Sigma_{p_j, f_j}$ be the identity map, and let $M_{\mathrm{Id}_{p_j, f_j}}$ be the total space of the abstract open-book $(\Sigma_{p_j, f_j}, \mathrm{Id}_{p_j, f_j})$. See Section 2.8, especially Example 2.16, for definitions and details regarding open-book decompositions.

Lemma 2.7   *There is a decomposition*

$$Y_{g,\boldsymbol{p},\boldsymbol{f}} = \Big(\overset{n}{\underset{j=1}{\#}} M_{\mathrm{Id}_{p_j, f_j}} \Big) \# (\#^m(S^1 \times S^2)),$$

such that Σ restricts to a page in each of the first n summands and to a Heegaard surface in the last summand. Moreover,

$$M_{\mathrm{Id}_{p_j, f_j}} \cong \#^{2p_j + f_j - 1}(S^1 \times S^2),$$

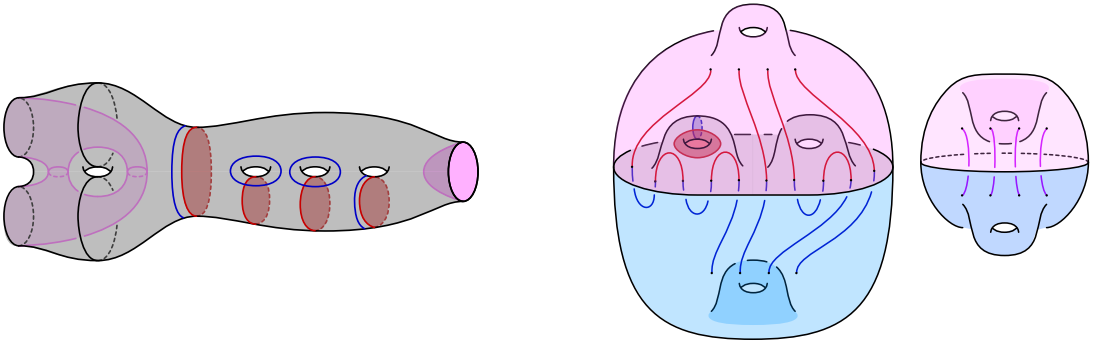*so $Y_{g,\boldsymbol{p},\boldsymbol{f}} \cong \#^k(S^1 \times S^2)$, with $k = 2p + f - n + m$.*

Figure 2: Left: a $(1, 2)$–standard Heegaard diagram for the standard Heegaard double $Y_{4,(1,0),(2,1)}$. Right: a schematic showing the standard Heegaard double $Y_{2,1,1}$, containing a $(3, 4)$–bridge splitting for an unlink; the unlink has no flat components and four vertical components.

Proof Consider the abstract open-book $(\Sigma_{p_j, f_j}, \mathrm{Id}_{p_j, f_j})$, and let $M_{\mathrm{Id}_{p_j, f_j}}$ denote the total space of this abstract open-book. Pick two pages, P_1 and P_2, of the open-book decomposition of $M_{\mathrm{Id}_{p_j, f_j}}$, and consider the two lensed cobordisms cobounded thereby. Each of these pieces is a handlebody of genus $2p_j + f_j - 1$, since it is diffeomorphic to H_{p_j, f_j}. A collection of arcs decomposing the page into a disk gives rise to a cut system for either handlebody, but these cut systems have the same boundary. The object described is a genus $2p_j + f_j - 1$ (symmetric) Heegaard splitting for $\#^{2p_j + f_j - 1}(S^1 \times S^2)$. The rest of the proof follows from Lemma 2.5. $\square$

Let Y be a standard Heegaard double. We consider the lensed compression bodies H_1 and H_2 as embedded submanifolds of Y in the following way, which is a slight deviation from the way they naturally embed in the Heegaard double. For $i = 1, 2$, let P_i^j denote the result of a slight isotopy of $\partial_- H_i^j$ into H_i along the product structure induced locally by the lensed cobordism structure of H_i. Let Y_1^j denote the lensed product cobordism cobounded by P_1^j and P_2^j. In this way, we think of the Heegaard double Y as divided into three regions: H_1, H_2, and $\bigsqcup_j Y_1^j$, each of whose connected components is a lensed compression body. The union of H_1 and H_2 along their common boundary, which we denote by Σ is a standard Heegaard splitting, and each Y_1^j is the product lensed cobordism H_{p_j, f_j}. See Figure 2, right, as well as Figure 5, for a schematic illustration of this structure. We call this decomposition a (*standard*) *Heegaard-page structure* and note that it is determined by the Heegaard splitting data (Σ, H_1, H_2), by Lemma 2.6.

2.5 Trivial tangles

A *tangle* is a pair $(H, \mathcal{T})$, where H is a compression body and $\mathcal{T}$ is a collection of neatly embedded arcs in H, called *strands*. Let Φ be a standard Morse function for H. After an ambient isotopy of $\mathcal{T}$ rel-∂, we can assume that Φ restricts to $\mathcal{T}$ to give a Morse function $\Phi|_{\mathcal{T}} \colon \mathcal{T} \to [-1, 3]$ such that each local maximum of $\mathcal{T}$ maps to $1 \in [-1, 3]$ and each local minimum maps to $0 \in [-1, 3]$. We have arranged that Φ be self-indexing on H and when restricted to $\mathcal{T}$.

A strand $\tau \subset \mathcal{T}$ is called *vertical* if τ has no local minimum or maximum with respect to $\Phi|_{\mathcal{T}}$, and is called *flat* if τ has a single local extremum, which is a maximum. Note that vertical strands have one boundary point in each of $\partial_+ H$ and $\partial_- H$, while flat strands have both boundary points in $\partial_+ H$. A tangle $\mathcal{T}$ is called *trivial* if it is isotopic rel-∂ to a tangle all of whose strands are vertical or flat. Such a tangle with b flat strands and v vertical strands is called an (b, v)–*tangle*, with the condition that it be trivial implicit in the terminology. More precisely, if $H \cong H_{g,\boldsymbol{p},\boldsymbol{f}}$, then we have an ordered partition of the vertical strands determined by which component Σ_{p_j, b_j} of $\partial_- H \cong \Sigma_{\boldsymbol{p},\boldsymbol{f}}$ contains the top-most endpoint of each vertical strand, and we can more meticulously describe $\mathcal{T}$ as an $(b, \boldsymbol{v})$–*tangle*. See Figure 1 for three examples of trivial tangles in lensed compression bodies.

Remark 2.8 In this paper, any tangle $(H, \mathcal{T})$ with $\partial_+ H$ disconnected will not contain flat strands. Moreover, such an H will always be a spread $(Y_i, \beta_i) \cong (\Sigma_{\boldsymbol{p},\boldsymbol{f}}, \boldsymbol{y}) \times I$, with β_i a geometric braid; see below. Therefore, we will never partition the flat strands of $\mathcal{T}$.

There is an obvious model tangle $(H_{g,\boldsymbol{p},\boldsymbol{f}}, \mathcal{T}_{b,\boldsymbol{v}})$ that is a lensed cobordism from $(\Sigma_{g,\boldsymbol{f}}, \boldsymbol{x}_{2b+v})$ to $(\Sigma_{\boldsymbol{p},\boldsymbol{f}}, \boldsymbol{y}_{\boldsymbol{v}})$ in which the first $2b$ points of $\boldsymbol{x}_{2b+v}$ are connected by slight push-ins of arcs in $\Sigma_{g,\boldsymbol{f}}$, and the final v rise vertically to $\Sigma_{\boldsymbol{p},\boldsymbol{f}}$, as prescribed by the standard height function on $H_{g,\boldsymbol{p},\boldsymbol{f}}$ and the ordered partitions. The points $\boldsymbol{x}_{2b+v}$ are called *bridge points*. A pair $(H, \mathcal{T})$ is determined up to diffeomorphism by the parameters g, b, $\boldsymbol{p}$, $\boldsymbol{f}$, and $\boldsymbol{v}$, and we refer to any tangle with these parameters as a $(g, b; \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–*tangle*. Note that this diffeomorphism can be assumed to be supported near $\partial_+ H$ and can be understood as a braiding of the bridge points $\boldsymbol{x}_{2b+v}$. For this reason, we consider trivial tangles up to isotopy rel-∂, and we think of each such tangle as having a fixed identification of the subsurface $(\Sigma_{g,b}, \boldsymbol{x}_{2b+v})$ of its boundary.

Let τ be a strand of a trivial tangle $(H, \mathcal{T})$. Suppose first that τ is flat. A *bridge semidisk for τ* is an embedded disk $D_\tau \subset H$ satisfying $\partial D_\tau = \tau \cup \tau^*$, where τ^* is an arc in $\partial_+ H$ with $\partial \tau^* = \partial \tau$, and $D_\tau \cap \mathcal{T} = \tau$. The arc τ^* is called a *shadow* for τ. Now suppose that τ is vertical. A *bridge triangle for τ* is an embedded disk $D_\tau \subset H$ satisfying $\partial D_\tau = \tau \cup \tau^* \cup \tau^-$, where τ^* (resp. τ^-) is an arc in $\partial_+ H$ (resp. $\partial_- H$) with one endpoint coinciding with an endpoint of τ and the other endpoint on $\partial(\partial_+ H)$, coinciding with the other endpoint of τ^- (resp. τ^*), and $D_\tau \cap \mathcal{T} = \tau$.

Remark 2.9 The existence of a bridge triangle for a vertical strand τ requires that $\partial_- H$ have boundary; there is no notion of a bridge triangle for a vertical strand in a compression body cobounded by closed surfaces. In this paper, if $\partial_+ H$ is ever closed, H will be a handlebody and will not contain vertical strands, so bridge semidisks and triangles will always exist for trivial tangles that we consider.

Given a trivial tangle $(H, \mathcal{T})$, a *bridge disk system for $\mathcal{T}$* is a collection Δ of disjoint disks in H, each component of which is a bridge semidisk or triangle for a strand of $\mathcal{T}$, such that Δ contains precisely one bridge semidisk or triangle for each strand of $\mathcal{T}$.

Lemma 2.10 *Let $(H, \mathcal{T})$ be a trivial tangle such that either $\partial_+ H$ has nonempty boundary or $\mathcal{T}$ contains no vertical strands. Then there is a bridge disk system Δ for $\mathcal{T}$.*

Proof There is a diffeomorphism from $(H, \mathcal{T})$ to $(H_{g,p,f}, \mathcal{T}_{b,v})$, as discussed above. This latter tangle has an obvious bridge disk system: the "slight push-in" of each flat strand sweeps out a disjoint collection of bridge semidisks for these strands, while the points $x \in x_{2b+v}$ corresponding to vertical strands can be connected to $\partial \Sigma_{g,f}$ via disjoint arcs, the vertical traces of which are disjoint bridge triangles for the vertical strands. Pulling back this bridge system to $(H, \mathcal{T})$ using the inverse diffeomorphism completes the proof.   □

We will refer to a $(0, v)$–tangle as a *vertical v–tangle* and to a $(b, 0)$–tangle as a *flat b–tangle*. In the case that $\mathcal{T}$ is a vertical tangle in a spread $H \cong H_{p,f}$, we call $\mathcal{T}$ a *v–thread* and call the pair $(H, \mathcal{T})$ a (p, f, v)–*spread*. Note that a (p, f, v)–spread is simply a lensed geometric (surface) braid; in particular, a $(0, 1, v)$–spread is a lensed geometric braid $(D^2 \times I, \beta)$.

2.6   Bridge splittings

Let K be a neatly embedded one-manifold in a three-manifold M. A *bridge splitting* of K is a decomposition
$$(M, K) = (H_1, \mathcal{T}_1) \cup_{(\Sigma, x)} \overline{(H_2, \mathcal{T}_2)},$$
where $(\Sigma; H_1, H_2)$ is a Heegaard splitting for M and $\mathcal{T}_i \subset H_i$ is a trivial tangle. If $\mathcal{T}_1$ is a trivial (b, v)–tangle, then we require that $\mathcal{T}_2$ be a trivial (b, v)–tangle, and we call the decomposition a $(g, p, f; b, v)$–*bridge splitting*. A one-manifold $K \subset M$ is in (b, v)–*bridge position* with respect to a Heegaard splitting of M if K intersects the compression bodies H_i as a (b, v)–tangle.

Remark 2.11 As we have assumed a correspondence between the components of the $\partial_- H_i$ (see Remark 2.4), we can require that the partitions of the vertical strands of the $\mathcal{T}_i$ respect this correspondence. This is the sense in which both $\mathcal{T}_i$ are (b, v)–tangles. This will be important when we turn a bridge splitting into a bridge-braid decomposition below.

More generally, we say that a bridge splitting is *standard* if the underlying Heegaard splitting
$$M = H_1 \cup_\Sigma \overline{H}_2$$
is standard (as defined in Section 2.4 above) and there are collections of bridge semidisks Δ_i for the flat strands of the tangles $\mathcal{T}_i$ whose corresponding shadows $\mathcal{T}_i^*$ have the property that $\mathcal{T}_1^* \cup_x \mathcal{T}_2^*$ is an embedded collection of polygonal arcs and curves. As a consequence, if (M, K) admits a standard bridge splitting, then K is the split union of an unlink (with one component corresponding to each polygonal curve of shadow arcs) with a braid (with one strand corresponding to each polygonal arc of shadows arcs). As described in Lemma 2.5, the ambient manifold M is a connected sum of copies of surfaces cross intervals and copies of $S^1 \times S^2$.

Consider the special case that M is the trivial lensed cobordism between $\partial_- H_1$ and $\partial_- H_2$ and $K \subset M$ is a v–braid — ie isotopic rel-∂ so that it intersects each level surface of the product lensed cobordism transversely. (Note that the $\partial_- H_i$ are necessarily connected, since Σ always is.) If $\Sigma = \partial_+ H_1$ defines a standard bridge splitting of (M, K), we refer to it as a *b–perturbing of a v–braid*.

Let $(H_1, \mathcal{T}_1)$ and $(H_2, \mathcal{T}_2)$ be two copies of the model trivial tangle $(H_{g,\boldsymbol{p},\boldsymbol{f}}, \mathcal{T}_{b,\boldsymbol{v}})$, and let

$$h \colon \partial_+(H_1, \mathcal{T}_1) \to \partial_+(H_2, \mathcal{T}_2)$$

be a diffeomorphism. Let (Y, L) be the pair obtained as the union of $(H_1, \mathcal{T}_1)$ and $(H_2, \mathcal{T}_2)$, where the boundaries $\partial_+(H_i, \mathcal{T}_i)$ are identified via h and the boundaries $\partial_-(H_i, \mathcal{T}_i)$ are identified via the identity map of $\partial_-(H_{g,\boldsymbol{p},\boldsymbol{f}}, \mathcal{T}_{b,\boldsymbol{v}})$. We call the pair (Y, L) a *bridge double* of $(H_{g,\boldsymbol{p},\boldsymbol{f}}, \mathcal{T}_{b,\boldsymbol{v}})$ along h. Note that a component of L can be referred to as *flat* or *vertical* depending on whether or not is disjoint from $\partial_- H_i$. We say that the bridge double is *standard* if:

(1) The bridge splitting $(H_1, \mathcal{T}_1) \cup_{(\Sigma, \boldsymbol{x})} \overline{(H_2, \mathcal{T}_2)}$ is standard.

(2) L has exactly v vertical components. In other words, each component of L hits $\partial_- H_i$ exactly once or not at all.

(3) L is an unlink.

Note that it follows that the vertical components of L are isotopic to meridians for the curve $\partial\Sigma$.

Let $(Y_{g,\boldsymbol{p},\boldsymbol{f}}, L_{b,\boldsymbol{v}})$ denote the bridge double of a standard bridge splitting with $(H_i, \mathcal{T}_i) \cong (H_{g,\boldsymbol{p},\boldsymbol{f}}, \mathcal{T}_{b,\boldsymbol{v}})$. The uniqueness of the *standard bridge double* $(Y_{g,\boldsymbol{p},\boldsymbol{f}}, L_{b,\boldsymbol{v}})$ is given by the following lemma, which generalizes Lemma 2.6 above.

Lemma 2.12 *Let* $(M, K) = (H_1, \mathcal{T}_1) \cup_{(\Sigma, \boldsymbol{x})} \overline{(H_2, \mathcal{T}_2)}$ *be a standard bridge splitting with* $(H_i, \mathcal{T}_i) \cong (H_{g,\boldsymbol{p},\boldsymbol{f}}, \mathcal{T}_{b,\boldsymbol{v}})$. *Then there is a unique (up to isotopy rel-∂) diffeomorphism*

$$\mathrm{Id}_{(M,K,\Sigma)} \colon \partial_-(H_1, \mathcal{T}_1) \to \partial_-(H_2, \mathcal{T}_2)$$

such that the identification space $(M, K)/_{x \sim \mathrm{Id}_{(M,K,\Sigma)}(x)}$, *where* $x \in \partial_-(H_1, \mathcal{T}_1)$, *is diffeomorphic to the standard bridge double* $(Y_{g,\boldsymbol{p},\boldsymbol{f}}, L_{b,\boldsymbol{v}})$.

Proof Let (M, K) be a standard bridge splitting. Suppose (Y, L) is the bridge double obtained via the gluing map $\mathrm{Id}_{(M,\Sigma)} \colon \partial_- H_1 \to \partial_- H_2$, which is determined uniquely up to isotopy rel-∂ by Lemma 2.6. The claim that must be justified is that $\mathrm{Id}_{(M,\Sigma)}$ is unique up to isotopy rel-∂ when considered as a map of pairs $\partial_-(H_1, \boldsymbol{y}_1) \to \partial_-(H_2, \boldsymbol{y}_2)$

Criterion (2) of a standard bridge double above states that K must close up to have v vertical components, where v is the number of vertical strands in the splitting (M, K). It follows that $\mathrm{Id}_{(M,\Sigma)}$ restricts to the identity permutation as a map $\boldsymbol{y}_1 \to \boldsymbol{y}_2$ — ie the end of a vertical strand in $\boldsymbol{y}_1$ must get matched with the end of the same strand in $\boldsymbol{y}_2$.

Let $(M, K)^\circ$ denote the pair obtained by deperturbing (in the classical, bridge-splitting-theoretic sense) the vertical arcs of K so that they have no local extrema, then removing tubular neighborhoods of them. Note that $(M, K)^\circ$ is a standard bridge splitting (of the flat components of K) of type $(g, \boldsymbol{p}, \boldsymbol{f}'; b', 0)$. The restriction $\mathrm{Id}^\circ_{(M, \Sigma)}$ to $(\partial_- H_1)^\circ$ is the identity on $\partial(\partial_- H_1)^\circ$, so we can apply Lemma 2.6 to conclude that $\mathrm{Id}^\circ_{(M, \Sigma)}$ is unique up to isotopy rel-∂. Since $\mathrm{Id}^\circ_{(M, \Sigma)}$ extends uniquely to a map $\mathrm{Id}_{(M, \Sigma, K)}$ of pairs, as desired, we are done. $\qquad\square$

Finally, consider a standard bridge double $(Y_{g, \boldsymbol{p}, \boldsymbol{f}}, L_{b, \boldsymbol{v}})$, and recall the Heegaard-page structure on $Y_{g, \boldsymbol{p}, \boldsymbol{f}}$. This induces a structure on L that we call a *bridge-braid structure*. In particular,

(1) $\mathcal{T}_i = L \cap H_i$ is a $(b, \boldsymbol{v})$–tangle, and

(2) $\beta_1^j = L \cap Y_1^j$ is a v_j–braid.

2.7 Disk-tangles

Let Z_k denote the four-dimensional 1–handlebody $\natural^k(S^1 \times B^3)$. Given nonnegative integers p, f, m, and n such that $k = 2p + f - 1 + m$ and ordered partitions $\boldsymbol{p}$ and $\boldsymbol{f}$ of p and f of length n, there is a natural way to think of Z_k as a lensed cobordism from the spread $Y_1 = H_{\boldsymbol{p}, \boldsymbol{f}}$ to the (m, n)–standard Heegaard splitting $(\Sigma; H_1, H_2) = (\Sigma_{g, f}; H_{g, f}, H_{g, f})$. Starting with $Y_1 \times [0, 1]$, attach $m + n - 1$ four-dimensional 1–handles to $Y_1 \times \{1\}$ so that the resulting four-manifold is connected. The three-manifold resulting from this surgery on $Y_1 \times \{1\}$ is $H_1 \cup_\Sigma \overline{H}_2$, and the induced structure on ∂Z_k is that of the standard Heegaard-page structure on $Y_{g; \boldsymbol{p}, \boldsymbol{f}}$. With this extra structure in mind, we denote this distinguished copy by Z_k by $Z_{g, k; \boldsymbol{p}, \boldsymbol{f}}$.

A *disk-tangle* is a pair $(Z, \mathscr{D})$ where $Z \cong Z_k$ and $\mathscr{D}$ is a collection of neatly embedded disks. A disk-tangle is called *trivial* if $\mathscr{D}$ can be isotoped rel-∂ to lie in ∂Z.

Proposition 2.13 *Let $\mathscr{D}$ and $\mathscr{D}'$ be trivial disk-tangles in Z. If $\partial \mathscr{D} = \partial \mathscr{D}'$, then $\mathscr{D}$ and $\mathscr{D}'$ are isotopic rel-∂ in Z.*

Proof Then case when $Z \cong B^4$ is a special case of a more general result of Livingston [24], and is also proved in [19]. See [28] for the general case. $\qquad\square$

A trivial disk-tangle $(Z, \mathscr{D})$ inherits extra structure along with $Z_{g, k; \boldsymbol{p}, \boldsymbol{f}}$, since we can identify $\partial \mathscr{D}$ with an unlink L in standard $(b, \boldsymbol{v})$–bridge position in $Y_{g; \boldsymbol{p}, \boldsymbol{f}}$. In this case, a disk $D \subset \mathscr{D}$ is called *vertical* (resp. *flat*) if it corresponds to a vertical (resp. flat) component of L. With this extra structure in mind, we call a trivial disk-tangle a $(c, \boldsymbol{v})$–*disk-tangle* and denote it by $\mathscr{D}_{c, \boldsymbol{v}}$, where c denotes the number of flat components of $\mathscr{D}$ and $\boldsymbol{v}$ denotes the partition numbers of vertical components. Note that $\mathscr{D}_{c, \boldsymbol{v}}$ is a tangle of $c + v$ disks. We call the pair $(Z_{g, k; \boldsymbol{p}, \boldsymbol{f}}, \mathscr{D}_{c; \boldsymbol{v}})$ a $(g, k, c; \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–*disk-tangle*. Note that Proposition 2.13 respects this extra structure, since part of the hypothesis was that the two disk systems have the same boundary. See Figure 3 for a schematic illustration.
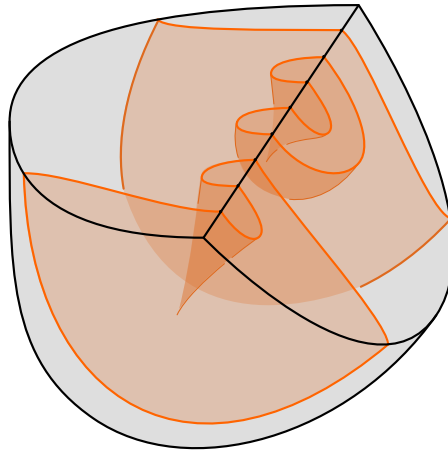
Figure 3: A schematic of the disk-tangle $\mathcal{D}_{1,2}$, which contains one flat component and two vertical components. Note that the 3–component unlink on the boundary is in $(3, 2)$–bridge position with respect to the standard Heegaard double $Y_{0,0,1}$ for the 3–sphere.

The special structure on $Z_{g,k;\boldsymbol{p},\boldsymbol{f}}$ described above induces a special Morse function $\Phi\colon Z \to \mathbb{R}$ with $m + n - 1$ critical points, all of which are index one. The next lemma characterizes trivial disk-tangles with respect to this standard Morse function.

Lemma 2.14 *Let $Z = Z_{g,k;\boldsymbol{p},\boldsymbol{f}}$, and let $\mathcal{D} \subset Z$ be a collection of neatly embedded disks with $\partial\mathcal{D} \cap Y_1$ a $\boldsymbol{v}$–thread. Suppose the restriction $\Phi_{\mathcal{D}}$ of Φ to $\mathcal{D}$ has c critical points, each of which is index zero. Then $\mathcal{D}$ is a $(c, \boldsymbol{v})$–disk-tangle for some ordered partition $\boldsymbol{v}$ of $v = |\mathcal{D}| - c$.*

Proof We parametrize $\Phi\colon Z \to \mathbb{R}$ so that $\Phi(Z) = [0, 1.5]$, $\Phi^{-1}(0) = Y_1 \setminus \nu(P_1 \cup P_2)$,

$$\Phi^{-1}(1.5) = (H_1 \cup_\Sigma \overline{H}_2) \setminus \nu(\overline{P}_1 \cup P_2),$$

and $\Phi(x) = 0.5$ for each critical point $x \in Z$ of Φ.

Let Γ denote the cores of the 1–handles of Z. By a codimension argument, we can assume, after a small perturbation of Φ that doesn't introduce any new critical points, that $\mathcal{D}$ is disjoint from a neighborhood $\nu(\Gamma) \cup Y_1 \times [0, 1]$. Thus, we can assume that $\Phi_{\mathcal{D}}(x) = 1.0$ for any critical point $x \in \mathcal{D}$ of $\Phi_{\mathcal{D}}$.

First, note that $0 \leq c \leq |\mathcal{D}|$; each connected component of $\mathcal{D}$ can have at most one minimum, since $\Phi_{\mathcal{D}}$ has no higher-index critical points. Let $\{D_i\}_{i=1}^c \subset \mathcal{D}$ denote the subcollection of disks in $\mathcal{D}$ that contain the index zero critical points of $\Phi_{\mathcal{D}}$. We claim that $D = \bigcup_{i=1}^c D_i$ is a $(c, 0)$–disk-tangle. We will now proceed to construct the required boundary-parallelism.

Consider the moving picture of the intersection $D_{\{t\}}$ of D with the cross-section $Z_{\{t\}} = \Phi^{-1}(1 + t)$ for $t \in [0, 0.5]$. This movie shows the birth of a c–component unlink L from c points at time $t = 0$, followed by an ambient isotopy of L as t increases. Immediately after the birth, say $t = \epsilon$, we have that the subdisks $D_{[1,1+\epsilon]} = D \cap \Phi^{-1}([1, 1 + \epsilon])$ of D are clearly boundary-parallel to a spanning collection

of disks E_ϵ for $L_\epsilon = D_{\{1+\epsilon\}}$. Now, we simply push this spanning collection of disks E_ϵ along through the isotopy taking L_ϵ to ∂D. Because this isotopy is ambient, the traces of the disks of E_ϵ are disjoint, thus they provide a boundary parallelism for D, as desired.

It remains to see that the collection D'' of disks in $\mathcal{D}$ containing no critical points of $\Phi_{\mathcal{D}}$ are also boundary parallel. Note however, that they will not be boundary parallel into $\Phi^{-1}(1.5)$, as before.

Let $\beta = D'' \cap Y_1$; by hypothesis, (Y_1, β) is a $(\boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–spread, ie Y_1 is a product lensed bordism (a spread) $H_{\boldsymbol{p},\boldsymbol{f}}$ and β is a vertical $\boldsymbol{v}$–tangle (a $\boldsymbol{v}$–thread) therein. Similar to before, we can assume that D'' is disjoint from a small neighborhood of the cores of the 1–handles.

Since D'' contains no critical points, it is vertical in the sense that we can think of it as the trace of an ambient isotopy of β in Y_1 as t increases from $t = 0$ to $t = 0.5$, followed by the trace of an ambient isotopy of β in $H_1 \cup_\Sigma \overline{H}_2$ between $t = 0.5$ and $t = 1.5$. The change in the ambient space is not a problem, since D'' is disjoint form the cores Γ of the 1–handles, hence these isotopies are supported away from the four-dimensional critical points.

If Δ is any choice of bridge triangles for β in Y_1, then the trace of Δ under this isotopy gives a boundary-parallelism of D'', as was argued above. We omit the details in this case. $\square$

Note that the assumption that β be a thread was vital in the proof, as it gave the existence of Δ. If β contained knotted arcs, the vertical disk sitting over such an arc would not be boundary parallel. Similarly, if β contained closed components, the vertical trace would be an annulus, not a disk. The converse to the lemma is immediate, hence it provides a characterization of trivial disk-tangles.

We next show how a standard bridge splitting can be uniquely extended to a disk-tangle. The following lemma builds on portions of [6, Section 4].

Lemma 2.15 *Let* $(M, K) = (H_1, \mathcal{T}_1) \cup_{(\Sigma, \boldsymbol{x})} \overline{(H_2, \mathcal{T}_2)}$ *be a standard* $(g, \boldsymbol{p}, \boldsymbol{f}; b, \boldsymbol{v})$–*bridge splitting.* *There is a unique (up to diffeomorphism rel-∂) pair* $(Z, \mathcal{D})$, *diffeomorphic to* $(Z_{g,k;\boldsymbol{p},\boldsymbol{f}}, \mathcal{D}_{c,\boldsymbol{v}})$, *such that the bridge double structure on* $\partial(Z, \mathcal{D})$ *is the bridge double of* (M, K).

Proof By Lemma 2.12, there is a unique way to close (M, K) up and obtain its bridge double (Y, L). By Laudenbach and Poénaru [23], there is a unique way to cap off $Y \cong \#^k(S^1 \times S^2)$ with a copy of Z of Z_k. By Proposition 2.13, there is a unique way to cap off L with a collection $\mathcal{D}$ of trivial disks. Since these choice are unique (up to diffeomorphism rel-∂ and isotopy rel-∂, respectively), the pair $(Z, \mathcal{D})$ inherit the correct bridge double structure on its boundary, as desired. $\square$

2.8 Open-book decompositions and braidings of links

We follow Etnyre's lecture notes [9] to formulate the definitions of this subsection. Let Y be a closed, orientable three-manifold. An *open-book decomposition* of Y is a pair (B, π), where B is a link in M (called the *binding*) and $\pi : Y \setminus B \to S^1$ is a fibration such that $P_\theta = \pi^{-1}(\theta)$ is a noncompact surface (called

the *page*) with $\partial P_\theta = B$. Note that it is possible for a given link B to be the binding of nonisotopic (even nondiffeomorphic) open-book decomposition of Y, so the projection data π is essential in determining the decomposition.

An *abstract open-book* is a pair (P, ϕ), where P is an oriented, compact surface with boundary, and $\phi: P \to P$ is a diffeomorphism (called the *monodromy*) that is the identity on a collar neighborhood of ∂P. An abstract open-book (P, ϕ) gives rise to a closed three-manifold, called the *model manifold*, with an open-book decomposition in a straightforward way. Define

$$Y_\phi = (P \times_\phi S^1) \cup \left(\bigsqcup_{|\partial P|} S^1 \times D^2 \right),$$

where $P \times_\phi S^1$ denotes the mapping torus of ϕ, and Y_ϕ is formed from this mapping torus by capping off each torus boundary component with a solid torus such that each $p \times_\phi S^1$ gets capped off with a meridional disk for each $p \in \partial P$. (Note that $p \times_\phi S^1 = p \times S^1$ by the condition on ϕ near the boundary of P.) Our convention is that $P \times_\phi S^1 = P \times [0, 1] / (x, 1) \sim (\phi(x), 0)$ for all $x \in P$.

If we let B_ϕ denote the cores of the solid tori used to form Y_ϕ, then we see that $Y_\phi \setminus B_\phi$ fibers over S^1, so we get an open-book decomposition (B_ϕ, π_ϕ) for Y_ϕ. Conversely, an open-book decomposition (B, π) of a three-manifold M gives rise to an abstract open-book (P_π, ϕ_π) in the obvious way such that $(Y_{\phi_\pi}, B_{\phi_\pi})$ is diffeomorphic to (M, B).

We now recall an important example which appeared in Lemma 2.7.

Example 2.16 Consider the abstract open-book (P, ϕ), where $P = \Sigma_{p,f}$ is a compact surface of genus p with f boundary components and $\phi: P \to P$ is the identity map. The total space Y_ϕ of this abstract open-book is diffeomorphic to $\#^{2p+f-1}(S^1 \times S^2)$. To see this, simply note that the union of half of the pages gives a handlebody of genus $2p + f - 1$; since the monodromy is the identity, Y_ϕ is the symmetric double of this handlebody.

Harer described a set of moves that suffice to pass between open-book decompositions on a fixed three-manifold [14]. These include Hopf stabilization and destabilization, as well as a certain double-twisting operation, which was known to be necessary in order to change the homotopy class of the associated plane field. (Harer's calculus was recently refined in [30].) In fact, Giroux and Goodman proved that two open-book decompositions on a fixed three-manifold have a common Hopf stabilization if and only if the associated plane fields are homotopic [12]. For a trisection-theoretic account of this story, see [7].

Having introduced open-book decompositions, we now turn our attention to braided links. Suppose that $\mathcal{L} \subset Y$ is a link and (B, π) is an open-book decomposition on Y. We say that $\mathcal{L}$ is *braided with respect to* (B, π) if $\mathcal{L}$ intersects each page of the open-book transversely. We say that $(Y, \mathcal{L})$ is equipped with the structure of an *open-book braiding*. The *index* of the braiding is the number of times that $\mathcal{L}$ hits a given page. By the Alexander theorem [1] and the generalization due to Rudolph [31], any link can be braided with respect to any open-book in any three-manifold.
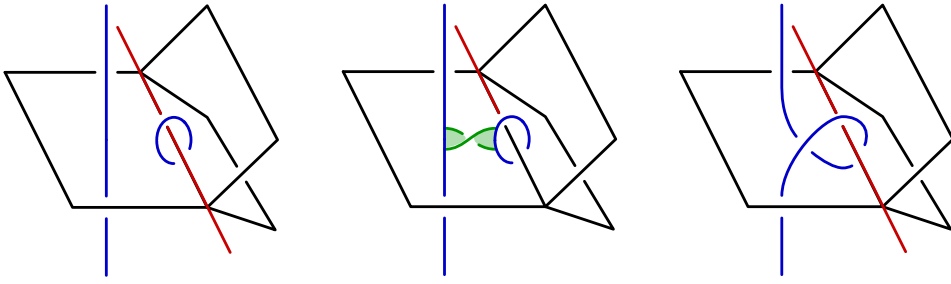
Figure 4: Markov stabilization, depicted as the banding of a braid to a meridian of the binding.

An *abstract open-book braiding* is a triple (P, y, ϕ), where P is an oriented, compact surface with boundary, $y \subset P$ is a collection of points, and $\phi \colon (P, y) \to (P, y)$ is a diffeomorphism. As with abstract open-books, this data gives rise to a manifold pair $(Y_\phi, \mathscr{L}_\phi)$, called the *model open-book braiding* of the abstract open-book braiding, where Y_ϕ has an open-book structure with binding B_ϕ and projection π_ϕ and $\mathscr{L}_\phi$ is braided with respect to (B_ϕ, π_ϕ). More precisely,

$$(Y_\phi, \mathscr{L}_\phi) = (P, y) \times_\phi S^1 = (P, y) \times [0, 1]/_{(x,0) \sim (\phi(x),1)}$$

for all $x \in P$. Conversely, a braiding of $\mathscr{L}$ about (B, π) gives rise in the obvious way to an abstract open-book braiding (P_π, ϕ_π) such that $(Y_{\phi_\pi}, \mathscr{L}_{\phi_\pi})$ is diffeomorphic to $(Y, \mathscr{L})$.

By the Markov theorem [25] or its generalization to closed 3–manifolds [32; 33], any two braidings of $\mathscr{L}$ with respect to a fixed open-book decomposition of Y can be related by an isotopy that preserves the braided structure, except at finitely many points in time at which the braiding is changed by a *Markov stabilization or destabilization*. We think of a Markov stabilization in the following way. Let J be a meridian for a component of the binding B of the open-book decomposition on Y, and let $\mathfrak{b}$ be a band connecting $\mathscr{L}$ to J such that the core of $\mathfrak{b}$ is contained in a page of the open-book decomposition and such that the link $\mathscr{L}' = \mathscr{L}_\mathfrak{b}$ resulting from the resolution of the band is braided about (B, π). We say that $\mathscr{L}'$ is obtained from $\mathscr{L}$ via a *Markov stabilization*, and we call the inverse operation *Markov destabilization*. (Markov destabilization can be thought of as attaching a vertical band to $\mathscr{L}'$ such that resolving the band has the effect of splitting off from $\mathscr{L}'$ a meridian for a binding component.) See Figure 4.

Suppose that $Y = Y^1 \sqcup \cdots \sqcup Y^n$ is the disjoint union of closed three-manifolds such that each Y^j is equipped with an open-book decomposition (B^j, π^j). Suppose that $\mathscr{L} = \mathscr{L}^1 \sqcup \cdots \mathscr{L}^n$ is a link such that $Y^j \subset Y^j$ is braided about (B^j, π^j). We say that $\mathscr{L}$ has *multiindex* $\boldsymbol{v} = (v^1, \ldots, v^n)$ if $\mathscr{L}^j$ has index v^j. We allow the possibility that $\mathscr{L}^j = \varnothing$ for any given j.

Remark 2.17 If Y is oriented, and we pick orientations on $\mathscr{L}$ and on a page P of (B, π), then we can associate a sign to each point of $\mathscr{L} \cap P$. By definition, if $\mathscr{L}$ is a knot, then each such point will have identical sign; more generally, connected components of $\mathscr{L}$ have this property. If the orientations of the points $\mathscr{L} \cap P$ all agree, then we say that the braiding is *coherently oriented*. If the orientations of these points disagree across components of $\mathscr{L}$, then we say that the braiding is *incoherently oriented*.

Our reason for considering incoherently oriented braidings is that sometimes a bridge trisection of a surface will induce a braiding of the boundary link that is incoherently oriented once the surface is oriented. A simple example of this, the annulus bounded by the $(2, 2n)$–torus link, will be explored in Examples 7.15 and 7.17. Even though some bridge trisections induce incoherently oriented braidings on the boundary link, it is always possible to find a bridge trisection of a surface such that the induced braiding is coherently oriented.

2.9 Formal definitions

Finally, we draw on the conventions laid out above to give formal definitions.

Definition 2.18 Let X be an orientable, connected four-manifold, and let

$$Y = \partial X = Y^1 \sqcup \cdots \sqcup Y^n,$$

where Y^j is a connected component of ∂X for each $j = 1, \ldots, n$. Let g, k^*, p, and f be nonnegative integers, and let $\boldsymbol{k}$, $\boldsymbol{p}$, and $\boldsymbol{f}$ be ordered partitions of type $(k^*, 3)$, (p, n), and $(b, n)^+$, respectively.

A $(g, \boldsymbol{k}; \boldsymbol{p}, \boldsymbol{f})$–*trisection* $\mathbb{T}$ of X is a decomposition $X = Z_1 \cup Z_2 \cup Z_3$ such that, for all $j = 1, \ldots, n$ and all $i \in \mathbb{Z}_3$,

(1) $Z_i \cong Z_{g, k_i; \boldsymbol{p}, \boldsymbol{f}}$,

(2) $Z_i \cap Z_{i+1} \cong H_{g; \boldsymbol{p}, \boldsymbol{f}}$,

(3) $Z_1 \cap Z_2 \cap Z_3 \cong \Sigma_{g, b}$, and

(4) $Z_i \cap Y^j \cong H_{\boldsymbol{p}, \boldsymbol{f}}$.

The four-dimensional pieces Z_i are called *sectors*, the three-dimensional pieces $H_i = Z_i \cap Z_{i-1}$ are called *arms*, and the central surface $\Sigma = Z_1 \cap Z_2 \cap Z_3$ is called the *core*. If $k_1 = k_2 = k_3 = k$, then $\mathbb{T}$ is described as a $(g, k; \boldsymbol{p}, \boldsymbol{f})$–*trisection* and is called *balanced*. Otherwise, $\mathbb{T}$ is called *unbalanced*. Similarly, if either of the ordered partitions $\boldsymbol{p}$ and $\boldsymbol{f}$ are balanced, we replace these parameters with the integers p/n and/or f/n, respectively. The parameter g is called the *genus* of $\mathbb{T}$. The surfaces $P_i^j = H_i \cap Y^j$ are called *pages*, and their union is denoted by P_i. The lensed product cobordisms $Y_i^j = Z_i \cap Y^j$ are called *spreads*, and their union is denoted by Y_i. The links $B^j = \Sigma \cap Y^j$ are called *bindings*, and their union is $B = \partial \Sigma$.

If X is oriented, we require that the orientation on Z_i induces the oriented decompositions

$$\partial Z_i = H_i \cup Y_i \cup \overline{H}_{i+1}, \quad \partial H_i = \Sigma \cup_B \overline{P}_i, \quad \partial Y_i = P_i \cup_B \overline{P}_{i+1}.$$

See Figure 5 (below) for a schematic illustrating these conventions.

Remarks 2.19   (1) If X is closed, then $n = 0$, $Y = \varnothing$, and $\mathbb{T}$ is a trisection as originally introduced by Gay and Kirby [10] and generalized slightly in [26].

(2) If X has a single boundary component, then $n = 1$, and $\mathbb{T}$ is a relative trisection as first described in [10] and later developed in [4], where gluing of such objects was studied, and in [6], where the diagrammatic aspect to the theory was introduced. The general case of multiple boundary components was recently developed in [5].

(3) Since $Y^j = Y_1^j \cup Y_2^j \cup Y_3^j$, with each $Y_i^j \cong H_{p_j, b_j}$, it follows that Y^j admits an open-book decomposition where P_i^j is a page for each $i \in \mathbb{Z}_3$ and B^j is the binding. This open-book decomposition is determined by $\mathbb{T}$, and the monodromy can be explicitly calculated from a relative trisection diagram [6].

(4) The triple (Σ, P_i, P_{i+1}) defines the standard Heegaard double structure on $\partial Z_i \cong Y_{g; p, f}$. It follows from Lemma 2.7 that $k_i = 2p + f - n + m_i$, where $(\Sigma; H_i, H_{i+1})$ is an (m_i, n)–standard Heegaard splitting. We call m_i the *interior* complexity of Z_i. Notice that g is bounded below by m_i and p, but not by f nor k_i.

Definition 2.20 Let $\mathbb{T}$ be a trisection of a four-manifold X. Let $\mathscr{F}$ be a neatly embedded surface in X. Let b, c^*, and v be nonnegative integers, and let $\boldsymbol{c}$ and $\boldsymbol{v}$ be ordered partitions of type $(c^*, 3)$ and (v, n), respectively. The surface $\mathscr{F}$ is in $(b, \boldsymbol{c}; \boldsymbol{v})$–*bridge trisected position* with respect to $\mathbb{T}$ (or is $(b, \boldsymbol{c}; \boldsymbol{v})$–*bridge trisected* with respect to $\mathbb{T}$) if, for all $i \in \mathbb{Z}_3$,

(1) $\mathscr{D}_i = Z_i \cap \mathscr{F}$ is a trivial $(c_i; \boldsymbol{v})$–disk-tangle in Z_i, and

(2) $\mathscr{T}_i = H_i \cap \mathscr{F}$ is a trivial $(b; \boldsymbol{v})$–tangle in H_i.

The disk components of the $\mathscr{D}_i$ are called *patches*, and the $\mathscr{T}_i$ are called *seams*. Let

$$\mathscr{L} = \partial \mathscr{F} = \mathscr{L}^1 \sqcup \cdots \sqcup \mathscr{L}^n,$$

where $\mathscr{L}^j = \mathscr{L} \cap Y^j$ is the link representing the boundary components of $\mathscr{F}$ that lie in Y^j. The pieces $\beta_i^j = \mathscr{L}^j \cap Z_i$ comprising the $\mathscr{L}_i$ are called *threads*.

If $\mathscr{F}$ is oriented, we require that the induced orientation of $\mathscr{D}_i$ induces the oriented decomposition

$$\partial \mathscr{D}_i = \mathscr{T}_i \cup \beta_i \cup \overline{\mathscr{T}}_{i+1}.$$

See Figure 5 (below) for a schematic illustrating these conventions.

The induced decomposition $\mathbb{T}_{\mathscr{F}}$ given by

$$(X, \mathscr{F}) = (Z_1, \mathscr{D}_1) \cup (Z_2, \mathscr{D}_2) \cup (Z_3, \mathscr{D}_3)$$

is called a $(g, \boldsymbol{k}, b, \boldsymbol{c}; \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–*bridge trisection* of $\mathscr{F}$ (or of the pair $(X, \mathscr{F})$). If $\mathbb{T}$ is balanced and $c_i = c$ for each $i \in \mathbb{Z}_3$, then $\mathbb{T}_{\mathscr{F}}$ is described as a $(g, k, b, c; \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–*bridge trisection* and is called *balanced*. Otherwise, $\mathbb{T}_{\mathscr{F}}$ is called *unbalanced*. Similarly, if the partition $\boldsymbol{v}$ is balanced, we replace this parameter with the integer v/n. The parameter b is called the *bridge number* of $\mathbb{T}_{\mathscr{F}}$.

Remarks 2.21   (1)   If X is a closed four-manifold, then $n = 0$, $\mathcal{L} = \varnothing$, and $\mathcal{F}$ is a closed surface in X. If $g = 0$, we recover the notion of bridge trisections originally introduced in [27], while the more general case of arbitrary g is treated in in [28].

(2)   If $\mathcal{L} \cap Y^j = \varnothing$ for some $j = 1, \ldots, n$, then $\mathcal{L}^j = \varnothing$. Equivalently, $v_j = 0$. If $\mathcal{L}^j$ is not empty, then

$$\mathcal{L}^j = \beta_1^j \cup \beta_2^j \cup \beta_3^j.$$

If follows that $\mathcal{L}^j$ is braided with index v_j with respect to the open-book decomposition (B^j, P_i^j) on Y^j induced by $\mathbb{T}$.

(3)   The link $L_i = \partial \mathcal{D}_i$ is in $(b, \boldsymbol{v})$–bridge position with respect to the standard Heegaard double structure on ∂Z_i.

(4)   The surface $\mathcal{F}$ has a cellular decomposition consisting of $(2b + 4v)$ 0–cells, $3v$ of which lie in the pages of ∂X; $(3b + 6v)$ 1–cells, $3v$ of which lie in the spreads of ∂X; and $(c_1 + c_2 + c_3 + 3v)$ 2–cells, $3v$ of which are vertical patches. It follows that the Euler characteristic of $\mathcal{F}$ is given as

$$\chi(\mathcal{F}) = c_1 + c_2 + c_3 + v - b.$$

(5)   Note that $c_i \geq b$, but that v is independent of b and the c_i.

We conclude this section with a key fact about bridge trisections. We refer to the union

$$(H_1, \mathcal{T}_2) \cup (H_2, \mathcal{T}_2) \cup (H_3, \mathcal{T}_3)$$

as the *spine* of the bridge trisection $\mathbb{T}$. Two bridge trisections $\mathbb{T}$ and $\mathbb{T}'$ for pairs $(X, \mathcal{F})$ and $(X, \mathcal{F}')$ are *diffeomorphic* if there is a diffeomorphism $\Psi: (X, \mathcal{F}) \to (X', \mathcal{F}')$ such that $\psi(Z_i, \mathcal{D}_i) = (Z'_i, \mathcal{D}'_i)$ for all $i \in \mathbb{Z}_3$. We consider spines up to diffeomorphism, and we note that such diffeomorphisms may induce braiding of the $\mathcal{T}_i$ near the P_i.

Proposition 2.22   *Two bridge trisections are diffeomorphic if and only if their spines are diffeomorphic.*

Proof   If Ψ is a diffeomorphism of bridge trisections $\mathbb{T}$ and $\mathbb{T}'$, then the restriction of Ψ to the spine of $\mathbb{T}$ is a diffeomorphism onto the spine of $\mathbb{T}'$. Conversely, suppose Ψ is a diffeomorphism from the spine of $\mathbb{T}$ to the spine of $\mathbb{T}'$ — ie $\Psi(H_i, \mathcal{T}_i) = (H'_i, \mathcal{T}'_i)$ for all $i \in \mathbb{Z}_3$. By Lemma 2.15, Ψ there is an extension of Ψ across $(Z_i, \mathcal{D}_i)$ that is uniquely determined up to isotopy fixing $(H_1, \mathcal{T}_i) \cup_{(\Sigma, \boldsymbol{x})} \overline{(H_{i+1}, \mathcal{T}_{i+1})}$ for each $i \in \mathbb{Z}_3$. It follows that Ψ extends to a diffeomorphism bridge trisections, as desired.   □

In light of this, we find that the four-dimensional data of a bridge trisection is determined by the three-dimensional data of its spine, a fact that will allow for the diagrammatic development of the theory in Sections 4 and 5.

Corollary 2.23   *A bridge trisection is determined uniquely by its spine.*

3 The four-ball setting

In this section, we restrict our attention to the study of surfaces in the four-ball. Moreover, we work relative to the standard genus zero trisection. These restrictions allow for a cleaner exposition than the general framework of Section 2 and give rise to a new diagrammatic theory for surfaces in this important setting.

3.1 Preliminaries and a precise definition

Here, we revisit the objects and notation introduced in Section 2 with the setting of B^4 in mind, culminating in a precise definition of a bridge trisection of a surface in B^4.

Let H denote the three-ball, and let B denote an equatorial curve on ∂H, which induces the decomposition

$$\partial H = \partial_+ H \cup_B \partial_- H$$

of the boundary sphere into two hemispheres. We think of H as being swept out by disks: smoothly isotope $\partial_+ H$ through H to $\partial_- H$. (Compare this description of H with the notion of a lensed cobordism from Section 2.2 and the development for a general compression body in Section 2.3.)

A trivial tangle is a pair $(H, \mathcal{T})$ such that H is a three-ball and $\mathcal{T} \subset H$ is a neatly embedded 1–manifold with the property that $\mathcal{T}$ can be isotoped until the restriction $\Phi_{\mathcal{T}}$ of the above Morse function to $\mathcal{T}$ has no minimum and at most one maximum on each component of $\mathcal{T}$. In other words, each component of $\mathcal{T}$ is a neatly embedded arc in H that is either *vertical* (with respect to the fibering of H by disks) or parallel into $\partial_+ H$. The latter arcs are called *flat*. We consider trivial tangles up to isotopy rel-∂. If $\mathcal{T}$ has v vertical strands and b flat strands, we call the pair $(H, \mathcal{T})$ a (b, v)–tangle. This is a special case of the trivial tangles discussed in Section 2.5.

Let H_1 and H_2 be three-balls, and consider the union $H_1 \cup_{\Sigma} \overline{H}_2$, where $\Sigma = \partial_+ H_1 = \partial_+ \overline{H}_2$. We consider this union of as a subset of the three-sphere Y so that $B = \partial \Sigma$ is an unknot and Σ, $\partial_- H_1$, and $\partial_- H_2$ are all disjoint disk fibers meeting at B. Let Y_1 denote

$$Y \setminus \mathrm{Int}(H_1 \cup_{\Sigma} \overline{H}_2),$$

and notice that Y_1 is simply an interval's worth of disk fibers for B, just like the H_i. We let Y denote the three-sphere with this extra structure, which we call the *standard Heegaard double* (see Section 2.4). Note that B can be thought of as the (unknotted) binding of an open-book decomposition of S^3 with disk page, with the pieces H_1, H_2, and Y_1 intersecting pairwise at pages and representing themselves lensed product cobordisms between these pages.

An unlink $L \subset Y$ is in (b, v)–*bridge position* with respect the standard Heegaard double structure if $L \cap H_i$ is a (b, v)–tangle, L is transverse to the disk fibers of Y_1, and each component of L intersects Y_1 in at most one arc. The v components of L that intersect Y_1 are called *vertical*, while the other b components are called *flat*.
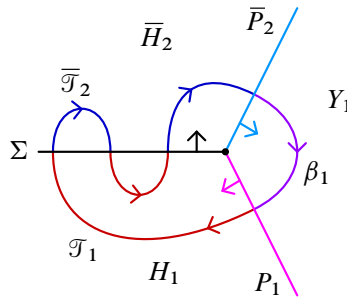
Figure 5: A schematic illustration of a standard Heegaard double, with orientation conventions for the constituent pieces of ∂Z_1 indicated.

Let Z denote the four-ball, with $\partial Z = Y$ regarded as the standard Heegaard double. A trivial disk-tangle is a pair $(Z, \mathcal{D})$ such that Z is a four-ball and $\mathcal{D}$ is a collection of neatly embedded disks, each of which is parallel into ∂Z. Note that the boundary $\partial \mathcal{D}$ is an unlink. If $\partial \mathcal{D}$ is in (b, v)–bridge position in $Y = \partial Z$, then the disk components of $\mathcal{D}$ are called *vertical* and *flat* in accordance with their boundaries. A (c, v)–*disk-tangle* is a trivial disk-tangle with c flat components and v vertical components.

Definition 3.1 Let $\mathscr{F}$ be a neatly embedded surface in B^4, and let $\mathbb{T}_0$ be the standard genus zero trisection of B^4. Let b and v be nonnegative integers, and let $\boldsymbol{c} = (c_1, c_2, c_3)$ be an ordered triple of nonnegative integers. The surface $\mathscr{F}$ is in $(b, \boldsymbol{c}; v)$–*bridge trisected position* with respect to $\mathbb{T}_0$ (or is $(b, \boldsymbol{c}; v)$–*bridge trisected* with respect to $\mathbb{T}_0$) if, for all $i \in \mathbb{Z}_3$,

(1) $\mathcal{D}_i = Z_i \cap \mathscr{F}$ is a trivial (c_i, v)–disk-tangle in the four-ball Z_i, and

(2) $\mathcal{T}_i = H_i \cap \mathscr{F}$ is a trivial (b, v)–tangle in the three-ball H_i.

The disk components of the $\mathcal{D}_i$ are called *patches*, and the $\mathcal{T}_i$ are called *seams*. Let $\mathcal{L} = \partial \mathscr{F}$. The braid pieces $\beta_i = \mathcal{L} \cap Z_i$ are called *threads*.

If $\mathscr{F}$ is oriented, we require that the induced orientation of $\mathcal{D}_i$ induces the oriented decomposition

$$\partial \mathcal{D}_i = \mathcal{T}_i \cup \beta_i \cup \overline{\mathcal{T}}_{i+1}.$$

The induced decomposition $\mathbb{T}_{\mathscr{F}}$ given by

$$(X, \mathscr{F}) = (Z_1, \mathcal{D}_1) \cup (Z_2, \mathcal{D}_2) \cup (Z_3, \mathcal{D}_3)$$

is called a $(b, \boldsymbol{c}, v)$–*bridge trisection* of $\mathscr{F}$ (or of the pair $(X, \mathscr{F})$). If $\mathbb{T}_{\mathscr{F}}$ is balanced and $c_1 = c_2 = c_3 = c$, then $\mathbb{T}_{\mathscr{F}}$ is a (b, c, v)–*bridge trisection* and is called *balanced*. Otherwise, $\mathbb{T}_{\mathscr{F}}$ is called *unbalanced*.

3.2 Band presentations

Let M be a three-manifold, and let J be a neatly embedded one-manifold in M. Let $\mathfrak{b}$ be a copy of $I \times I$ embedded in M, and denote by $\partial_1 \mathfrak{b}$ and $\partial_2 \mathfrak{b}$ the portions of $\partial \mathfrak{b}$ corresponding to $I \times \{-1, 1\}$ and

$\{-1, 1\} \times I$, respectively. We call such a $\mathfrak{b}$ a *band* for J if $\mathrm{Int}(\mathfrak{b}) \subset M \setminus J$ and $\partial \mathfrak{b} \cap J = \partial_1 \mathfrak{b}$. The arc of $\mathfrak{b}$ corresponding to $\{0\} \times I$ is called the *core* of $\mathfrak{b}$.

Let $J_{\mathfrak{b}}$ denote the one-manifold obtained by *resolving* the band $\mathfrak{b}$,

$$J_{\mathfrak{b}} = (J \setminus \partial_1 \mathfrak{b}) \cup \partial_2 \mathfrak{b}.$$

The band $\mathfrak{b}$ for J gives rise to a *dual band* $\mathfrak{b}^*$ that is a band for $J_{\mathfrak{b}}$, so $\partial_1 \mathfrak{b}^* = \partial_2 \mathfrak{b}$ and $\partial_2 \mathfrak{b}^* = \partial_1 \mathfrak{b}$. Note that, as embedded squares in M, we have $\mathfrak{b} = \mathfrak{b}^*$, though their cores are perpendicular. More generally, given a collection $\mathfrak{b}$ of disjoint bands for J, we denote by $J_{\mathfrak{b}}$ the *resolution* of all the bands in $\mathfrak{b}$. As above, the collection $\mathfrak{b}^*$ of dual bands is a collection of bands for $J_{\mathfrak{b}}$.

Definition 3.2 (band presentation) A *band presentation* is a 2–complex in S^3 defined by a triple $(\mathscr{L}, U, \mathfrak{b})$ as follows:

(1) $\mathscr{L} \subset S^3$ is a link;

(2) U is a split unlink in $S^3 \setminus \nu(\mathscr{L})$; and

(3) $\mathfrak{b}$ is a collection of bands for $\mathscr{L} \sqcup U$ such that $U' = (\mathscr{L} \sqcup U)_{\mathfrak{b}}$ is an unlink.

If U is the empty link, then we write $(\mathscr{L}, \mathfrak{b})$ and call the encoded 2–complex in S^3 a *ribbon presentation*.

We consider two band presentations to be *equivalent* if they are ambient isotopic as 2–complexes in S^3. Given a fixed link $\mathscr{L} \subset S^3$, two band presentations $(\mathscr{L}, U_1, \mathfrak{b}_1)$ and $(\mathscr{L}, U_2, \mathfrak{b}_2)$ are *equivalent rel-$\mathscr{L}$* if they are equivalent via an ambient isotopy that preserves $\mathscr{L}$ setwise. (In other words, $\mathscr{L}$ is fixed, although the attaching regions of $\mathfrak{b}$ are allowed to move along $\mathscr{L}$.)

Band presentations encode smooth, compact, neatly embedded surfaces in B^4 in a standard way. Before explaining this, we first fix some conventions that will be useful later. (Here, we follow standard conventions, as in [20; 21; 27; 28].)

Let $h \colon B^4 \to [0, 4]$ be a standard Morse function on B^4 — ie h has a single critical point, which is definite of index zero and given by $h^{-1}(0)$, while $h^{-1}(4) = \partial B^4 = S^3$. For any compact submanifold X of B^4 and any $0 \le t < s \le 4$, let $X_{[t,s]}$ denote $X \cap h^{-1}([t, s])$ and let $X_{\{t\}} = X \cap h^{-1}(t)$. For example, $B^4_{[t,s]} = h^{-1}[t, s]$. Similarly, for any compact submanifold Y of $B^4_{\{t\}}$ and any $0 \le r < s \le 4$, let $Y[r, s]$ denote the vertical cylinder obtained by pushing Y along the gradient flow across the height interval $[r, s]$, which we call a *gradient product*. We extend these notions in the obvious way to open intervals and singletons in $[0, 4]$.

Now we will show how, given a band presentation $(\mathscr{L}, U, \mathfrak{b})$, we can construct the *realizing surface* $\mathscr{F}_{(\mathscr{L}, U, \mathfrak{b})}$: a neatly embedded surface in B^4 with boundary $\mathscr{L}$. Start by considering $(\mathscr{L}, U, \mathfrak{b})$ as 2–complex in $B^4_{\{2\}} \cong S^3$, and consider the surface $\mathscr{F}$ with the properties

(1) $\mathscr{F}_{(3,4]} = \mathscr{L}(3, 4]$;

(2) $\mathscr{F}_{\{3\}} = \mathscr{L}\{3\} \sqcup D$, where D is a collection of spanning disks for the unlink $U\{3\} \subset B^4_{\{3\}} \cong S^3$;

(3) $\mathcal{F}_{(2,3)} = (\mathcal{L} \sqcup U)(2,3)$;

(4) $\mathcal{F}_{\{2\}} = (\mathcal{L} \sqcup U) \cup \mathfrak{b}$;

(5) $\mathcal{F}_{(1,2)} = U'(1,2)$;

(6) $\mathcal{F}_{\{1\}} = D'$, where D' is a collection of spanning disks for the unlink $U' \subset B_{\{1\}}^4 \cong S^3$; and

(7) $\mathcal{F}_{[0,1)} = \varnothing$.

Note that $\sqcup$ represents the split union, and we assume that D is contained in a three-ball B that is disjoint from $\mathcal{L}\{3\}$. Any two such choices of spanning disks D and D' are isotopic after perturbation into $B(3, 3+\epsilon)$ and $B_{(1,1-\epsilon)}^4$, respectively, by Proposition 2.13. Note also that $\partial\mathcal{F} = \mathcal{F} \cap B_{\{4\}}^4 = \mathcal{L}\{4\}$.

Proposition 3.3 *Every neatly embedded surface $\mathcal{F}$ with $\partial\mathcal{F} = \mathcal{L}$ is isotopic rel-∂ to a realizing surface $\mathcal{F}_{(\mathcal{L},U,\mathfrak{b})}$ for some band presentation $(\mathcal{L}, U, \mathfrak{b})$. If $\mathcal{F}$ has a handle-decomposition with respect to the standard Morse function on B^4 consisting of c_1 cups, n bands, and c_3 caps, then $(\mathcal{L}, U, \mathfrak{b})$ can be assumed to satisfy $|U| = c_3$, $|\mathfrak{b}| = n$, and $|U'| = c_1$.*

Proof Given $\mathcal{F}$, we can assume after a minor perturbation that the restriction $h_\mathcal{F}$ of a standard height function $h: B^4 \to [0, 4]$ is Morse. After reparametrizing the codomain of h, we can assume that the critical points of $h_\mathcal{F}$ are contained in $h^{-1}((1.5, 2.5))$. For each index zero critical point x of $h_\mathcal{F}$, we choose a vertical strand ω connecting x to $B_{\{1\}}^4$. (Here, vertical means that $\omega_{\{t\}}$ is a point or empty for each $t \in [1, 2.5]$.) By a codimension count, ω is disjoint from $\mathcal{F}$, except at x. We can use a small regular neighborhood of ω to pull x down to $B_{\{1\}}^4$. Repeating, we can assume that the index zero critical points of $h_\mathcal{F}$ lie in $B_{\{1\}}^4$. By a similar argument, we achieve that the index two critical points of $h_\mathcal{F}$ lie in $B_{\{3\}}^4$ and that the index one critical points of $h_\mathcal{F}$ lie in $B_{\{2\}}^4$.

Next, we perform the standard flattening of the critical points: for each critical point x of index i, find a small disk neighborhood N of x in $\mathcal{F}$, and isotope $\mathcal{F}$ so that N lies flat in $B_{\{i+1\}}^4$. Near critical points of index zero or two, $\mathcal{F}$ now resembles a flat-topped or flat-bottomed cylinder; for index one critical points, N is now a flat square. Let $\mathfrak{b}'$ denote the union of the flat, square neighborhoods of the index one critical points in $B_{\{2\}}^4$.

So far, we have achieved properties (2), (4), (6), and (7) of a realizing surface. Properties (1), (3), and (5) say that $\mathcal{F}$ should be a gradient product on the intervals $(3, 4]$, $(2, 3)$, and $(1, 2)$, respectively. The products $\mathcal{F}_{(3,4]}$ and $\mathcal{L}(3, 4]$ (for example) agree at $\mathcal{F}_{\{4\}} = \mathcal{L}\{4\}$, but may disagree in $B_{\{t\}}^4$ for $t \in (3, 4)$. This issue can be addressed by a "combing-out" process.

For each $t \in [1, 4]$, we can choose ambient isotopies $G_t: [0, 1] \times B_{\{t\}}^4 \to B_{\{t\}}^4$ such that

(1) $G_4(s, x) = x$ for all $s \in [0, 1]$ and $x \in B_{\{4\}}^4$;

(2) $G_t(0, x) = x$ for all $t \in [1, 4]$ and $x \in B_{\{t\}}^4$;

(3) $G_t(1, \mathcal{F}_{\{t\}}) = \mathcal{L}\{t\}$ for all $t \in (3, 4]$, where we now let $\mathcal{L} = \mathcal{F}_{\{4\}}$;
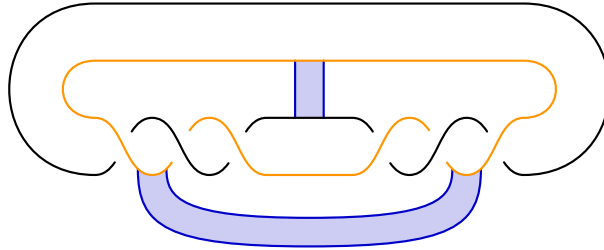
Figure 6: A band presentation for the punctured spun trefoil, considered as a neatly embedded disk in B^4 with unknotted boundary.

(4) $G_t(1, \mathscr{F}_{\{t\}}) = (\mathscr{L} \sqcup U)\{t\}$ for all $t \in (2, 3)$, where we now let $\mathscr{L} \sqcup U = G_3(\mathscr{F}_{\{3\}} \setminus \mathrm{Int}\, D)$;

(5) $G_t(1, \mathscr{F}_{\{t\}}) = U'\{t\}$ for all $t \in (1, 2)$, where we now let $U' = G_2(\partial \mathscr{F}_{[0,2)})$; and

(6) G_t is smoothly varying in t.

After applying the family G_t of ambient isotopies to $\mathscr{F}_{[1,4]}$, we have properties (1), (3), and (5), as desired. However, the ambient isotopies G_t have now altered $\mathscr{F}_{\{t\}}$ for $t = 1, 2, 3$. For example, the disks D and D' have been isotoped around in their respective level sets; but, clearly, properties (2), (4), (6), and (7) are still satisfied. We remark that, if desired, we can choose G_t so that

(a) the disks of D end up contained in small, disjoint 3–balls and either

(b) the disks of D' have the same property or

(c) the bands $\mathfrak{b}$ have the same property.

However, we cannot always arrange (a), (b), *and* (c) if we want $\mathscr{F}_{(1,2)}$ to be a gradient product.

With a slight abuse of notation, we now let $\mathscr{L} = \mathscr{L}\{2\}$, $U = U\{2\}$, and $\mathfrak{b} = G_2(\mathfrak{b}')$. (The only abuse is which level set of the now-gradient-product portion $\mathscr{L}[2,4]$ of $\mathscr{F}$ should be denoted by $\mathscr{L}$.) In the end, we have that $\mathscr{F}$ is the realizing surface of the band presentation $(\mathscr{L}, U, \mathfrak{b})$.

With regards to the second claim of the proposition, assume that $\mathscr{F}$ has c_1 cups, n bands, and c_3 caps once it is in Morse position. Each cap gives rise to a component of U, while each cup gives rise to a component of U'. The numbers of bands, cups, and caps are constant throughout the proof.   □

Examples of a band presentations are shown below in Figures 8(a), 10(a), and 13(g). However, each of these is a ribbon presentation. Throughout the rest of the paper, we will work almost exclusively with ribbon presentations. To emphasize the generality of Definition 3.2, we give in Figure 6 a nonribbon band presentation, where the black unknot is $\mathscr{L}$ and the orange unknot is U. Note that a nonribbon band presentation $(\mathscr{L}, U, \mathfrak{b})$ for a surface $\mathscr{F}$ can always be converted to a ribbon presentation $(\mathscr{L}', \mathfrak{b})$ for a surface $\mathscr{F}'$ by setting $\mathscr{L}' = \mathscr{L} \sqcup U$. The ribbon surface $\mathscr{F}'$ is obtained from the nonribbon surface $\mathscr{F}$ by puncturing at each maxima and dragging the resulting unlink to the boundary.

3.3 Bridge-braiding band presentations

Recall the standard Heegaard-double decomposition $Y = Y_{0,0,1}$ of S^3 that was introduced in Section 2.4 and revisited in Section 3.1, which is a decomposition of S^3 into three trivial lensed cobordisms (three-balls), H_1, H_3, and Y_3, which meet along disk pages $H_1 \cap \overline{H}_3 = \Sigma$ and $H_i \cap Y_3 = P_i$ whose boundary is the unknotted braid axis B in S^3. The choice to use H_3 instead of H_2 will ensure that the labelings of our pieces agree with our conventions for the labeling of the pieces of a bridge trisection, as in the proof of Proposition 3.12 below.

Definition 3.4 (bridge-braided) A band presentation $(\mathscr{L}, U, \mathfrak{b})$, considered with respect to the standard Heegaard-page decomposition $Y_{0,0,1}$ of S^3, is called $(b, \mathbf{c}; v)$–*bridge-braided* if

(1) $\beta_3 = \mathscr{L} \cap Y_3$ is a v–braid;

(2) $\mathscr{L} \cap (H_1 \cup_\Sigma \overline{H}_3)$ is a b'–perturbing of a v–braid;

(3) U is in b''–bridge position with respect to Σ;

(4) $\mathfrak{b} \cap \Sigma$ is precisely the cores y_* of $\mathfrak{b}$, which are embedded in Σ;

(5) there is a bridge system Δ for the trivial tangle $\mathscr{T}_3 = H_3 \cap (\mathscr{L} \cup U)$ whose shadows Δ_* have the property that $\Delta_* \cup y_*$ is a collection of embedded arcs in Σ; and

(6) $U' = (\mathscr{L} \cup U)_\mathfrak{b}$ is a (c_1+v)–component unlink that is in standard (b, v)–bridge position with respect to $Y_{0,0,1}$ (hence, U' consists of c_1 flat components and v vertical components).

Here, $b = b' + b''$, $c_3 = |U|$, $c_2 = b - |\mathfrak{b}|$, and $c_1 = |U'| - v$. Let $\hat{\beta}$ denote the index v braiding of $\mathscr{L}$ given by $\beta_3 \cup \mathscr{T}_1 \cup \overline{\mathscr{T}}_3$. In reference to this added structure, we denote the bridge-braided band presentation by $(\hat{\beta}, U, \mathfrak{b})$. If $U = \varnothing$, so $(\mathscr{L}, \mathfrak{b})$ is a ribbon presentation, we denote the corresponding bridge-braiding by $(\hat{\beta}, \mathfrak{b})$.

We say that a band in $\mathfrak{b}$ is *dualized* by the bridge disk in Δ whose shadow is adjacent to the band's core in the embedded polygonal arc.

Proposition 3.5 *Let $\mathscr{F} \subset B^4$ be a surface with $\partial \mathscr{F} = \mathscr{L}$, and let $\hat{\beta}$ be an index v braiding of $\mathscr{L}$. There is a bridge-braided band presentation $(\hat{\beta}, U, \mathfrak{b})$ such that $\mathscr{F} = \mathscr{F}_{(\hat{\beta}, U, \mathfrak{b})}$. If $\mathscr{F}$ has a handle-decomposition with respect to the standard Morse function on B^4 consisting of c_1 cups, n bands, and c_3 caps, then $(\hat{\beta}, U, \mathfrak{b})$ can be assumed to be $\bigl(b, (c_1, b-(n+v), c_3); v\bigr)$–bridge-braided for some $b \in \mathbb{N}$.*

Proof Consider $\mathscr{F} \subset B^4$ with $\partial \mathscr{F} = \mathscr{L}$. By Proposition 3.3, we can assume (after an isotopy rel-∂) that $\mathscr{F} = \mathscr{F}_{(\mathscr{L}, U, \mathfrak{b})}$ for some band presentation $(\mathscr{L}, U, \mathfrak{b}')$. We assume that $|U| = c_3$, $|\mathfrak{b}'| = n$, and $|(\mathscr{L} \sqcup U)_{\mathfrak{b}'}| = c_1$. By Alexander's theorem [1], there is an ambient isotopy $G_4 \colon I \times B^4_{\{4\}} \to B^4_{\{4\}}$ taking $\partial \mathscr{F}$ to $\hat{\beta}$. As in the proof of Proposition 3.3, there is a family G_t of ambient isotopies extending G_4 across B^4. This results in the "combing-out" of Alexander's isotopy G_4, with the final effect that $\mathscr{F}$ is the realizing surface of the (not-yet-bridge-braided) band presentation $(\hat{\beta}, U, \mathfrak{b}')$. Henceforth, we consider the 2–complex corresponding to $(\hat{\beta}, U, \mathfrak{b}')$ to be living in $B^4_{\{2\}}$, as in Proposition 3.3.

We have already obtained properties (1) and (2) towards a bridge-braided band presentation; although, presently $b' = 0$. (This will change automatically once we begin perturbing the bridge surface Σ relative to $\hat{\beta}$ and U.) By an ambient isotopy of $B^4_{\{2\}}$ that is the identity in a neighborhood of $\hat{\beta}$, we can move U to lie in bridge position with respect to Σ, realizing property (3). (Again, the bridge index b'' of this unlink will change during what follows.) Since this ambient isotopy was supported away from $\hat{\beta}$ it can be combed-out (above and below) via a family of isotopies that are supported away from the gradient product $\hat{\beta}[2, 4]$; so $\mathcal{F}$ is still the realizing surface.

Next, after an ambient isotopy that fixes $\hat{\beta} \sqcup U$ setwise (and pointwise near Σ), we can arrange that $\mathfrak{b}'$ lies in $H_1 \cup_\Sigma \overline{H}_3$. (Think of the necessity of sliding the ends of $\mathfrak{b}'$ along β_3 to extract it from Y_3, while isotoping freely the unattached portion of $\mathfrak{b}'$ to the same end.) This time, we need only comb-out towards $h^{-1}(0)$. Using the obvious Morse function associated to $(H_1 \cup_\Sigma \overline{H}_3) \setminus \nu(B)$, we can flow $\mathfrak{b}'$, in the complement of $\hat{\beta} \sqcup U$, so that the cores of the bands lie as an immersed collection of arcs y in $\Sigma \setminus \nu(x)$. At this point, we can perturb the bridge surface Σ relative to $\hat{\beta} \sqcup U$ to arrange that the cores y be embedded in Σ. For details as to how this is achieved, we refer the reader to Figure 10 (and the corresponding discussion starting on page 17) of [27]. Now that the cores y_* of $\mathfrak{b}'$ are embedded in Σ, we can further perturb Σ relative to $\hat{\beta} \sqcup U$ (as in Figure 11 of [27]) to achieve that $\mathfrak{b}' \cap \Sigma$ is precisely the cores of $\mathfrak{b}'$. Thus, we have that the bands $\mathfrak{b}'$ satisfy property (4). A further perturbation of Σ relative to $\hat{\beta} \sqcup U$ produces, for each band υ of $\mathfrak{b}'$, a dualizing bridge disk Δ_υ, as required by property (5). (See Figure 12 of [27].)

However, at this point it is possible that the c_1–component unlink $U'' = (\hat{\beta} \sqcup U)_{\mathfrak{b}'}$ is *not* in standard (b, v)–bridge position; more precisely, it is possible that components of U'' intersect Y_3 in more than one strand. On the other hand, we automatically have that $U'' \cap Y_3$ is a v–braid, since the band resolutions changing $\mathcal{L} \cup U$ into U'' were supported away from Y_3. Moreover, we know that $U'' \cap H_i$ is a (b, v)–tangle; this follows from the proof of [27, Lemma 3.1].

Thus, we must modify U'' in order to obtain an unlink in standard position. To do so, we will produce a new collection $\mathfrak{b}''$ of bands such that $U' = U''_{\mathfrak{b}''}$ is a (c_1+v)–component unlink in (b, v)–bridge position. We call the bands $\mathfrak{b}''$ *helper bands*. We will then let $\mathfrak{b} = \mathfrak{b}' \sqcup \mathfrak{b}''$, and the proof will be complete.

Since (Y_3, β_3) is a v–braid, there is a collection of bridge triangles Δ for β_3. Let $\omega = \Delta \cap (P_1 \cup_B \overline{P}_3)$. Let $\mathfrak{b}''$ denote the collection of v bands whose core are the arcs ω and that are framed by the two-sphere $P_1 \cup_B \overline{P}_3$. By a minor isotopy that fixes U'' setwise (and pointwise away from a neighborhood of $\partial\omega$), we consider $\mathfrak{b}''$ as lying in the interior of $H_1 \cup_\Sigma \overline{H}_3$. Thus, $\mathfrak{b}''$ is a collection of bands for $\mathcal{T}_1 \cup_x \overline{\mathcal{T}}_3$. See Figure 7 for two simple examples.

Let $U' = U''_{\mathfrak{b}''}$. Let J denote the components of U' containing the strands of β_3. Since the helper bands $\mathfrak{b}''$ were created from the bridge triangles of Δ, we find that J bounds a collection of v disjoint meridional disks for B. In particular, J is a v–component unlink in v–braid position with respect to B. Let $K = U' \setminus J$, and note that K is isotopic (disregarding the Heegaard double structure) to the unlink U''. It follows that
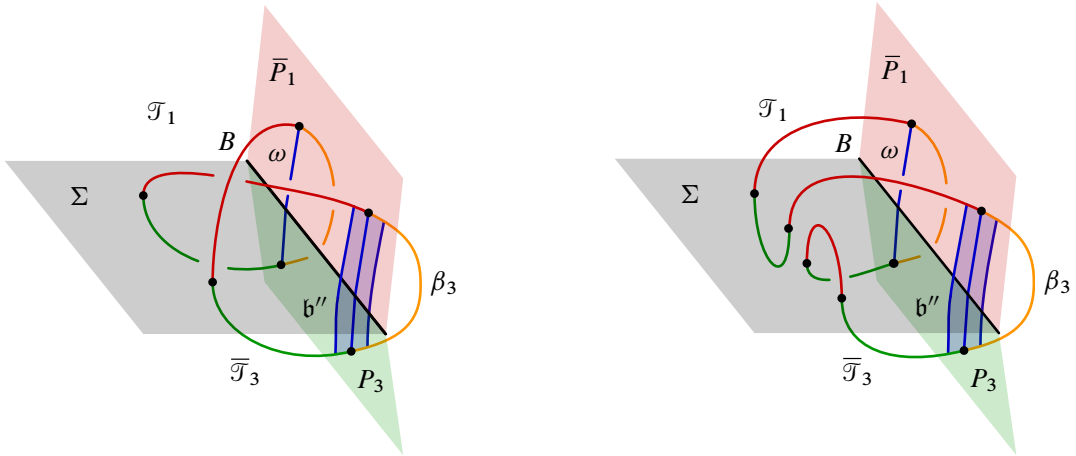
Figure 7: Adding extra bands to ensure that U' is in standard (b, v)–bridge position.

K is a c_1–component unlink in bridge position with respect to Σ. Therefore, U' is a (c_1+v)–component unlink in standard (b, v)–bridge position, as required by property (6) of Definition 3.4.

Now, to wrap up the construction, we let $\mathfrak{b} = \mathfrak{b}' \cup \mathfrak{b}''$. While we have arranged the bands of $\mathfrak{b}'$ are in the right position with respect to the Heegaard splitting, we must now repeat the process of perturbing the bridge splitting in order to level the helper bands $\mathfrak{b}''$. The end result is that the bands of $\mathfrak{b}$ satisfy properties (4) and (5) of Definition 3.4. In the process, we have not changed the fact that properties (1)–(3) and (6) are satisfied, though we may have further increased the parameters b' and b'' (and, thus, $b = b' + b''$) during this latest bout of perturbing.

We complete the proof by noting that $|U| = c_3$, $|U'| = c_1 + v$, and $|\mathfrak{b}| = n + v$. □

Remark 3.6 A key technical step in the proof of Proposition 3.5 was the addition of the so-called *helper bands* $\mathfrak{b}''$ to the original set $\mathfrak{b}'$ of bands that were necessary to ensure that U' was in standard position. In the proof, $\mathfrak{b}''$ consisted of v bands; in practice, one can make do with a subset of these v bands. This can be seen in the two simple examples of Figure 7, where the addition of only one band (in each example) suffices to achieve standard bridge position. In Figure 7, left, the addition of the single band shown transforms an unknot component of U'' that is in 2–braid position into a pair of 1–braids (one of which is perturbed) in the link U'. In Figure 7, right, an unknot component that is not braided at all is transformed to the same result. In each of these examples, the addition of a second band corresponding to the second arc of ω would be superfluous.

From a Morse-theoretic perspective, the helper bands correspond to canceling pairs of minima and saddles: the minima are the meridional disks bounded by J. Using more bands from $\mathfrak{b}''$ than is strictly necessary results in a surface with more minima (and bands) than are actually required to achieve the desired bridge-braided band presentation. Below, when we convert the bridge-braided band presentation

to a bridge trisection, we will see that the superfluous bands and minima have the effect that the bridge trisection produced is perturbed — see Section 9. Another way of thinking about the helper bands is that they ensure that the trivial disk-tangle $\mathscr{D}_1$ in the resulting bridge trisection has enough vertical patches.

We require that each vertical component of U' intersect Y_3 in a single thread so that the corresponding patch will be vertical. If some wound twice around as a braid, it would bound a patch in Z_3 that is not vertical with respect to the relevant Morse function on Z_3; see the proof of Proposition 3.12 below.

Before proving that a bridge-braided band presentation can be converted to a bridge trisection, we pause to give a few examples illustrating the process of converting a band presentation into a bridge-braided band presentation.

Example 3.7 (figure-8 knot Seifert surface) Figure 8(a) shows a band presentation for the genus one Seifert surface for the figure-8 knot, together with a gray dot representing an unknotted curve about which the knot will be braided; this braiding is shown in Figure 8(b). Note that the resolution of the bands at this point would yield a unknot (denoted U'' in the proof of Proposition 3.5) that is in 3–braid position. Thus, at least two helper bands are need. In Figure 8(c) we have attached three helper bands, as described in the proof of Proposition 3.5. Note that the cores of these bands are simultaneously parallel to the arcs one would attach to form the braid closure, and the disks exhibiting this parallelism correspond to the bridge triangles in the proof. In Figure 8(d), all five bands have been leveled so that they are framed by the bridge sphere, intersecting it only in their cores. In addition, each band is dualized by a bridge disk for $\mathcal{T}_3$. Three of these bridge disks are obvious. The remaining two are only slightly harder to visualize; one can choose relatively simple disks corresponding to any two of the three remaining flat arcs.

Figure 8(e) shows a tri-plane diagram for the bridge trisection that can be obtained from the bridge-braided band presentation given in Figure 8(d) according to Proposition 3.12. (See Section 4 for precise details regarding tri-plane diagrams.) Figure 8(f) shows the pairwise unions of the seams of this bridge trisection. Relevant to the present discussion is the fact that the second two unions each contain a closed, unknotted component. The fact that the red-blue union contains such a component is related to the fact that we chose to use three helper bands, when two would suffice. The fact that the green-blue union contains such a component is related to the fact that the bridge splitting in Figure 8(d) is excessively perturbed. We leave it as an exercise to the reader to deperturb the bridge splitting of Figure 8(d) to obtain a simpler bridge-braided band presentation.

Example 3.8 (figure-8 knot Seifert surface redux) As discussed in Remark 3.6, it is often not necessary to append v helper bands. The frames of Figure 9 are analogous to those of Figure 8, with the main change being that only two of the three helper bands are utilized. The two innermost bands from Figure 8(c) have been chosen, and they have each been slid once over the original bands from Figure 9(b) to make the subsequent picture slightly simpler.
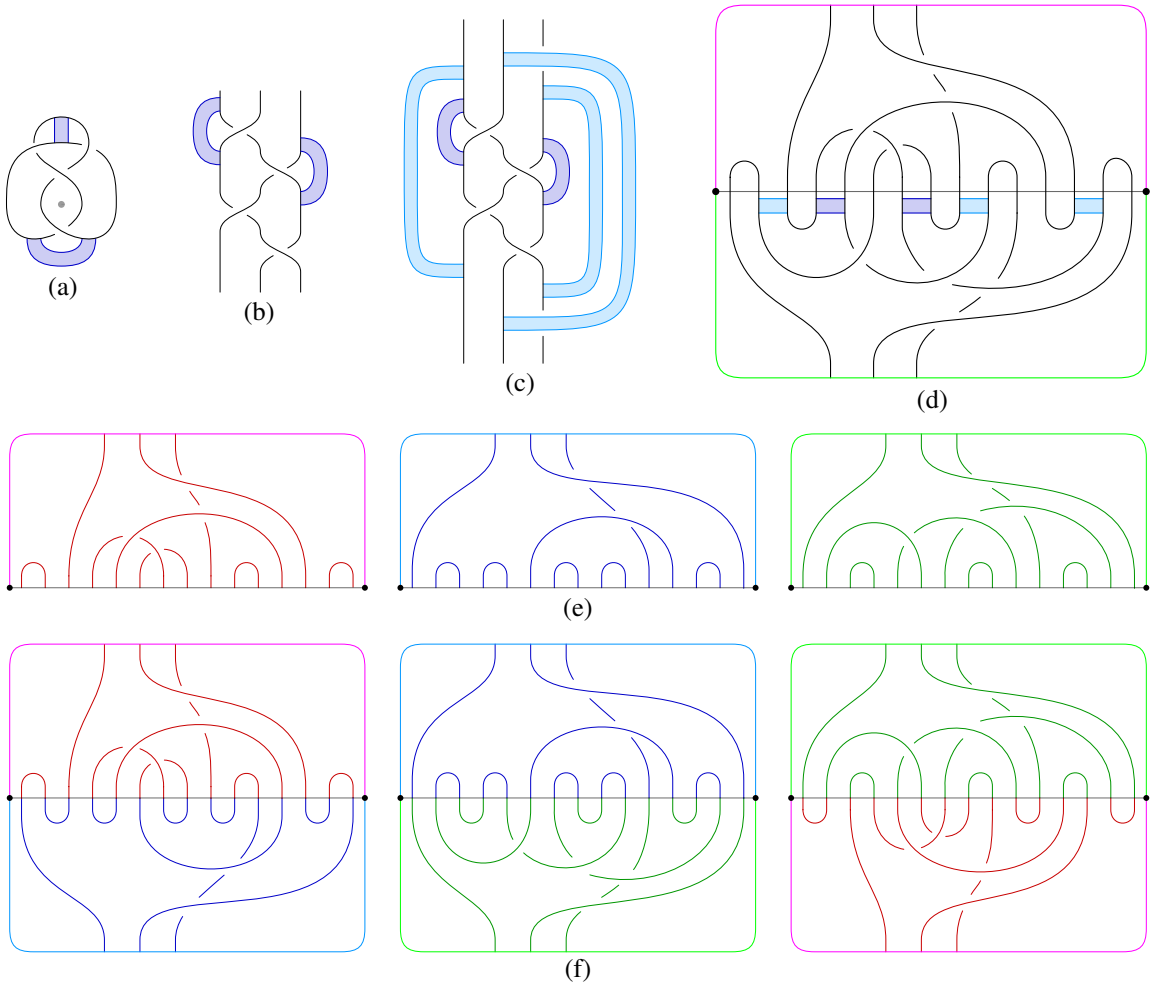
Figure 8: Top row: the process of converting a band presentation for the genus one Seifert surface for the figure-8 knot into a bridge-braided band presentation. Middle row: a tri-plane diagram corresponding to the bridge-braided band presentation of (d). See Figure 9 for a second instantiation of this example.

Since fewer bands are included, the bridge splitting required to level and dualize them is simpler. In this case, the perturbing in Figure 9(d) is minimal. In light of these variations, we see in Figure 9(f) that the pairwise unions of the seams of the bridge trisection contain no closed components, implying the bridge trisection is not perturbed — see Section 9.

Example 3.9 (stevedore knot ribbon disk) Figure 10(a) shows a band presentation for a ribbon disk for the stevedore knot, together with a gray dot representing an unknotted curve about which the knot is braided in Figure 10(b). Note that the result of resolving the band in Figure 10(b) is a 4–braiding of the 2–component unlink, with each component given by a 2–braid. Thus, at least two helper bands are
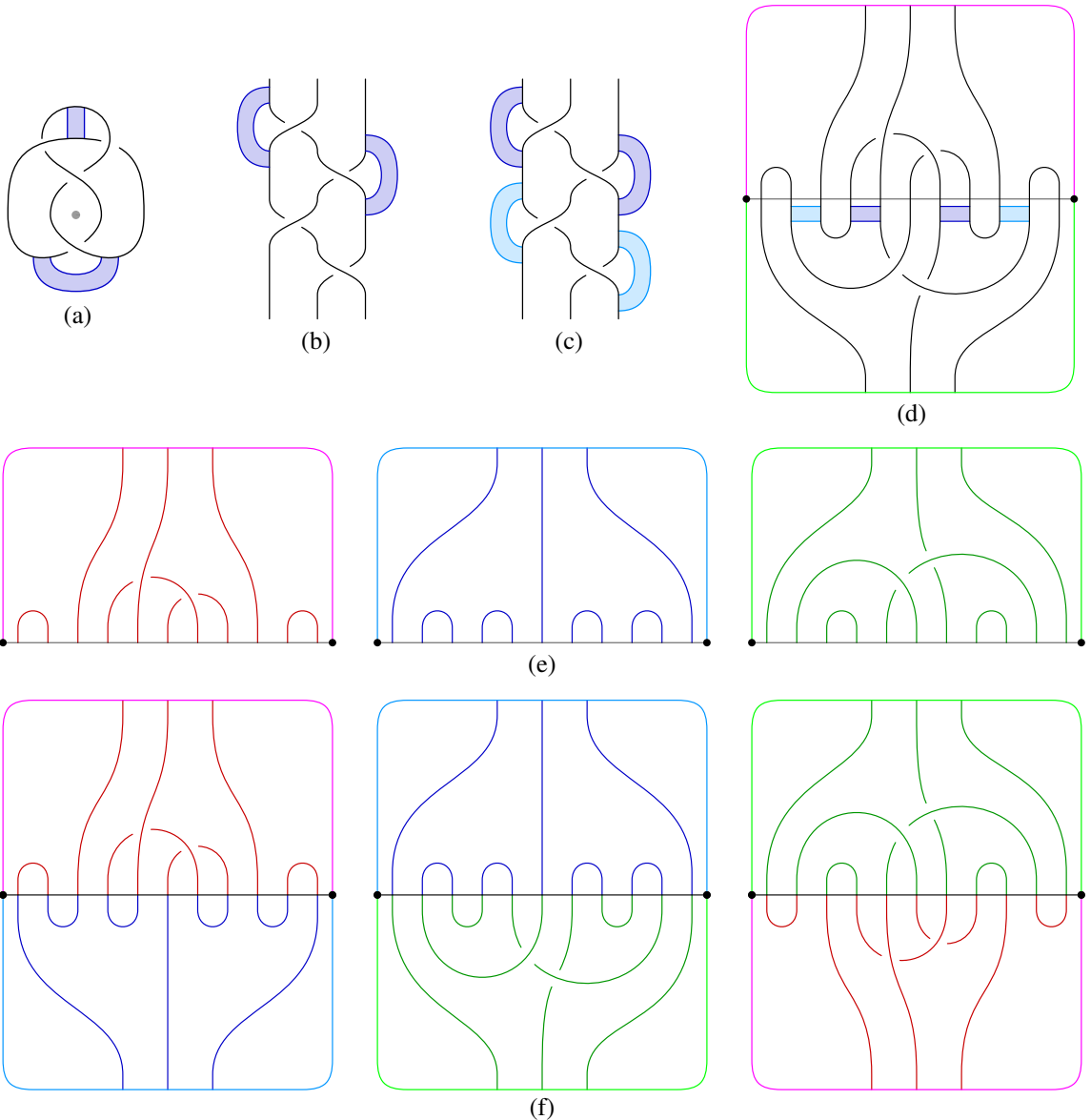
Figure 9: Top row: the process of converting a band presentation for the genus one Seifert surface for the figure-8 knot into a bridge-braided band presentation. Middle row: a tri-plane diagram corresponding to the bridge-braided band presentation of (d). See Figure 8 for another instantiation of this example.

required to achieve bridge-braided band position in this example; Figure 10(c) shows two such bands that suffice. (See Remark 3.10 below.)

Figure 10(d) gives a bridge-braided band presentation for the ribbon disk, with the caveat that the helper bands do not appear to be leveled as shown. However, we claim that such a leveling is possible: First, note
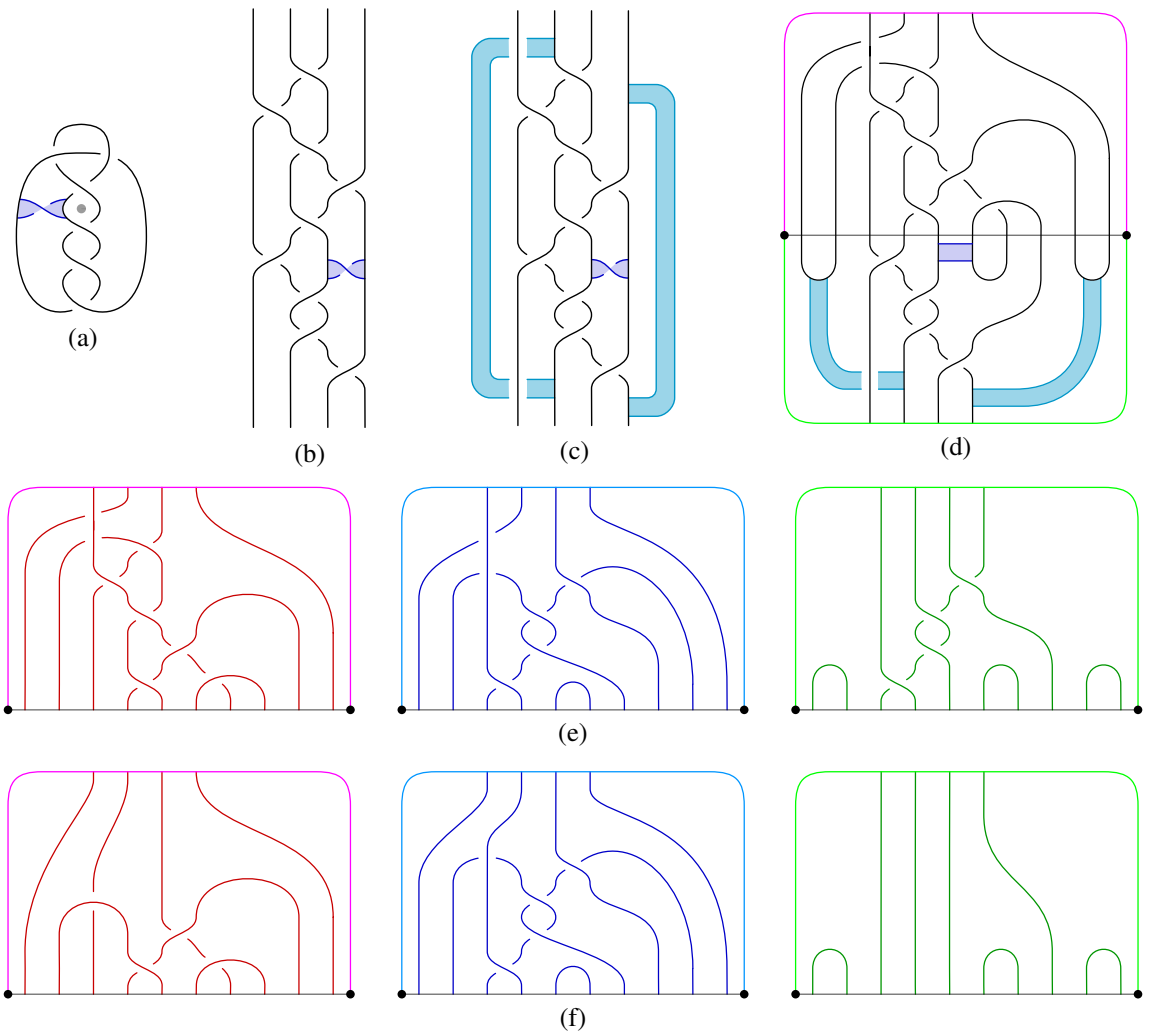
Figure 10: Top row: the process of converting a band presentation for a ribbon disk for the steve-dore knot into a bridge-braided band presentation. Middle row: a tri-plane diagram corresponding to the bridge-braided band presentation of (d). Bottom row: a second tri-plane diagram, obtained from the first via a sequence of tri-plane moves.

that the left helper band can be isotoped so that its core lies in the bridge sphere without self-intersection. Depending on how one chooses to do this, the core may intersect the core of the dark blue band (the original fission band for the ribbon disk). However, since this latter band is dualized by a bridge disk for $\mathcal{T}_3$, there is an isotopy pushing the helper band off the fission band. At this point, the left helper band and the fission band are both level, disjoint, and dualized by bridge disks. Now, we note that the right helper band can be isotoped so that its core lies in the bridge sphere without self-intersection. To do this, however, we must slide the right helper band over the fission band so that their endpoints (attaching regions) are disjoint. Again, the core may intersect the cores of the other two bands, but since the other

two bands are each dualized by bridge disks, we may push the core of the right helper band off the cores of the other two bands. The end result is that all three bands lies in the required position.

Figure 10(e) shows a tri-plane diagram for the bridge trisection corresponding to the bridge-braided band position from Figure 10(d). It is worth observing that it was not necessary to carry out the leveling of the bands described in the previous paragraph; it suffices simply to know that it can be done. Had we carried out the leveling described above, the result would have been a tri-plane diagram that could be related to the one given by a sequence of interior Reidemeister moves. Figure 10(f) shows a tri-plane diagram that is related to the tri-plane diagram of Figure 10(e) by tri-plane moves. See Section 4 for details regarding these moves.

Remark 3.10 There is a subtle aspect to Figure 10(c) that is worth pointing out. Suppose instead that the left helper band were chosen to cross over the braid in the two places where it crosses under. It turns out that this new choice is still a helper band but would fail to result in a bridge-braided band position. To be precise, let $\mathcal{T}$ denote the braid in Figure 10(c), which we think of as a 4–stranded tangle, and let $\mathfrak{b}$ denote this new choice of bands — ie three bands that are identical to the ones shown in Figure 10(c), except that the left helper band passes above $\mathcal{T}$ in two places, rather than under. The resolution $\mathcal{T}_{\mathfrak{b}}$ is a new 4–stranded tangle. Regardless of any concerns about bridge position that could be alleviated by perturbing $\mathcal{T}$, it is necessary that $\mathcal{T}_{\mathfrak{b}}$ be a 4–braid. However, this is not the case in this example. In fact, $\mathcal{T}_{\mathfrak{b}}$ is not even a trivial tangle! The reader can check that $\mathcal{T}_{\mathfrak{b}}$ is the split union of two trivial arcs, together with a 2–stranded tangle $\mathcal{T}'$ that has a closure to the square knot.

So, the "helper bands" of the $\mathfrak{b}$ presently being considered are not actually helper bands in the sense that they don't transform U'' into an unlink U' in standard position, as required. Of course, by the proof of Proposition 3.5, we know that we can augment $\mathfrak{b}$ by adding two more helper bands, resulting in a total of five bands, so that the result can be bridge-braided. On the other hand, Figure 10 shows that it is possible to achieve a bridge-braided band position with fewer than four helper bands; comparison of Figures 8 and 9 gives another example of this. Precisely when this is possible and precisely how one chooses a more efficient set of helper bands of this sort is not clear; we pose the following question.

Question 3.11 Does there exist a surface $\mathcal{F}$ in B^4 such that every (b, v)–bridge braided band presentation of $\mathcal{F}$ requires v helper bands?

Such a surface would have the property that every bridge trisection contains some flat patches. For this reason, it cannot be ribbon, due to the results of Section 3.4 below.

Having discussed in detail the above examples, we now return our attention to the goal of bridge trisecting surfaces.

Proposition 3.12 Let $\mathcal{F} \subset B^4$ be the realizing surface for a $(b, \mathbf{c}; v)$–bridge-braided band presentation $(\hat{\beta}, U, \mathfrak{b})$. Then $\mathcal{F}$ admits a $(b, \mathbf{c}; v)$–bridge trisection $\mathcal{T}_{(\hat{\beta}, U, \mathfrak{b})}$.
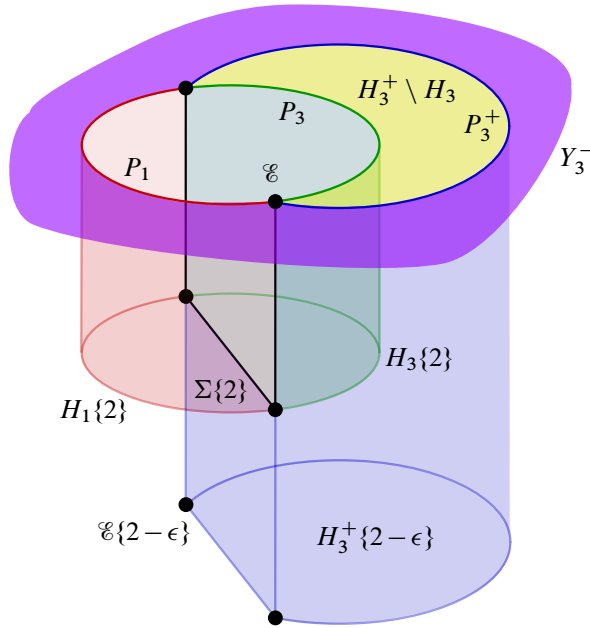
Figure 11: A schematic illustrating how to obtain a bridge trisection from a bridge-braided band presentation; codimension two objects are not shown.

Proof As in Proposition 3.3, we imagine that the 2–complex $\mathscr{L} \cup U \cup \mathfrak{b}$ corresponding to the bridge-braided band presentation $(\hat{\beta}, \mathfrak{b}, U)$ is lying in the level set $B^4_{\{2\}}$, which inherits the Heegaard double structure (H_1, H_3, Y_3). Assume that $\mathscr{F}$ is the corresponding realizing surface. We modify this 2–complex so that the bands $\mathfrak{b}$ lie in the interior of H_3, rather than centered on Σ.

Let $\epsilon > 0$, and assume that the resolution of the bands $\mathfrak{b}$ for $\mathscr{L} \cup U$ occurs in $H_3(2-\epsilon, 2)$. So $\mathscr{F}\{2\} = \mathscr{L} \cup U$, while $\mathscr{F}\{2-\epsilon\} = U'$. Let (P_3^+, x_3^+) denote a slight push-off of (P_3, x_3) into $(H_3, \mathcal{T}_3)$. Let (H_{13}^-, β_{13}^-) denote the corresponding contraction of (Y_3, β_3), and let $(H_3^+, \mathcal{T}_3^+)$ denote the corresponding expansion of $(H_3, \mathcal{T}_3)$. In other words, we remove a (lensed) collar of P_3 from Y_3 and add it to H_3.

We will now describe the pieces of a bridge trisection for $\mathscr{F}$. Figure 11 serves as a guide to the understanding these pieces. Define

(1) $(\Sigma', x') = (\Sigma, x)\{2\} \cup B[2,4]$;

(2) $(H_1', \mathcal{T}_1') = (H_1, \mathcal{T}_1)\{2\} \cup (P_1, x_1)[2,4]$;

(3) $\overline{(H_2', \mathcal{T}_2')} = (\Sigma, x)[2-\epsilon, 2] \cup (H_3^+, \mathcal{T}_3^+)\{2-\epsilon\} \cup (P_3^+, x_3^+)[2-\epsilon, 4]$;

(4) $(H_3', \mathcal{T}_3') = (H_3, \mathcal{T}_3)\{2\} \cup (P_3, x_3)[2,4]$;

(5) $(Z_1', \mathscr{D}_1') = (B^4, \mathscr{F})_{[0,2-\epsilon]} \cup ((H_1, \mathcal{T}_1)[2-\epsilon, 2]) \cup (Y_3^-, \beta_3^-)[2-\epsilon, 2]$;

(6) $(Z_2', \mathscr{D}_2') = ((B^4, \mathscr{F})_{[2-\epsilon, 2]} \cap H_3^+[2-\epsilon, 2]) \cup ((Y_3 \setminus \text{Int}(Y_3^-), \beta_3 \setminus \text{Int}(\beta_3^-))[2,4])$; and

(7) $(Z_3', \mathscr{D}_3') = (B^4, \mathscr{F})_{[2,4]} \cap (H_1 \cup_\Sigma \overline{H}_3)[2,4]$.
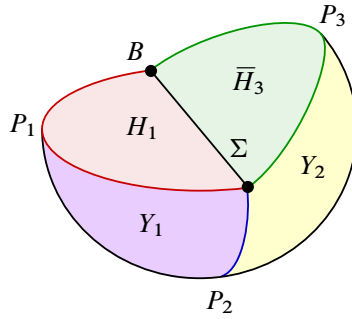
Figure 12: Two-thirds of a trisection, with induced orientations on the boundary.

It is straightforward to verify that the pairs (1)–(7) have the right topology, except in the case of (3) and (6), where slightly more care is needed. For (3), the claim is that $(H_2', \mathcal{T}_2') \cong (H_2, (\mathcal{T}_2)_\flat)$ is a trivial (b, v)–tangle. For (6), the claim is that the trace $(Z_2', \mathcal{D}_2')$ of this band attachment is a trivial (c_2, v)–disk-tangle. Both of these claims follow from the fact that each band of $\flat$ is dualized by a bridge disk for $\mathcal{T}_3$; this is essentially [27, Lemma 3.1]. Finally, it only remains to verify that the pieces (1)–(7) intersect in the desired way. This is straightforward to check, as well.   □

Remark 3.13 Care has been taken to track the orientations throughout this section so that the orientations of the pieces of the bridge trisection produced in Proposition 3.12 agree with the orientation conventions given in Section 2.9. For example, the union $H_1 \cup_\Sigma \overline{H}_3$ appearing in the bridge-braided band presentation set-up of Definition 3.4 gets identified with a portion of $B^4\{2\}$ in the proof of Proposition 3.12, where it is oriented as the boundary of $B^4[0, 2]$. This agrees with the convention that $\partial Z_1 = H_3 \cup_\Sigma H_1 \cup Y_3$, so $\partial(Z_2 \cup Z_3) = Y_1 \cup H_1 \cup_\Sigma \overline{H}_3 \cup Y_2$. See Figure 12.

Proposition 3.14 *If $\mathcal{F}$ admits a $(b, c; v)$–bridge trisection, then $\mathcal{F} = \mathcal{F}_{(\hat{\beta}, U, \flat)}$ for some $(b, c; v)$–bridge-braided band presentation $(\hat{\beta}, U, \flat)$.*

Proof Suppose $\mathcal{F}$ is in bridge position with respect to $\mathbb{T}_0$. Consider the link $L_3 = \beta_3 \cup \mathcal{T}_3 \cup \overline{\mathcal{T}}_1 = \partial \mathcal{D}_3$. Let $\overline{L}$ denote the vertical components of $L_3 \setminus \text{Int}(\beta_3) = \mathcal{T}_3 \cup_x \overline{\mathcal{T}}_1$, and let U denote the flat components. Then we have $\partial \mathcal{D}_3 = \overline{L} \cup \beta_3 \cup U$; in particular, $\overline{L}$ is parallel to β_3 (as oriented tangles) through the vertical disks of $\mathcal{D}_3$. Let $\mathcal{L}$ be the closed one-manifold given by

$$\beta_1 \cup \beta_2 \cup \overline{L}.$$

By the above reasoning, $\mathcal{L}$ is boundary parallel to the boundary braid $\beta_1 \cup \beta_2 \cup \beta_3 = \hat{\beta} = \partial \mathcal{F}$ via the vertical disks of $\mathcal{D}_3$.

Let $Y = Y_1 \cup H_1 \cup \overline{H}_3 \cup Y_2$ and note that Y has the structure of a standard Heegaard-double decomposition $(H_1, H_3, Y_1 \cup Y_2)$ on $S^3 = \partial(Z_1 \cup Z_2)$ and is oriented as the boundary of $Z_1 \cup Z_2$, which induces the opposite orientations on the 3–balls H_1 and H_3 as does Z_3. See Figure 12. It will be with respect to this

structure that we produce a bridge-braided band presentation for $\mathcal{F}$. Note that $\mathcal{L} \cap (Y_1 \cup Y_2)$ is already a v–braid, giving condition (1) of the definition of a bridge-braided band presentation. Similarly, conditions (2) and (3) have been met given the position of $\bar{L} \cup U$ with respect to the Heegaard splitting $H_1 \cup_\Sigma \bar{H}_3$.

Next, we must produce the bands $\mathfrak{b}$. This is done in the same way as in [27, Lemma 3.3]. We consider the bridge splitting $(H_2, \mathcal{T}_2) \cup_{(\Sigma, \boldsymbol{x})} \overline{(H_3, \mathcal{T}_3)}$, which is standard—ie the union of a perturbed braid and a bridge splitting of an unlink. Choose shadows $\mathcal{T}_2^*$ and $\mathcal{T}_3^*$ on Σ for these tangles. Note that we choose shadows only for the flat strands in each tangle, not for the vertical strands. Because the splitting is standard, we may assume that $\mathcal{T}_2^* \cup \mathcal{T}_3^*$ is a disjoint union of c_2 simple closed curves $C_1, \dots, C_{c_2}$, together with some embedded arcs, in the interior of Σ. For each closed component C_i, choose a shadow $\bar{\tau}_i^* \subset (\mathcal{T}_2^* \cap C_i)$. Let

$$\omega^* = \mathcal{T}_2^* \setminus \left(\bigcup_{i=1}^{c_2} \bar{\tau}_i^* \right).$$

In other words, ω^* consists of the shadow arcs of $\mathcal{T}_2^*$, less one arc for each closed component of $\mathcal{T}_2^* \cup \mathcal{T}_3^*$. Note that $|\omega^*| = b - c_2$.

The arcs of ω^* will serve as the cores of the bands $\mathfrak{b}$ as follows. Let $\mathfrak{b} = \omega^* \times I$, where the interval is in the vertical direction with respect to the Heegaard splitting $H_1 \cup_{\bar{\Sigma}} \bar{H}_3$. In other words, $\mathfrak{b}$ is a collection of rectangles with vertical edges lying on $\bar{L} \cup U$ and a horizontal edge in each of H_1 and $\bar{H}_3$ that is parallel through $\mathfrak{b}$ to ω^*. We see that condition (4) is satisfied.

Note that the arcs ω^* came from chains of arcs in $\mathcal{T}_2^* \cup \mathcal{T}_3^*$, so each one is adjacent to a shadow arc in $\mathcal{T}_3^*$. This is obvious in the case of the closed components, since each such component must be an even length chain of shadows alternating between $\mathcal{T}_2^*$ and $\mathcal{T}_3^*$. Similarly, each nonclosed component consists of alternating shadows. This follows from the fact that these arcs of shadows correspond to vertical components of $\bar{L}$, each of which must have the same number of bridges on each side of Σ. These adjacent shadow arcs in $\mathcal{T}_3^*$ imply that $\mathfrak{b}$ is dual to a collection of bridge disks for $\mathcal{T}_3$, as required by condition (5).

Finally, let $U' = \mathcal{L}_\mathfrak{b}$, which should be thought of as lying in $H_1 \cup \bar{H}_2 \cup \beta_1$. In fact, $U' = \mathcal{T}_1 \cup \overline{\mathcal{T}}_2 \cup \beta_1$, so it is the standard link $L_{c_1,w}$ in the standard Heegaard-double structure on ∂Z_1. Thus, (6) is satisfied, and the proof is complete. $\square$

The following example illustrates the proof of Proposition 3.14.

Example 3.15 (square knot disk) Figure 13(a) shows a tri-plane diagram for a surface that we will presently determine to be the standard ribbon disk for the square knot, as described by the band presentation in Figure 13(g). The first step to identifying the surface is to identify the boundary braid. In the proof of Proposition 3.14, this was done by considering the union $\beta_1 \cup \beta_2 \cup \bar{L}$. Diagrammatically, this union can be exhibited by the following three part process:
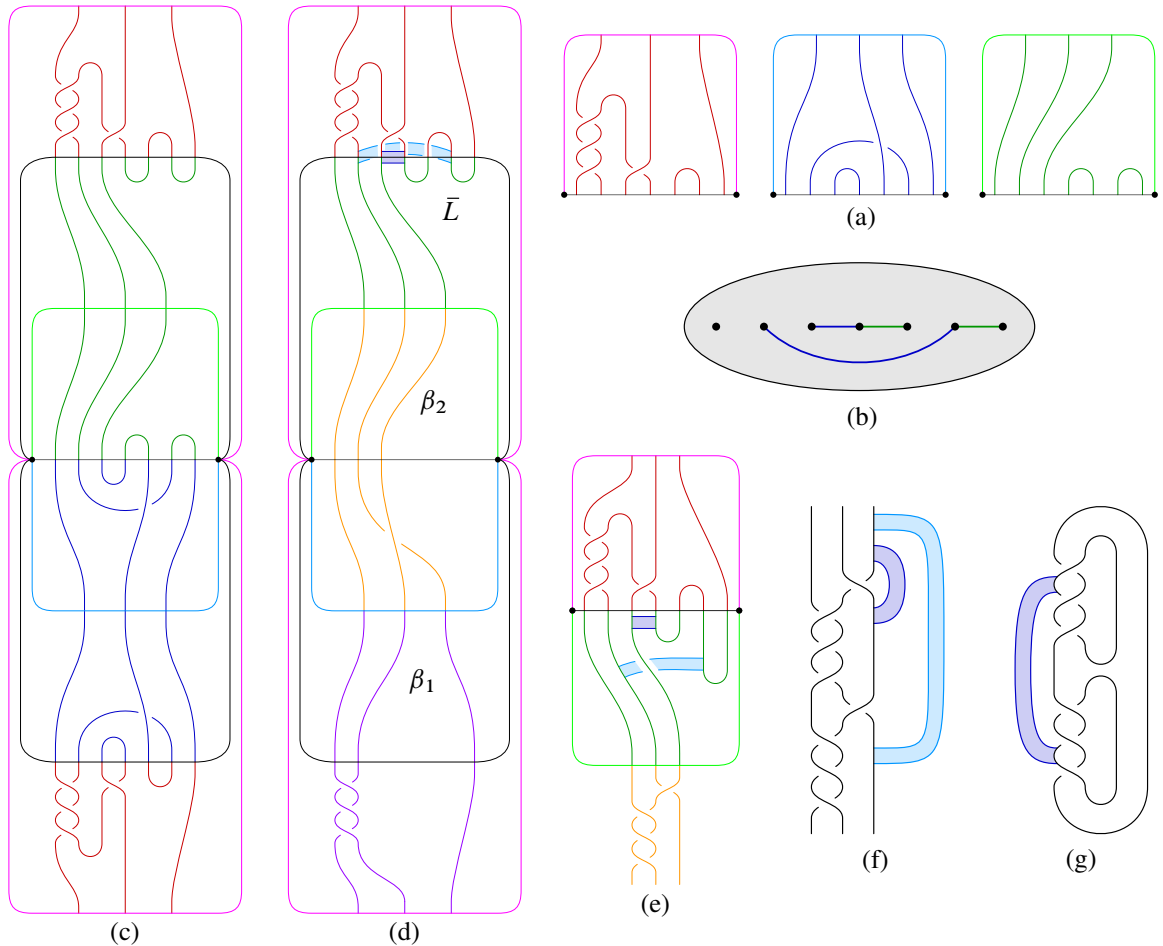
Figure 13: The process of converting the tri-plane diagram (a) into a bridge-braided band presentation (e) in order to identify the underlying surface, which in this case can be seen to be the standard ribbon disk for the square knot (g).

(1) Start with the cyclic union $\mathcal{T}_1 \cup \overline{\mathcal{T}}_3 \cup \mathcal{T}_3 \cup \overline{\mathcal{T}}_2 \cup \mathcal{T}_2 \cup \overline{\mathcal{T}}_1$ of the seams of the bridge trisection; see Figure 13(c).

(2) Discard any components that are not braided; there are no such components in the present example, though there would be if this process were repeated with the tri-plane diagram in Figure 8(e) — a worthwhile exercise.

(3) Straighten out (deperturb) near the intersections $\mathcal{T}_3 \cap \overline{\mathcal{T}}_2$ and $\mathcal{T}_2 \cap \overline{\mathcal{T}}_1$; see Figure 13(d).

If we continued straightening out near $\mathcal{T}_1 \cup \overline{\mathcal{T}}_3$, we would obtain a braid presentation for the boundary link; see Section 4.1 for a discussion relating to this point. Presently, however, it suffices to consider the 1–manifold $\beta_1 \cup \beta_2 \cup \bar{L}$ shown in Figure 13(d), which we know to be isotopic (via the deperturbing near $\mathcal{T}_1 \cap \overline{\mathcal{T}}_3$) to the boundary braid.
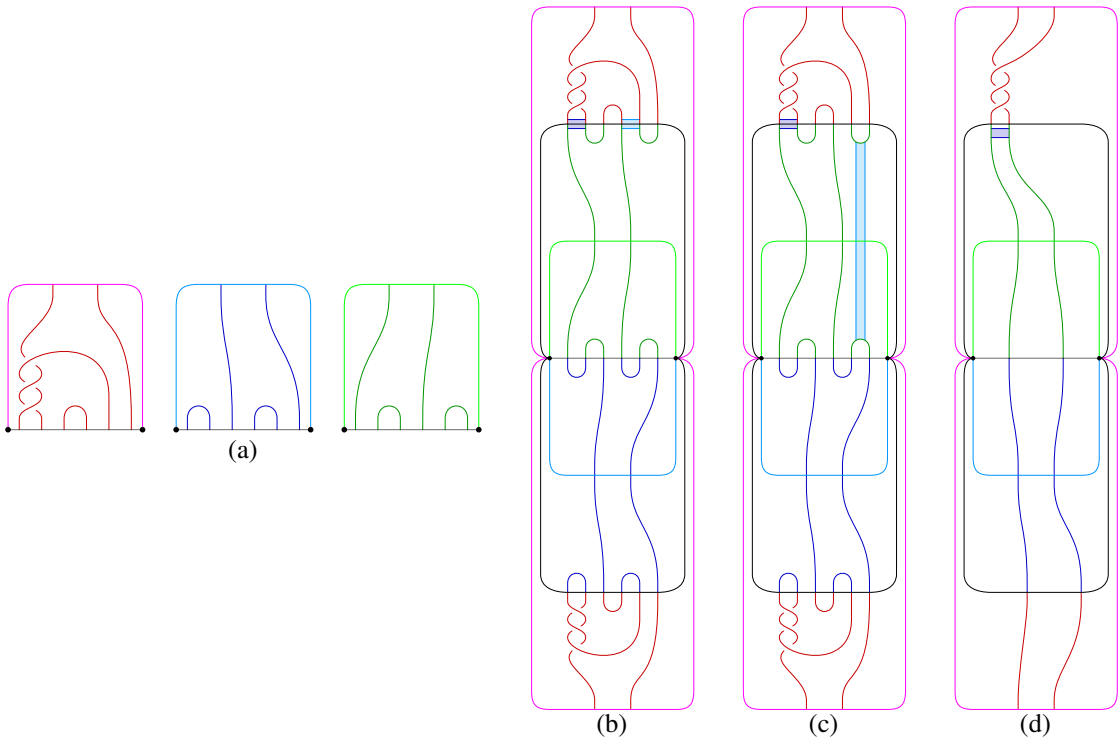
Figure 14: Recovering the boundary braid (d) from a tri-plane diagram (a), with bands tracked. The surface described is the Möbius band bounded by the right-handed trefoil in S^3.

Having identified the boundary braid, we must identify a set of bands that will exhibit a bridge-braided band presentation corresponding to the original bridge trisection. Following the proof of Proposition 3.14, these bands will come from a subset of the shadows $\mathcal{T}_2^*$. To this end, shadows for the tangles $\mathcal{T}_2$ and $\mathcal{T}_3$ are shown in Figure 13(b). If there are closed components, one shadow of $\mathcal{T}_2^*$ is discarded from each such component. In the present example, this step is not necessary; again, consider repeating this exercise with the tri-plane diagram from Figure 8(e). So, the set ω_* of the cores of the bands we are looking for, is precisely the blue shadows of Figure 13(b). In Figure 13(d) these shadows have been thickened vertically into bands that are framed by the bridge sphere $\mathcal{T}_1 \cap \overline{\mathcal{T}}_3$. In Figure 13(e), this picture has been simplified, and the bands have been perturbed into $\overline{\mathcal{T}}_3$. In Figure 13(f), the bridge splitting structure has been forgotten, and the boundary braid is clearly visible. At this point, we see that one band (light blue) is a helper band and can be discarded. At last, Figure 13(g), we recover an efficient band presentation for the surface originally described by the tri-plane diagram of Figure 13(a).

A large family of ribbon disks for the square knot that are pairwise nonisotopic rel-boundary was introduced in [29]; it would be interesting to have bridge trisections for these disks.

Example 3.16 (2–stranded torus links) Figure 14(a) shows a tri-plane diagram corresponding to a bridge trisection of the Möbius band bounded in S^3 by the $(2, 3)$–torus knot; see Figure 14(d) for the

band presentation. However, this example could be generalized by replacing the four half-twists in the first diagram $\mathbb{P}_1$ with n half-twists for any $n \in \mathbb{Z}$, in which case the surface described would be the annulus (respectively, the Möbius band) bounded by the $(2, n)$–torus link when n is even (respectively, the $(2, n)$–torus knot when n is odd).

In any event, Figure 14(b)–(d) gives cross-sections of the bridge trisected surface with concentric shells of B^4, as described in Example 4.5 below. In this example, we also track the information about bands encoded in the tri-plane diagram; cf Figure 13 and Example 3.15. In slight contrast to the square knot examples, the shadows of $\mathcal{T}_2$ are quite simple, so the bands are easy to include. In Figure 14(c), it becomes apparent that the right band (light blue) is a helper band and can be disregarded.

A shadow diagrammatic analysis of this example is given in Example 5.10.

Theorem 3.17 *Let* $\mathbb{T}_0$ *be the standard trisection of* B^4, *and let* $\mathcal{F} \subset B^4$ *be a neatly embedded surface with* $\mathcal{L} = \partial \mathcal{F}$. *Fix an index* v *braiding* $\hat{\beta}$ *of* $\mathcal{L}$. *Suppose* $\mathcal{F}$ *has a handle decomposition with* c_1 *cups,* n *bands, and* c_3 *caps. Then, for some* $b \in \mathbb{N}_0$, $\mathcal{F}$ *can be isotoped to be in* $(b, \boldsymbol{c}; v)$–*bridge trisected position with respect to* $\mathbb{T}_0$, *such that* $\partial \mathcal{F} = \hat{\beta}$, *where* $c_2 = b - n$.

Proof By Proposition 3.5, $\mathcal{F} = \mathcal{F}_{(\hat{\beta}, U, \mathfrak{b})}$ for some bridge-braided band presentation $(\hat{\beta}, U, \mathfrak{b})$ of type $(b, \boldsymbol{c}; v)$. By Proposition 3.12, $\mathcal{F}$ admits a bridge trisection of the same type. $\square$

3.4 Bridge-braided ribbon surfaces

By construction, a $(b, \boldsymbol{c}; v)$–bridge-braided ribbon presentation $(\hat{\beta}, \mathfrak{b})$ will have $c_3 = 0$. The next lemma shows that this fact can be used to systematically decrease the number c_1 of components of the unlink U', at the expense of increasing the index v of the braid $\hat{\beta}$.

Lemma 3.18 *If* $\mathcal{F}$ *is the realizing surface for a* $(b, (c_1, c_2, 0); v)$–*bridge-braided ribbon presentation* $(\hat{\beta}, \mathfrak{b})$ *with* $c_1 > 0$, *then* $\mathcal{F}$ *is the realizing surface for a* $(b, (c_1-1, c_2, 0); v+1)$–*bridge-braided ribbon presentation* $(\hat{\beta}^+, \mathfrak{b})$, *where* $\hat{\beta}^+$ *is a Markov perturbation of* $\hat{\beta}$. *The Markov perturbation can be assumed to be positive.*

Proof Suppose that $(\hat{\beta}, \mathfrak{b})$ is a bridge-braided ribbon presentation with respect to the standard Heegaard double structure (H_1, H_3, Y_3) on S^3, as in Definition 3.4. We orient $\hat{\beta}$ so that it winds counterclockwise about the braid axis $B = \partial \Sigma$. This induces an orientation on the arcs of $\bar{L} = \mathcal{T}_1 \cup_x \overline{\mathcal{T}}_3$, which induces an orientation on the bridge points x: a bridge point $x \in x$ is *positive* if an oriented arc of $\bar{L}$ passes from H_1 to $\overline{H}_3$ through x. Since $c_3 = 0$, every point of x can be oriented in this way.

Recall from the proof of Proposition 3.12 that we can perturb the bands of $\mathfrak{b}$, which originally intersect Σ in their core arcs, into the interior of H_3 so that they may be thought of as bands for the tangle $\mathcal{T}_3$. Let $\mathcal{T}_2 = (\mathcal{T}_3)_\mathfrak{b}$, and let $L' = \mathcal{T}_1 \cup_x \overline{\mathcal{T}}_2$.
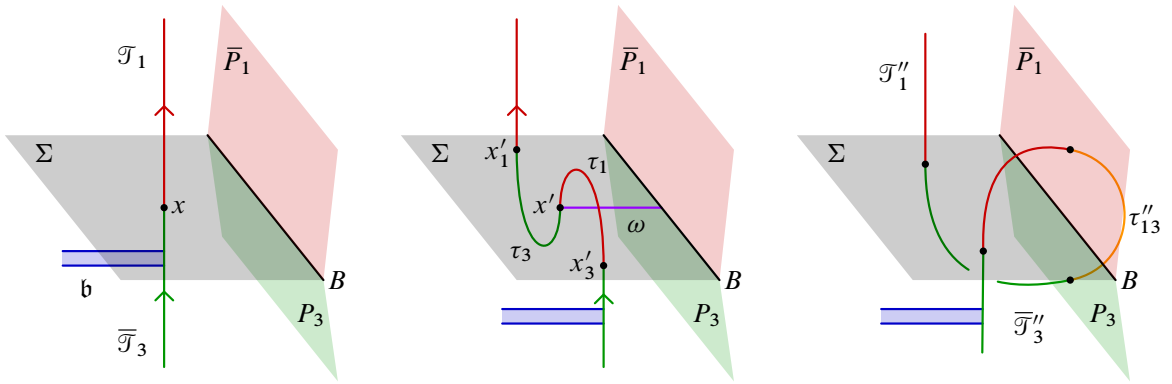
Figure 15: Modifying a bridge trisection of a ribbon surface to remove a flat patch at the expense of Markov-stabilizing the boundary braid.

Utilizing the assumption that $c_1 > 0$, let J be a flat component of U'. Let x be a positive point of $L \cap \Sigma$ so that $x \in J$. See Figure 15, left. Such a point exists, since J contains a flat arc of $\mathscr{T}_1$, and the endpoints of this arc have differing signs. We perturb Σ at x to produce a new bridge splitting $\mathscr{T}_1' \cup_{x'} \overline{\mathscr{T}}_3'$, which we consider as $\mathscr{T}_i' = \mathscr{T}_i \cup \tau_i$, where τ_i is the new flat strand near x. If Δ_i was a bridge system for $\mathscr{T}_i$, then $\Delta_i' = \Delta_i \cup D_i$ is a bridge system for $\mathscr{T}_i'$, where D_i is a bridge semidisk for τ_i. See Figure 15, middle, and note that there may or may not be a band attached to $\mathscr{T}_3$ near x.

Now, we have that $x' = \tau_1 \cap \tau_3$ is negative. Let $x_i' = \partial \tau_i \setminus x$ denote the positive points introduced by this perturbation. Let $\lambda = \tau_1 \cup_x \tau_3$. Note that we can assume there is no band of $\mathfrak{b}$ incident to either τ_i. The bridge splitting $\mathscr{T}_1' \cup_x \overline{\mathscr{T}}_3'$ is perturbed at x'. We will swap this perturbation for a Markov perturbation by dragging the point x' towards and through the boundary B of Σ. Let ω be an embedded arc in Σ connecting x' to B such that $\mathrm{Int}(\omega) \cap x = \varnothing$. Since ω is dualized by each of the two small bridge semidisks $D_i \subset \Delta_i'$, we can assume that $\mathrm{Int}(\omega) \cap \Delta_i' = \varnothing$.

Change $(\hat{\beta}, \mathfrak{b})$ by an ambient isotopy that is supported in a tubular neighborhood of ω and that pushes x' along ω towards and past B. This is a finger move of λ along ω. (Note that the surface $\mathscr{F}$ is locally a product of λ near x'.) Let λ' denote the end result of this finger move; ie a portion of λ has been pushed out of $H_1 \cup_\Sigma \overline{H}_3$ into Y_3. Let $\tau_i'' = \lambda' \cap H_i$. Let $\tau_{13}'' = \lambda' \cap Y_3$. Let D_i'' denote the bridge triangle resulting from applying the ambient isotopy to D_i. We see immediately that τ_i'' are vertical, and that $\Delta_i'' = (\Delta_i' \setminus D_i) \cup D_i''$ is a bridge system for $\mathscr{T}_i'' = (\mathscr{T}_i' \setminus \tau_i) \cup \tau_i''$. It's also clear that τ_{13}'' is a vertical strand in Y_3. We make the following observations, with an eye towards Definition 3.4:

(1) $\beta_3'' = \beta_3 \cup \tau_{13}''$ is a $(v+1)$–braid.

(2) $L'' = \mathscr{T}_1'' \cup_{x''} \overline{\mathscr{T}}_3''$ is a perturbing of a $(v+1)$–braid.

(3) We still have $c_1 = 0$; the $\mathscr{T}_i''$ are $(b-v-1)$–perturbings of $(v+1)$–braids.

(4) The bands $\mathfrak{b}$ can still be isotoped to intersect Σ in their cores.

(5) The bride disks Δ_3'' dualize the bands $\mathfrak{b}$.

(6) $(\hat{\beta})_{\mathfrak{b}}$ has one fewer flat component.

Thus, we have verified that conditions (1)–(6) of Definition 3.4 are still satisfied, with the only relevant differences being that each tangle has an additional vertical strand and the flat component J of U' is now vertical. It follows that we have produced a bridge-banded ribbon presentation $(\hat{\beta}^+, \mathfrak{b})$ for $\mathscr{F}$, where $\hat{\beta}^+$ is a Markov perturbation of $\hat{\beta}$. $\square$

Remark 3.19 The hypothesis that $c_3 = 0$ in the above lemma was necessary to ensure that the process described in the proof resulted in a Markov perturbation of the boundary. If $c_3 > 0$, then it is possible that each point $x \in \boldsymbol{x} \cap J$ lies on a (flat) component of U. If the proof were carried out in this case, it would have the effect of changing the link type from $\mathscr{L}$ to the split union of $\mathscr{L}$ with an unknot on the boundary of $\mathscr{F}$. This is reflective of the general fact that a nonribbon $\mathscr{F}$ with boundary $\mathscr{L}$ can be thought of as a ribbon surface for the split union of $\mathscr{L}$ with an unlink.

Recall that $\boldsymbol{c}$ is an ordered partition of type $(c, 3)$ for some $c \in \mathbb{N}_0$; in particular, $c = c_1 + c_2 + c_3$.

Lemma 3.20 *If $\mathscr{F}$ is the realizing surface for a $(b, \boldsymbol{c}; v)$–bridge-braided band presentation $(\hat{\beta}, U, \mathfrak{b})$ with $c_i = 0$ for some i, then $\mathscr{F}$ is the realizing surface for a $(b, 0; v+c)$–bridge-braided ribbon presentation $(\hat{\beta}^{++}, \mathfrak{b}'')$, where $\hat{\beta}^{++}$ is a Markov perturbation of $\hat{\beta}$.*

Proof Suppose $\mathscr{F}$ is the realizing surface for a $(b, \boldsymbol{c}; v)$–bridge-braided band presentation $(\hat{\beta}, U, \mathfrak{b})$ with $c_i = 0$ for some i. By Proposition 3.12, $\mathscr{F}$ admits a $(b, (c_1, c_2, c_3); v)$–bridge trisection filling $\hat{\beta}$. By relabeling the pieces, we can assume that $c_3 = 0$. By Proposition 3.14, this gives us a $(b, (c_1, c_2, 0); v)$–bridge-braided ribbon presentation $(\hat{\beta}, \mathfrak{b}')$. Note that while the braid type $\hat{\beta}$ hasn't changed, the bands $\mathfrak{b}$ may have, and the intersection of $\hat{\beta}$ with the pieces of the standard Heegaard-double decomposition may have as well. Nonetheless, we can apply Lemma 3.18 iteratively to decrease c_1 to zero, at the cost of Markov-perturbing $\hat{\beta}$ into a $(v+c_1)$–braid $\hat{\beta}^+$.

Passing back to a $(b, (0, c_2, 0); v+c_1)$–bridge trisection filling $\hat{\beta}^+$ via Proposition 3.12, relabeling, and applying Proposition 3.14, we extract a $(b, (c_2, 0, 0); v+c_1)$–bridge-braided ribbon presentation $(\hat{\beta}^+, \mathfrak{b}'')$. Again, the bands and the precise bridge splitting may have changed. However, a second application of Lemma 3.18 allows us to decrease c_2 to zero, at the cost of Markov perturbing $\hat{\beta}^+$ into a $(v+c_1+c_2)$–braid $\hat{\beta}^{++}$. Note that we have Markov perturbed a total of $c = c_1 + c_2$ times. $\square$

Theorem 3.21 *Let $\mathbb{T}_0$ be the standard trisection of B^4, and let $\mathscr{F} \subset B^4$ be a neatly embedded surface with $\mathscr{L} = \partial\mathscr{F}$. Let $\hat{\beta}$ be an index v braiding of $\mathscr{L}$. Then the following are equivalent:*

(1) *$\mathscr{F}$ is ribbon.*

(2) *$\mathscr{F}$ admits a $(b, \boldsymbol{c}; v)$–bridge trisection filling $\hat{\beta}$ with $c_i = 0$ for some i.*

(3) *$\mathscr{F}$ admits a $(b, 0; v+c)$–bridge trisection filling a Markov perturbation $\hat{\beta}^+$ of $\hat{\beta}$.*

Proof Assume (1). Since $\mathscr{F}$ is ribbon, it admits a $(b, (c_1, c_2, 0); v)$–bridge-braided ribbon presentation $(\hat{\beta}, \mathfrak{b})$, by Proposition 3.5. By Proposition 3.12, this can be turned into a $(b, (c_1, c_2, 0); v)$–bridge trisection filling $\hat{\beta}$, which implies (2).

Assume (2). The bridge trisection filling $\hat{\beta}$ with $c_i = 0$ for some i gives a bridge-braided ribbon presentation $(\hat{\beta}, \mathfrak{b}')$ with $c_i = 0$ for the same i. By Lemma 3.20, there is a $(b, 0; v+c)$–bridge-braided ribbon presentation $(\hat{\beta}^{++}, \mathfrak{b}'')$ for $\mathscr{F}$, where $\hat{\beta}^{++}$ is a Markov perturbation of $\hat{\beta}$. By Proposition 3.12, this gives a $(b, 0; v+c)$–bridge trisection of $\mathscr{F}$ filling $\hat{\beta}^{++}$. This implies (3), where $\hat{\beta}^{++}$ is denoted by $\hat{\beta}^+$ for simplicity.

Assume (3). The $(b, 0; v+c)$–bridge trisection filling $\hat{\beta}^+$ gives rise to a bridge-braided ribbon presentation $(\hat{\beta}^+, \mathfrak{b}'')$ of the same type, by Proposition 3.14, such that $\mathscr{F} = \mathscr{F}_{(\hat{\beta}^+, \mathfrak{b}'')}$. However, a band presentation of a surface is precisely a handle-decomposition of the surface with respect to the standard Morse function on B^4. It follows that $\mathscr{F}$ can be built without caps; hence, $\mathscr{F}$ is ribbon, and (1) is implied.

Note for completeness that (2) can be seen to imply (1) by the argument immediately above, and that (3) implies (2) trivially. □

4 Tri-plane diagrams

A significant feature of the theory of trisections (broadly construed) is that it gives rise to new diagrammatic representations for four-dimensional objects (manifolds and knotted surfaces therein). In this section, we describe the diagrammatic theory for bridge trisections of surfaces in the four-ball. Recall the notational set-up of Section 3.1.

Let $(H, \mathscr{T})$ be a tangle with $H \cong B^3$. Let $E \subset H$ be a neatly embedded disk with $\partial \mathscr{T} \subset \partial E$. By choosing a generic projection of H onto E, we can represent $(H, \mathscr{T})$ by a *tangle diagram*. In the case that $H \cong B^3$, the lensed cobordism structure on $(H, \mathscr{T})$ discussed in Section 2.3 can be thought of as inducing the hemispherical decomposition of $\partial H \cong S^2$. So, we refer to $\partial_+ H$ and $\partial_- H$ as the *southern* and *northern* boundaries. This induces a decomposition of ∂E into a *northern arc* and a *southern arc*. See Figure 16 for examples of $(1, 2)$–tangle diagrams.

Definition 4.1 A $(b, c; v)$–*tri-plane diagram* is a triple $\mathbb{P} = (\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3)$ such that $\mathbb{P}_i$ is a (b, v)–tangle diagram and the union $\mathbb{P}_i \cup \overline{\mathbb{P}}_i$ is a tangle diagram for a split union of a v–braid with a c_i–component unlink. (Note that $\overline{\mathbb{P}}_i$ is the diagram $\mathbb{P}_i$ with crossing information reversed.) The southern arcs (and the $2b + v$ points $\boldsymbol{x}$ that they contain) are assumed to be identified. We denote the v points contained in the northern arc of $\mathbb{P}_i$ by $\boldsymbol{y}_i$; the three northern arcs are not identified.

A tri-plane diagram describes a bridge trisected surface in the following way. Let $(H_i, \mathscr{T}_i)$ be tangles corresponding to the tangle diagrams $\mathbb{P}_i$. Then the triple of tangle diagrams can be thought of as describing the union

$$(H_1, \mathscr{T}_1) \cup (H_2, \mathscr{T}_2) \cup (H_3, \mathscr{T}_3)$$

of these tangles, where $(H_i, \mathcal{T}_i) \cap \overline{(H_{i+1}, \mathcal{T}_i)} = (\Sigma, \boldsymbol{x})$. This explains the identification of the southern portions of the tangle diagrams in the definition. Now, by definition, each union $(H_i, \mathcal{T}_i) \cup \overline{(H_{i+1}, \mathcal{T}_{i+1})}$ is the split union of a braid with an unlink of c_i components inside a 3–ball. By Lemma 2.15, there is a unique way to glom on to this 3–ball a (c_i, v)–disk-tangle $(Z_i, \mathcal{D}_i)$, where $Z_i \cong B^4$. Therefore, the union

$$(Z_1, \mathcal{D}_1) \cup (Z_2, \mathcal{D}_2) \cup (Z_3, \mathcal{D}_3)$$

is a bridge trisected surface in B^4. The following is a corollary to Theorem 3.17, which showed that surface in B^4 admit bridge trisections.

Corollary 4.2 *Every neatly embedded surface in B^4 can be described by a tri-plane diagram.*

Proof By Theorem 3.17, every such surface in B^4 can be put in bridge position with respect to the genus zero trisection $\mathbb{T}_0$. The corresponding bridge trisection is determined by its spine

$$(H_1, \mathcal{T}_1) \cup (H_2, \mathcal{T}_2) \cup (H_3, \mathcal{T}_3).$$

This spine can be represented by a tri-plane diagram by choosing a triple of disks $E_i \subset H_i$ whose boundaries agree and choosing generic projections $H_i \twoheadrightarrow E_i$ that induce tangle diagrams for the $\mathcal{T}_i$. $\square$

The union $E_i \cup E_2 \cup E_3$ of disks that appeared in the proof above is called a *tri-plane* for the bridge trisection. We consider bridge trisections up to ambient isotopy, and an ambient isotopy of a bridge trisection can change the induced tri-plane diagram. These changes can manifest in following three ways, which we collectively call *tri-plane moves*. See Figure 16 for an illustration of each move.

An *interior Reidemeister move* on $\mathbb{P}$ is a Reidemeister move that is applied to the interior of one of the tangle diagrams $\mathbb{P}_i$. Interior Reidemeister moves correspond to ambient isotopies of the surface that are supported away from ∂B^4 and away from the core surface Σ. They also reflect the inherent indeterminacy of choosing a tangle diagram to represent a given tangle.

A *core (braid) transposition* is performed as follows: Pick a pair of adjacent bridge points $x, x' \in \boldsymbol{x}$, recalling that x and x' are (identified) points in the southern arc of each of the three tangles diagram. Apply a braid transposition to all three tangle diagrams that exchanges x and x'. This introduces a crossing in each tangle diagram; the introduced crossing should have the same sign in each diagram. Bridge sphere braiding corresponds to ambient isotopies of the surface that are supported in a neighborhood of the core surface Σ. Note that this gives an action of the braid group $\mathcal{M}(D^2, \boldsymbol{x})$ on the set of tri-plane diagrams.

A *page (braid) transposition* is performed as follows: Pick a pair of adjacent points $y, y' \in \boldsymbol{y}_i$ in the northern arc of one of the tangle diagrams. Apply a braid transposition to this tangle diagram that exchanges y and y'. In contrast to a core transposition, the braid transposition is only applied simultaneously to one diagram. Page transpositions correspond to ambient isotopies of the surface that are supported near ∂B^4.

Interior Reidemeister moves and core transpositions featured in the theory of bridge trisections of closed surfaces in the four-sphere described in [27]. See, in particular, [27, Lemma 7.4] for more details.
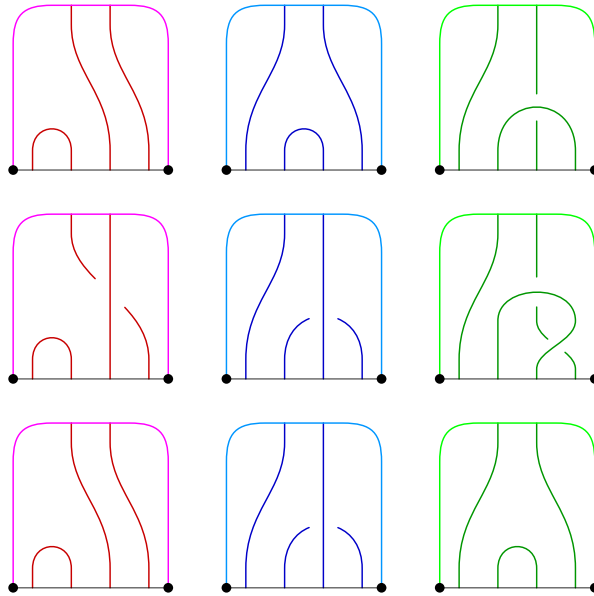
Figure 16: Top: a tri-plane diagram. Middle: the result of applying to the top a bridge sphere braid transposition at the third and fourth bridge points. Bottom: the result of applying to the middle a page braid transposition in the first tangle and a Reidemeister move in the third tangle.

Proposition 4.3 *Suppose* $\mathbb{P}$ *and* $\mathbb{P}'$ *are tri-plane diagrams corresponding to isotopic bridge trisections. Then* $\mathbb{P}$ *and* $\mathbb{P}'$ *can be related by a finite sequence tri-plane moves.*

Proof As in the proof of [27, Lemma 7.4], it suffices to assume that we have a fixed $\mathcal{E} = E_1 \cup E_2 \cup E_3$ within $\mathcal{H} = H_1 \cup H_2 \cup H_3$ and that we have two sets of seams $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2 \cup \mathcal{T}_3$ and $\mathcal{T}' = \mathcal{T}'_1 \cup \mathcal{T}'_2 \cup \mathcal{T}'_3$ determining a pair of isotopic spines in B^4.

Note that the southern endpoints of the $\mathcal{T}_i$ and the $\mathcal{T}'_i$ are both contained in the southern arc $\partial E_i \cap \mathrm{Int}(B^4)$, while all the northern endpoints are contained in the northern arc $\partial E_i \cap \partial B^4$. Without loss of generality, we assume the northern (resp. southern) endpoints of $\mathcal{T}_i$ agree with the northern (resp. southern) endpoints of $\mathcal{T}'_i$ for each i.

As in the proof of [27, Lemma 7.4], if f_t is an ambient isotopy of $\mathcal{H}$ such that f_0 is the identity and $f_1(\mathcal{T}) = \mathcal{T}'$, then f_t induces a loop in the configuration space of the bridge points $x = \mathcal{T} \cap \Sigma$. In this setting, f_t also induces, for each $i \in \mathbb{Z}_3$, a loop in the configuration space of the points $y \in \mathcal{T}_i \cap \partial_- H_i$ in the disk $\partial_- H_i$.

We write f_t as $f_t^\Sigma \cup f_t^1 \cup f_t^2 \cup f_t^3 \cup f_t'$, where f_t^Σ agrees with f_t in a small neighborhood of Σ and is the identity outside of a slightly larger neighborhood of Σ; f_t^i agrees with f_t in a small neighborhood of $\partial_- H_i$ and is the identity outside a slightly larger neighborhood of $\partial_- H_i$; and f_t' is supported away from the small neighborhoods of Σ and $\partial_- H_i$. Since these can be isolated to a single region near $\partial_- H_i$ for some i, they are independent of each other.

Since f_t^Σ corresponds to a braiding of the bridge points $\mathbf{x}$, there are tri-plane diagrams $\mathbb{P}$ and $\mathbb{P}^\Sigma$ corresponding to $\mathcal{T}$ and $\mathcal{T}^\Sigma = f_1^\Sigma(\mathcal{T})$ that are related by a sequence of core transpositions. Continuing, there is a tri-plane diagram $\mathbb{P}''$ corresponding to $\mathcal{T}'' = (f_1^1 \cup f_1^2 \cup f_1^3)(\mathcal{T}^\Sigma)$ that is related to $\mathbb{P}^\Sigma$ by a sequence of interior Reidemeister moves. Finally, the tri-plane diagram $\mathbb{P}'$ corresponds to $f_1'(\mathcal{T}'')$ and is related to $\mathbb{P}''$ by a sequence of page transpositions. In total, $\mathbb{P}$ and $\mathbb{P}'$ are related by a sequence of tri-plane moves, as desired. $\qquad\square$

4.1 Recovering the boundary braid from a tri-plane diagram

We now describe how to recover the boundary braid $(S^3, \mathcal{L}) = \partial(B^4, \mathcal{F})$ from the data of a tri-plane diagram for $(B^4, \mathcal{F})$. This process is illustrated in the example of the Seifert surface for the figure-8 knot in Figure 17; see Figure 8 for more details regarding this example. See also Figure 13 for another example.

Let $\mathbb{P} = (\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3)$ be a tri-plane diagram for a surface $(B^4, \mathcal{F})$. Let $\mathcal{E} = (E_1, E_2, E_3)$ denote the underlying tri-plane. Let $\partial_- E_i$ and $\partial_+ E_i$ denote the northern and southern boundary arcs of these disks, respectively, and let $S_i^0 = \partial_- E_i \cap \partial_+ E_i$ their 0–sphere intersections. Recall that, diagrammatically, the arcs $\partial_+ E_i$ correspond to the core surface Σ of the trisection, which is a disk, and the 0–spheres S_i^0 correspond to the unknot $B = \partial\Sigma$, which we think of as the binding of an open-book decomposition of S^3 with three disk pages given by the P_i. Recall that Σ is isotopic rel-∂ to each of the $\mathbb{P}_i$ via the arms H_i.

With this in mind, consider the planar link diagram $\circ\widehat{\mathbb{P}}$ obtained as follows. First, form the cyclic union

$$\overline{\mathbb{P}}_3 \cup \mathbb{P}_3 \cup \overline{\mathbb{P}}_2 \cup \mathbb{P}_2 \cup \overline{\mathbb{P}}_1 \cup \mathbb{P}_1,$$

where $\mathbb{P}_{i+1}$ and $\overline{\mathbb{P}}_i$ are identified along their southern boundaries, and $\overline{\mathbb{P}}_i$ and $\mathbb{P}_i$ are identified along their northern boundaries. Note that the cyclic ordering here is the opposite of what one might expect. This important subtlety is explained in the proof of Proposition 4.4 below. The corresponding union of the disks of the tri-plane

$$\overline{E}_3 \cup E_3 \cup \overline{E}_2 \cup E_2 \cup \overline{E}_1 \cup E_1$$

is topologically a two-sphere S^2. In particular, the 0–spheres S_i^0 have all been identified with a single 0–sphere S^0, which we think of as poles of the two-sphere. We represent this two-sphere in the plane by cutting open along $\partial_- E_1$ and embedding the resulting bigon so that the E_i and $\overline{E}_i$ lie in the yz–plane with $E_3 \cap \overline{E}_2$ on the y–axis. See Figure 17(b). In this way, the diagram $\circ\widehat{\mathbb{P}}$ encodes a link in a three-sphere. The unknotted binding B in S^3 can be thought of as the unit circle in the xy–plane. (The positive x–axis points out of the page.) Each longitudinal arc on S^2, including the northern and southern arcs of each E_i, corresponds to a distinct page, given six in all. However, the ambient three-sphere in which this link lives is not $S^3 = \partial B^4$, as the proof of Proposition 4.4 will make clear.

Note that the diagram $\circ\widehat{\mathbb{P}}$ will have only two types of connected components:

(1)   components that meet each disk E_i and are homotopically essential in $S^2 \setminus \nu(S^0)$, and

(2)   components that are null-homotopic and are contained in some pair $E_{i+1} \cup \overline{E}_i$.

Components of type (1) will correspond to the boundary link $(S^3, \mathscr{L})$, while components of the second kind will correspond to split unknots. The components of type (1) are not braided in the sense of being everywhere transverse to the longitudinal arcs of S^2 but, as we shall justify below, they become braided after a sequence of Reidemeister moves and isotopies that are supported away from S^0. Define $\circ\mathbb{P}$ to be the result of discarding all components of type (2) from $\circ\widehat{\mathbb{P}}$, then straightening out the arcs of type (1) until they give a braid diagram in the sense that they are everywhere transverse to the longitudinal arcs connecting the poles $S^0 \subset S^2$.

Proposition 4.4 *Suppose* $\mathbb{P} = (\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3)$ *is a tri-plane diagram for* $(B^4, \mathscr{F})$. *Then the diagram* $\circ\mathbb{P}$ *is a braid diagram for the boundary link* $(S^3, \mathscr{L}) = \partial(B^4, \mathscr{F})$.

Proof Consider the spine $H_1 \cup H_2 \cup H_3$ of the genus zero trisection $\mathbb{T}_0$ of B^4. Let N be a small lensed neighborhood of this spine inside B^4. Here, the qualifier "lensed" has the effect that $N \cap \partial B^4$ is unchanged,

$$N \cap \partial B^4 = P_1 \sqcup P_2 \sqcup P_3.$$

We can decompose ∂N into six pieces,

$$\partial N = H_1^+ \cup H_3^- \cup H_3^+ \cup H_2^- \cup H_2^+ \cup H_1^-,$$

where the pieces intersect cyclically in the following manner: the $H_{i+1}^+ \cap H_i^- = \Sigma_i^-$ are the three obvious push-offs of Σ into ∂N, and $H_i^- \cap H_i^+ = P_i$. Because $B = \partial\Sigma = \partial\Sigma_i = \partial P_i$, it follows that ∂N is a closed 3–manifold. In fact, there is an obvious "radial" diffeomorphism $N \to B^4$ that pushes $H_{i+1}^+ \cup H_i^-$ onto Y_i in an *orientation-preserving* way. To unpack this last statement, recall that Z_i induces an orientation on its boundary such that

$$\partial Z_i = H_i \cup_\Sigma \overline{H}_{i+1} \cup Y_i.$$

In ∂N, we have corresponding pieces $H_{i+1}^+ \cup_{\Sigma_i^-} H_i^+$, but the correspondences

$$H_i \leftrightarrow H_i^-, \quad \overline{H}_{i+1} \leftrightarrow H_{i+1}^+, \quad \Sigma \leftrightarrow \Sigma_i^-$$

all reverse orientation. This is because the outward normal to N points into Z_i. Figure 36, left, provides a potentially helpful schematic.

Bringing the surface $\mathscr{F}$ into the picture, we have the identification

$$\partial N \cap \mathscr{F} = (H_1, \mathscr{T}_1) \cup \overline{(H_3, \mathscr{T}_3)} \cup (H_3, \mathscr{T}_3) \cup \overline{(H_2, \mathscr{T}_2)} \cup (H_2, \mathscr{T}_2) \cup \overline{(H_1, \mathscr{T}_1)}.$$

If $\mathscr{E} = E_1 \cup E_2 \cup E_3$ was our original tri-plane, then there are obvious disks $E_i^\pm \subset H_i^\pm$ onto which $\partial N \cap \mathscr{F}$ can be projected. As discussed in the text preceding this proposition, the union of the $E_i^\pm$ is a two-sphere, which can be identified with the plane, as discussed. Adopting this identification, we find that the induced diagram $\circ\mathbb{P}$ is a planar diagram for $\partial N \cap \mathscr{F}$.

Recall that, by definition, $\mathbb{P}_{i+1} \cup \overline{\mathbb{P}}_i$ is a diagram for (the mirror of) a split union of a braid with an unlink. Thus, the total union $\circ\mathbb{P}$ is (currently) a diagram for a split union of a closed braid and three unlinks.
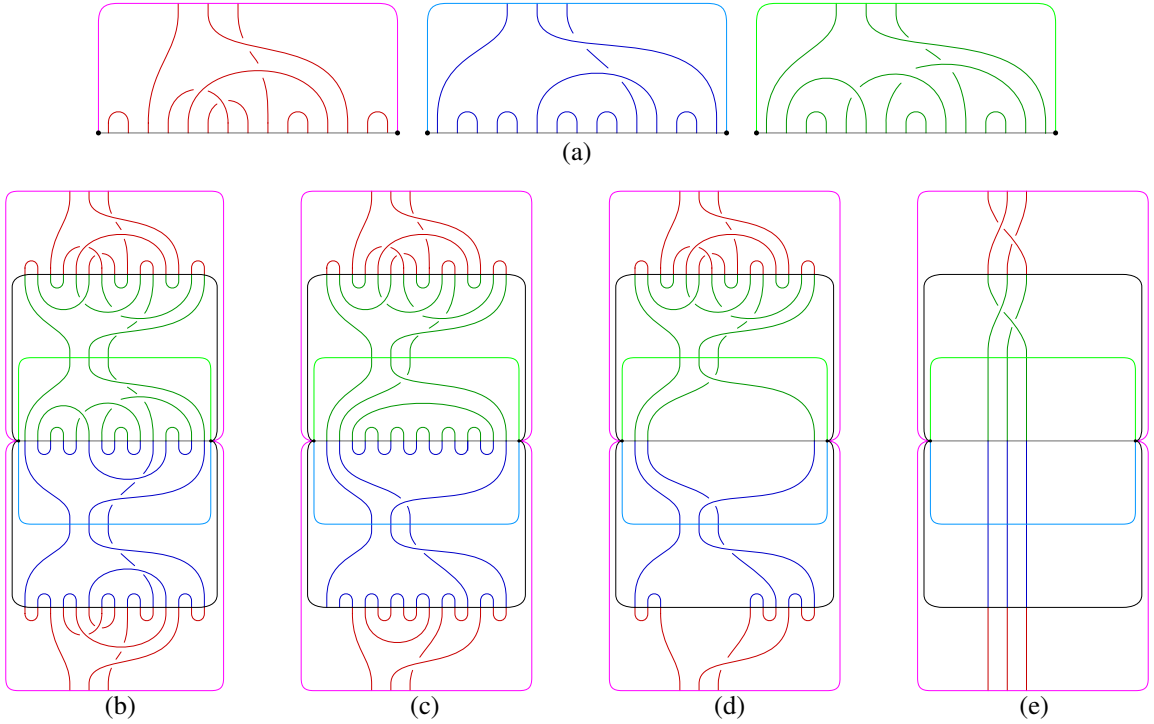
Figure 17: Recovering the boundary braid (e) from a tri-plane diagram (a). Compare with Figure 8.

Note that although the diagram describes a closed (geometric) braid, the diagram may not be braided. See Figure 17(b).

It remains to observe how this diagram changes as the neighborhood N is enlarged until it fills up all of B^4 and ∂N coincides with $S^3 = \partial B^4$. Two things happen in the course of this. First, the unlinks will shrink to points and disappear as the neighborhood N is enlarged to encompass the flat patches of the trivial disk-tangles that cap them off. Second, the portions of the diagram corresponding to the closed braid will "straighten out", meaning they will deperturb until the diagram is an honest braid diagram. Finally, the neighborhood N will coincide with all of B^4, the union of the $E_i^{\pm}$ will live in S^3, and the diagram $\circ \mathbb{P}$ will correspond to a braid diagram for $\mathscr{L} = \partial \mathscr{F}$, as desired.   □

Example 4.5 Figure 17(b) shows the diagram $\circ \widehat{\mathbb{P}}$ corresponding to the tri-plane diagram in Figure 17(a). (This tri-plane diagram corresponds to the Seifert surface for the figure-8 knot; see Figure 8 for more details.) The two black dots represent the braid axis, and each arc connecting the these dots corresponds to a disk page of the braid axis.

As described in the proof of Proposition 4.4, the sequence in Figure 17(b)–(e) can be thought of as describing the cross-section of the bridge trisected surface with concentric shells in B^4, starting with the boundary of a regular neighborhood of the spine of the trisection of B^4 and terminating in the boundary of B^4. Moving from (b) to (c) in Figure 17, the cross-section changes only by isotopy, revealing clearly

the presence of two unknotted, type (2) components. In the transition to Figure 17(d), these components cap off and disappear. In the transition to Figure 17(e), the flat structure is forgotten as we deperturb. The end result is the boundary of the surface, described by a braid.

For more examples, see Figures 13 and 14, which were discussed in Examples 3.15 and 3.16, respectively.

5 Shadow diagrams

The previous section developed a diagrammatic representation and calculus for bridge trisections in B^4 that made use of the fact that B^4 admits a genus zero trisection. In this section, we switch to an analysis of diagrams for bridge trisection of surfaces in general four-manifolds. Here, the tri-plane-diagrammatic approach is not possible, so we work instead with objects called shadow diagrams.

Consider a $(g, b; \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–tangle $(H, \mathcal{T})$. Let Δ be a bridge disk system for $\mathcal{T}$. We now fix some necessary notation.

- Let $\Sigma = \partial_+ H$.
- Let $\alpha \subset \Sigma$ be a defining set of curves for H, disjoint from Δ.
- Let $\mathfrak{a}$ denote a collection of neatly embedded arcs, disjoint from Δ and α such that surgering Σ along α and $\mathfrak{a}$ results in a disjoint union of disks. We assume $|\mathfrak{a}|$ is minimized.
- Let $\mathcal{T}^*$ denote the shadows of the flat strands of $\mathcal{T}$ — ie those coming from the bridge semidisks.
- Let $\mathscr{A}^*$ denote the shadows for the vertical strands — ie those coming from the bridge triangles.
- Let $\boldsymbol{x} = \mathcal{T} \cap \Sigma$.

The tuple $(\Sigma, \alpha, \mathcal{T}^*, \boldsymbol{x})$ is called a *tangle shadow* for the pair $(H, \mathcal{T})$. The tuple $(\Sigma, \alpha, \mathfrak{a}, \mathcal{T}^*, \mathscr{A}^*, \boldsymbol{x})$ is called an *augmented tangle shadow* for the pair $(H, \mathcal{T})$. We will say that an augmented tangle shadow is an *augmenting* of the underlying tangle shadow. Figure 18 shows a pair of augmented tangle shadows: one is found by considering the red, pink, and orange arcs and curves, while the other is found by considering the dark blue, light blue, and orange arcs and curves. Note that we consider (augmented) tangle shadows up to isotopy rel-∂.

Lemma 5.1 *A tangle shadow determines a tangle $(H, \mathcal{T})$ up to an isotopy fixing $\Sigma = \partial_+ H$.*

Note that a tangle shadow cannot detect braiding of $(H, \mathcal{T})$ supported near $(P, \boldsymbol{y}) = \partial_-(H, \mathcal{T})$; augmenting the shadow diagram does not solve this problem.

Proof Given a shadow diagram $(\Sigma, \alpha, \mathcal{T}^*, \boldsymbol{x})$, let H be the lensed cobordism obtained from the spread $H \times [0, 1]$ by attaching 3–dimensional 2–handles along the curves $\alpha \times \{1\}$. Let $\mathcal{T} \subset H$ be obtained by perturbing the interiors of the shadows $\mathcal{T}^* \times \{0\}$ into the interior of H to obtain the flat strands of $\mathcal{T}$ and
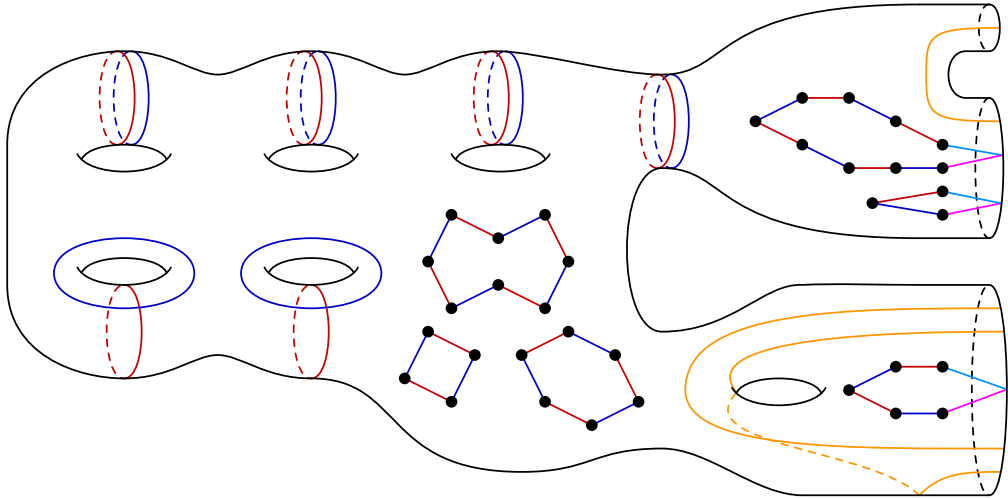
Figure 18: A pair of (augmented) tangle shadows that, taken together, give a standard (augmented) splitting shadow. The relevant parameters for each handlebody are $g = 6$, $n = 2$, $m = 3$, $\boldsymbol{p} = (0, 1)$, and $\boldsymbol{f} = (2, 1)$. The relevant parameters for each tangle are $b = 16$ and $\boldsymbol{v} = (0, 2, 1)$. The arcs and curves of the α_i and $\mathcal{T}_i^*$ for $i = 1, 2$ are shown in red and blue, respectively, while the arcs of the $\mathcal{A}_i$ are shown in pink and light blue, respectively, and the arcs of $\mathfrak{a}_1 = \mathfrak{a}_2$ are shown in orange.

extending the marked points $\boldsymbol{x}$ to vertical arcs $\boldsymbol{x} \times [0, 1]$ using a product structure of the spread to obtain the vertical strands of $\mathcal{T}$. Such a product structure is unique up to diffeomorphism of $(P, \boldsymbol{y}) = \partial_-(H, \mathcal{T})$, so the resulting tangle is determined up to braiding near $(P, \boldsymbol{y})$; any two tangles differing thusly are isotopic via an isotopy supported away from Σ. $\square$

As a matter of convention, we have assumed without loss of generality that the curves and arcs of $\alpha \cup \mathfrak{a} \cup \mathcal{T}^* \cup \mathcal{A}^*$ are all pairwise disjoint; it is not strictly necessary, for example, to assume $\alpha \cap \mathcal{T}^* = \varnothing$, but this can always be achieved. Given a tangle shadow $(\Sigma, \alpha, \mathcal{T}^*, \boldsymbol{x})$, we recall two standard moves: Let α_1 and α_2 be two curves in α, and let ω be a embedded arc in Σ connecting α_1 to α_2 such that $\text{Int}(\omega) \cap (\alpha \cup \mathcal{T}^* \cup \boldsymbol{x}) = \varnothing$. Then $N = \nu(\alpha_1 \cup \omega \cup \alpha_2)$ is a pair of pants. Let α_1' be the boundary component of N not parallel to α_1 or α_2. Then $\alpha' = \alpha \setminus \{\alpha_1\} \cup \{\alpha_1'\}$ is a new defining set of curves for H. We say that α' is obtained from α by a *curve slide* of α_1 over α_2 along ω. Now let τ_1^* be an arc of $\mathcal{T}^*$ and let α_2 be a curve in α (resp. the boundary of a regular neighborhood of another arc τ_2^* of $\mathcal{T}^*$). Let ω be a embedded arc in Σ connecting τ_1^* to α_2 such that $\text{Int}(\omega) \cap (\alpha \cup \mathcal{T}^* \cup \boldsymbol{x}) = \varnothing$. Let $(\tau_1^*)'$ denote the arc obtained by banding τ_1^* to α_2 using the surface-framed neighborhood of ω. Then $(\mathcal{T}^*)' = \mathcal{T}^* \setminus \tau_1^* \cup (\tau_1^*)'$ is a new collection of shadows for the flat strands of $\mathcal{T}$. We say that $(\mathcal{T}^*)'$ is obtained from $\mathcal{T}^*$ by an *arc slide* of τ_1^* over α_2 (resp. τ_2^*) along ω. Two shadow diagrams for $(H, \mathcal{T})$ are called *slide-equivalent* if they can be related by a sequence of curve slides and arc slides.

Given an augmented tangle shadow $(\Sigma, \alpha, \mathfrak{a}, \mathcal{T}^*, \mathcal{A}^*, \boldsymbol{x})$, we have further moves. Similar to above, we have arc slide moves that allow one to slide arcs of $\mathfrak{a}$ or $\mathcal{A}^*$ over arcs and curves of α and $\mathcal{T}^*$. Note that

we do not allow an arc or curve of any type to slide over an arc of $\mathfrak{a}$ nor $\mathscr{A}^*$. Two (augmented) shadow diagrams that are related by a sequence of these two types of moves are called *slide-equivalent*. The following is a generalization of a foundational result of Johansson [17], and follows from a standard argument, which we sketch.

Proposition 5.2 *Two tangle shadows for a given tangle are slide-equivalent.*

Proof Let $(\Sigma, \alpha, \mathscr{T}^*, \boldsymbol{x})$ and $(\Sigma, \beta, \mathscr{S}^*, \boldsymbol{x})$ be two shadow diagrams that define the same tangle $(H, \mathscr{T})$. Assume these diagrams have been isotoped to intersect minimally. We will show that there is a sequence of isotopies and slides among the arcs and curves of $\alpha \cup \mathscr{T}^*$ that result in these arcs and curves agreeing with those of $\beta \cup \mathscr{S}^*$.

Choose cut disks $\mathscr{D}(\alpha)$ and $\mathscr{D}(\beta)$ in H, so $\partial \mathscr{D}(\alpha) = \alpha$ and $\partial \mathscr{D}(\beta) = \beta$. Choose bridge disks $\Delta(\mathscr{T}^*)$ and $\Delta(\mathscr{S}^*)$, so $\partial \Delta(\mathscr{T}^*) = \mathscr{T}^* \cup_{\boldsymbol{x}} \mathscr{T}$ and $\partial \Delta(\mathscr{S}^*) = \mathscr{S}^* \cup_{\boldsymbol{x}} \mathscr{T}$. Assume that $\mathscr{D}(\alpha) \cap \Delta(\mathscr{T}^*) = \mathscr{D}(\beta) \cap \Delta(\mathscr{S}^*) = \varnothing$, and $\mathscr{D}(\alpha) \cap \mathscr{T} = \mathscr{D}(\beta) \cap \mathscr{T} = \varnothing$.

We can assume there are no closed curves of intersection between the collections of disks as follows. Suppose, for example, that $\mathscr{D}(\alpha) \cap \Delta(\mathscr{S}^*)$ contains a closed curve component. Choose one such component that is innermost in $\Delta(\mathscr{S})$, bounding a disk $D \subset \Delta(\mathscr{S}^*)$ with $\text{Int}(D) \cap (\mathscr{D}(\alpha) \cup \Delta(\mathscr{T}^*)) = \varnothing$. Surger $\mathscr{D}(\alpha)$ along D, discarding the sphere component to get a new cut system filling α. Repeating, we can arrange via surgery, that there are no curves of intersection among any of the disks.

It follows that every component of $(\mathscr{D}(\alpha) \cup \Delta(\mathscr{T}^*)) \cap (\mathscr{D}(\beta) \cup \Delta(\mathscr{S}^*))$ is an arc that is neatly embedded in each of the two disk coinciding along it. There are three of cases to consider, based on whether this arc intersects $\mathscr{T}$ at (i) both endpoints, (ii) one endpoint, or (iii) no endpoints.

Choose an arc a of intersection of type (i) that is outermost in $\Delta(\mathscr{S}^*)$, so it cobounds and embedded disk (a bigon) $D \subset \Delta(\mathscr{S}^*)$ with an arc b of $\mathscr{T}$. The arc a also cobounds a disk E in $\Delta(\mathscr{T}^*)$ with the arc b. A slight push-off of $D \cup E$ is an embedded two-sphere in $H \setminus \nu(\mathscr{D}(\alpha) \cup \Delta(\mathscr{T}^*))$, which is homeomorphic to $P \times I$, so it bounds a three-ball. (Here, $P = \partial_- H$.) Note that this three-ball might intersect $\mathscr{D}(\beta) \cup \Delta(\mathscr{S}^*)$, but this is of no concern. The three-ball guides and isotopy of E rel-boundary until it agrees with D; then, E can be isotoped rel-$\mathscr{T}$ off D, to remove the arc a of intersection. This reduces the number of arcs of intersection of type (i), and can be repeated until none remain.

Next, choose an arc a of intersection of type (iii) that is outermost in $\mathscr{D}(\beta) \cup \Delta(\mathscr{S}^*)$, so it cobounds an embedded disk (a bigon) $D \subset (\mathscr{D}(\beta) \cup \Delta(\mathscr{S}^*))$ with an arc b of $\beta \cup \mathscr{S}^*$. The arc a also cobounds a disk E in $\mathscr{D}(\alpha) \cup \Delta(\mathscr{T}^*)$ with some arc $b' \subset \alpha \cup \mathscr{T}^*$. Let Σ' denote the surface obtained from Σ by surgering along $\mathscr{D}(\alpha)$, excluding (if applicable) a disk containing a. We think of Σ' as an embedded submanifold of H that agrees with Σ away from the curves of surgery, so $b \cup b'$ is a curve in Σ'. Let H' denote the compression body cobounded by Σ' and P, and note that H' is either $P \times I$ or $P \times I$, plus a 1–handle whose belt-sphere is a curve of α containing a. In either event, $b \cup b'$ bounds the disk $D \cup E$ in H'.

First, if $D \cup E$ is boundary parallel (into Σ'), then there is a disk in Σ' bounded by $b \cup b'$. We can isotope b' across this disk to make it agree with b, then we can push it off b. During this isotopy, we might push b' over scars of the surgery and over shadow arcs of $\mathcal{T}^*$. In this case, there is a sequence of isotopies and slides that move b' to b on Σ, and the arc a of intersection is removed.

Second, if $D \cup E$ is not boundary-parallel in H', then it must be isotopic to a disk of $\mathcal{D}(\alpha)$ containing a. In this case, let b'' be the arc of α such that $b' \cup b''$ is the curve of α containing a. It follows that b'' is isotopic to b in Σ', so we can proceed as above to move b'' and remove the arc a of intersection. In this way, we can assume that, after some slides and isotopy, there are no arcs of intersection of type (iii).

Arcs of intersection of type (ii) can be removed in a similar way, combining aspects of the first two arguments. The result is that slides (among the curves and arcs of $\alpha \cup \mathcal{T}^*$) can be performed to achieve that $\alpha \cup \mathcal{T}^*$ is disjoint from $\beta \cup \mathcal{S}^*$. Surger Σ along the curves of β to get Σ'. In Σ', the curves of α all bound disks, so they can be isotoped to agree with the scars of the β curves. These isotopies might move α curves across each other and over arcs of $\mathcal{T}^*$, and these occurrences correspond to slides. Similarly, in Σ' the arcs of $\mathcal{T}^*$ are isotopic rel-boundary to those of $\mathcal{S}^*$, with these isotopies potentially involving slides over each other and over the scars of the curves of α. The end result is that $\alpha \cup \mathcal{T}^* = \beta \cup \mathcal{S}^*$, as desired. $\square$

A tuple $(\Sigma, \alpha_1, \alpha_2, \mathcal{T}_1^*, \mathcal{T}_2^*, \boldsymbol{x})$ is called a *splitting shadow* if each tuple $(\Sigma, \alpha_i, \mathcal{T}_i^*, \boldsymbol{x})$ is a tangle shadow. A splitting shadow gives rise to a bridge splitting of pair (M, K) in the same way that a tangle shadow gives rise to a tangle (see Lemma 5.1); in particular, K is determined only up to braiding supported near ∂M. Recall the notion of a standard bridge splitting of (M, K) from Section 2.6. If a splitting shadow corresponds to a standard bridge splitting, then the tangle shadows $(\Sigma, \alpha_i, \mathcal{T}_i^*, \boldsymbol{x})$ are (for $i = 1, 2$, respectively) slide-equivalent to tangle shadows $(\Sigma, \alpha_i', (\mathcal{T}_i^*)', \boldsymbol{x})$ such that $(\Sigma, \alpha_1', \alpha_2')$ is a standard Heegaard diagram (Section 2.4) and $(\mathcal{T}_1^*)' \cup (\mathcal{T}_2^*)'$ is a neatly embedded collection of polygonal arcs and curves such that the polygonal curves bound disjointly embedded disks. We call such a splitting shadow $(\Sigma, \alpha_1, \alpha_2, \mathcal{T}_1^*, \mathcal{T}_2^*, \boldsymbol{x})$ *standard*. Figure 18 shows a standard splitting shadow (ignore the pink, light blue, and orange arcs for now). Two splitting shadows are called *slide-equivalent* if the two pairs of corresponding tangle shadows are slide-equivalent.

Definition 5.3 A $(g, \boldsymbol{k}, f, \boldsymbol{c}; \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–*shadow diagram* is a tuple $(\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, \boldsymbol{x})$, such that the tuple $(\Sigma, \alpha_i, \alpha_{i+1}, \mathcal{T}_i^*, \mathcal{T}_{i+1}^*, \boldsymbol{x})$ is slide-equivalent to a standard splitting shadow for each $i \in \mathbb{Z}_3$.

Two shadow diagrams are called *slide-equivalent* if the three pairs of corresponding tangle shadows are slide-equivalent.

Figure 19 shows a shadow diagram corresponding to the bridge trisection of the ribbon disk for the stevedore knot described in Figure 10. Note the orientation convention: the shadow diagram surface Σ is oriented positively as the boundary of each arm of the spine. So, we should rotate each tangle represented
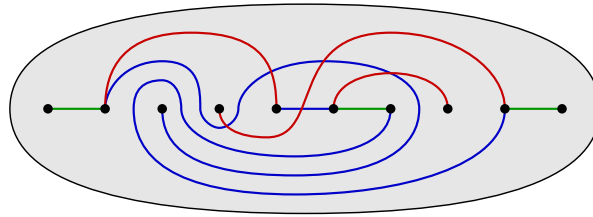
Figure 19: A shadow diagram for the bridge trisection given in Figure 10, which corresponds a ribbon disk for the stevedore knot.

in Figure 10(f) 90° backwards into the plane of the page, so that we are viewing Σ from above in order to arrive at the correct shadow diagram. This subtlety is the source of some confusion in the literature; see [18, Remark 2.10] for a related discussion.

Proposition 5.4 *A $(g, k, f, c; p, f, v)$–shadow diagram determines the spine of a $(g, k, f, c; p, f, v)$–bridge trisection uniquely. Any two shadow diagrams for a fixed bridge trisection are slide-equivalent.*

Proof First, note that a shadow diagram determines the spine of a bridge trisection. This follows immediately from the definition of a shadow diagram, Lemma 5.1, and the definition of a spine; see Proposition 2.22. The first claim follows from the fact that a bridge trisection is determined up to diffeomorphism by its spine, by Proposition 2.22. The second claim follows from Proposition 5.2. □

Since bridge trisections are determined by their spines (Corollary 2.23), we find that any surface $(X, \mathscr{F})$ can be described by a shadow diagram.

Corollary 5.5 *Let X be a smooth, orientable, compact, connected four-manifold, and let $\mathscr{F}$ be a neatly embedded surface in X. Then $(X, \mathscr{F})$ can be described by a shadow diagram.*

5.1 Recovering the boundary braid from a shadow diagram

We now see how to recover the information about the boundary of a bridge trisected pair $(X, \mathscr{F})$. By augmenting a shadow diagram for the bridge trisection, we will recover this information in the form of an abstract open-book braiding, as defined in Section 2.8. What follows is based on the monodromy algorithm described by Castro, Gay, and Pinzón-Caicedo in [6] and is closely related to the notion of an arced relative trisection diagram, as described in [11].

To start, we return our attention to pairs of augmented tangle shadows. A tuple

$$(\Sigma, \alpha_1, \alpha_2, \mathfrak{a}_1, \mathfrak{a}_2, \mathscr{T}_1^*, \mathscr{T}_2^*, \mathscr{A}_1^*, \mathscr{A}_2^*, x)$$

is called a *standard augmented splitting shadow* if

- for each $i = 1, 2$, $(\Sigma, \alpha_i, \mathfrak{a}_i, \mathscr{T}_i^*, \mathscr{A}_i^*, x)$ is a augmented tangle shadow;
- $(\Sigma, \alpha_1, \alpha_2, \mathscr{T}_1^*, \mathscr{T}_2^*, x)$ is a standard splitting shadow;

- the components of $\mathcal{T}_1^* \cup \mathcal{T}_2^* \cup \mathcal{A}_1^* \cup \mathcal{A}_2^*$ intersecting $\partial\Sigma$ bound disjointly embedded polygonal disks, each of which intersects $\partial\Sigma$ in a single point; and

- $\mathfrak{a}_1 = \mathfrak{a}_2$.

See Figure 18 for an example of a standard augmented splitting shadow.

Definition 5.6 (augmented shadow diagram) An *augmented* $(g, \boldsymbol{k}, f, \boldsymbol{c}; \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–*shadow diagram* is a tuple $(\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathfrak{a}_1, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, \mathcal{A}_1^*, \boldsymbol{x})$, such that the tuple $(\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, \boldsymbol{x})$ is a shadow diagram, and $(\Sigma, \alpha_1, \mathfrak{a}_1, \mathcal{T}_1^*, \mathcal{A}_1^*, \boldsymbol{x})$ is an augmented tangle shadow.

A *fully augmented* $(g, \boldsymbol{k}, f, \boldsymbol{c}; \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{v})$–*shadow diagram* is a tuple

$$(\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathfrak{a}_1, \mathfrak{a}_2, \mathfrak{a}_3, \mathfrak{a}_4, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, \mathcal{A}_1^*, \mathcal{A}_2^*, \mathcal{A}_3^*, \mathcal{A}_4^*, \boldsymbol{x})$$

such that the tuple $(\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, \boldsymbol{x})$ is a shadow diagram, the tuples $(\Sigma, \alpha_1, \mathfrak{a}_1, \mathcal{T}_1^*, \mathcal{A}_1^*, \boldsymbol{x})$ and $(\Sigma, \alpha_1, \mathfrak{a}_4, \mathcal{T}_1^*, \mathcal{A}_4^*, \boldsymbol{x})$ are augmented tangle shadows for the same tangle, and:

(1) For $i = 1, 2$, the diagram

$$(\Sigma, \alpha_i, \alpha_{i+1}, \mathfrak{a}_i, \mathfrak{a}_{i+1}, \mathcal{T}_i^*, \mathcal{T}_{i+1}^*, \mathcal{A}_i^*, \mathcal{A}_{i+1}^*, \boldsymbol{x})$$

is slide-equivalent to a standard augmented splitting shadow

$$(\Sigma, \alpha_i', \alpha_{i+1}', \mathfrak{a}_i', \mathfrak{a}_{i+1}', (\mathcal{T}_i^*)', (\mathcal{T}_{i+1}^*)', (\mathcal{A}_i^*)', (\mathcal{A}_{i+1}^*)', \boldsymbol{x}).$$

(2) The diagram

$$(\Sigma, \alpha_3, \alpha_1, \mathfrak{a}_3, \mathfrak{a}_4, \mathcal{T}_3^*, \mathcal{T}_1^*, \mathcal{A}_3^*, \mathcal{A}_4^*, \boldsymbol{x})$$

is slide-equivalent to a standard augmented splitting shadow

$$(\Sigma, \alpha_3'', \alpha_1'', \mathfrak{a}_3'', \mathfrak{a}_4'', (\mathcal{T}_3^*)'', (\mathcal{T}_1^*)'', (\mathcal{A}_3^*)'', (\mathcal{A}_4^*)'', \boldsymbol{x}).$$

We say that an augmented shadow diagram is an *augmenting* of the underlying shadow diagram and that a fully augmented shadow diagram is a *full-augmenting* of the underlying (augmented) shadow diagram.

We now describe how the data of an augmented shadow diagram allows us to recover the boundary open-book braiding $(Y, \mathcal{L})$ of the corresponding bridge trisected pair $\partial(X, \mathcal{F})$. First, we note the following crucial connection between augmented shadow diagrams and fully augmented shadow diagrams.

Proposition 5.7 *There is an algorithmic way to complete an augmented shadow diagram to a fully augmented shadow diagram, which is unique up to slide-equivalence.*

Proof Start with an augmented shadow diagram $(\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathfrak{a}_1, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, \mathcal{A}_1^*, \boldsymbol{x})$. Restrict attention to the splitting shadow $(\Sigma, \alpha_1, \alpha_2, \mathcal{T}_1^*, \mathcal{T}_2^*, \boldsymbol{x})$. By definition, this diagram is slide-equivalent to a standard splitting shadow $(\Sigma, \alpha_1', \alpha_2', (\mathcal{T}_1^*)', (\mathcal{T}_2^*)', \boldsymbol{x})$. Choose a sequence of arc and curve slides
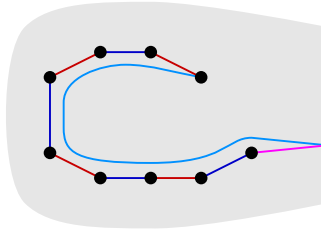
Figure 20: Obtaining $\mathcal{A}_2^*$ from $(\mathcal{A}_1^*)'$.

realizing this equivalence. Whenever a slide involving the arcs and curves of $\alpha_1 \cup \mathcal{T}_1^*$ is performed along an arc ω that intersects $\mathfrak{a}_1 \cup \mathcal{A}_1^*$, first slide the offending arcs of $\mathfrak{a}_1 \cup \mathcal{A}_1^*$ out of the way using the same slide-arc ω. Now the splitting shadow has been standardized, but the arcs of $\mathfrak{a}_1 \cup \mathcal{A}_1^*$ may intersect the curves and arcs of $\alpha_2' \cup (\mathcal{T}_2^*)'$. Intersections of $\mathfrak{a}_1 \cup \mathcal{A}_1^*$ with the curves of α_2' can be removed via slides over the curves of α_1' dual to curves of α_2'. Recall that the closed components of $(\mathcal{T}_1^*) \cup (\mathcal{T}_2^*)'$ are embedded polygonal curves, while the nonclosed components are embedded polygonal arcs. Moreover, the arcs of $\mathcal{A}_1^*$ connect one end of each polygonal arc to $\partial\Sigma$. Intersections of (the interior of) $\mathfrak{a}_1 \cup \mathcal{A}_1^*$ with the polygonal curves of $(\mathcal{T}_1^*) \cup (\mathcal{T}_2^*)'$ can be removed via slides over the arcs of $(\mathcal{T}_1^*)'$ included in these polygonal curves. Intersections of (the interior of) $\mathfrak{a}_1 \cup \mathcal{A}_1^*$ with the polygonal arcs of $(\mathcal{T}_1^*) \cup (\mathcal{T}_2^*)'$ can be removed via slides over the arcs of $(\mathcal{T}_1^*)'$ included in these polygonal arc, provided one is careful to slide towards the end of the polygonal arc that is not attached to $\mathcal{A}_1^*$.

Once the described slides have all been carried out, the collections $\mathfrak{a}_1$ and $\mathcal{A}_1^*$ of arcs will have been transformed into new collections, which we denote by $\mathfrak{a}_1'$ and $(\mathcal{A}_1^*)'$, respectively. The key fact is that $\mathfrak{a}_1'$ and $(\mathcal{A}_1^*)'$ are disjoint (in their interiors) from the arcs and curves of $\alpha_2' \cup (\mathcal{T}_2^*)'$. Set $\mathfrak{a}_2 = \mathfrak{a}_1'$, and note that $\mathfrak{a}_2$ has the desired property of being (vacuously) slide-equivalent to $\mathfrak{a}_2' = \mathfrak{a}_1'$. To define $\mathcal{A}_2^*$, note that at this point the union of the polygonal arcs of $(\mathcal{T}_1^*)' \cup (\mathcal{T}_2^*)'$ with $(\mathcal{A}_1^*)'$ is a collection of embedded "augmented" polygonal arcs each of which intersects $\partial\Sigma$ in a single point. Let $\mathcal{A}_2^*$ be the collection of arcs obtained by pushing each augmented polygonal arc off itself slightly, while preserving its endpoint that lies in the interior of Σ. See Figure 20. This can be thought of as sliding the endpoint of $(\mathcal{A}_1^*)'$ that lies in the interior of Σ along the polygonal arc of $(\mathcal{T}_1^*)' \cup (\mathcal{T}_2^*)'$ that it intersects until it reaches the end. Having carried out these steps, we have that $(\Sigma, \alpha_1', \alpha_2', \mathfrak{a}_1', \mathfrak{a}_2, (\mathcal{T}_1^*)', (\mathcal{T}_2^*)', (\mathcal{A}_1^*)', \mathcal{A}_2^*, \boldsymbol{x})$ is a standard augmented splitting shadow, as desired.

Next, we repeat the process outlined in the first two paragraph, starting this time with the splitting shadow $(\Sigma, \alpha_2', \alpha_3, (\mathcal{T}_2^*)', \mathcal{T}_3^*, \boldsymbol{x})$: Standardize the splitting shadow, and include the arcs of $\mathfrak{a}_2 \cup \mathcal{A}_2^*$ in the slides when necessary. Perform additional slides to obtain the new collection of arcs $\mathfrak{a}_2'$, and $(\mathcal{A}_2^*)'$ whose interiors are disjoint from all other arcs and curves. Let $\mathfrak{a}_3 = \mathfrak{a}_2'$, and obtain $\mathcal{A}_3^*$ from $(\mathcal{A}_2^*)'$ in the same way as before, so that the new diagram $(\Sigma, \alpha_2'', \alpha_3', \mathfrak{a}_2', \mathfrak{a}_3, (\mathcal{T}_2^*)'', (\mathcal{T}_3^*)', (\mathcal{A}_2^*)', \mathcal{A}_3^*, \boldsymbol{x})$ is a standard augmented splitting shadow, as desired. Note that $(\Sigma, \alpha_2'', (\mathcal{T}_2^*)'', \boldsymbol{x})$ is slide-equivalent to the original diagram $(\Sigma, \alpha_2, \mathcal{T}_2^*, \boldsymbol{x})$.

Finally, repeat the process once more, starting with the splitting shadow $(\Sigma, \alpha_3', \alpha_1', (\mathcal{T}_3^*)', (\mathcal{T}_1^*)', \boldsymbol{x})$ and performing slides until we can obtain new collections $\mathfrak{a}_4$ and $\mathcal{A}_4^*$ from the modified collections $\mathfrak{a}_3'$ and $(\mathcal{A}_3^*)'$, as before. At this point, there is a minor wrinkle. We are not finished once we set $\mathfrak{a}_4 = \mathfrak{a}_3'$ and obtain $\mathcal{A}_4^*$ from $(\mathcal{A}_3^*)'$ as before. The reason is that these choices for $\mathfrak{a}_4$ and $\mathcal{A}_4$ might not be compatible with the original tangle shadow $(\Sigma, \alpha_1, \mathcal{T}_1^*, \boldsymbol{x})$, rather these choices are compatible with the slide-equivalent tangle shadow $(\Sigma, \alpha_1'', (\mathcal{T}_1^*)'', \boldsymbol{x})$. To remedy this issue, we perform the slides to change this latter tangle shadow to the former one, and we carry $\mathfrak{a}_4$ and $\mathcal{A}_4^*$ with us along the way, sliding them over arcs and curves when necessary. In abuse of notation, we denote the results of this transformation $\mathfrak{a}_4$ and $\mathcal{A}_4^*$.

In summary, we have produce the collections of arcs $\mathfrak{a}_2, \mathfrak{a}_3, \mathfrak{a}_4, \mathcal{A}_2^*, \mathcal{A}_3^*$, and $\mathcal{A}_4^*$ required to fully augment the original augmented shadow diagram.

To establish uniqueness, suppose $(\mathfrak{a}_2, \mathfrak{a}_3, \mathfrak{a}_4, \mathcal{A}_2^*, \mathcal{A}_3^*, \mathcal{A}_4^*)$ and $(\bar{\mathfrak{a}}_2, \bar{\mathfrak{a}}_3, \bar{\mathfrak{a}}_4, \overline{\mathcal{A}}_2^*, \overline{\mathcal{A}}_3^*, \overline{\mathcal{A}}_4^*)$ are two sets of full-augmentation arcs for the given augmented shadow diagram. By surgering Σ along the corresponding arcs and curves of $\alpha_i \cup \mathcal{T}_i^*$, we can regard the augmentation arcs as lying on P_i. By definition, there is a vertical isotopy taking $\mathfrak{a}_i \cup \mathcal{A}_i^*$ on P_i to $\mathfrak{a}_{i+1} \cup \mathcal{A}_{i+1}^*$ on P_{i+1} through $H_i \cup H_{i+1}$. The same is true for $\bar{\mathfrak{a}}_{i+1} \cup \overline{\mathcal{A}}_{i+1}^*$, so it follows that $\mathfrak{a}_{i+1} \cup \mathcal{A}_{i+1}^*$ and $\bar{\mathfrak{a}}_{i+1} \cup \overline{\mathcal{A}}_{i+1}^*$ can be isotoped to agree on P_2 via a vertical isotopy in H_{i+1}. Working sequentially, it follows that the two collections of full-augmenting arcs are slide-equivalent, as claimed. □

Following Castro, Gay, and Pinzón-Caicedo, we refer to the above algorithm as the *monodromy algorithm*. What follows a is generalization of the discussion of [11, Section 3]; see also [6, Section 4; 8, Section 2].

Given an augmented shadow diagram $\mathfrak{D} = (\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathfrak{a}_1, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, \mathcal{A}_1^*, \boldsymbol{x})$, let $(H, \mathcal{T})$ denote the tangle determined by the tangle shadow $(\Sigma, \alpha_1, \mathcal{T}_1^*, \boldsymbol{x})$. Let $(P, \boldsymbol{y})_{\mathfrak{D}} = \partial_-(H, \mathcal{T})$. We call $(P, \boldsymbol{y})_{\mathfrak{D}}$ the *page* of the shadow diagram. Fix an identification $\mathrm{Id}\colon (P, \boldsymbol{y})_{\mathfrak{D}} \to (\Sigma_{\boldsymbol{p}, \boldsymbol{f}}, \boldsymbol{x}_{\boldsymbol{p}, \boldsymbol{f}})$. We use the standard Morse structure on H to consider $\mathfrak{a}_1$ and $\mathcal{A}_1^*$ as lying in P. Consider the arcs $\mathfrak{a} = \mathrm{Id}(\mathfrak{a}_1)$, which cut the standard surface into a collection of disks, and the arcs $\mathcal{A}^* = \mathrm{Id}(\mathcal{A}_1^*)$, which connect the marked points to the boundary in the standard pair.

Let $\mathfrak{D}^+ = (\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathfrak{a}_1, \mathfrak{a}_2, \mathfrak{a}_3, \mathfrak{a}_4, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, \mathcal{A}_1^*, \mathcal{A}_2^*, \mathcal{A}_3^*, \mathcal{A}_4^*, \boldsymbol{x})$ be a full-augmenting of $\mathfrak{D}$. We consider the arcs $\mathfrak{a}_4$ and $\mathcal{A}_4^*$ as lying in P, as well. Consider the arcs $\mathfrak{a}' = \mathrm{Id}(\mathfrak{a}_4)$ and the arcs $(\mathcal{A}^*)' = \mathrm{Id}(\mathcal{A}_4^*)$. Let $\phi_{\mathfrak{D}}$ be the automorphism of $(\Sigma_{\boldsymbol{p}, \boldsymbol{f}}, \boldsymbol{x}_{\boldsymbol{p}, \boldsymbol{f}})$ satisfying $\phi_{\mathfrak{D}}(\mathfrak{a} \cup \mathcal{A}^*) = \mathfrak{a}' \cup (\mathcal{A}^*)'$, noting that $\phi_{\mathfrak{D}}$ is unique up to isotopy. We call $\phi_{\mathfrak{D}}$ the *monodromy* of the shadow diagram.

Lemma 5.8 *The monodromy $\phi_{\mathfrak{D}}$ is determined up to conjugation by the shadow diagram $\mathfrak{D}$.*

Proof Proposition 5.7 shows that the arcs $\mathfrak{a}_4 \cup \mathcal{A}_4^*$ are uniquely determined (up to slide-equivalence) by the choice of augmentation arcs $\mathfrak{a}_1 \cup \mathcal{A}_1^*$. This means that the arcs $\mathfrak{a}_4 \cup \mathcal{A}_4^*$ are determined uniquely up to isotopy when considered relative to $\mathfrak{a}_1 \cup \mathcal{A}_1^*$ on $(P, \boldsymbol{y})$. Now, the choice of $\mathfrak{a}_1 \cup \mathcal{A}_1^*$ determines a

parametrization of (P, y), and this choice is equivalent to a choice of product structure on $(H, \mathcal{T})$ near (P, y). The important thing is that this product structure is fixed by the choice of $\mathfrak{a}_1 \cup \mathscr{A}_1^*$, and $\mathfrak{a}_4 \cup \mathscr{A}_4^*$ is considered relative to this choice. So, if a different choice of $\mathfrak{a}_1 \cup \mathscr{A}_1^*$ were made, there would be a diffeomorphism of (P, y) between the two choices, and this diffeomorphism would also relate the corresponding choices for $\mathfrak{a}_4 \cup \mathscr{A}_4^*$. Therefore, the monodromy is determined up to conjugation by $\mathfrak{D}$. $\square$

The relevance of $\phi_{\mathfrak{D}}$ is given in the following proposition; we refer the reader to Section 2.8 for relevant notation and terminology regarding open-book decompositions and braidings. The following is a generalization of [6, Theorem 5] and [11, Lemma 3.1]. Note that up to this point, we have neglected the fact that, as oriented manifolds, $\partial(H_i, \mathcal{T}_i) = (\Sigma, x) \cup \overline{(P_i, y_i)}$, while $\partial(Y_i, \beta_i) = (P_i, y_i) \cup \overline{(P_{i+1}, y_{i+1})}$. This fact manifests importantly in the next theorem, where we relate the monodromy of a shadow diagram to the monodromy of the boundary braiding of a trisection; care is taken with orientations here.

Proposition 5.9 *Suppose that $\mathfrak{D}$ is a shadow diagram for a bridge trisection $\mathbb{T}$ of a pair $(X, \mathscr{F})$. Let $\phi_{\mathfrak{D}}$ denote the monodromy of the shadow diagram, and let $(Y_{\phi_{\mathfrak{D}}}, \mathscr{L}_{\phi_{\mathfrak{D}}})$ denote the model open-book braiding corresponding to the abstract open-book braiding $(\Sigma_{p, f}, x_{p, f}, \phi_{\mathfrak{D}})$. Then there is an orientation-preserving diffeomorphism*

$$\psi_{\mathfrak{D}} : \partial(X, \mathscr{F}) \to (Y_{\phi_{\mathfrak{D}}}, \mathscr{L}_{\phi_{\mathfrak{D}}}).$$

Proof Let $(H_1, \mathcal{T}_1) \cup (H_2, \mathcal{T}_2) \cup (H_3, \mathcal{T}_3)$ denote the spine of the bridge trisection determined by the diagram $\mathfrak{D}$; recalling Propositions 2.22 and 5.4. Fix an identification $\psi : (P_1, y_1) \to (\Sigma_{p, f}, x_{p, f})$ and regard this latter pair as a page $(P, y) \times \{0\}$ in the model open-book braiding $(Y_{\phi_{\mathfrak{D}}}, \mathscr{L}_{\phi_{\mathfrak{D}}})$, which we think of as $(P, y) \times_{\phi_{\mathfrak{D}}} S^1$. Note that $(Y_{\phi_{\mathfrak{D}}}, \mathscr{L}_{\phi_{\mathfrak{D}}})$ is well-defined, because $\phi_{\mathfrak{D}}$ is determined by $\mathfrak{D}$ up to conjugation.

Choose an augmenting of $\mathfrak{D}$ by picking arcs $\mathfrak{a}_1$ and $\mathscr{A}_1^*$, which we consider as having been isotoped vertically to lie in (P_1, y_1). Let $\mathfrak{a} \times \{0\}$ and $\mathscr{A}^* \times \{0\}$ denote the arcs on $(P, y) \times \{0\}$ that are the images of $\mathfrak{a}_1$ and $\mathscr{A}_1^*$ under ψ. Apply the monodromy algorithm of Proposition 5.9 to obtain a full-augmenting of $\mathfrak{D}$. Consider the arcs $\mathfrak{a}_1'$, $(\mathscr{A}_1^*)'$, and $(\mathscr{A}_2^*)'$ coming from the standard augmented splitting diagram for

$$(M_1, K_1) = (H_1, \mathcal{T}_1) \cup_{(\Sigma, x)} \overline{(H_2, \mathcal{T}_2)},$$

noting that, regarded as arcs in (P_1, y_1), $\mathfrak{a}_1$ and $\mathfrak{a}_1'$ are isotopic rel-∂, as are $\mathscr{A}_1^*$ and $(\mathscr{A}_1^*)'$. These arcs determine the identity map $\mathrm{Id}_{(M_1, K_1, \Sigma)}$ described in Lemma 2.12. In particular, this gives a unique extension of ψ to a diffeomorphism from the spread (Y_1, β_1) in $\partial(X, \mathscr{F})$ to the spread $(P, y) \times \left[\frac{2}{3}, 1\right]$ in $(Y_{\phi_{\mathfrak{D}}}, \mathscr{L}_{\phi_{\mathfrak{D}}})$. The strange parametrization of the interval is due to the fact that (P_1, y_1) is positively oriented in $\partial(Y_1, \beta_1)$, so we match it to the positively oriented end $(P, y) \times \{1\}$.

Repeating the step described above $(i = 1)$ for $i = 2$ and $i = 3$ — using intervals $\left[\frac{1}{3}, \frac{2}{3}\right]$ and $\left[0, \frac{1}{3}\right]$ — allows us to extend ψ_1 to a map $\psi_{\mathfrak{D}}$ whose domain is the entire boundary

$$\partial(X, \mathscr{F}) = (Y_1, \beta_1) \cup (Y_2, \beta_2) \cup (Y_3, \beta_3)$$

and whose codomain is $(P, y) \times [0, 1]$, equipped with the identification $(x, 1) \sim (\phi'(x), 0)$, where ϕ' must take the arcs $\mathfrak{a}_1 \cup \mathscr{A}_1^*$ to the arcs $\mathfrak{a}_4 \cup \mathscr{A}_4^*$, in order for $\psi_{\mathfrak{D}}$ to be continuous. However, this implies that ϕ' is isotopic rel-∂ to $\phi_{\mathfrak{D}}$, by definition, and we have that $\psi_{\mathfrak{D}}$ respects the original identification space structure on $(Y_{\phi_{\mathfrak{D}}}, \mathscr{L}_{\phi_{\mathfrak{D}}})$, hence is a diffeomorphism, as desired. $\qquad\square$

Example 5.10 (Möbius band for the trefoil) Figure 21(a) shows a shadow diagram corresponding to the bridge trisection of the Möbius band bounded by the right-handed trefoil in S^3 that was discussed in Example 3.16; cf Figure 14. Since this is a $(2; 0, 2)$–bridge trisection, we have that $(P, y) = \partial_-(H_1, \mathscr{T}_1)$ is a disk with two distinguished points in its interior. This pair is shown in Figure 21(d), together with a pair of arcs that connect the points y to ∂P. Using the Morse function on $(H_1, \mathscr{T}_1)$, these arcs can be flowed rel-∂ to lie in Σ, as shown in Figure 21(e). Note that H_1 induces opposite orientations on P_1 and Σ, hence the indicated reflection between (c) and (d) of Figure 21. In Figure 21(f), the shadows for $(H_2, \mathscr{T}_2)$ have been added, making an splitting shadow for (M_1, K_1), which is a geometric 2–braid in $D^2 \times I$, one component of which is twice-perturbed, while the other is not perturbed. In Figure 21(g), a slide of an arc of $\mathscr{A}_1^*$ has been performed to arrange that all arcs are disjoint in their interiors, and the arcs of $\mathscr{A}_2^*$ have been obtained, as described in the proof of Proposition 5.9; this is an augmented splitting shadow for (M_1, K_1). Figure 21(h) shows a splitting shadow for (M_2, K_2), with $\mathscr{A}_2^*$ remembered, and since all arcs are disjoint in their interiors, the arcs of $\mathscr{A}_3^*$ have been derived. Figure 21(i) shows a splitting shadow for (M_3, K_3), with the arcs of $\mathscr{A}_3^*$ remembered, and Figure 21(j) is obtained from this diagram by arc slides of arcs from $\mathscr{T}_3^* \cup \mathscr{A}_3^*$, before $\mathscr{A}_4^*$ is obtained. In Figure 21(k), the arcs of $\mathscr{A}_1^*$ and $\mathscr{A}_4^*$ are shown with the arcs of $\mathscr{T}_1^*$ in Σ. Figure 21(l) shows the result of flowing $\mathscr{A}_1^* \cup \mathscr{A}_4^*$ up to the page (P, y).

Figure 21(l) allows us to see that the braiding induced on the boundary of the bridge trisection is diffeomorphic to the abstract open-book (P, y, σ_1^3), where P is a disk, y is two points, and σ_1 is a positive braid transposition of the two points of y. This derivation is a shadow diagram version of the calculation of this braiding given in Example 3.16 and Figure 14.

Example 5.11 (disk for the trefoil in $(\mathbb{CP}^2)^\circ$) Figure 22(a) shows a shadow diagram corresponding to a bridge trisection of a disk bounded by the right-handed trefoil in $(\mathbb{CP}^2)^\circ$, the result of removing a neighborhood of a point from $\mathbb{CP}^2$. The two circles represent the foot of a handle for the surface Σ and are identified via vertical reflection. If one forgets the bridge points x and all shadow arcs, one obtains a $(1, 0; 0, 1)$–trisection diagram for this four-manifold. The bridge trisection itself is type $(2, (0, 1, 0); 2)$; the union of the blue and green shadows includes a bigon. As in the previous example, we have that $(P, y) = \partial_-(H_1, \mathscr{T}_1)$ is a disk with two distinguished points in its interior. This pair is shown in Figure 22(d), together with a pair of arcs that connect the points y to ∂P. Using the Morse function on $(H_1, \mathscr{T}_2)$, these arcs can be flowed rel-∂ to lie in Σ, as shown in Figure 22(e). In Figure 22(f), the shadows for $(H_2, \mathscr{T}_2)$ have been added, giving a splitting shadow for (M_1, K_1), which is a geometric 2–braid in $D^2 \times I$, one component of which is twice-perturbed with respect to the once-stabilized Heegaard splitting of this spread. In Figure 22(g), a number of arc slides of $\mathscr{T}_1^* \cup \mathscr{A}_1^*$ have been performed to arrange
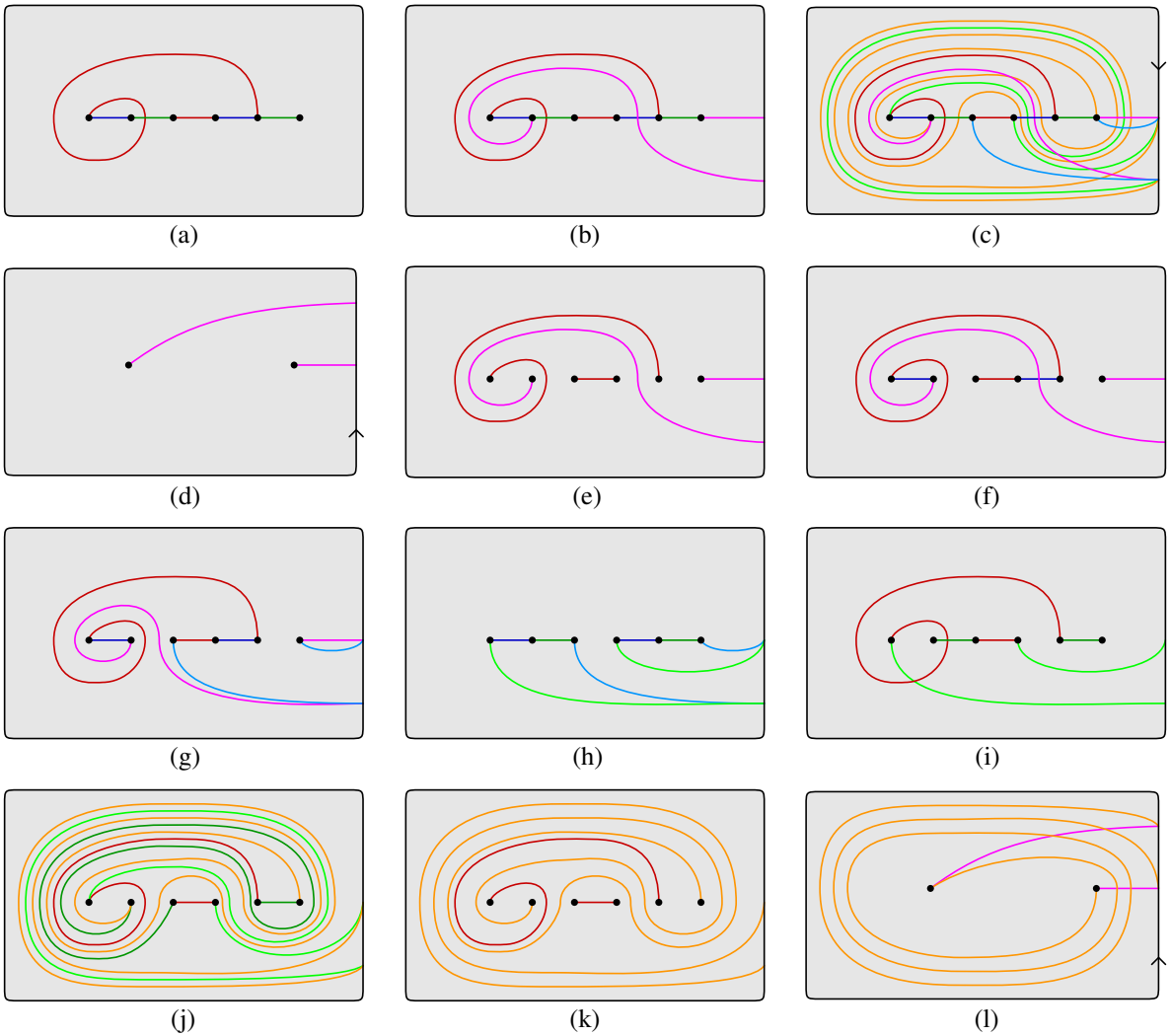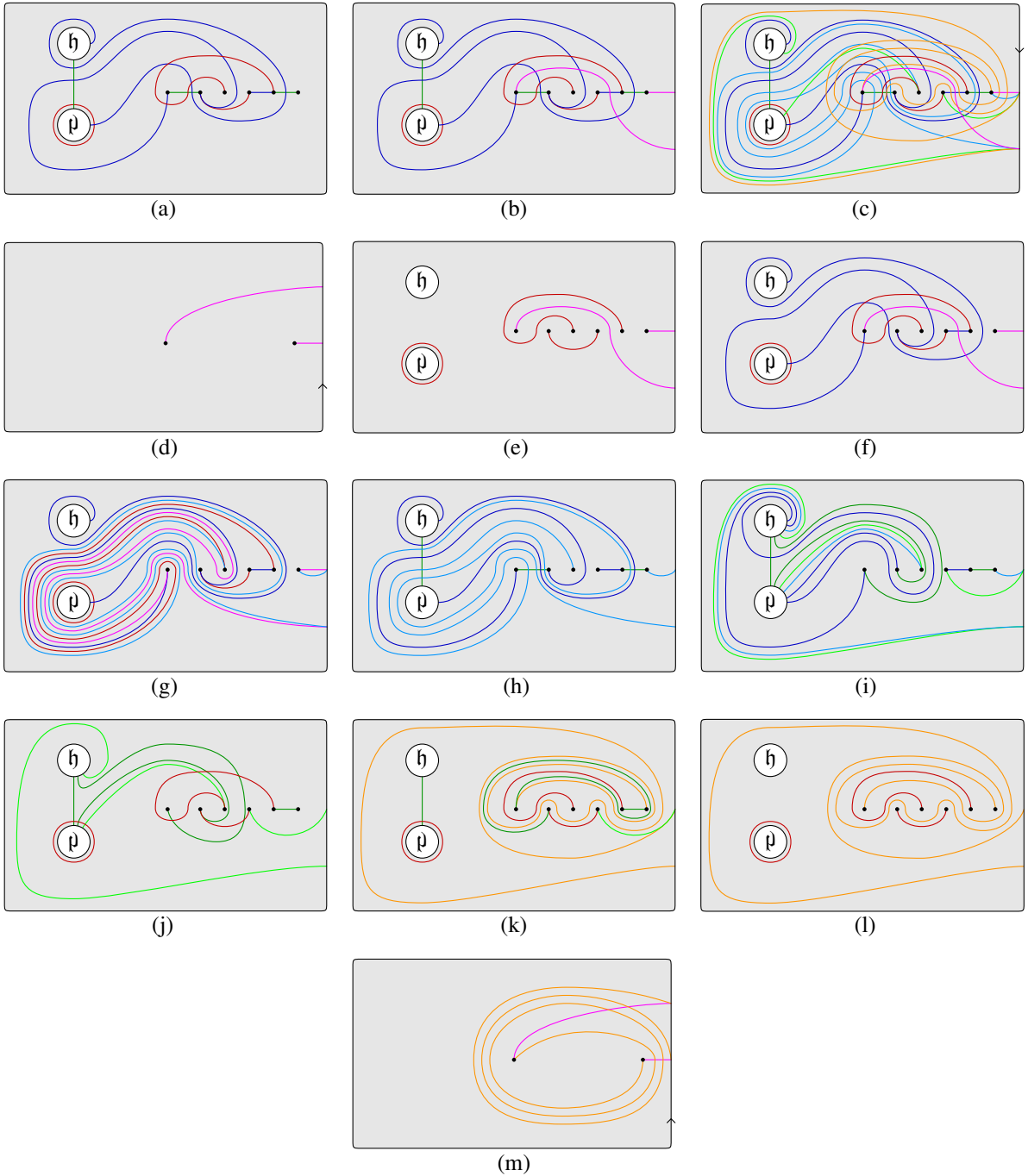
Figure 21: A shadow diagram (a), an augmented shadow diagram (b), and a fully augmented shadow diagram (c) for a bridge trisection for the Möbius band bounded by the right-handed trefoil in S^3. Diagrams (e)–(k) illustrate the process described by the monodromy algorithm of Proposition 5.9, used to find the full-augmenting (c) of the augmented shadow diagram (b). We recover the braiding induced on the boundary of the bridge trisection by studying (l), which shows the arcs $\mathfrak{a}$ and $\mathfrak{a}'$ in the page (P, Y).

that all arcs and curves are disjoint in their interiors, save the standard curve pair $\alpha_1 \cup \alpha_2$. From this standard splitting shadow, the arcs of $\mathscr{A}_2^*$ have been obtained, as described in the proof of Proposition 5.9. Figure 22(h) shows a splitting shadow for (M_2, K_2), with $\mathscr{A}_2^*$ remembered. Figure 22(i) shows the standard augmented splitting shadow resulting from a number or arc slides, together with the arcs of $\mathscr{A}_3^*$. Figure 22(j) shows a splitting shadow for (M_3, K_3), with the arcs of $\mathscr{A}_2^*$ remembered, and Figure 22(k) shows a slide-equivalent standard splitting shadow, with $\mathscr{A}_\circ^*$ derived. In Figure 22(l), the arcs of $\mathscr{A}_1^*$ and

Figure 22: A shadow diagram (a), an augmented shadow diagram (b), and a fully augmented shadow diagram for a bridge trisection for the disk bounded by the right-handed trefoil in $(\mathbb{CP}^2)^\circ$. Diagrams (d)–(m) illustrate the process described by the monodromy algorithm of Proposition 5.9, used to find a full-augmenting of a shadow diagram and recover the braiding induced on the boundary of the bridge trisection.
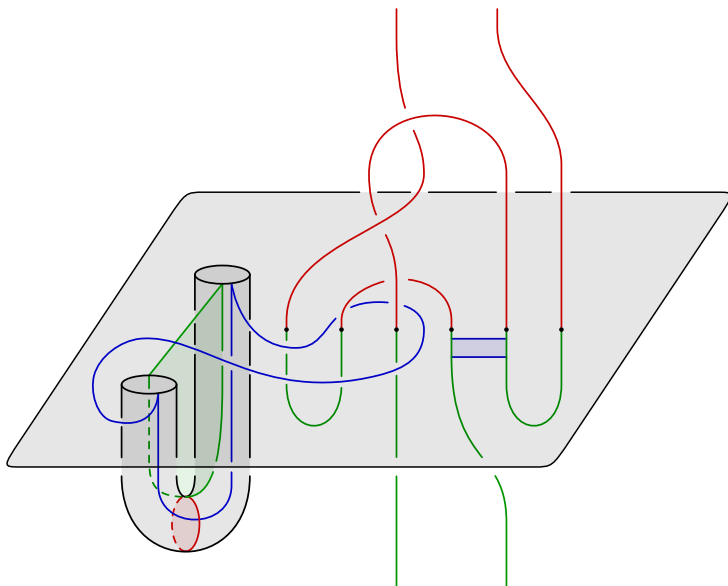
Figure 23: A three-dimensional rendering of the shadow diagram in Figure 22(a) corresponding to the disk bounded by the right-handed trefoil in $(\mathbb{CP}^2)^\circ$.

$\mathscr{A}_\circ^*$ are shown with the arcs and curves of the original tangle shadow for $(H_1, \mathscr{T}_1)$ in Σ. Figure 22(m) shows the result of flowing $\mathscr{A}_1^* \cup \mathscr{A}_\circ^*$ up to the page $(P, \boldsymbol{y})$.

Figure 22(m) allows us to see that the braiding induced on the boundary of the bridge trisection is diffeomorphic to the abstract open-book $(P, \boldsymbol{y}, \sigma_1^3)$, where P is a disk, $\boldsymbol{y}$ is two points, and σ_1 is a right-handed braid transposition of the two points of $\boldsymbol{y}$. This proves that this bridge trisection corresponds to a surface bounded by the right-handed trefoil in $(\mathbb{CP}^2)^\circ$. From the bridge trisection parameters, we conclude that the surface is a disk, since it has Euler characteristic one and is connected.

A three-dimensional rendering for this example is given in Figure 23. The ambient 3–manifold is $S^3 = \partial(\mathbb{CP}^2)^\circ$, equipped with the Heegaard-page structure coming from the compression body $H_{1,0,1}$. The right-handed trefoil is in 2–braid position, and perturbed twice with respect to the genus one Heegaard surface Σ. (Note that Σ is oriented as ∂H_1.) The closed curve shown in blue is the belt-sphere for the 2–handle that is attached to a 0–cell B^4 to build $(\mathbb{CP}^2)^\circ$. The curve lies on Σ with surface-framing -1. This reflects the fact that $(\mathbb{CP}^2)^\circ$ can be thought of as being built from $\bar{S}^3 \times [-1, 0]$ by attaching a $(+1)$–framed 2–handle along the corresponding curve in the mirror manifold $\bar{S}^3 \times \{-1\}$, before capping off with a 0–handle below. A single band is shown for the boundary knot, but this band is a helper-band in the sense of Remarks 3.6 and 3.10 and Section 3.3 more generally. In fact, relative to the Morse function on $(\mathbb{CP}^2)^\circ$, the disk bounded by the trefoil can be (and has been) assumed to have no saddle points, just a single minimum. However, the Morse function on $(\mathbb{CP}^2)^\circ$ coming from the bridge trisection will require the disk to be built from a pair of vertical disks (since we require a 2–braid on the boundary), and the helper-band joins these disks together. Compare with the Morse-theoretic proof of Theorem 8.1.

6 Gluing bridge trisected surfaces and shadow diagrams

In this section, we describe how to glue bridge trisected surfaces along portions of their boundary in a way that respects the bridge trisection structure. The gluing of trisections was first discussed by Castro [5], with further development given by Castro and Ozbagci [8] and by the author and Gay [11]. We conclude this section with some examples of simple gluings of bridge trisected pairs with disconnected boundary, as well as a more complicated example involving the surfaces bounded by the right-handed trefoil discussed above. We refer the reader to Section 5 for necessary concepts related to shadow diagrams.

The development below is a generalization of previous developments to the setting of bridge trisections for four-manifold pairs and is complicated by the fact that we allow the four-manifolds being glued to have multiple boundary components and for the gluings to involve proper submanifolds of these boundaries. To account for this, we will allow our gluing maps to be *partial diffeomorphisms*, which means that they may be defined on proper subsets of their domain. This subset is called the *domain of definition* of the map; the image of the domain of definition is called the *range*, and may be a proper subset of the codomain. The domain of definition and range of our partial diffeomorphisms will always be closed submanifolds of the domain and codomain, respectively.

Let $\mathbb{T}$ be a bridge trisection of a pair $(X, \mathscr{F})$, and let $\mathfrak{D}$ be a shadow diagram for $\mathbb{T}$. Let $(P, y) = \partial_-(H_1, \mathscr{T}_1)$, and let $\phi_{\mathfrak{D}} \colon (P, y) \to (P, y)$ be the monodromy automorphism determined by $\mathfrak{D}$ according to Proposition 5.7. Let $\psi_{\mathfrak{D}} \colon \partial(X, \mathscr{F}) \to (Y_{\phi_{\mathfrak{D}}}, \mathscr{L}_{\phi_{\mathfrak{D}}})$ be the diffeomorphism given by Proposition 5.9, where $(Y_{\phi_{\mathfrak{D}}}, \mathscr{L}_{\phi_{\mathfrak{D}}})$ is the model pair of the abstract open-book $(P, y, \phi_{\mathfrak{D}})$. We note that both $\phi_{\mathfrak{D}}$ and $\psi_{\mathfrak{D}}$ depend on the underlying bridge trisection $\mathbb{T}$, and are determined up to postcomposing with an automorphism of (P, y). Thus, we might as well denote these maps by $\phi_{\mathbb{T}}$ and $\psi_{\mathbb{T}}$; we will adopt either decoration, depending on whether we wish to emphasize the shadow diagram or the underlying bridge trisection.

We work in the generality of bridge trisected pairs with disconnected boundary, so we emphasize the decomposition

$$(Y, \mathscr{L}) = (Y^1, \mathscr{L}^1) \sqcup \cdots \sqcup (Y^n, \mathscr{L}^n)$$

of $(Y, \mathscr{L}) = \partial(X, \mathscr{F})$ into connected components of Y; for any connected component Y^j of Y, we may have $\mathscr{L}^j$ disconnected — ie a link. Thus, we have corresponding decomposition of the pairs (P, y), $(P_{\phi_{\mathbb{T}}}, y_{\phi_{\mathbb{T}}})$, and $(Y_{\phi_{\mathbb{T}}}, \mathscr{L}_{\phi_{\mathbb{T}}})$, and of the maps $\phi_{\mathbb{T}}$ and $\psi_{\mathbb{T}}$.

Our first result is that bridge trisections that induce diffeomorphic braidings on some portion of their boundaries can be glued along those boundaries to obtain a new bridge trisection. By a *diffeomorphism of open-book braidings* we mean a diffeomorphism of three-manifold pairs that restricts to a diffeomorphism of pages (hence, commutes with the monodromies).

Proposition 6.1 *Let $\mathbb{T}'$ and $\mathbb{T}''$ be bridge trisections for pairs $(X', \mathscr{F}')$ and $(X'', \mathscr{F}'')$. Suppose we have an orientation-reversing partial diffeomorphism of open-book braidings $\Psi \colon \partial(X', \mathscr{F}') \to \partial(X'', \mathscr{F}'')$. Then the pair $(X, \mathscr{F}) = (X', \mathscr{F}') \cup_{\Psi} (X'', \mathscr{F}'')$ inherits a canonical bridge trisection $\mathbb{T} = \mathbb{T}' \cup_{\Psi} \mathbb{T}''$.*

Proof Let $(Y', \mathscr{L}')$ and $(Y'', \mathscr{L}'')$ denote the domain of definition and range of Ψ, respectively, noting that these are closed (possibly proper) submanifolds of $\partial(X', \mathscr{F}')$ and $\partial(X'', \mathscr{F}'')$, respectively.

After potentially changing Ψ by an isotopy through diffeomorphisms of open-book braidings, we can assume that $\Psi(P_i', y_i') = (P_i'', y_i'')$ for each $i \in \mathbb{Z}_3$. We will verify that gluing the various corresponding pieces of $\mathbb{T}'$ and $\mathbb{T}''$ together according to Ψ results in a collection of pieces giving a bridge trisection of $(X, \mathscr{F})$.

Consider the restriction of Ψ to the binding B' of the open-book decomposition of $(Y', \mathscr{L}')$, recalling that $B' = \partial(\Sigma', x')$ and $B'' = \Psi(B') = \partial(\Sigma'', x'')$. Let $(\Sigma, x) = (\Sigma', x') \cup_{\Psi} (\Sigma'', x'')$, which is simply the union of two surfaces with marked points and boundary along closed subsets of their respective boundaries, hence a new surface with marked points and (possibly empty) boundary.

Consider the restriction of Ψ to the pages P_i' for each $i \in \mathbb{Z}_3$, recalling that $(P_i', y_i') = \partial(H_i', \mathcal{T}_i')$ and $(P_i'', y_i'') = \Psi(P_i', y_i') = \partial(H_i'', \mathcal{T}_i'')$. Let $(H_i, \mathcal{T}_i) = (H_i', \mathcal{T}_i') \cup_{\Psi_{(P_i', y_i')}} (H_i'', \mathcal{T}_i'')$, noting that

$$\partial(H_i, \mathcal{T}_i) = (\Sigma, x) \cup_B \left((\partial_-(H_i', \mathcal{T}_i') \setminus (P_i', y')) \sqcup (\partial_-(H_i'', \mathcal{T}_i'') \setminus (P_i'', y'')) \right).$$

(A word of caution regarding notation: The fact that we are considering gluings along potentially strict subsets of the boundaries complicates the exposition notationally. For example, earlier in the paper, we would have written $(P_i', y_i') = \partial_-(H_i', \mathcal{T}_i')$, but here we regard $(P_i', y_i') \subset \partial_-(H_i', \mathcal{T}_i')$ as the portion of $\partial_-(H_i', \mathcal{T}_i')$ lying in the domain of definition.)

For each $i \in \mathbb{Z}_3$, let $\mathfrak{a}_i'$ be a neatly embedded collection of arcs in $P_i' \setminus y_i'$ such that surgery along the arcs reduces P_i' to a collection of disks with the number of connected components as P_i'. Moreover, we require that $\mathfrak{a}_i'$ and $\mathfrak{a}_{i+1}'$ be isotopic rel-∂ in $Y' \setminus \mathscr{L}'$ via an isotopy that is monotonic with respect to the open-book structure. Let $\mathfrak{a}_i'' = \Psi(\mathfrak{a}_i')$. For each $i \in \mathbb{Z}_3$, let $\mathscr{A}_i$ be an embedded collection of arcs connecting the points of y_i' to $\partial P_i'$, and assume, as before, that $\mathscr{A}_i'$ and $\mathscr{A}_{i+1}'$ are isotopic via an isotopy that fixes $\mathscr{A}_i' \cap \partial P_i'$ and is monotonic with respect to the open-book-braiding structure; the free endpoints of $\mathscr{A}_i'$ will move along $\mathscr{L}'$. Let $\mathscr{A}_i'' = \Psi(\mathscr{A}_i')$.

Using the Morse structure on $(H_i', \mathcal{T}_i')$, flow the arcs of $\mathfrak{a}_i'$ and $\mathscr{A}_i'$ down to Σ', and denote the results $(\mathfrak{a}_i^*)'$ and $(\mathscr{A}_i^*)'$, respectively. Let E_i' and T_i' denote the traces of the respective isotopies, noting that the E_i' are compression disks for the H_i', and that the T_i' are bridge triangles for the vertical strands $y_i' \times [0, 1]$. Do the same for $\mathfrak{a}_i''$ and $\mathscr{A}_i''$ to obtain $(\mathfrak{a}_i^*)''$ and $(\mathscr{A}_i^*)''$ on Σ'', with corresponding traces E_i'' and T_i''.

Let D_i' and D_i'' be collections of neatly embedded disks in H_i' and H_i'', respectively, such that surgery along D_i' and D_i'' reduces H_i' and H_i'', respectively, to spreads $\partial_-(H_i', \mathcal{T}_i') \times [0, 1]$ and $\partial_-(H_i'', \mathcal{T}_i'') \times [0, 1]$.

For each connected component of (P_i', y'), pick a disk of D_i' adjacent to that component in the sense that one of the two scars resulting from surgery along the chosen disk lies in the corresponding component of $(P_i', y') \times [0, 1]$. (Equivalently, the chosen disk is the cocore of a 1–handle connecting the component of $(P_i', y') \times [0, 1]$ to another component of the spread obtained by surgery.) Let $F_i' \subset D_i'$ denote the chosen disks. Then, we claim that

$$D_i = (D_i' \setminus F_i') \sqcup (E_i' \cup_\Psi E_i'') \sqcup D_i''$$

is a collection of compression disks in H_i such that surgery along D_i reduces H_i to

$$(\partial_-(H_i') \setminus P_i') \sqcup (\partial_-(H_i'') \setminus P_i'').$$

To see that this is the case, note that the result of surgering H_i along $D_i \sqcup F_i'$ is precisely

$$\left((\partial_-(H_i') \setminus P_i') \times [0, 1]\right) \sqcup \left(\bigsqcup_{m'} D^2 \times [0, 1]\right) \sqcup \left((\partial_-(H_i'') \setminus P_i'') \times [0, 1]\right),$$

where m' is the number of connected components of Y_i', P_i', and F_i'. The effect of removing the disks of F_i' from this collection of compression disk is to attach 1–handles, one for each $D^2 \times [0, 1]$ in the above decomposition, connecting the m' copies of $D^2 \times [0, 1]$ to the rest of the spread. It follows that H_i is a compression body with $\partial_+ H_i = \Sigma$ and $\partial_-(H_i) = (\partial_-(H_i') \setminus P_i') \sqcup (\partial_-(H_i'') \setminus P_i'')$, as desired.

Moreover, let Δ_i' and Δ_i'' be bridge disks for the flat strands of $\mathcal{T}_i'$ and $\mathcal{T}_i''$, respectively. Then,

$$\Delta_i = \Delta_i' \sqcup (T_i' \cup_\Psi T_i'') \sqcup \Delta_i''$$

is a collection of bridge semidisks and triangles for the strands of $\mathcal{T}_i' \cup_\Psi \mathcal{T}_i''$ in H_i. The key thing to note here is that the bridge triangles T_i' for the vertical strands $y_i' \times [0, 1]$ glue to the corresponding bridge triangles T_i'' for the vertical strands of $y_i'' \times [0, 1]$ along the identified arcs $\mathcal{A}_i' \cup_\Psi \mathcal{A}_i''$ to give bridge disks for the new flat strands $(y_i' \times [0, 1]) \cup_\Psi (y_i'' \times [0, 1])$.

Finally, consider the restriction of Ψ to the spreads (Y_i', β_i') cobounded by (P_i', y_i') and (P_{i+1}', y_{i+1}') in $(Y', \mathcal{L})$, recalling that $(Y_i', \beta_i') = (Z_i', \mathcal{D}_i') \cap \partial(X', \mathcal{F}')$, and noting that $\Psi(Y_i', \beta_i') = (Y_i'', \beta_i'')$. Let $(Z_i, \mathcal{D}_i) = (Z_i', \mathcal{D}_i') \cup_\Psi (Z_i'', \mathcal{D}_i'')$ for each $i \in \mathbb{Z}_3$. We claim that the fact that the $(Z_i, \mathcal{D}_i)$ are trivial disk-tangles follows easily from the detailed argument just given that the $(H_i, \mathcal{T}_i)$ are trivial tangles. The reason is that a trivial disk-tangle $(Z, \mathcal{D})$ can be naturally viewed as the lensed product $(H, \mathcal{T}) \times [0, 1]$ such that the decomposition of $\partial(H, \mathcal{T}) = (S, x) \cup_{\partial S} (P, y)$ gives rise to a bridge-braid structure on $\partial(Z, \mathcal{D})$. Precisely, the lensed product $(H_{g,p,f}, \mathcal{T}_{b,v}) \times [0, 1]$ is $(Z_{g,k;p,f}, \mathcal{D}_{c;v})$, where $k = g + p + f - n$ and n is the length of the partition p. The structure on the boundary is that of a symmetric Heegaard double. Moreover, we have that $\partial_-(Z, \mathcal{D}) = \partial_-(H, \mathcal{T}) \times [0, 1]$, so gluing two trivial disk-tangles along a portion of their negative boundaries is the same as gluing the corresponding trivial tangles (of which the trivial disk-tangles are lensed products) along the corresponding portions of their negative boundaries, then taking the product with the interval. Succinctly, the gluings along portions of the negative boundaries commute with the taking of the products with the interval. Therefore, the $(Z_i, \mathcal{D}_i)$ are trivial disk-tangles, as desired.

It remains only to verify that $(Z_i, \mathcal{D}_i) \cap (Z_{i-1}, \mathcal{D}_{i-1}) = (H_i, \mathcal{T}_i)$ and $(H_i, \mathcal{T}_i) \cap (H_{i+1}, \mathcal{T}_{i+1}) = \Sigma$, but this is immediate. $\qquad\qquad\square$

Remark 6.2 Proposition 6.1 holds in the case that $\mathbb{T}' = \mathbb{T}''$ and Ψ is a (partial) *self*-gluing! See Example 6.6 below.

Having established how to glue bridge trisections from the vantage point of bridge trisected pairs, we now turn our attention to understanding gluings diagrammatically. Suppose that $\mathbb{T}'$ and $\mathbb{T}''$ are bridge trisections of pairs $(X', \mathcal{F}')$ and $(X'', \mathcal{F}'')$ with augmented shadow diagrams $\mathfrak{D}'$ and $\mathfrak{D}''$, respectively. Let $f: \partial(\Sigma, \mathfrak{a}_1', (\mathscr{A}_1^*)') \to \partial(\Sigma', \mathfrak{a}_1', (\mathscr{A}_1^*)')$ be an orientation-reversing *partial* diffeomorphism. We call $\mathfrak{D}'$ and $\mathfrak{D}''$ *gluing compatible* if there is an orientation-reversing *partial* diffeomorphism

$$\psi_f(\mathfrak{D}', \mathfrak{D}''): (P_1', y_1') \to (P_1'', y_1'')$$

that extends f and commutes with the monodromies of the diagrams — ie $\psi_f(\mathfrak{D}', \mathfrak{D}'') \circ \phi_{\mathfrak{D}'} = \phi_{\mathfrak{D}''}$ — where this composition is defined. In this case, we call f a *compatible* (*partial*) *gluing*.

The map $\psi_f(\mathfrak{D}', \mathfrak{D}'')$ determines an orientation-reversing (partial) diffeomorphism

$$\Upsilon_f(\mathfrak{D}', \mathfrak{D}''): (Y_{\phi_{\mathfrak{D}'}}, \mathscr{L}_{\phi_{\mathfrak{D}'}}) \to (Y_{\phi_{\mathfrak{D}''}}, \mathscr{L}_{\phi_{\mathfrak{D}''}})$$

of abstract open-book braidings. So, we can define a (partial) gluing map

$$\Psi_f(\mathfrak{D}', \mathfrak{D}''): \partial(X', \mathcal{F}') \to \partial(X'', \mathcal{F}'')$$

of the bridge trisected pairs by

$$\Psi_f(\mathfrak{D}', \mathfrak{D}'') = \psi_{\mathfrak{D}''}^{-1} \circ \Upsilon_f(\mathfrak{D}', \mathfrak{D}'') \circ \psi_{\mathfrak{D}'}.$$

Again, we are interested in partial boundary-gluings, so we reiterate that the above caveats regarding the domain and codomain apply to $\Psi_f(\mathfrak{D}', \mathfrak{D}'')$. Given this set-up, we can now describe how gluing shadow diagrams corresponds to gluing bridge trisected four-manifold pairs.

Proposition 6.3 *Suppose that $\mathbb{T}'$ and $\mathbb{T}''$ are bridge trisections of four-manifold pairs $(X', \mathcal{F}')$ and $(X'', \mathcal{F}'')$, respectively, and that the corresponding fully augmented shadow diagrams $\mathfrak{D}'$ and $\mathfrak{D}''$ admit a compatible gluing f. Let $\mathfrak{D} = \mathfrak{D}' \cup_f \mathfrak{D}''$, and let $(X, \mathcal{F}) = (X', \mathcal{F}') \cup_{\Psi_f(\mathfrak{D}', \mathfrak{D}'')} (X'', \mathcal{F}'')$. Then $\mathfrak{D}$ is a fully augmented shadow diagram for the bridge trisection on $(X, \mathcal{F})$ given in Proposition 6.1, once it is modified in the following ways:*

(1) *The arcs of $(\mathfrak{a}_4)' \sqcup (\mathscr{A}_4^*)'$ and $(\mathfrak{a}_4)'' \sqcup (\mathscr{A}_4^*)''$ whose endpoints lie in the domain of definition and range of f should be deleted.*

(2) *If $\partial X''$ is disconnected, then, for each component Y'' of the range of $\Psi_f(\mathfrak{D}', \mathfrak{D}'')$ there is a subcollection of curves of α_i'', for each $i \in \mathbb{Z}_3$, that separate the components of $\partial \Sigma''$ corresponding to Y'' from the other components of $\partial \Sigma''$. Throw out one curve from the subcollection of curves corresponding to each connected component of the range of $\Psi_f(\mathfrak{D}', \mathfrak{D}'')$.*

(3) *If $\partial X''$ is connected but $\partial X'$ is disconnected, then, for each component Y' of the domain of definition of $\Psi_f(\mathfrak{D}', \mathfrak{D}'')$ there is a subcollection of curves of α_i', for each $i \in \mathbb{Z}_3$, that separate the components of $\partial \Sigma'$ corresponding to Y' from the other components of $\partial \Sigma'$. Throw out one curve from the subcollection of curves corresponding to each connected component of the domain of definition of $\Psi_f(\mathfrak{D}', \mathfrak{D}'')$.*

Proof The first modifications required above is a minor issue. If this is not done, then the would-be-deleted arcs give rise to extra shadows and curves that are redundant in the encoding of the trivial tangle $(H_1, \mathcal{T}_1)$. The next two modifications are more serious, and are required to ensure that the resulting diagram is a shadow diagram. The rationale was made clear in the proof of Proposition 6.1, where this precise discarding was carried out at the level of compression disks. Note that only one of the final two modification will need to be made in practice.

The rest of the proof follows from the proof of Proposition 6.1, as applied to the gluing $\Psi_f(\mathfrak{D}', \mathfrak{D}'')$. $\square$

We conclude this section with some examples illustrating gluings of bridge trisected four-manifold pairs.

Example 6.4 First, we recall the bridge trisected surfaces bounded by the right-handed trefoil discussed in Examples 5.10 and 5.11. Let $\mathfrak{D}'$ denote the fully augmented shadow diagram in Figure 24, top left, which corresponds to a bridge trisection of the pair $(X', \mathscr{F}')$, where $\mathscr{F}'$ is a disk bounded by the right-handed trefoil in $X' = (\mathbb{CP}^2)^\circ$. Let $\mathfrak{D}''$ denote the fully augmented shadow diagram in Figure 24, top right, which corresponds to the pair $(X'', \mathscr{F}'')$, where $\mathscr{F}''$ is the Möbius band bounded by the left-handed trefoil in S^3, which we imagine as being perturbed so that its interior lies in $X'' = B^4$. Note that $\mathfrak{D}''$ is the mirror of the diagram shown in Figure 21(c). Orientations for the boundaries of the diagrams are shown.

These bridge trisections induce open-book braidings on the boundaries of their corresponding manifold pairs that are orientation-reversing diffeomorphic. Both open-book braidings have disk page and boundary link in 2–braid position: For $\mathfrak{D}'$, the monodromy is three *positive* half-twists about the two braid points. This was described in Example 5.11 and Figure 22. However, for $\mathfrak{D}''$, the half-twists are *negative*, since $\mathfrak{D}''$ is the mirror of the diagram discussed in Example 5.10 and Figure 21.

Let $f : \partial \mathfrak{D}' \to \partial \mathfrak{D}''$ be the orientation-reversing diffeomorphism that matches the endpoints of the arcs $(\mathscr{A}_1^*)'$ with those of $(\mathscr{A}_1^*)''$. There is an orientation-reversing diffeomorphism

$$\psi_f(\mathfrak{D}', \mathfrak{D}'') : (P_1', y_1') \to (P_2'', y_2'')$$

that extends f; simply pick the obvious diffeomorphism relating the pair in Figure 22(d) to the mirror of the pair in Figure 21(d). It follows that is a compatible gluing corresponding to an orientation-reversing diffeomorphism $\Psi_f(\mathfrak{D}', \mathfrak{D}'')$.

Let $(X, \mathscr{F}) = (X', \mathscr{F}') \cup_{\Psi_f(\mathfrak{D}', \mathfrak{D}'')} (X'', \mathscr{F}'')$. By Proposition 6.3, the $\mathfrak{D} = \mathfrak{D}' \cup_f \mathfrak{D}''$ shown in Figure 24, bottom left, is a shadow diagram for $(X, \mathscr{F})$. Observe how the arcs $(\mathscr{A}_4^*)'$ and $(\mathscr{A}_4^*)''$ have been discarded according with the first modification required by Proposition 6.3. (The second and third modification are
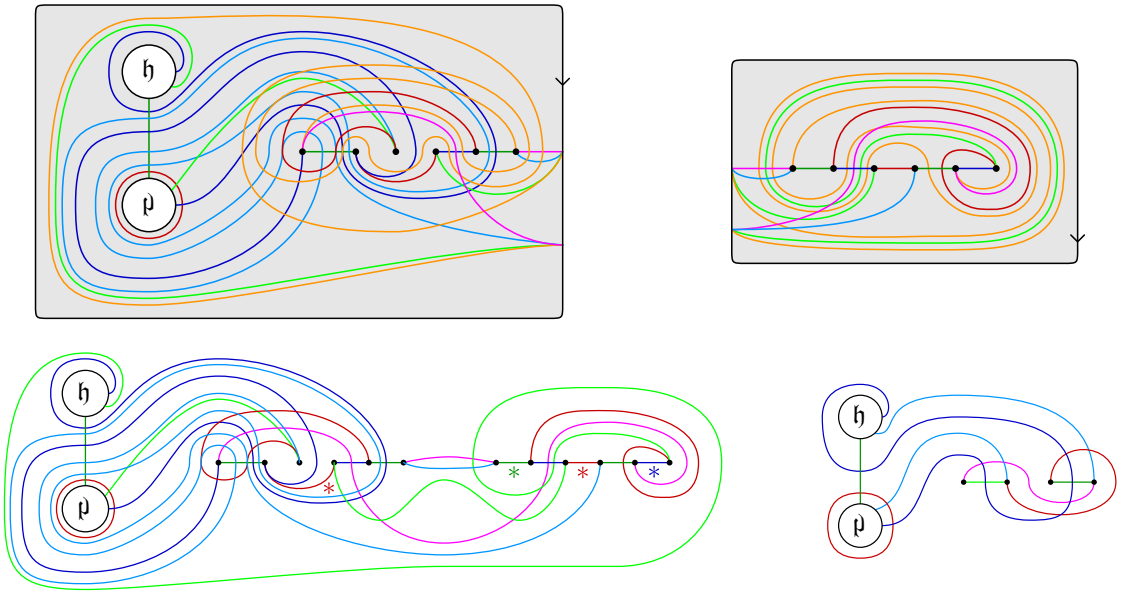
Figure 24: Top left: a shadow diagram for the disk bounded by the right-handed trefoil in $(\mathbb{CP}^2)^\circ$. Top right: a shadow diagram for the Möbius band bounded by the right-handed trefoil in B^4. Bottom left: the result of gluing these diagrams via the unique compatible gluing — a shadow diagram for a projective plane in $\mathbb{CP}^2$. The bottom right is obtained from the bottom left by deperturbing along the indicated shadows; see Section 9.2 for relevant definitions.

not necessary in this example, since $\partial X'$ and $\partial X''$ are connected.) A brief examination reveals that this diagram can be deperturbed three times, using the indicated shadows. (See Section 9 for details about perturbation.) Doing so produces the diagram of Figure 24, bottom right.

We have that $X \cong \mathbb{CP}^2$ and $\mathscr{F} \cong \mathbb{RP}^2$, but it is not true that $(X, \mathscr{F}) \cong (\mathbb{CP}^2, \mathbb{RP}^2)$, where the latter pair is the projectivization of the standard pair $(\mathbb{C}^3, \mathbb{R}^3)$. The standard projective pair $(\mathbb{CP}^2, \mathbb{RP}^2)$ is depicted in [28, Figure 2]. One way to distinguish these two pairs is to note that $\mathscr{F}$ has normal Euler number $+6$, while $\mathbb{RP}^2$ has normal Euler number $+2$. Moreover, $\pi_1(X \setminus \nu(\mathscr{F})) \cong \mathbb{Z}/2\mathbb{Z}$, while $\pi_1(\mathbb{CP}^2 \setminus \nu(\mathbb{RP}^2)) \cong 1$. These facts are left as exercises to the reader.

Example 6.5 Consider the shadow diagram $\mathfrak{D}'$ shown in Figure 25, top left, which corresponds to a bridge trisection of the cylinder pair $(X', \mathscr{F}') = (S^3 \times I, S^1 \times I)$. The underlying trisection of $S^3 \times I$ can be thought of as follows. If one "trisects" S^3 into three three-balls, which meet pairwise along disk pages of the open-book decomposition with unknotted boundary — so the triple intersection of the three-balls is this binding — then the trisection of $S^3 \times I$ can be thought of as the product of this "trisection" of S^3 with the interval, and the core Σ is simply the product of the binding with the interval. So, the diagram $\mathfrak{D}'$ can be thought of as a bridge trisection for a copy $\mathscr{F}$ of Σ. To carry this out, the copy $\mathscr{F}$ of the annular core must be perturbed relative the original copy Σ of the core. We leave it as an exercise to the reader to verify that $\mathfrak{D}'$ describes the cylinder pair, as claimed.
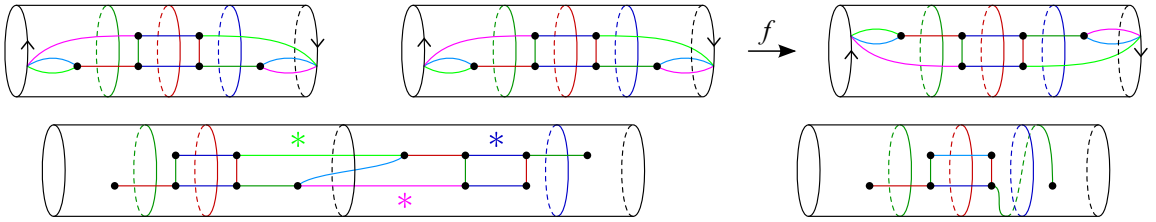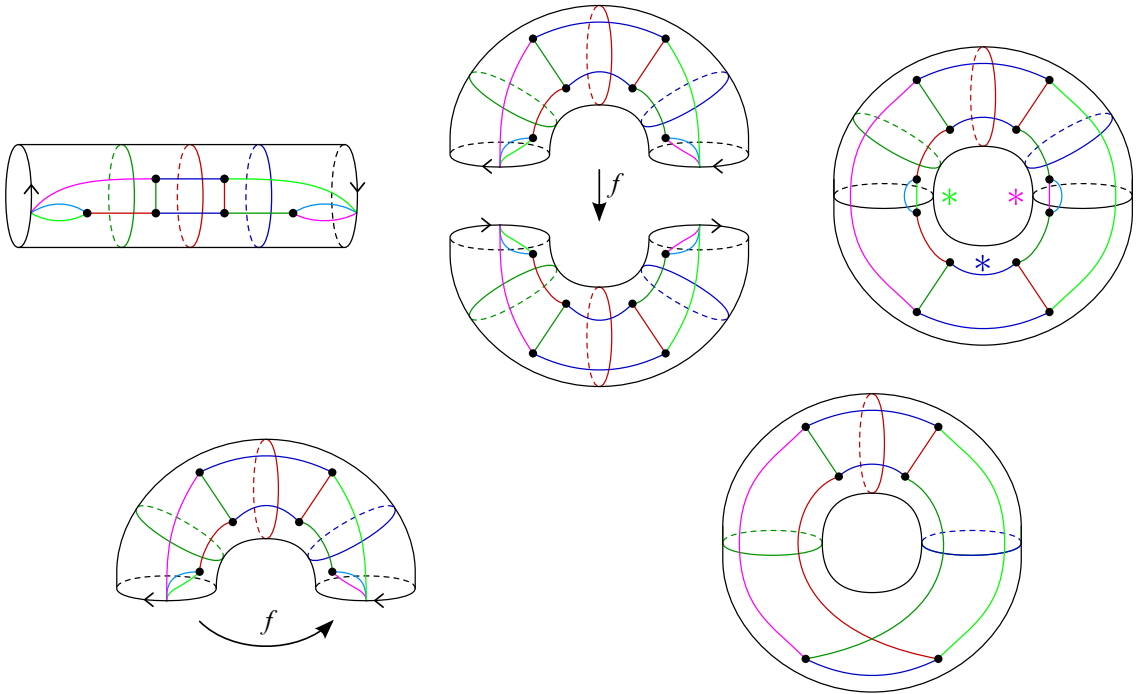
Figure 25: Top left: a shadow diagram for $S^3 \times I$. Top right: A copy of this diagram and a copy of its mirror, with compatible gluing f indicated. Bottom left: the result of the gluing, $S^3 \times I$. The bottom right is obtained from the bottom left by deperturbing along the indicated shadows.

Now, let $\mathfrak{D}''$ denote a mirror copy of $\mathfrak{D}'$ that corresponds to a second copy of cylinder pair,

$$(X'', \mathscr{F}'') = (S^3 \times I, S^1 \times I).$$

Each of the two boundary components of both $(X', \mathscr{F}')$ and $(X'', \mathscr{F}'')$ have induced open-book braidings with page a disk with one braid point. Let $f : \partial\mathfrak{D}' \to \partial\mathfrak{D}''$ be the orientation-reversing partial diffeomorphism shown in Figure 25, top right — ie f maps the boundary component $S^1 \times \{1\}$ of $\mathfrak{D}'$ to the boundary component $S^1 \times \{0\}$ of $\mathfrak{D}''$. Trivially, f extends to an orientation-reversing partial diffeomorphism $\psi_f(\mathfrak{D}', \mathfrak{D}'') : (P_1', y_1') \to (P_2'', y_2'')$ between the page pairs corresponding to the boundary components of the chosen boundary components of $\mathfrak{D}'$ and $\mathfrak{D}''$. Thus, we have an orientation-reversing partial diffeomorphism $\Psi_f(\mathfrak{D}', \mathfrak{D}'') : \partial(X', \mathscr{F}') \to \partial(X'', \mathscr{F}'')$.

Let $(X, \mathscr{F}) = (X', \mathscr{F}') \cup_{\Psi_f(\mathfrak{D}', \mathfrak{D}'')} (X'', \mathscr{F}'')$. By Proposition 6.3, the diagram $\mathfrak{D} = \mathfrak{D}' \cup_f \mathfrak{D}''$ shown in Figure 25, bottom left, is a shadow diagram for $(X, \mathscr{F})$. Note that one curve of each color has been discarded in accordance with modification (2). As before, the diagram obtained from gluing can be deperturbed. (This is a common phenomenon when gluing shadow diagrams.) The diagram obtained after deperturbing (and performing slides), shown in Figure 25, bottom right, is diffeomorphic to the original diagram $\mathfrak{D}'$. Of course, $(X, \mathscr{F}) \cong (S^3 \times I, S^1 \times I)$.

In this example, modification (1) of Proposition 6.3 is implicit; the arcs $\mathfrak{a}_4'$, $(\mathscr{A}_4^*)'$, $\mathfrak{a}_4''$, and $(\mathscr{A}_4^*)''$ were never drawn and were never needed. More interestingly, we see how modification (2) is required. The curves of $\mathfrak{D}''$ have been discarded upon gluing. Had this not been done, there would have been parallel curves in α_i for each $i \in \mathbb{Z}_3$. This would imply that $P_i = \partial_- H_i$ would have a two-sphere component, which is not allowed.

Example 6.6 Finally, we consider two more compatible gluings involving $\mathfrak{D}'$. First, let $\mathfrak{D}''$ denote a mirror copy of $\mathfrak{D}'$, and let $f : \partial\mathfrak{D}' \to \partial\mathfrak{D}''$ be the compatible gluing shown in Figure 26, top middle. This compatible gluing is similar to the one explored in Example 6.5, but this time f is not a partial diffeomorphism. The induced gluing $\Psi_f(\mathfrak{D}', \mathfrak{D}'')$ matches the two boundary components of $(X, \mathscr{F})$ with the corresponding components of $(X'', \mathscr{F}'')$. As a result, $(X, \mathscr{F}) = (X', \mathscr{F}') \cup_{\Psi_f(\mathfrak{D}', \mathfrak{D}'')} (X'', \mathscr{F}'')$ is the closed four-manifold pair $(S^3 \times S^1, S^1 \times S^1)$, and the diagram $\mathfrak{D} = \mathfrak{D}' \cup_f \mathfrak{D}''$ for this pair is shown in

Figure 26: Top left: a shadow diagram for $S^3 \times I$. Top middle: a copy of this diagram and a copy of its mirror, with compatible gluing f indicated. Top right: the result of the gluing, $S^3 \times S^1$. Bottom left: a compatible self-gluing of the diagram. Bottom right: the result of the self gluing, $S^3 \times S^1$. The bottom right is obtained from the top right by deperturbing along the indicated shadows.

Figure 26, top right. As in Example 6.5, the redundant arcs have been suppressed, and the curves α_i'' have been discarded upon gluing. Also, we can again deperturb, arriving at the diagram of Figure 26, bottom right.

Now, let f denote the compatible self-gluing shown in Figure 26, bottom left. The induced self-map of $(S^3 \times I, S^1 \times I)$ is $\Psi_f(\mathfrak{D}') \colon (S^3 \times \{0\}, S^1 \times \{0\}) \to (S^3 \times \{1\}, S^1 \times \{1\})$. The diagram resulting from the compatible self-gluing f is the diagram of Figure 26, bottom right, which describes $(S^3 \times S^1, S^1 \times S^1)$, as noted before.

7 Classification and examples

In this section, we classify $(b, \boldsymbol{c}; v)$–bridge trisections in the trivial cases where one or more of the parameters is sufficiently small. Then, we present families of examples representing more interesting choices of parameters and pose questions about further possible classification results. To get started, we discuss the connected sum and boundary connected sum operations, then we introduce some notions of reducibility for bridge trisections.

7.1 Connected sum of bridge trisections

Given trisections $\mathbb{T}'$ and $\mathbb{T}''$ for four-manifolds X' and X'', it is straightforward to see that there is a trisection $\mathbb{T} = \mathbb{T}' \# \mathbb{T}''$ describing $X' \# X''$. Let $\varepsilon \in \{', ''\}$. All that needs to be done is to choose the points $x^\varepsilon \in X^\varepsilon$ that determine the connected sum to lie on the respective cores. Having done so, the pieces of the trisection $\mathbb{T}$ can be described by $\Sigma = \Sigma' \# \Sigma''$, $H_i = H_i' \natural H_i''$, and $Z_i = Z_i' \natural Z_i''$. Note that $\mathbb{T}$ is independent of the choice of points made above.

Remark 7.1 The connected sum operation, as described, is a very simple example of a gluing of trisections, as described in detail in Section 6. Each of $\mathbb{T}^\varepsilon \setminus \nu(x^\varepsilon)$ is automatically a trisection with one new boundary component diffeomorphic to S^3. If $\mathfrak{D}^\varepsilon$ is a shadow diagram for $\mathbb{T}^\varepsilon$, then $\mathfrak{D}^\varepsilon \setminus \nu(x^\varepsilon)$ is a diagram for $\mathbb{T}^\varepsilon \setminus \nu(x^\varepsilon)$ after a simple modification is made in the case that $\partial X \neq \varnothing$: in this case, a curve δ must be added to each of the α_i that is parallel to the curve $\partial \nu(x^\varepsilon)$ where Σ' and Σ'' were glued together (this is a separating reducing curve in the sense of Definition 7.6, below).

There is a complication in extending this interpretation to connected sum of bridge trisections with boundary that was not present in discussions of the connected sum of *closed* bridge trisections elsewhere in the literature. The naïve idea is to simply choose the connected sum points x^ε to be bridge points. This works for closed bridge trisections, because every bridge point is incident to a flat strand in each of the three trivial tangles. This is not the case for bridge trisections with boundary. To convince oneself of the problem, try to form the connect sum of two bridge trisections, each of which is a copy of the bridge trisection described in Figure 31, top left, which corresponds to the standard positive Möbius band. It is simply not possible: the removal of an open neighborhood around any bridge point has the effect that one of the trivial tangles will no longer be trivial, since it will have a strand with no endpoints on Σ.

One might think that perturbing the bridge trisection (see Section 9.2) would fix the problem by creating a bridge point that is incident to flat strands in each arm; however, the problem persists due to consideration of the vertical patches. Since vertical patches are only allowed to be incident to one component of ∂X, we cannot puncture our bridge trisection at a bridge point that is incident to a vertical patch.

The next lemma makes precise when puncturing a bridge trisection at a bridge point produces a new bridge trisection and indicates how to form the connected sum of bridge trisections.

Lemma 7.2 *Let $\mathbb{T}$ be a bridge trisection for a pair $(X, \mathscr{F})$, and let x be a bridge point. Then $\mathbb{T} \setminus \nu(x)$ is a bridge trisection for the pair $(X \setminus \nu(x), \mathscr{F} \setminus \nu(x))$ if and only if x is incident to a flat patch of $\mathscr{D}_i$ for each $i \in \mathbb{Z}_3$.*

If $\mathfrak{D} = (\Sigma, \alpha_1, \alpha_2, \alpha_3, \mathcal{T}_1^*, \mathcal{T}_2^*, \mathcal{T}_3^*, x)$ is a shadow diagram for $\mathbb{T}$, then a shadow diagram for $\mathbb{T} \setminus \nu(x)$ can be obtained as follows: Let $\delta = \partial \nu(x)$ in $\mathfrak{D}$. For each arc τ_i^* of $\mathcal{T}_i^*$ that is incident to x, choose a

Figure 27: Left: a shadow diagram for a bridge trisection of (B^4, D^2). Right: the diagram obtained by puncturing at the bridge point x.

neighborhood $\nu(\tau_i^*) \supset \nu(x)$ and let $\delta_i = \partial\nu(\tau_i^*)$. Let $\Sigma' = \Sigma \setminus \nu(x)$, $\alpha_i' = \alpha_i \cup \delta_i$, $(\mathcal{T}_i^*)' = \mathcal{T}_i^* \setminus \tau_i^*$, and $\mathbf{x}' = \mathbf{x} \setminus \{x\}$. Then there are two cases: If $\partial X = \varnothing$, then

$$\mathfrak{D} = (\Sigma', \alpha_1, \alpha_2, \alpha_3, (\mathcal{T}_1^*)', (\mathcal{T}_2^*)', (\mathcal{T}_3^*)', \mathbf{x}')$$

is a shadow diagram for $T \setminus \nu(x)$. If $\partial X \neq \varnothing$, then

$$\mathfrak{D} = (\Sigma', \alpha_1', \alpha_2', \alpha_3', (\mathcal{T}_1^*)', (\mathcal{T}_2^*)', (\mathcal{T}_3^*)', \mathbf{x}')$$

is a shadow diagram for $T \setminus \nu(x)$.

Proof If x is incident to a flat patch of $\mathcal{D}_i$ for each $i \in \mathbb{Z}_3$, then it is straightforward to verify that the pieces of $\mathbb{T} \setminus \nu(x)$ form a bridge trisection. The main substantive changes are that

(1) the number of components of ∂X, $\partial\Sigma$, and $\partial_- H_i$ all increase by one; and

(2) for each $i \in \mathbb{Z}_3$, the flat strand of $\mathcal{T}_i$ becomes a vertical strand and the flat patch of $\mathcal{D}_i$ incident to x becomes a vertical patch.

Conversely, if x is incident to a vertical patch $D \subset \mathcal{D}_i$ for some $i \in \mathbb{Z}_3$, then $\mathcal{D}_i \setminus \nu(x)$ is no longer a trivial disk-tangle, since $D \setminus \nu(x)$ is neither vertical nor flat, as it intersects multiple components of ∂X.

If $\partial X = \varnothing$, then the H_i are handlebodies and the H_i' are compression bodies with $\partial_- H_i' \cong D^2$. In this case, the curves α_i still encode H_i' without modification. If $\partial X \neq \varnothing$, then $\partial_- H_i' \cong \partial_- H_i \sqcup D^2$. In this case, δ must be added to α_i in order to encode the fact that the new component of $\partial_- H_i'$ is disjoint from the original ones. As curves in a defining set, δ and δ_i serve the same role, since they are isotopic. The only reason for pushing δ off τ_i^* is to satisfy our convention that the shadow arcs be disjoint from the defining set of curves for the handlebody. The shadow arcs τ_i^* are deleted regardless of whether ∂X is empty, since these shadows correspond to flat strands that become vertical strands upon removal of $\nu(x)$. $\square$

Example 7.3 Consider the shadow diagram $\mathfrak{D}$ shown in Figure 27, left, which corresponds to a bridge trisection of the trivial disk in the four-ball. Figure 27, right, shows the diagram corresponding to the bridge trisection $\mathbb{T}' = \mathbb{T} \setminus \nu(x)$ for $(X', \mathcal{F}') = (B^4 \setminus \nu(x), D^2 \setminus \nu(x))$. Note that this diagram is equivalent to that of Figures 25, top left, and 26, top left.

In light of this lemma, it is clear that we can obtain a bridge trisection for the connected sum of surfaces by choosing the connected sum points to be bridge points incident only to flat patches. Though such

bridge points need not always exist (see the Möbius band example reference above), they can be created via interior perturbation — at most one in each direction. The punctured trisections $\mathbb{T}^\varepsilon \setminus \nu(x^\varepsilon)$ can be canonically glued along the novel boundary components (which are three-sphere-unknot pairs), according to the techniques of Section 6. Note that in the case that at least one of X' and X'' have boundary, then at least one of the curves δ_i' or the curves δ_i'' should be discarded upon gluing, as dictated by Propositions 6.1 and 6.3. Compare Example 7.3 to Example 6.5.

So far, we have viewed the connected sum of bridge trisections as a special case of gluing bridge trisections, and it has been noted that, for this approach to work, we must form the connected sum at bridge points that are incident to flat patches in each disk-tangle. However, it is possible to work in a slightly more general way so that the punctured objects need not be bridge trisections themselves, but their union will be a bridge trisection of the connected sum.

Lemma 7.4 *Let* $\mathbb{T}'$ *and* $\mathbb{T}''$ *be bridge trisections for pairs* $(X', \mathscr{F}')$ *and* $(X'', \mathscr{F}'')$, *respectively, and let* x' *and* x'' *be bridge points such that, for each* $i \in \mathbb{Z}_3$, *one of* x' *or* x'' *is incident to a flat patch in* $\mathbb{T}^\varepsilon$. *Then the result*

$$\mathbb{T} = (\mathbb{T}' \setminus \nu(x')) \cup (\mathbb{T}'' \setminus \nu(x''))$$

obtained by removing open neighborhoods of the x^ε *from the* $\mathbb{T}^\varepsilon$ *and gluing along resulting boundaries so that the corresponding trisection pieces are matched is a bridge trisection for* $(X, \mathscr{F}) = (X', \mathscr{F}') \# (X'', \mathscr{F}'')$.

Proof Let D_i^ε be the patch of $\mathscr{D}_i^\varepsilon$ containing x^ε for each $i \in \mathbb{Z}_3$ and each $\varepsilon \in \{', ''\}$. Let $D_i = D_i' \cup_{\partial \nu(x^\varepsilon)} D_i''$. Then

$$\mathscr{D}_i = \mathscr{D}_i' \cup_{\partial \nu(y^\varepsilon)} \mathscr{D}_i'' = (\mathscr{D}_i' \setminus D_i') \sqcup (\mathscr{D}_i'' \setminus D_i'') \sqcup D_i.$$

For each $i \in \mathbb{Z}_3$, one of the D_i^ε will be flat, so D_i will be flat or vertical, according to whether the other of the D_i^ε is flat or vertical. In any event, each disk of $\mathscr{D}_i$ has at most one critical point, and we have a trivial disk-tangle, since the boundary sum of trivial disk-tangles is a trivial disk-tangle.

A similar argument shows that the arms of $\mathbb{T}$ are just the boundary sum of the arms of the $\mathbb{T}^\varepsilon$ and that each strand is vertical or flat, as desired. The details are straightforward to check.   □

Note that while the parameters g and $\boldsymbol{k}$ are additive under connected sum, the parameters b and $\boldsymbol{c}$ are (-1)–subadditive (eg $b = b' + b'' - 1$). In the case that the $(X^\varepsilon, \mathscr{F}^\varepsilon)$ have nonempty boundary, the boundary parameters $\boldsymbol{p}$, $\boldsymbol{f}$, $\boldsymbol{v}$, and n are all additive, since we are discussing connected sum at an *interior point* of the pairs. Unlike the case of the connected sum of two four-manifold trisections, here, the resulting bridge trisection is highly dependent on the choice of bridge points made above.

7.2 Boundary connected sum of bridge trisections

Now consider the operation of boundary connected sum of four-manifolds. We start with the set-up as above, but now we choose the summation points to be points y^ε lying in components K^ε of the

bindings $\partial\Sigma^\varepsilon$ for each $\varepsilon \in \{',''\}$. In this case, the pieces of the trisection $\mathbb{T} = \mathbb{T}' \natural \mathbb{T}''$ can be described as $\Sigma = \Sigma' \natural \Sigma''$, $H_i = H_i' \natural H_i''$, $Z_i = Z_i' \natural Z_i''$, $B = B' \# B''$, $P_i = P_i' \natural P_i''$, and $Y_i = Y_i' \natural Y_i''$. And in this case, g, $\boldsymbol{k}$, and $\boldsymbol{p}$ are additive, while $\boldsymbol{f}$ and n are (-1)–subadditive, and $\mathbb{T}$ is highly dependent on the choice of binding component K^ε made above.

The situation becomes more complicated when we consider the boundary connected sum of bridge trisected pairs. The issue here is that $\mathscr{F}^\varepsilon \cap \partial\Sigma^\varepsilon = \varnothing$, so we cannot choose the y^ε to lie simultaneously on Σ^ε and on $\mathscr{F}^\varepsilon$. Our approach is to first perform the boundary connected sum of the ambient four-manifolds, as just described, then consider the induced bridge trisection of the split union $(X, \mathscr{F}' \sqcup \mathscr{F}'')$ of surface links. We now describe a modification of this bridge trisection that will produce a bridge trisection of $(X, \mathscr{F}' \natural \mathscr{F}'')$.

Suppose that we would like to form the boundary connected sum of $(X', \mathscr{F}')$ with $(X'', \mathscr{F}'')$ at points $y^\varepsilon \in \partial\mathscr{F}^\varepsilon$. Without loss of generality, we can assume that $y^\varepsilon \in \mathscr{F}^\varepsilon \cap P_i^\varepsilon$; in relation to the open-book structure on (the chosen component of) ∂X^ε, we assume that y^ε lies on the page P_i^ε. Henceforth, our model is dependent on the choice of $i \in \mathbb{Z}_3$.

Choose arcs ω^ε connecting the points y^ε to the chosen binding components $K^\varepsilon \subset B^\varepsilon$. Let z^ε denote the points of $\omega^\varepsilon \cap K^\varepsilon$. Form the boundary connected sum of the ambient four-manifolds at the points z^ε, as described above, so that $\mathscr{F}' \sqcup \mathscr{F}''$ is in bridge position with respect to $\mathbb{T}$. Note that the arcs ω^ε give rise to an arc ω in the page of P_i connecting the points y^ε.

Use the height function on H_i to flow ω down to the core Σ. Let Q represent the square traced out by this isotopy, and let $\omega_* = Q \cap \Sigma$. Let N be a regular neighborhood of Q in X. We will change $\mathscr{F}' \sqcup \mathscr{F}''$ to $\mathscr{F}' \natural \mathscr{F}''$ in a way that will produce a bridge trisection for the latter from the bridge trisection of the former, and this change will be supported inside N. See Figure 28, top left, for a (faithful) schematic of this set-up. The figures depict the case of $i = 1$.

Proposition 7.5 *A bridge trisection for $(X, \mathscr{F}) = (X' \natural X'', \mathscr{F}' \natural \mathscr{F}'')$ can be obtained from the bridge trisection of $(X, \mathscr{F}' \sqcup \mathscr{F}'')$ described above by replacing the local neighborhood N of Q shown in Figure 28, top left, with the local neighborhood N' shown in Figure 28, top right. The replacement can be seen in a shadow diagram as the local replacement of the portion of the diagram supported near ω_* shown in Figure 28, bottom left, with the portion shown in Figure 28, bottom right.*

Proof Near ω_*, the neighborhood N is precisely the $(0; 0, 2)$–bridge trisection of two copies of the trivial disk in B^4. To recover all of N, we extend upward along Q. Because ω was lowered to Σ along a pair of vertical strands of $(H_i, \mathscr{T}_i' \sqcup \mathscr{T}_i'')$, we see that the entirety of N is still just the 2–bridge trisection of two copies of the trivial disk. In other words, N is isolating, in a bridge-trisected way, a small disk from each of the $\mathscr{F}^\varepsilon$.

Now, to perform the (ambient) boundary connected sum of the $\mathscr{F}^\varepsilon$ at the points y^ε, we must attach a half-twisted band $\mathfrak{b}$ connecting these points. (It should be half-twisted because $\partial\mathscr{F}'$ and $\partial\mathscr{F}''$ are braided

The neighborhood N.

The neighborhood N'.

Local shadow diagram before...

...and after boundary connected sum.
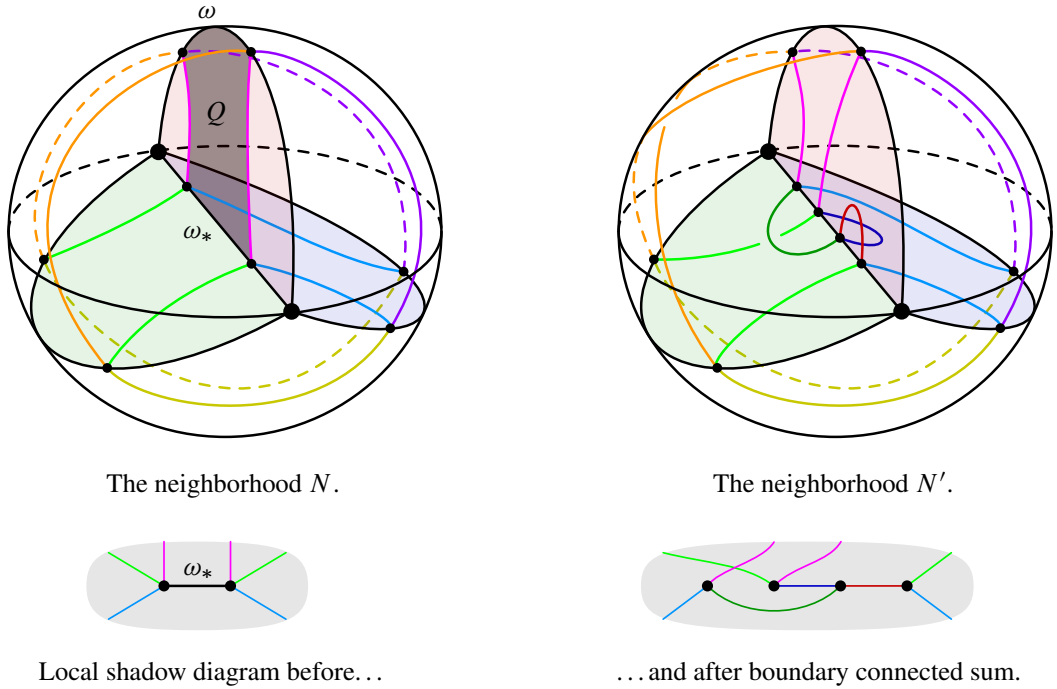
Figure 28: The trisected local neighborhood in the top left is exchanged for the trisected local neighborhood in the top right to carry out an ambient boundary connected sum of surface-links. The local change is depicted with shadow diagrams in the change from the bottom left to the bottom right. Note that, globally, the pink shadow arcs necessarily correspond to vertical strands of $\mathcal{T}_1$, while the light blue and light green shadow arcs may correspond (globally) to either flat or vertical strands.

about B; the half-twist will ensure that the result $\partial\mathcal{F}' \# \partial\mathcal{F}''$ is still braided about B.) We also assume that the core of $\mathfrak{b}$ lies in P_i. The change affected by attaching the half-twisted band is localized to the neighborhood N. Therefore, it suffices to understand how N is changed.

Although we are describing an ambient boundary connected sum of surfaces in a four-manifold X that may be highly nontrivial, the neighborhood N is a four-ball, so it makes sense to import the bridge-braided band presentation technology from Section 3. Figure 29, left, shows a bridge-braided ribbon presentation for N, together with the half-twisted band $\mathfrak{b}$. Figure 29, middle, shows the effect of attaching the band, together with the dual band; this is a ribbon presentation for the boundary connected sum of the two disks in N. Figure 29, right, shows a bridge-braided ribbon presentation for this object, which we denote by N'. Note that the boundaries of N and N' are both 2–braids and are identical, except where they differ by a half-twist. As stated before, we assume this difference is supported near P_i. (Note that in the schematic of Figure 28, top right, the half-twist is shown in the spread Y_{i-1}, rather than in P^i, due the reduction in dimension. Similarly, in the frames of Figure 29, left and middle, the band $\mathfrak{b}$ and the crossing are similarly illustrated away from P_i.)
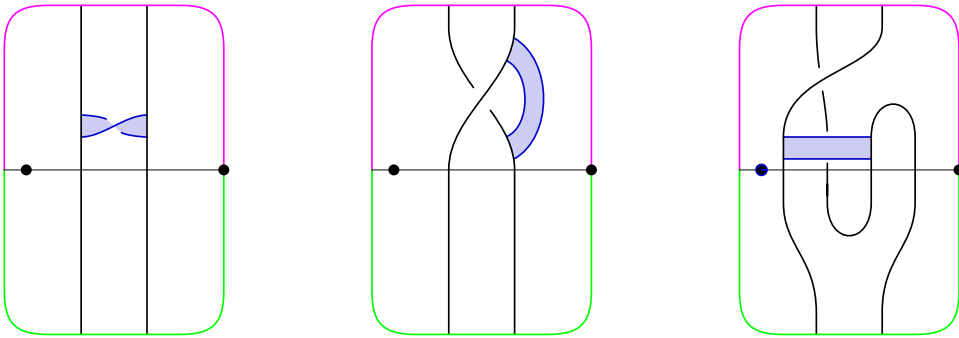
Figure 29: Left: a ribbon presentation for N, together with the band $\mathfrak{b}$ realizing the boundary connected sum. Middle: a ribbon presentation for N', the result of the boundary connected sum. Right: a bridge-braided ribbon presentation for N'.

The neighborhood N' is the $(1; 0, 2)$–bridge trisection of the spanning disk for the unknot that induces the braiding of the unknot as a $(2, 1)$–curve in the complement of the (unknotted) binding. The corresponding bridge-braided ribbon presentation has one band, which is a helper band in the sense of Remarks 3.6 and 3.10. This helper band is the dual band to $\mathfrak{b}$.

Because ∂N and $\partial N'$ are identical away from a neighborhood of ω, we can cut N out and glue in N' to realize the attaching of $\mathfrak{b}$; ie to realize the ambient boundary connected sum. □

7.3 Notions of reducibility

We now discuss three notions of reducibility for trisections of pairs that we will show correspond with the connected sum and boundary connected sum operations discussed above. These properties are distinct from, but related to, the properties of being stabilized or perturbed, which are discussed in Section 9.

Definition 7.6 Let $\mathbb{T}$ be a bridge trisection for a pair $(X, \mathcal{F})$. Let $\delta \subset \Sigma \setminus \nu(\boldsymbol{x})$ be an essential simple closed curve.

(1) The curve δ is called a *reducing curve* if, for each $i \in \mathbb{Z}_3$, there exists a disk $E_i \subset H_i \setminus \nu(\mathcal{T}_i)$ with $\partial E_i = \delta$.

(2) The curve δ is called a *decomposing curve* if, for each $i \in \mathbb{Z}_3$, there exists a disk $E_i \subset H_i$ with $\partial E_i = \delta$ and with $|E_i \cap \mathcal{T}_i| = 1$. A decomposing curve is called *trivial* if it bounds a disk in Σ containing a single bridge point.

(3) An embedded three-sphere $S \subset X$ is a *trisected reducing sphere* if $Z_i \cap S$ is a three-ball and $H_i \cap S$ is a disk for each $i \in \mathbb{Z}_3$, and $\Sigma \cap S$ is a reducing curve.

(4) An embedded three-sphere-unknot pair $(S, K) \subset (X, \mathcal{F})$ is a *(nontrivial) trisected decomposing sphere pair* if
$$(Z_i \cap S, \mathcal{D}_i \cap S) \cong (B^3, I)$$

is a trivial 1–strand tangle in a three-ball for each $i \in \mathbb{Z}_3$, and $\Sigma \cap S$ is a (nontrivial) decomposing curve.

(5) A trisection is *reducible* (resp. *decomposable*) if it admits a reducing curve (resp. a nontrivial decomposing curve).

Let $\eta \subset \Sigma \setminus \nu(\boldsymbol{x})$ be an essential, neatly embedded arc.

(6) The arc η is called a *reducing arc* if, for each $i \in \mathbb{Z}_3$, there exists a neatly embedded arc $\eta_i \subset P_i$ and a disk $E_i \subset H_i \setminus \nu(\mathcal{T}_i)$ with $\partial E_i = \eta \cup \eta_i$.

(7) A neatly embedded three-ball $B \subset X \setminus \mathscr{F}$ is a *trisected boundary-reducing ball* if, for all $i \in \mathbb{Z}_3$, $Z_i \cap B$ is a three-ball and $H_i \cap B$ is a disk, and $\Sigma \cap B$ is a reducing arc.

(8) A trisection is *boundary-reducible* if it admits a reducing arc.

Lemma 7.7 *If a trisection $\mathbb{T}$ is reducible, decomposable, or boundary-reducible, then $\mathbb{T}$ admits, respectively, a trisected reducing sphere, a nontrivial trisected decomposing sphere pair, or a trisected boundary-reducing ball.*

Proof What follows is closely based on the proof of Proposition 3.5 from [26], where reducing curves are assumed (implicitly) to be separating, and some clarification is lacking. Here, we give added detail and address the latter two conditions, which are novel.

Suppose $\mathbb{T}$ is either reducible or decomposable, with reducing or decomposing curve δ bounding disks E_i in the H_i. Let $R_i = E_i \cup_\delta \overline{E}_{i+1}$ be the given two-sphere in $H_i \cup_\Sigma \overline{H}_{i+1} \subset \partial Z_i$. Recall (Section 2.7) that Z_i is built by attaching 4–dimensional 1–handles the lensed product $Y_i \times [0, 1]$ along $Y_i \times \{1\}$. A priori, the R_i may not be disjoint from the belt spheres of the 1–handles in Z_i; however, by [22], it can be arranged via handleslides and isotopies of the 1–handles that R_i is disjoint from the belt spheres. Thus, we can assume that either (1) R_i is parallel to a belt sphere, or (2) R_i is contained in $Y_i \times \{1\}$. These cases correspond to whether δ is nonseparating or separating, respectively. In case (1), R_i bounds the cocore of the 1–handle, which is a three-ball in Z_i. In case (2), since Y_i is irreducible, R_i bounds a three-ball in Y_i whose interior can be perturbed into Z_i. In either case, we get a three-ball B_i in Z_i whose boundary is $E_i \cup_\delta \overline{E}_{i+1}$, and the union $S_\delta = B_1 \cup B_2 \cup B_3$ gives a trisected three-sphere.

In the case that δ is reducing, we are done: S_δ is a trisected reducing sphere. In the case that δ is a decomposing curve, it remains to show that $S_\delta \cap \mathscr{F}$ is unknotted and $B_i \cap \mathscr{F}$ is a trivial arc; the former is implied by the latter, which we now show. Note that B_i and $\mathscr{D}_i$ are both neatly embedded in Z_i and that $\mathscr{D}_i$ is boundary parallel. Using the boundary parallelism of $\mathscr{D}_i$, we can arrange that a component D of $\mathscr{D}_i$ intersects B_i if and only if D intersects $R_i = \partial B_i$. It follows that there is a unique component $D \subset \mathscr{D}_i$ that intersects B_i. If we isotope D to a disk $D_* \subset \partial Z_i$, then we find that $D_* \cap R_i$ consists of an arc and some number of simple close curves. By an innermost curve argument, we may surgery D_* to obtain a new disk D'_* such that $D'_* \cap R_i$ consists solely of an embedded arc. Since D'_* and D_* have the same boundary, they are isotopic rel-∂ in Z_i by Proposition 2.13. Reversing this ambient isotopy,

we can arrange that $B_i \cap \mathscr{D} = B_i \cap D$ consists of a single arc. Moreover, this arc is trivial, since it is isotopic to the arc $R_i \cap D_*$ in ∂Z_i, and R_i is a decomposing sphere for either the unknot ∂D or the unknotted, vertical strand $\mathscr{D} \cap H_i \cup_\Sigma \overline{H}_{i+1}$. Either way, R_i cuts off an unknotted arc. Thus, (S_δ, K) can be constructed to be a decomposing sphere for the trisection, as desired, where K is the three-fold union of the trivial arcs $B_i \cap \mathscr{F}$.

Now suppose that $\mathbb{T}$ is boundary-reducible, with reducing arc η and arcs η_i such that $\eta \cup \eta_i$ bounds a disk $E_i \subset H_i$. Consider the neatly embedded 2–disk $R_i = E_i \cup_\eta \overline{E}_{i+1}$ in $H_i \cup_\Sigma \overline{H}_{i+1} \subset \partial Z_i$. Let B_i be the trace of a small isotopy that perturbs the interior of R_i into Z_i. Then the union $B_\eta = B_1 \cup B_2 \cup B_3$ is a trisected three-ball. If η is a reducing arc, we are done. $\qquad\square$

Remark 7.8 (regarding nonseparating curves) Reducing curves are almost always separating in the following sense. Suppose that δ is a nonseparating reducing curve. Then there is a curve $\eta \subset \Sigma$ that is dual to δ. Let $\delta' = \partial\nu(\delta \cup \eta)$. Then δ' is a separating reducing curve, unless it is inessential (ie parallel to a boundary component of Σ or null-homotopic in Σ). This only occurs if Σ is the core of the genus one trisection for $S^1 \times S^3$ or for its puncture, $(S^1 \times S^3)^\circ$. In any event, the neighborhood $\nu(S_\delta \cup \eta)$, where S_δ is the reducing sphere corresponding to δ as in Lemma 7.7, is diffeomorphic to $(S^1 \times S^3)^\circ$.

If δ is a nonseparating decomposing curve with corresponding decomposing pair (S_δ, K_δ), then K_δ can be separating or nonseparating as a curve in $\mathscr{F}$. If K_δ is nonseparating, then we can surger $(X, \mathscr{F})$ along the pair (S, K) to obtain a new pair $(X', \mathscr{F}')$. That the surgery of $\mathscr{F}$ along K can be performed ambiently uses the fact that K is an unknot in S, hence bounds a disk in $X \setminus \mathscr{F}$. Working backwards, there is an $S^0 \subset \mathscr{F}' \subset X$ along which we can surger $(X', \mathscr{F}')$ to obtain $(X, \mathscr{F})$. It follows that $X = X' \# (S^1 \times S^3)$ and $\mathscr{F}$ is obtained from $\mathscr{F}'$ by tubing. Diagrammatically, the surgery from $(X, \mathscr{F})$ to $(X', \mathscr{F}')$ is realized by surgering Σ along δ. Note that this tubing is not necessarily trivial in the sense that it may or may not be true that $(X, \mathscr{F}) = (X', \mathscr{F}') \# (S^1 \times S^3, S^1 \times S^1)$.

A bridge trisection satisfying one of the three notions of reducibility decomposes in a natural way. See Section 7.1 for a detailed discussion of connected sum and boundary connected sum operations. For example, presently, we let $\mathbb{T}' \# \mathbb{T}''$ denote the connected sum of trisections, regardless of whether the connected summing point is a bridge point or not.

Proposition 7.9 *Let $\mathbb{T}$ be a bridge trisection for a pair $(X, \mathscr{F})$.*

(1) *If $\mathbb{T}$ admits a separating reducing curve, then there exist pairs $(X', \mathscr{F}')$ and $(X'', \mathscr{F}'')$ with trisections $\mathbb{T}'$ and $\mathbb{T}''$ such that $\mathbb{T} = \mathbb{T}' \# \mathbb{T}''$ and*

$$(X, \mathscr{F}) = (X' \# X'', \mathscr{F}' \sqcup \mathscr{F}'').$$

(2) *If $\mathbb{T}$ admits a nontrivial, separating decomposing curve, then there exist pairs $(X', \mathscr{F}')$ and $(X'', \mathscr{F}'')$ with trisections $\mathbb{T}'$ and $\mathbb{T}''$ such that $\mathbb{T} = \mathbb{T}' \# \mathbb{T}''$ and*

$$(X, \mathscr{F}) = (X' \# X'', \mathscr{F}' \# \mathscr{F}'').$$

(3) *If $\mathbb{T}$ admits a separating reducing arc, then there exist pairs $(X', \mathscr{F}')$ and $(X'', \mathscr{F}'')$ with trisections $\mathbb{T}'$ and $\mathbb{T}''$ such that $\mathbb{T} = \mathbb{T}' \natural \mathbb{T}''$ and*

$$(X, \mathscr{F}) = (X' \natural X'', \mathscr{F}' \sqcup \mathscr{F}'').$$

Proof If $\mathbb{T}$ admits a separating reducing curve δ, then it admits a separating trisected reducing sphere S_δ, by Lemma 7.7. Cutting open along S_δ and capping off the two resulting three-sphere boundary components with genus zero trisections of B^4 results in two new trisections $\mathbb{T}'$ and $\mathbb{T}''$ for pairs $(X', \mathscr{F}')$ and $(X'', \mathscr{F}'')$, as desired in part (1). For part (2), we proceed as above, except we cap off with two genus zero 0–bridge trisections of (B^4, D^2) to achieve the desired result. (If any of the disks E_i bounded by δ in the H_i intersect vertical strands τ_i, then we can perturb to make these intersecting strands flat. If such perturbations are performed before cutting, they can be undone with deperturbation after gluing. This is related to the discussion immediately preceding Lemma 7.4.)

If $\mathbb{T}$ admits a separating reducing arc η, then it admits a separating trisected reducing ball B_η, by Lemma 7.7. Cutting open along B_η results in two new trisections $\mathbb{T}'$ and $\mathbb{T}''$ for pairs $(X', \mathscr{F}')$ and $(X'', \mathscr{F}'')$, as desired in part (3). □

Remark 7.10 (boundary-decomposing arcs) Conspicuously absent from the above notions of reducibility is a characterization of what might be referred to as boundary-decomposability — in other words, a characterization of when we have

$$(X, \mathscr{F}) = (X' \natural X'', \mathscr{F}' \natural \mathscr{F}'').$$

The obvious candidate for such a notion would be the existence of a neatly embedded, essential arc $\eta \subset \Sigma$, similar to the one involved in the notion of boundary-reducibility, but where the disks E_i each intersect the respective $\mathscr{T}_i$ in precisely one point. However, a lengthy examination of such arcs reveals that they rarely correspond to surfaces that are boundary connected sums in the desired way. To the point, many of the examples given later in this section admit such arc, but are not boundary-connected sums of bridge trisected surfaces. We have been unable to find a satisfying characterization of when this occurs.

7.4 Classification for small parameters

As a first example, consider the $(4, (2, 4, 2); 3)$–bridge trisection shown in Figure 30, which is the boundary sum of a 1–bridge trisection, a 3–bridge trisection that is perturbed, and three 0–bridge trisections and corresponds to $(B^4, S^2 \sqcup S^2 \sqcup D^2 \sqcup D^2 \sqcup D^2)$. (The perturbation is a finger perturbation; see Definition 9.9.) It turns out that such a bridge trisection is obtained whenever $c_i = b$ for some $i \in \mathbb{Z}_3$. (Recall that ∂D_i contains a flat b–bridge c_i component unlink, so $b \geq c_i$ for all $i \in \mathbb{Z}_3$.)

Proposition 7.11 *Let $\mathbb{T}$ be a $(b, c; v)$–bridge trisection for a surface $(B^4, \mathscr{F})$. If $b = c_i$ for some $i \in \mathbb{Z}_3$, then $c_{i+1} = c_{i+2} = c$ and*

$$(B^4, \mathscr{F}) = \left(B^4, \left(\bigsqcup_c S^2 \right) \sqcup \left(\bigsqcup_v D^2 \right) \right),$$
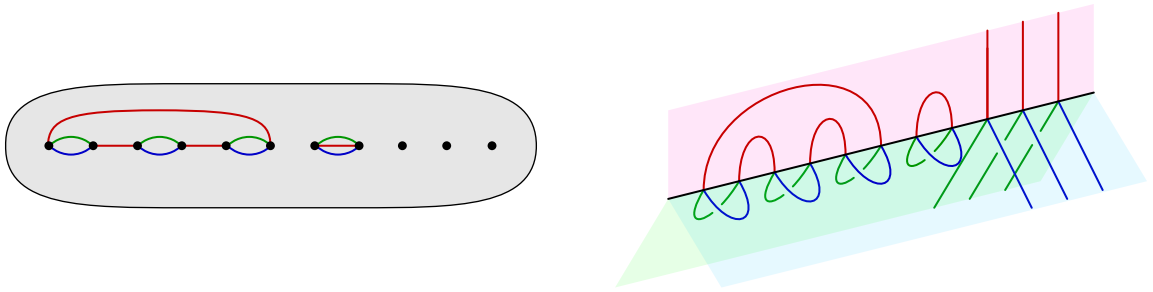
Figure 30: A shadow diagram, left, and schematic tri-plane diagram, right, for the unique $(4, (2, 4, 2); 3)$–bridge trisection, which is totally reducible.

and $\mathbb{T}$ is the boundary sum of c genus zero bridge trisections of (B^4, S^2), each of which is a finger perturbation of the 1–bridge trisection, and v genus zero 0–bridge trisections of (B^4, D^2).

Proof Suppose without loss of generality that $c_2 = b$. By Proposition 3.14, $\mathscr{F}$ admits a $(b, \boldsymbol{c}; v)$–bridge-braided band presentation. In particular, $\mathscr{F}$ can be built with $n = b - c_2 = 0$ bands. It follows that $c_1 = c_3$. It also follows that the flat disks of $(Z_2, \mathscr{D}_2)$ are given as products on the b flat strands of $(H_2, \mathscr{T}_2)$.

We can assume that the union of the red and blue shadow arcs is a collection of c_1 embedded polygons in Σ, since they determine a b–bridge c_1–component unlink in $H_1 \cup_\Sigma \overline{H}_2$. We can also assume that the green shadow arcs coincide with the blue shadow arcs, due to the product structure on the flat disks of $\mathscr{D}_2$. See Figure 30, left.

Let δ be a simple closed curve in $\Sigma \setminus \nu(\boldsymbol{x})$ that separates the red/blue polygons from the bridge points that are adjacent to no shadow arc. (Note that, here, every bridge point is adjacent to either 0 or 3 shadow arcs by the above considerations.) Then δ is a reducing curve for $\mathbb{T}$ such that $\mathbb{T} = \mathbb{T}^1 \# \mathbb{T}^2$, where $\mathbb{T}^1$ is a $(b, \boldsymbol{c})$–bridge trisection for a pair $(S^4, \mathscr{F}^1)$ and $\mathbb{T}^2$ is a $(0, 0; v)$–bridge trisection for a pair $(B^4, \mathscr{F}^2)$.

Because the blue and green shadow arcs coincide, each polygon is a finger perturbation of the 1–bridge splitting of (S^4, S^2), and $\mathscr{F}^1 = \bigsqcup_c S^2$. Moreover, $\mathbb{T}^1$ admits $c - 1$ reducing curves that completely separate the polygons. It follows that $\mathbb{T}^1$ is connected sum of perturbations of the 1–bridge trisection of (S^4, S^2), as desired. Finally, the bridge trisection $\mathbb{T}^2$ admits $v - 1$ reducing arcs that cut it up into v copies of the genus zero 0–bridge trisection of (B^4, D^2), as desired. $\square$

Having dispensed of the case when $c_i = b$ for some $i \in \mathbb{Z}_3$, we consider the case when $b = 1$ and, in light of the above, $c_i = 0$ for all $i \in \mathbb{Z}_3$. Two simple examples of such bridge trisections are given in Figure 31.

For a more interesting family of examples, consider the $(2, 4)$–torus link $T_{2,4}$, which bounds the union of the trivial Möbius band M^2 and the trivial disk D^2. (Imagine Figure 32, left, with the three parallel circles replaced with a single circle.) Now, consider the surface $\mathscr{F}_v$ obtained by replacing the D^2 with $v - 1$ parallel, trivial disks; Figure 32, left, shows the case of $v = 4$. A $(1, 0; v)$–bridge trisection $\mathbb{T}_v$ for
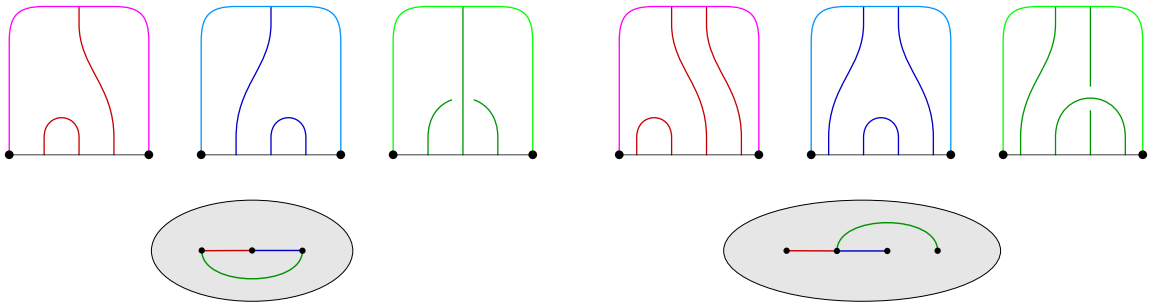
Figure 31: Top: the $(1, 0; 1)$–bridge trisection corresponding to the standard (positive) Möbius band (B^4, M^2). Bottom: the $(1, 0; 2)$–bridge trisection corresponding to the unknotted disk (B^4, D^2) with (positive) Markov stabilized, unknotted boundary.

$(B^4, \mathscr{F}_v)$ is shown in Figure 32, right. Note that when $v = 1$, $\mathbb{T}_v$ corresponds the trivial (positive) Möbius band with unknotted boundary and was given diagrammatically in Figures 31, top left and bottom left.

One can check using the techniques of Section 4.1 that the bridge trisection $\mathbb{T}_v$ induces the v–braiding of $\partial \mathscr{F}_v$ given in Artin generators by

$$(\sigma_1 \sigma_2 \cdots \sigma_{v-2} \sigma_{v-1}^2 \sigma_{v-2} \cdots \sigma_2 \sigma_1)^2.$$

In other words, one strand wraps twice around the other $v - 1$ strands. The link $\partial \mathscr{F}_v$ can be thought of as taking the $(v-1, 0)$–cable of one component of $T_{2,4}$.

Proposition 7.12 *The bridge trisection $\mathbb{T}_v$ is the unique (up to mirroring) irreducible $(1, 0; v)$–bridge trisection.*

Proof Suppose that $\mathbb{T}$ is an irreducible $(1, 0; v)$–bridge trisection, and consider a shadow diagram for $\mathbb{T}$. Since $b = 1$, there is a unique shadow arc of each of color. Since $c = 0$, the union of any two of these
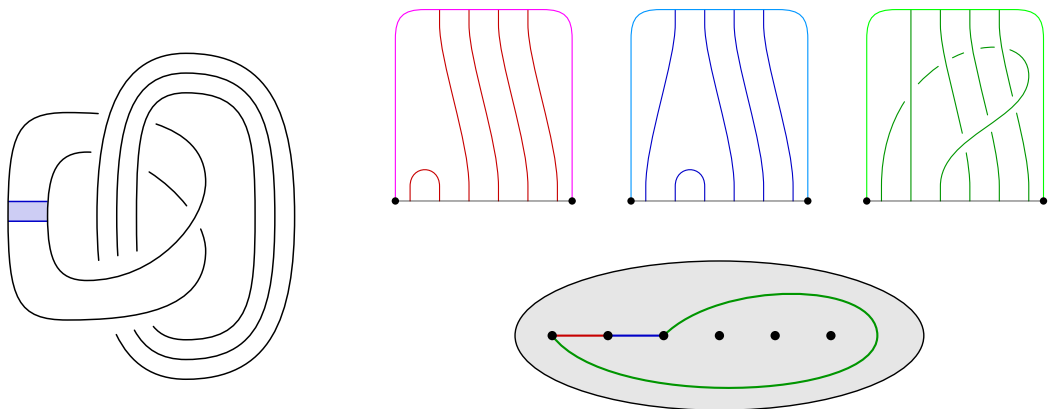
Figure 32: Left: a ribbon presentation for a nontrivial linking of a Möbius band with trivial disks in B^4. Such surfaces admit 1–bridge trisections, diagrams for which are shown to the right.

shadow arcs is a connected, embedded, polygonal arc in Σ, by Proposition 5.2 (no slides are possible, only isotopies). There are two cases: either the union of the three shadow arcs is a circle, or the union of the three shadow arcs is a Y–shaped tree.

Suppose the union is a Y–shaped tree. Let η be an arc connecting the tree to $\partial\Sigma$, and let ω be the arc boundary of a neighborhood of the union of η and the tree. In other words, ω is a neatly embedded arc in $\Sigma \setminus \nu(\boldsymbol{x})$ that separates the tree from the rest of the diagram. If the rest of the diagram is nonempty, then δ is a reducing arc for the bridge trisection, and we have $\mathbb{T} = \mathbb{T}^1 \natural \mathbb{T}^2$, where $\mathbb{T}^1$ is a $(1, 0; 2)$–bridge trisection (with Y–shaped shadow diagram) and $\mathbb{T}^2$ is a $(0, 0; v)$–bridge trisection, with $v > 0$. This contradicts the assumption that $\mathbb{T}$ was irreducible. If $v = 0$ (ie the rest of the diagram is empty), then $\mathbb{T} = \mathbb{T}^1$ is the Markov perturbation of the genus zero 0–bridge trisection and is shown in Figure 31, bottom right, so $\mathbb{T}$ is reducible, another contradiction.

Now suppose that the union of the three shadow arcs is a circle, and let $D \subset \Sigma$ denote the disk the union bounds. Suppose there is a bridge point in $\Sigma \setminus D$. Then there is a reducing arc separating the bridge point from D, so $\mathbb{T}$ is boundary reducible, a contradiction. So, the $v - 1$ bridge points that are not adjacent to a shadow arc are contained in D. Therefore, the shadow diagram is the one given in Figure 32, bottom right, or, in the case that $v = 1$, in Figure 31, bottom left. □

Having walked through these modest classification results, we now present some families of examples, as well as some questions and conjectures about further classification results.

Example 7.13 Consider the three $(2, 0; 1)$–bridge trisections shown in Figure 33, which correspond to the punctured torus and two different Klein bottles. All three surfaces are isotopic into S^3 and are bounded by the unknot. The two Klein bottles decompose as boundary connected sums of Möbius bands bounded by the unknot in S^3. The Klein bottle depicted in Figure 33, middle, is the boundary connected sum of two positive Möbius bands; and the Klein bottle depicted in Figure 33, bottom, is the boundary connected sum of a positive and a negative Möbius bands

These three bridge trisections can be obtained by taking the three unique $(3, 1)$–bridge trisections [27, Section 4.5] of closed surfaces in S^4 and puncturing at a bridge point.

Conjecture 7.14 *There are exactly three (up to mirroring) irreducible $(2, 0; 1)$–bridge trisections.*

Example 7.15 Consider the $(2, 0; 2)$–bridge trisection shown in Figure 34, left, which corresponds the annulus S^3 bounded by the $(2, 4)$–torus link. Compare with Example 3.16 and Figure 14(a). Replacing the three positive half-twists with n half-twists for some $n \in \mathbb{Z}$ gives a surface in S^3 bounded by the $(2, n)$–torus link that is a Möbius band if n is odd and an annulus if n is even.

One interesting aspect of the case when n is even relates to the orientation of the boundary link. The boundary link, which is the $(2, n)$–torus link, inherits an orientation as a 2–braid. It also inherits an orientation from the spanning annulus that the bridge trisection describes. These orientations don't agree!
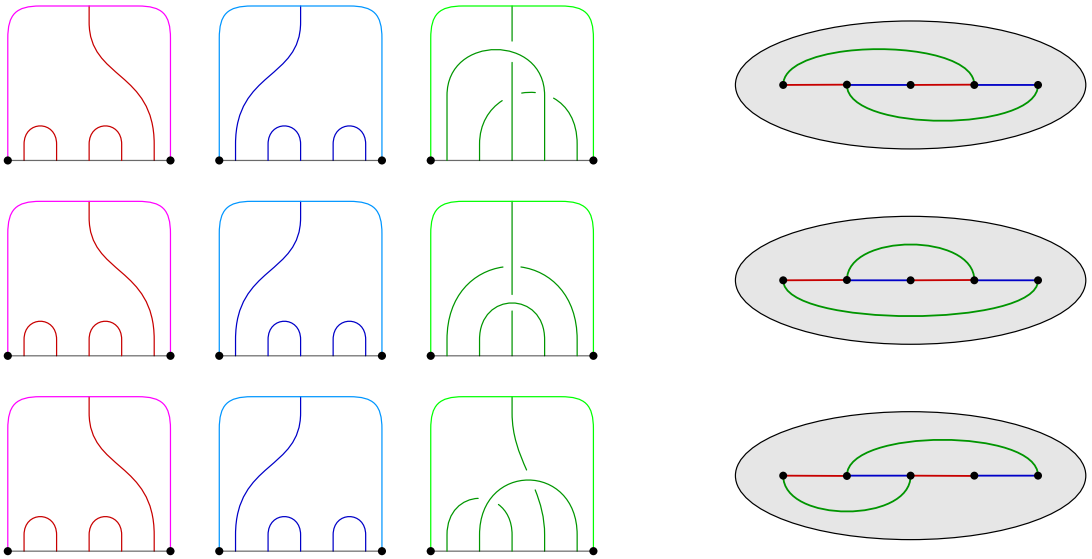
Figure 33: Three $(2, 0; 1)$–bridge trisections for surfaces bounded by the unknot and isotopic in S^3. The top row describes a punctured torus; the middle row describes the boundary connected sum of two positive Möbius bands; and the bottom row describes the boundary connected sum of a positive and a negative Möbius band.

In other words, the bridge trisections of the spanning annuli for these links induce a braiding of the links, but this braiding is not coherent with respect to the orientation of the links induced by the annuli. Compare with Example 7.17 below.

Conjecture 7.16 *Every $(2, 0; 2)$–bridge trisection is diffeomorphic to one described in Example 7.15 and in Figure 34.*

Example 7.17 Figure 35, left, gives a $(3, 0; 3)$–bridge trisection for the annulus in S^3 bounded by the $(2, 4)$–torus link. In contrast to the bridge trisection for this surface discussed in Example 7.15 and illustrated in Figure 34, this bridge trisection induces a coherent 3–braiding of the boundary link. This example could be generalized to give an $(n+1, 0; n+1)$–bridge trisection for the annulus bounded by the $(2, n)$–torus link for any even $n \in \mathbb{Z}$.
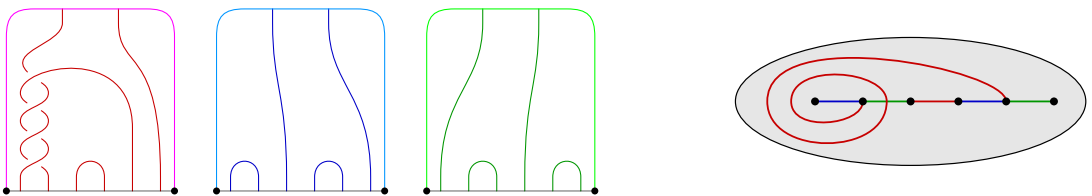
Figure 34: Diagrams for a $(2, 0; 2)$–bridge trisection of the planar surface bounded by the $(2, n)$–torus link in S^3; shown is $n = 4$.
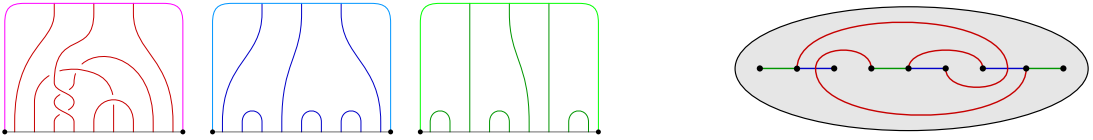
Figure 35: Diagrams for a $(3, 0; 3)$–bridge trisection of the planar surface bounded by the $(2, n)$–torus link in S^3; shown is $n = 4$.

8 Proof of Theorem 8.1

We now make use of the general framework outlined in Section 2 to give a proof of Theorem 8.1, which we restate for convenience. We adopt the notation and conventions of Definition 2.18.

Theorem 8.1 *Let $\mathbb{T}$ be a trisection of a four-manifold X with $\partial X = Y$, and let (B, π) denote the open-book decomposition of Y induced by $\mathbb{T}$. Let $\mathscr{F}$ be a neatly embedded surface in X; let $\mathscr{L} = \partial \mathscr{F}$; and fix a braiding $\hat{\beta}$ of $\mathscr{L}$ about (B, π). Then $\mathscr{F}$ can be isotoped to be in bridge trisected position with respect to $\mathbb{T}$ such that $\partial \mathscr{F} = \hat{\beta}$. If $\mathscr{L}$ already coincides with the braiding β, then this isotopy can be assumed to restrict to the identity on Y.*

Note that if X is closed, then Theorem 8.1 is equivalent to [28, Theorem 1]. For this reason, we assume henceforth that $Y = \partial X \neq \varnothing$. We will prove Theorem 8.1 using a sequence of lemmata. Throughout, we will disregard orientations. All isotopies are assumed to be smooth and ambient. First, we describe the existence of a Morse function $\Phi_{\mathbb{T}}$ on (most of) X that is well-adapted to the trisection $\mathbb{T}$. We will want to think of X as a lensed cobordism from Y_1 to $Y_2 \cup_{P_3} Y_3$.

Lemma 8.2 *There is a self-indexing Morse function*

$$\Phi_{\mathbb{T}} : X \setminus \nu(P_1 \cup_B P_2 \cup_B P_3) \to [0, 4]$$

such that

(1) *$\Phi_{\mathbb{T}}$ has no critical points of index zero or four;*

(2) *$Y_1 \setminus \nu(P_1 \cup_B \overline{P}_2) = \Phi_{\mathbb{T}}^{-1}(0)$;*

(3) *$(H_1 \cup_{\Sigma} \overline{H}_2) \setminus \nu(P_1 \cup_B \overline{P}_2) = \Phi_{\mathbb{T}}^{-1}(1.5)$;*

(4) *$\Phi_{\mathbb{T}}(H_3 \setminus \nu(P_3)) \subset [1.5, 2.5)$;*

(5) *$Y_3 \setminus \nu(P_3 \cup_B \overline{P}_1) = \Phi_{\mathbb{T}}^{-1}(4)$; and*

(6) *the index j critical points of $\Phi_{\mathbb{T}}$ are contained in $\mathrm{Int}(Z_j)$.*

Note that if $\Phi_{\mathbb{T}}(x) \geq 2.5$, then $x \in Z_3$.

Proof The existence of the Morse function and property (1) are standard consequences of the cobordism structure. The other properties are easy and commonly discussed within the theory of trisections; see [10], for example. The set-up is made evident by the schematics of Figure 36. □
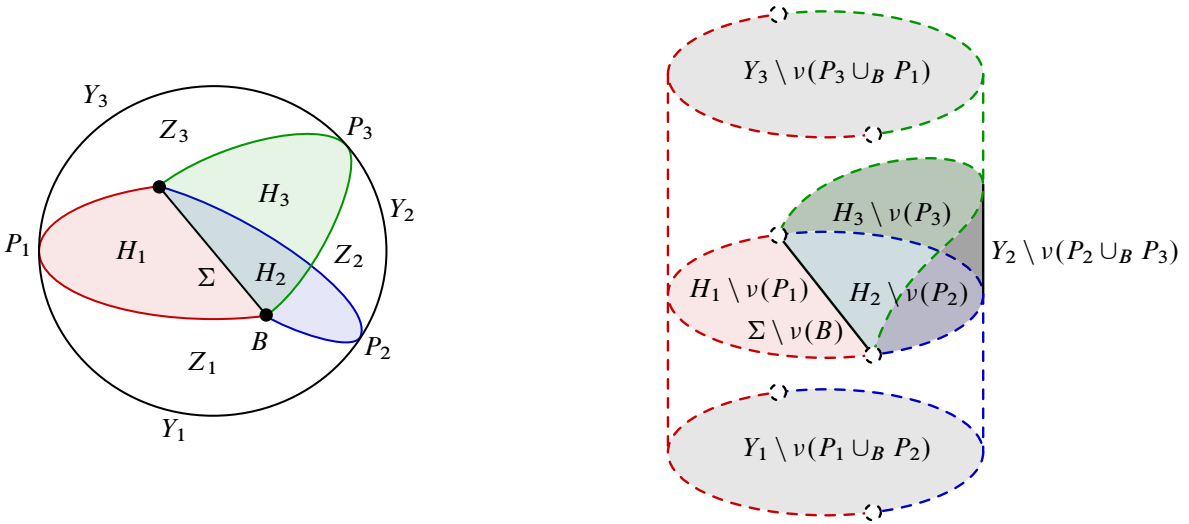
Figure 36: Passing from a trisection to a natural Morse function on $X \setminus \nu(P_1 \cup P_2 \cup P_3)$.

Now, Z_1 is the result of attaching four-dimensional 1–handles to the lensed product $Y_1 \times I$. The core Σ can be assumed to satisfy $\Phi_{\mathbb{T}}(\Sigma \setminus \nu(B)) = 1.5$, and, together with P_1 and P_2, it gives a standard Heegaard double decomposition of ∂Z_1. The attaching circles of the four-dimensional 2–handles are assumed to be contained in a (1–complex) spine of the compression body H_2, with the result of Dehn surgery thereupon being H_3. The trace of this 2–handle attachment is Z_2, and Z_3 is the (lensed cobordism) trace of attaching four-dimensional 3–handles to $H_3 \cup_\Sigma \overline{H}_1$, the result of which is Y_3. (Note that Z_2 is not quite a lensed cobordism from this perspective, since Y_2 is a vertical portion of its boundary $\partial Z_2 = H_2 \cup Y_2 \cup \overline{H}_3$.)

For the remainder of the section, we let $\Phi = \Phi_{\mathbb{T}}$. Let $\Phi_i = \Phi|_{Z_i}$ for $i = 1, 2, 3$. Recall the standard Morse function on $Z_i \cong Z_{g,k_i,\boldsymbol{p},\boldsymbol{f}}$ that was discussed in Section 2.7. By the above discussion, we have the following consequence of Lemma 8.2:

Corollary 8.3 *If $i = 1$ or $i = 3$, then Φ_i is a standard Morse function on $Z_i \cong Z_{g,k_i,\boldsymbol{p},\boldsymbol{f}}$.*

Presently, we will begin to isotope $\mathscr{F}$ to lie in bridge trisected position with respect to $\mathbb{T}$.

Lemma 8.4 *After an isotopy of $\mathscr{F}$ that is supported near ∂X, we can assume that $\mathscr{L} = \hat{\beta}$.*

Proof By the Alexander theorem [1] or the generalization due to Rudolph [31], $\mathscr{L}$ can be braided with respect to the open-book decomposition (B, π). By the Markov theorem [25] or its generalization to closed 3–manifolds [32; 33], any two braidings of $\mathscr{L}$ with respect to (B, π) are isotopic. Thus, by an isotopy of $\mathscr{F}$ that is supported near Y, we can assume that $\mathscr{L}$ is given by the braiding to $\hat{\beta}$. $\square$

Any modifications made to $\mathscr{F}$ henceforth will be isotopies that restrict to the identity on Y. Let $\Phi_{\mathscr{F}}$ denote the restriction of Φ to $\mathscr{F}$. (Note that by choosing a small enough collar $\nu(Y)$ in X, we can assume that $\mathscr{F} \cap \nu(Y) = \mathscr{L} \times I$. By a small isotopy of $\mathscr{F}$ rel-∂, we can assume that $\Phi_{\mathscr{F}}$ is Morse.)

Lemma 8.5 *After an isotopy of $\mathscr{F}$ rel-∂, we can assume that $\Phi_{\mathscr{F}} \colon \mathscr{F} \to \mathbb{R}$ is Morse and that*

(1) *the minima of $\Phi_{\mathscr{F}}$ occur in Z_1,*

(2) *the saddles of $\Phi_{\mathscr{F}}$ occur in $\Phi^{-1}(1.5)$, and*

(3) *the maxima of $\Phi_{\mathscr{F}}$ occur in Z_3.*

Proof That the critical points can be rearranged as desired follows from an analysis of their various ascending and descending manifolds. A detailed analysis of this facet of (embedded) Morse theory can be found in [3]. Here, we simply make note of the key points.

The ascending (unstable) membrane of a maximum of $\Phi_{\mathscr{F}}$ is one-dimensional; think of a vertical arc emanating from the maximum and terminating in Y_3. (Vertical means the intersection with each level set is either a point or empty.) Generically, such an arc will be disjoint from $\mathscr{F}$ and will be disjoint from the descending spheres of the critical points of Φ (which have index one, two, or three) in each level set. Thus, the gradient flow of Φ can be used to push the maxima up (and the minima down), and we obtain that the minima lie below $\Phi^{-1}(1.5)$ (ie in Z_1) and that the maxima lie above $\Phi^{-1}(2.5)$ (ie in Z_3). Having arranged the extrema in this way, we move on to consider the saddles.

The ascending membranes of the saddles of $\Phi_{\mathscr{F}}$ are two-dimensional, while the descending spheres of the index one critical points of Φ are zero-dimensional. Thus, we can flow the saddles up past the index one critical points of Φ, until they lie in $\Phi^{-1}(1.5)$. Symmetrically, we can flow saddles down past the index three critical points of Φ to the same result. □

Let $\mathscr{D}_i = \mathscr{F} \cap Z_i$ for $i = 1, 2, 3$. Assume that $\hat{\beta}$ is a braiding of $\mathscr{L}$ of multiindex $\boldsymbol{v}$.

Lemma 8.6 *If $\Phi_{\mathscr{F}}$ has c_1 minima and c_3 maxima, then $\mathscr{D}_1$ is a $(c_1, \boldsymbol{v})$–disk-tangle, and $\mathscr{D}_3$ is a $(c_3, \boldsymbol{v})$–disk-tangle.*

Proof By Corollary 8.3, Φ_1 is a standard Morse function on Z_i. By Lemma 2.14, since $(\Phi_1)|_{\mathscr{D}_1}$ has c_1 minima and no other critical points, and since $\mathscr{F} \cap Y_1 = \hat{\beta} \cap Y_1$ is a $\boldsymbol{v}$–thread, this implies that $\mathscr{D}_1$ is a $(c, \boldsymbol{v})$–disk-tangle. The corresponding result holds for $\mathscr{D}_3$, after turning Φ_3 and $(Z_3, \mathscr{D}_3)$ upside down. □

Next, we see that the trisection $\mathbb{T}$ can be isotoped to ensure the intersections $\mathscr{T}_i = \mathscr{F} \cap H_i$ are trivial tangles for $i = 1, 2, 3$.

Lemma 8.7 *After an isotopy of $\mathbb{T}$, we can assume that each $\mathscr{T}_i$ is a $(b, \boldsymbol{v})$–tangle for some $b \geq 0$.*

Proof The level set $\Phi^{-1}(1.5)$ is simply $M = (H_1 \cup_\Sigma \overline{H}_2) \setminus \nu(\overline{P}_1 \cup_B P_2)$. The intersection $\mathscr{F} \cap \Phi^{-1}(1.5)$ is a 2–complex $L \cup \mathfrak{b}$, where L is a neatly embedded one-manifold L, and $\mathfrak{b}$ is a collection of bands. Here, we are employing the standard trick of flattening $\mathscr{F}$ near each of the saddle points of $\Phi_{\mathscr{F}}$. (See Section 3.2 for a precise definition of a band.)

We have a Heegaard splitting $(\Sigma; H_1, H_2)$ that induces a Morse function $\Psi \colon \Phi^{-1}(1.5) \to \mathbb{R}$. In what follows, we will perturb this splitting (ie homotope this Morse function) to improve the arrangement of

the 2–complex $L \cup \mathfrak{b}$. First, we perturb Σ so that it becomes a bridge surface for L. At this point, we have arranged that $\mathcal{T}_1$ and $\mathcal{T}_2$ are $(b', \boldsymbol{v})$–tangles, for some value b' that will likely be increased by what follows.

Next, we can perturb Σ until the bands $\mathfrak{b}$ can be isotoped along the gradient flow of Ψ so that their cores lie in Σ. We can further perturb Σ until $\mathfrak{b} \cap \Sigma$ consists solely of the cores of $\mathfrak{b}$, which are embedded in Σ; said differently, the bands of $\mathfrak{b}$ are determined by their cores in Σ, together with the surface-framing given by the normal direction to Σ in $\Psi^{-1}(1.5)$. Finally, we can further perturb Σ until each band is dualized by a bridge semidisk for $\mathcal{T}_2$. The details behind this approach were given in the proof of Theorem 1.3 (using Figures 10–12) of [27] and discussed in [28].

Finally, we isotope Σ so that $\mathfrak{b}$ is contained in H_2; in other words, we push the bands slightly into H_2 so as to be disjoint from Σ. Since each band of $\mathfrak{b}$ is dualized by a bridge semidisk for $\mathcal{T}_2$, the result $\mathcal{T}_3 = (\mathcal{T}_2)_{\mathfrak{b}}$ of resolving $\mathcal{T}_2$ using the bands of $\mathfrak{b}$ is a new trivial tangle. The proof of this claim is explained in detail in [27, Lemma 3.1 and Figure 8]. (Though it is not necessary, we can even perturb Σ so that $\mathfrak{b}$ is dualized by a bridge disk at *both* of its endpoints, as in the aforementioned [27, Figure 8].)

Note that all of the perturbations of Σ were supported away from $\nu(P_1 \cup_B P_2)$, so each of the $\mathcal{T}_i$ contained precisely $\boldsymbol{v}$ vertical strands throughout. In the end, each is a $(b, \boldsymbol{v})$–tangle for some $b \geq 0$. $\square$

Finally, we verify that $\mathcal{D}_2$ is a trivial disk-tangle in Z_2.

Lemma 8.8 *If $c_2 = b - |\mathfrak{b}|$, then $\mathcal{D}_2$ is a $(c_2, \boldsymbol{v})$–disk-tangle.*

Proof As in the preceding lemma, this follows exactly along the lines of [27, Lemma 3.1], with only slight modification to account for the vertical strands. This is particularly easy to see if one assumes that $\mathfrak{b}$ meets dualizing disks at each of its endpoints, as in [27, Figure 8]. $\square$

Thus, we arrive at a proof of Theorem 8.1.

Proof of Theorem 8.1 After performing the isotopies of $\mathcal{F}$ and $\mathbb{T}$ outlined in the lemmata above, we have arranged that, for $i = 1, 2, 3$, the intersection $\mathcal{D}_i = \mathcal{F} \cap Z_i$ is a $(c_i, \boldsymbol{v})$–disk-tangle in Z_i (Lemmata 8.6 and 8.8) and the intersection $\mathcal{T}_i = \mathcal{F} \cap H_i$ is a $(b, \boldsymbol{v})$–tangle (Lemma 8.7). Thus, $\mathcal{F}$ is in $(b, \boldsymbol{c}; \boldsymbol{v})$–bridge trisected position with respect to $\mathbb{T}$, where $\boldsymbol{c} = (c_1, c_2, c_3)$, and the ordered partition $\boldsymbol{v}$ comes from the multiindex $\boldsymbol{v}$ of the braiding $\hat{\beta}$ of $\mathcal{L} = \partial \mathcal{F}$. $\square$

9 Stabilization operations

In this section we describe various stabilization and perturbation operations that can be used to relate two bridge trisections of a fixed four-manifold pair. We encourage the reader to refer back to the discussion of connected sums and boundary connected sums of bridge trisections presented in Section 7.

9.1 Stabilization of four-manifold trisections

First, we'll recall the original stabilization operation of Gay and Kirby [10], as developed in [26].
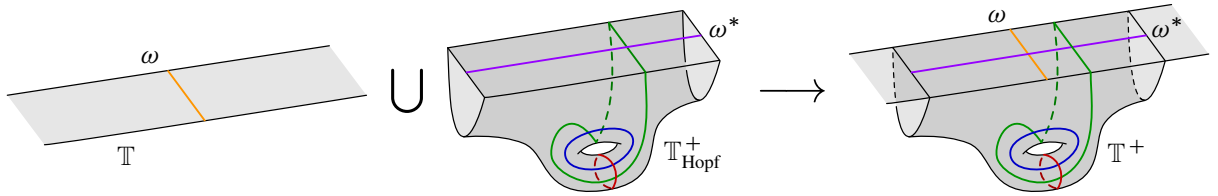
Figure 37: The positive Hopf stabilization $\mathbb{T}^+$ of a trisection $\mathbb{T}$ along an arc ω in the core of $\mathbb{T}$.

Definition 9.1 (core stabilization) Let $\mathbb{T}$ be a $(g, \boldsymbol{k}; \boldsymbol{p}, \boldsymbol{f})$–trisection for a four-manifold X, and let ω be an arc in Int(Σ). Fix an $i \in \mathbb{Z}_3$. Perturb the interior of ω into $H_{i+1} = Z_i \cap Z_{i+1}$, and let Σ' denote the surface obtained by surgering Σ along ω. Then, Σ' is the core of a $(g+1, \boldsymbol{k}'; \boldsymbol{p}, \boldsymbol{f})$–trisection $\mathbb{T}'$ for X, where $\boldsymbol{k}' = \boldsymbol{k}$, except that $k_i' = k_i + 1$, which is called the *core i–stabilization* of $\mathbb{T}$.

The importance of this operation rests in the following result of Gay and Kirby.

Theorem 9.2 [10] *Suppose that $\mathbb{T}$ and $\mathbb{T}'$ are two trisections of a fixed four-manifold X, and assume that either $\partial X = \varnothing$ or $\mathbb{T}$ and $\mathbb{T}'$ induce isotopic open-book decomposition on each connected component of ∂X. Then $\mathbb{T}$ and $\mathbb{T}'$ become isotopic after they are each core stabilized some number of times.*

Performing a core i–stabilization is equivalent to forming the (interior) connected sum with a simple trisection of S^4. Let $\mathbb{T}_i$ denote the genus one trisection of S^4 with $k_i = 1$. See [26] for details.

Proposition 9.3 $\mathbb{T}'$ *is a core i–stabilization of $\mathbb{T}$ if and only if $\mathbb{T}' = \mathbb{T} \# \mathbb{T}_i$.*

Next, we recall the stabilization operation for trisections that corresponds to altering the induced open-book decomposition on the boundary by the plumbing of a Hopf band. Let $\mathbb{T}_{\text{Hopf}}^+$ (resp. $\mathbb{T}_{\text{Hopf}}^-$) denote the genus one trisection of B^4 that induces the open-book decomposition on S^3 with binding the positive (resp. negative) Hopf link.

Definition 9.4 (Hopf stabilization) Let $\mathbb{T}$ be a $(g, \boldsymbol{k}; \boldsymbol{p}, \boldsymbol{f})$–trisection for a four-manifold X. Let $\omega \subset (\Sigma \setminus \alpha_i)$ be a neatly embedded arc, which we consider in P_i. Let $\mathbb{T}^\pm$ denote the trisection obtained by plumbing $\mathbb{T}$ to $\mathbb{T}_{\text{Hopf}}^\pm$ along the projection of ω, as in Figure 37. We call $\mathbb{T}^\pm$ the *positive/negative Hopf (i, j)–stabilization* of $\mathbb{T}$ along ω.

By a *plumbing* of trisections, we mean a plumbing of pages along the projection of arcs to the pages. Diagrammatically, this is represented by plumbing the relative trisection diagrams along the corresponding arcs in the core surface, as in Figure 37. This induces boundary connected sums at the level of the three-dimensional and four-dimensional pieces of the trisections and plumbing at the level of the core surfaces and pages. Hopf stabilization was first studied in the setting of trisections by Castro [4] and Castro, Gay and Pinzón-Caicedo [6]. We rephrase their main result in the more general setting of the present article.

Proposition 9.5 [6, Corollary 17] *Let* $\mathbb{T}$ *be a* $(g, \boldsymbol{k}; \boldsymbol{p}, \boldsymbol{f})$–*trisection for a four-manifold* X *inducing an open-book decomposition* (B, π) *on* ∂X. *Then a positive (resp. negative) Hopf stabilization* $\mathbb{T}^{\pm}$ *is a* $(g+1, \boldsymbol{k}; \boldsymbol{p}', \boldsymbol{f}')$–*trisection of* X *inducing a positive (resp. negative) Hopf stabilization of* (B, π), *where* $\boldsymbol{f}'$ *is obtained from* $\boldsymbol{f}$ *by either increasing or decreasing the value of* f_j *by one, and* $\boldsymbol{p}'$ *is obtained from* $\boldsymbol{p}$ *by either decreasing or increasing the value of* p_j *by one, according with, in each case, whether or not* ω *spans distinct boundary components of* P_i^j *or not.*

The upshot of this proposition is that, to the extent that open-book decompositions of three-manifolds are related by Hopf stabilization and destabilization, any two trisections of a compact four-manifold can be related by a sequence of Hopf stabilizations and core stabilizations. Giroux and Goodman proved that two open-book decompositions on a fixed three-manifold have a common Hopf stabilization if and only if the associated plane fields are homotopic [12], answering a question of Harer [14]. From this, together with Theorem 9.2, we can state the following.

Corollary 9.6 *Suppose that* $\mathbb{T}$ *and* $\mathbb{T}'$ *are two trisections of a fixed four-manifold* X. *Assume that* $\partial X \neq \varnothing$ *and that for each component of* ∂X, *the open-book decompositions induced by* $\mathbb{T}$ *and* $\mathbb{T}'$ *have associated plane fields that are homotopic. Then* $\mathbb{T}$ *and* $\mathbb{T}'$ *become isotopic after they are each core stabilized and Hopf stabilized some number of times.*

Recently, Piergallini and Zuddas showed there is a complete set of moves (including a *double-twist* move) that suffice to relate any two open-book decompositions on a given three-manifold [30]. By giving trisection-theoretic versions of each move, Castro, Islambouli, Miller, and Tomova were able to prove a strengthening of the original Gay–Kirby uniqueness theorem for trisected manifolds with boundary [7].

9.2 Interior perturbation of bridge trisections

Having overviewed stabilization operations for four-manifold trisections, we now discuss the analogous operations for bridge trisections. To avoid confusion, we will refer to these analogous operations as *perturbation operations*; they will generally correspond to perturbing the bridge trisected surface relative to the core surface. Throughout, the obvious inverse operation for a perturbation will be referred to as a *deperturbation*.

We begin by recalling the perturbation operation for bridge trisections first introduced in [27] and invoked in [28]. This perturbation operation requires the existence of a flat disk in $\mathcal{D}_i$. To distinguish this operation from the subsequent one, we append the adjective "Whitney".

Definition 9.7 (Whitney perturbation) Let $\mathcal{F}$ be a neatly embedded surface in a four-manifold X such that $\mathcal{F}$ is in $(\boldsymbol{b}, \boldsymbol{c}; \boldsymbol{v})$–bridge trisected position with respect to a trisection $\mathbb{T}$ of X. Let $D \subset \mathcal{D}_i$ be a flat disk, and let $D_* \subset Y_i$ be a disk that has no critical points with respect to the standard Morse function on Y_i and that is isotopic rel-∂ to D, via a three-ball B. Let Δ be a neatly embedded disk in B that
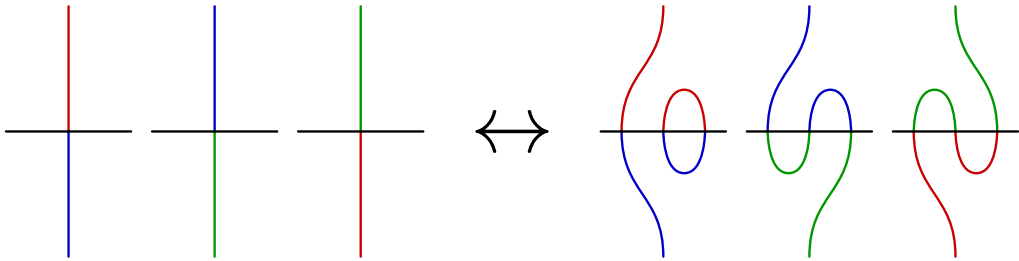
Figure 38: A local picture corresponding to a finger 1–perturbation.

intersects D_* in a vertical strand. Let $\mathscr{F}'$ denote the surface obtained by isotoping $\mathscr{F}$ via a Whitney move across Δ. Then $\mathscr{F}'$ is in $(b+1, c'; v)$–bridge trisected position with respect to $\mathbb{T}$, where $c' = c$, except that $c'_i = c_i + 1$. This Whitney move is called an *Whitney i–perturbation*.

See [28, Figures 14 and 23] for a visualization of a Whitney perturbation. The usefulness of Whitney perturbations is made clear by the following result, which was proved in [27] in the case that $\mathbb{T}$ has genus zero (so $X = S^4$) and in [16] in the general case.

Theorem 9.8 [16; 27] *Fix a four-manifold X and a trisection $\mathbb{T}$ of X. Let $\mathscr{F}, \mathscr{F}' \subset X$ be isotopic closed surfaces, and suppose $\mathbb{T}_{\mathscr{F}}$ and $\mathbb{T}_{\mathscr{F}'}$ are bridge trisections of $\mathscr{F}$ and $\mathscr{F}'$ induced by $\mathbb{T}$. Then there is a sequence of interior (Whitney) perturbations and deperturbations relating $\mathbb{T}_{\mathscr{F}}$ and $\mathbb{T}_{\mathscr{F}'}$*

Even without the presence of a flat disk, there is still a perturbation operation available. Despite being called a "finger" perturbation, the following perturbation is not an inverse to the Whitney perturbation. The adjective "Whitney" and "finger" are simply descriptive of how the surface is isotoped relative to the core to achieve the perturbation. However, it is true that the inverse to a Whitney perturbation (or a finger perturbation) is a finger deperturbation.

Definition 9.9 (finger perturbation) Let $\mathscr{F}$ be a neatly embedded surface in a four-manifold X such that $\mathscr{F}$ is in $(b, c; v)$–bridge trisected position with respect to a trisection $\mathbb{T}$ of X. Fix a bridge point $x \in \boldsymbol{x}$, and let N be a small neighborhood of x, so $N \cap \mathscr{F}$ is a small disk. Let $\omega \subset \partial N$ be a trivial arc connecting $\mathcal{T}_i$ to Σ. Perform a finger-move of $\mathscr{F}$ along ω, isotoping a small bit of $\mathscr{F}$ toward and through Σ, as in Figure 38. Let $\mathscr{F}'$ denote the resulting surface. Then, $\mathscr{F}'$ is in $(b+1, c'; v)$–bridge position with respect to $\mathbb{T}$, where $c' = c$, except that $c'_i = c_i + 1$. This finger move is called an *finger i–perturbation*.

Note that the disk of the disk-tangle $\mathcal{D}_i$ containing the bridge point x is neither required to be flat nor vertical in the definition of a finger perturbation. However, if this disk *is* flat, then the operation is the simplest form of a Whitney perturbation, corresponding to the case where the vertical strand in D_* is boundary parallel through vertical strands. The simplicity of the finger perturbation operation is expressed by the following proposition. Let $\mathbb{T}^i_{S^2}$ denote the 2–bridge trisection of the unknotted two-sphere satisfying $c_i = 2$.

Proposition 9.10 *If the bridge trisection* $\mathbb{T}'_{\mathscr{F}}$ *is obtained from the bridge trisection* $\mathbb{T}_{\mathscr{F}}$ *via a finger* i*–perturbation, then* $\mathbb{T}'_{\mathscr{F}} = \mathbb{T}_{\mathscr{F}} \# \mathbb{T}^i_{S^2}$.

The proof is an immediate consequence of how bridge trisections behave under connected sum. Note that a Whitney perturbation corresponds to a connected sum as in the proposition if and only if it is a finger perturbation; in general, a Whitney perturbation cannot be described as the result of a connected sum of bridge trisections. For example, the unknotted two-sphere admits a $(4, 2)$–bridge trisection that is not a connected sum of (nontrivial) bridge trisections, even though it is (Whitney) perturbed.

9.3 Markov perturbation of bridge trisections

Let $\mathbb{T}_{D^2}$ denote the 0–bridge trisection of the unknotted disk D^2 in B^4.

Definition 9.11 (Markov perturbation)   Let $\mathbb{T}'$ be a $(b, \boldsymbol{c}; \boldsymbol{v})$–bridge trisection of a neatly embedded surface $(X', \mathscr{F}')$, and let $\mathbb{T}''$ be the 0–bridge trisection of (B^4, D^2). Choose points $y^\varepsilon \in \mathcal{T}^\varepsilon_i \cap P^\varepsilon_i$ for $\varepsilon \in \{\prime, \prime\prime\}$. Let $(X, \mathscr{F}) = (X', \mathscr{F}') \natural (B^4, D^2)$, and let $\mathbb{T} = \mathbb{T}' \natural \mathbb{T}''$. Then $\mathbb{T}$ is a $(b+1, \boldsymbol{c}; \boldsymbol{v}')$–bridge trisection of $(X, \mathscr{F}) = (X', \mathscr{F}')$, where $\boldsymbol{v} = \boldsymbol{v}'$, except that $v^j = (v^j)' + 1$, where $y^1 \in \mathscr{L}^j$. The bridge trisection $\mathbb{T}'$ is called the *Markov i–perturbation* of $\mathbb{T}$.

In justification of this definition: That $\mathbb{T}'$ is a new bridge trisection follows from Proposition 7.5. That $\mathscr{F}'$ is isotopic to $\mathscr{F}$ follows from the fact that we are forming the boundary connected sum with a trivial disk. That $\mathscr{L}'$ is obtained from $\mathscr{L}$ via a Markov perturbation follows from our understanding of a Markov perturbation as the trivial connected sum of a braided link with a meridian of a component of the binding — see Section 2.8. Note that the left-most blue and green arcs of Figure 39 are shown in light blue and light green to indicate that they might correspond to flat or vertical strands. The pink arcs correspond to vertical strands.

The importance of this operation is due to the generalized Markov theorem, which states that any two braidings of a given link with respect to a fixed open-book decomposition can be related by an isotopy that preserves the braided structure, except at finitely many points in time at which the braiding is changed by a Markov stabilization or destabilization [25; 32; 33]. See Section 2.8.

Taken together, the stabilization and perturbation moves described in this section should suffice to relate any two bridge trisections of a fixed four-manifold pair. Compare with the known uniqueness results [7; 10; 16; 27].
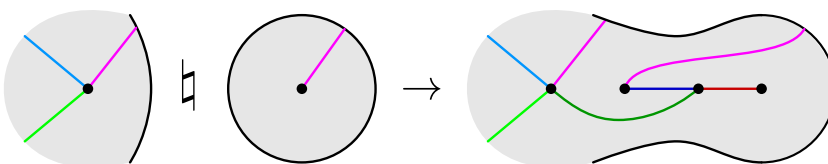
Figure 39: Shadow diagrams depicting the local process of Markov 3–perturbation.

Conjecture 9.12 *Let* $\mathbb{T}_1$ *and* $\mathbb{T}_2$ *be bridge trisections of a given surface* $(X, \mathscr{F})$ *that are diffeomorphic as trisections of* X. *Then there are diffeomorphic bridge trisections* $\mathbb{T}_1'$ *and* $\mathbb{T}_2'$ *such that* $\mathbb{T}_\varepsilon'$ *is obtained from* $\mathbb{T}_\varepsilon$ *via a sequence of moves, each of which is of one of the following types*:

(1) *a core stabilization*,

(2) *a Hopf stabilization*,

(3) *a relative double twist*,

(4) *an interior perturbation/deperturbation*,

(5) *a Markov perturbation/deperturbation*.

To prove this conjecture, it should suffice to carefully adapt the techniques of [16] from the setting of isotopy of closed four-manifold pairs equipped with Morse functions to the setting of isotopy rel-∂ of four-manifold pairs with boundary. The following is a diagrammatic analog to this conjecture.

Conjecture 9.13 *Suppose that* $\mathfrak{D}^1$ *and* $\mathfrak{D}^2$ *are shadow diagrams for a fixed surface-link* $(X, \mathscr{F})$. *Then* $\mathfrak{D}^1$ *and* $\mathfrak{D}^2$ *can be related by a finite sequence of moves, each of which is of one of the following types*:

(1) *a core stabilization/destabilization*,

(2) *a Hopf stabilization/destabilization*,

(3) *a relative double twist*,

(4) *an interior perturbation/deperturbation*,

(5) *a Markov perturbation/deperturbation*,

(6) *an arc or curve slide*,

(7) *an isotopy rel-∂*.

References

[1] **J W Alexander**, *Note on Riemann spaces*, Bull. Amer. Math. Soc. 26 (1920) 370–372 MR Zbl

[2] **R Blair**, **M Campisi**, **S A Taylor**, **M Tomova**, *Kirby–Thompson distance for trisections of knotted surfaces*, J. Lond. Math. Soc. 105 (2022) 765–793 MR Zbl

[3] **M Borodzik**, **M Powell**, *Embedded Morse theory and relative splitting of cobordisms of manifolds*, J. Geom. Anal. 26 (2016) 57–87 MR Zbl

[4] **N A Castro**, *Relative trisections of smooth 4–manifolds with boundary*, PhD thesis, University of Georgia (2016) Available at https://getd.libs.uga.edu/pdfs/castro_nickolas_a_201605_phd.pdf

[5] **N A Castro**, *Trisecting smooth 4–dimensional cobordisms*, preprint (2017) arXiv 1703.05846

[6] **N A Castro**, **D T Gay**, **J Pinzón-Caicedo**, *Diagrams for relative trisections*, Pacific J. Math. 294 (2018) 275–305 MR Zbl

[7] **N A Castro**, **G Islambouli**, **M Miller**, **M Tomova**, *The relative $\mathcal{L}$–invariant of a compact 4–manifold*, Pacific J. Math. 315 (2021) 305–346 MR Zbl

[8] **N A Castro**, **B Ozbagci**, *Trisections of 4–manifolds via Lefschetz fibrations*, Math. Res. Lett. 26 (2019) 383–420 MR Zbl

[9] **J B Etnyre**, *Lectures on open book decompositions and contact structures*, from "Floer homology, gauge theory, and low-dimensional topology" (D A Ellwood, P S Ozsváth, A I Stipsicz, Z Szabó, editors), Clay Math. Proc. 5, Amer. Math. Soc., Providence, RI (2006) 103–141 MR Zbl

[10] **D Gay**, **R Kirby**, *Trisecting 4–manifolds*, Geom. Topol. 20 (2016) 3097–3132 MR Zbl

[11] **D Gay**, **J Meier**, *Doubly pointed trisection diagrams and surgery on 2–knots*, Math. Proc. Cambridge Philos. Soc. 172 (2022) 163–195 MR Zbl

[12] **E Giroux**, **N Goodman**, *On the stable equivalence of open books in three-manifolds*, Geom. Topol. 10 (2006) 97–114 MR Zbl

[13] **R E Gompf**, **M Scharlemann**, **A Thompson**, *Fibered knots and potential counterexamples to the property 2R and slice-ribbon conjectures*, Geom. Topol. 14 (2010) 2305–2347 MR Zbl

[14] **J Harer**, *How to construct all fibered knots and links*, Topology 21 (1982) 263–280 MR Zbl

[15] **M W Hirsch**, *Differential topology*, Graduate Texts in Math. 33, Springer (1976) MR Zbl

[16] **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

[17] **K Johannson**, *Topology and combinatorics of 3–manifolds*, Lecture Notes in Math. 1599, Springer (1995) MR Zbl

[18] **J Joseph**, **J Meier**, **M Miller**, **A Zupan**, *Bridge trisections and classical knotted surface theory*, Pacific J. Math. 319 (2022) 343–369 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] **A Kawauchi**, *A survey of knot theory*, Birkhäuser, Basel (1996) 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] **F Laudenbach**, *Sur les 2–sphères d'une variété de dimension 3*, Ann. of Math. 97 (1973) 57–81 MR Zbl

[23] **F Laudenbach**, **V Poénaru**, *A note on 4–dimensional handlebodies*, Bull. Soc. Math. France 100 (1972) 337–344 MR Zbl

[24] **C Livingston**, *Surfaces bounding the unlink*, Michigan Math. J. 29 (1982) 289–298 MR Zbl

[25] **A A Markov**, *Über die freie Äquivalenz geschlossener Zöpfe*, Mat. Sb. 43 (1936) 73–78 Zbl

[26] **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

[27] **J Meier**, **A Zupan**, *Bridge trisections of knotted surfaces in S^4*, Trans. Amer. Math. Soc. 369 (2017) 7343–7386 MR Zbl

[28] **J Meier**, **A Zupan**, *Bridge trisections of knotted surfaces in 4–manifolds*, Proc. Natl. Acad. Sci. USA 115 (2018) 10880–10886 MR Zbl

[29] **J Meier**, **A Zupan**, *Generalized square knots and homotopy 4–spheres*, J. Differential Geom. 122 (2022) 69–129 MR Zbl

[30] **R Piergallini**, **D Zuddas**, *Special moves for open book decompositions of 3–manifolds*, J. Knot Theory Ramifications 27 (2018) art. id. 1843008 MR Zbl

[31] **L Rudolph**, *Constructions of quasipositive knots and links*, *I*, from "Knots, braids and singularities" (C Weber, editor), Monogr. Enseign. Math. 31, Enseign. Math., Geneva (1983) 233–245 MR Zbl

[32] **R K Skora**, *Closed braids in 3–manifolds*, Math. Z. 211 (1992) 173–187 MR Zbl

[33] **P A Sundheim**, *The Alexander and Markov theorems via diagrams for links in 3–manifolds*, Trans. Amer. Math. Soc. 337 (1993) 591–607 MR Zbl

Department of Mathematics, Western Washington University
Bellingham, WA, United States

jeffrey.meier@wwu.edu

http://jeffreymeier.org

msp

Equivariantly slicing strongly negative amphichiral knots

Keegan Boyle

Ahmad Issa

We prove obstructions to a strongly negative amphichiral knot bounding an equivariant slice disk in the 4–ball using the determinant, Spin^c–structures and Donaldson's theorem. Of the 16 slice strongly negative amphichiral knots with 12 or fewer crossings, our obstructions show that 8 are not equivariantly slice, we exhibit equivariant ribbon diagrams for 5 others, and the remaining 3 are unknown. Finally, we give an obstruction to a knot being strongly negative amphichiral in terms of Heegaard Floer correction terms.

57K10, 57M60

1 Introduction

A *strongly negative amphichiral* knot (K, σ) is a smooth knot $K \subset S^3$ along with a smooth (orientation-reversing) involution $\sigma \colon S^3 \to S^3$ such that $\sigma(K) = K$ and σ has exactly two fixed points, both of which lie on K; see Figure 1. A knot $K \subset S^3$ is *slice* if it bounds a smooth disk (the *slice disk*) properly embedded in B^4. Our main goal is to study when there exists an equivariant slice disk for a strongly negative amphichiral knot (K, σ). Specifically, we are interested in the following property:

Definition 1.1 A strongly negative amphichiral knot (K, σ) is *equivariantly slice* if there is a smooth slice disk D and a smooth involution $\sigma' \colon B^4 \to B^4$ with $\sigma'(D) = D$ which restricts to σ on $\partial B^4 = S^3$.

Figure 1 gives an example of a strongly negative amphichiral diagram, that is, a knot diagram with the strongly negative amphichiral symmetry given by π–rotation around an axis perpendicular to the page followed by reflection across the plane of the diagram. Furthermore, the knot in Figure 1 is equivariantly slice. The slice disk is given by performing the pair of equivariant band moves shown in red, then equivariantly capping off the resulting 3–component unlink in B^4. Among nontrivial prime knots with 12 or fewer crossings, there are 16 slice strongly negative amphichiral knots. For five of them, namely 8_9, 10_{99}, $12a_{819}$, $12a_{1269}$ and $12n_{462}$, we found similar equivariant ribbon diagrams; see the table in Section 7.

Strongly negative amphichiral knots, and in particular the equivariant surfaces they bound in the 4–ball, have been studied less than their more popular orientation-preserving cousins: strongly invertible knots (see for example Boyle and Issa [2] and Sakuma [23]) and periodic knots (see for example [2], Cha and Ko [5], Davis and Naik [6], and Grove and Jabuka [14] among others). Many of the obstructions used
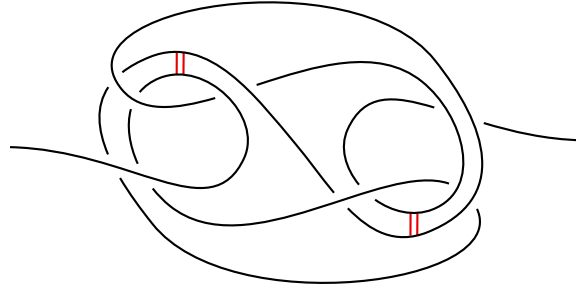
Figure 1: A strongly negative amphichiral diagram for 8_9. The symmetry is given by π–rotation around an axis perpendicular to the page, followed by a reflection across the plane of the diagram. An equivariant slice disk can be seen by performing the band moves shown in red.

in the strongly invertible and periodic settings do not adapt to the strongly negative amphichiral case. In fact, even showing that the (nonequivariant) 4–genus for strongly negative amphichiral knots can be arbitrarily large was only recently accomplished by Miller [20].

Our first equivariant slice obstruction comes from studying the knot determinant. It was shown by Goeritz [10] that the determinant of an amphichiral knot is the sum of two squares (see also Friedl, Miller and Powell [9] for a partial generalization and Stoimenow [24] for the converse). We prove the following strengthening of this determinant condition in the case that K bounds an equivariant slice disk:

Theorem 1.2 *If K is an equivariantly slice strongly negative amphichiral knot, then $\det(K)$ is the square of a sum of two squares.*

Theorem 1.2 shows that the six slice strongly negative amphichiral knots 10_{123}, $12a_{435}$, $12a_{990}$, $12a_{1019}$, $12a_{1225}$ and $12n_{706}$ are not equivariantly slice.

Our second obstruction, which applies to knots with an alternating strongly negative amphichiral diagram, comes from applying Donaldson's theorem [8]. Donaldson's theorem can often be used to obstruct the existence of slice disks (see for example Lisca [18]). More recently, it has also been used to obstruct equivariant slice disks for strongly invertible and periodic knots [2]. A key ingredient in that obstruction is the existence of an invariant definite spanning surface for the knot. In contrast, strongly negative amphichiral knots do not bound invariant spanning surfaces in S^3. Instead, we use the fact that, if K bounds an equivariant slice disk D, then the subset S of Spin^c–structures on the double branched cover $Y = \Sigma(S^3, K)$ that extend over $\Sigma(B^4, D)$ is $\tilde{\sigma}$–invariant, where $\tilde{\sigma}$ is a lift of the symmetry σ to Y; see Proposition 4.1 and the discussion following its proof. Donaldson's theorem can be used to obtain restrictions on S. Using the interplay between the pair of checkerboard surfaces exchanged by the symmetry, we carefully keep track of Spin^c–structures, allowing us to compute the $\tilde{\sigma}$–action on Spin$^c(Y)$. This results in a nice combinatorial description of the $\tilde{\sigma}$–action on Spin$^c(Y)$ in terms of the oriented incidence matrices of the checkerboard graphs for an alternating symmetric diagram. Specifically, we prove the following theorem:

Theorem 1.3 *Let (K, σ) be a knot with an alternating strongly negative amphichiral diagram and let $Y = \Sigma(S^3, K)$. Let $F_\pm$ be the positive and negative definite checkerboard surfaces, let $J_\pm^*$ be compatible oriented incidence matrices with a row removed[1] for the checkerboard graphs of $F_\pm$, and let $A_\pm = J_\pm^*(J_\pm^*)^\top \in M_n(\mathbb{Z})$ be the Goeritz matrices for $F_\pm$. Then there is a lift $\tilde{\sigma} : Y \to Y$ for which the map $\tilde{\sigma}^* : \mathrm{Spin}^c(Y) \to \mathrm{Spin}^c(Y)$ is determined by*

$$\tilde{\sigma}^*[J_+^* v] = [J_-^* v] \quad \text{for all } v \in \mathbb{Z}^{2n} \text{ with } v \equiv (1, 1, \ldots, 1)^\top \in (\mathbb{Z}/2\mathbb{Z})^{2n},$$

where $\mathrm{Spin}^c(Y) \cong \mathrm{Char}(\mathbb{Z}^n, A_+)/\mathrm{im}(2A_+)$. Moreover, if K is equivariantly slice, then there is a lattice embedding $A : (\mathbb{Z}^n, A_+) \to (\mathbb{Z}^n, \mathrm{Id})$ such that

$$S = \{[u] \in \mathrm{Spin}^c(Y) \mid u = A^\top v \text{ for some } v \in \mathbb{Z}^n \text{ with } v \equiv (1, 1, \ldots, 1)^\top \in (\mathbb{Z}/2\mathbb{Z})^n\}$$

is $\tilde{\sigma}^*$–invariant.

Using Theorem 1.3, we show that $12a_{1105}$ and $12a_{1202}$ are not equivariantly slice (see Section 5), even though they satisfy the determinant condition in Theorem 1.2 as $\det(12a_{1105}) = 17^2 = (4^2 + 1^2)^2$ and $\det(12a_{1202}) = 13^2 = (3^2 + 2^2)^2$. Of the slice strongly negative amphichiral knots with 12 or fewer crossings, this leaves only $12a_{458}$, $12a_{477}$ and $12a_{887}$ for which equivariant sliceness is unknown. See Section 7 for a table of equivariant knot diagrams for these knots.

Our analysis of the $\tilde{\sigma}$–action on $\mathrm{Spin}^c(\Sigma(S^3, K))$ also leads us to the following obstruction to strongly negative amphichirality in terms of Heegaard Floer correction terms.

Theorem 1.4 *Let (K, σ) be a strongly negative amphichiral knot and let $\tilde{\sigma}$ be a lift of σ to $Y := \Sigma(S^3, K)$ (see Proposition 2.1). Then the orbits of $\mathrm{Spin}^c(Y)$ under the action of $\tilde{\sigma}$ take the following form:*

(1) *There is exactly one orbit $\{\mathfrak{s}_0\}$ of order 1 with $d(Y, \mathfrak{s}_0) = 0$.*

(2) *All other orbits $\{\mathfrak{s}, \tilde{\sigma}(\mathfrak{s}), \tilde{\sigma}^2(\mathfrak{s}), \tilde{\sigma}^3(\mathfrak{s})\}$ have order 4 and*

$$d(Y, \tilde{\sigma}^i(\mathfrak{s})) = (-1)^i d(Y, \mathfrak{s}) \quad \text{for all } i.$$

For example, the figure eight knot 4_1 is strongly negative amphichiral and $\Sigma(S^3, 4_1) = L(5, 2)$, which has correction terms $\{0, \frac{2}{5}, -\frac{2}{5}, \frac{2}{5}, -\frac{2}{5}\}$. We checked that, for all 2–bridge knots with 12 or fewer crossings, the d–invariants have this structure precisely when the knot is strongly negative amphichiral, leading us to the following conjecture:

Conjecture 1.5 *Let $p, q \in \mathbb{N}$ with p odd and $(p, q) = 1$. The following are equivalent:*

(1) *The Heegaard Floer correction terms of the lens space $L(p, q)$ can be partitioned into multisets, each of the form $\{r, -r, r, -r\}$ for some $r \in \mathbb{Q}$, and a single set $\{0\}$.*

---

[1] See Definition 4.6. Here $J_\pm^*$ is an n by $2n$ matrix.

(2) *The 2–bridge knot $K(p/q)$ is amphichiral.*

(3) *There is an orientation-reversing self-diffeomorphism of $L(p, q)$.*

(4) $q^2 \equiv -1 \pmod{p}$.

We note that (2), (3) and (4) are known to be equivalent (see for example Bonahon [1, Theorem 3], Hodgson and Rubinstein [15, Corollary 4.12] and Stoimenow [24, Section 4]). Theorem 1.4 shows that (2) implies (1), since $\Sigma(S^3, K(p/q)) = L(p, q)$ and a 2–bridge knot is amphichiral if and only if it is strongly negative amphichiral. Thus Conjecture 1.5 is equivalent to showing that (1) implies any of the other conditions.

1.1 Open questions

We conclude the introduction with a list of interesting open questions for further exploration.

Question 1.6 Is there a nonslice strongly negative amphichiral knot with equivariant 4–genus larger than its 4–genus?

Question 1.7 Is there a strongly negative amphichiral knot which is topologically equivariantly slice but not smoothly equivariantly slice?

Question 1.8 Is every strongly negative amphichiral knot with Alexander polynomial 1 topologically equivariantly slice?

Question 1.9 If a strongly negative amphichiral knot is smoothly equivariantly slice, then must the knot admit an equivariant ribbon diagram, as in Figure 1?

Acknowledgments

2 Lifting the action to the double branched cover

In this section we show that the strongly negative amphichiral involution σ on S^3 lifts to the double branched cover $\Sigma(S^3, K)$. Since we are interested in equivariant slice disks for K, we also show that this lift $\tilde{\sigma}$ can be extended to $\Sigma(B^4, S)$ for any equivariant surface $S \subset B^4$ with $\partial S = K$. Specifically, we have the following proposition, which is similar to [2, Proposition 12]. However, in our situation there are no fixed points disjoint from the branch set; the amphichiral involution lifts to an order-4 symmetry on the double branched cover.

Proposition 2.1 *Let $S \subset S^4$ be a closed connected smoothly embedded surface and let $\sigma: (S^4, S) \to (S^4, S)$ be a smooth involution with nonempty fixed-point set contained in S. Let $p: \Sigma(S^4, S) \to S^4$ be the projection map from the double branched cover and let $\tau: \Sigma(S^4, S) \to \Sigma(S^4, S)$ be the nontrivial*

deck transformation map. Then there is a lift $\tilde{\sigma}\colon \Sigma(S^4, S) \to \Sigma(S^4, S)$ such that the following diagram commutes:

$$\begin{array}{ccc} \Sigma(S^4, S) & \xrightarrow{\ \tilde{\sigma}\ } & \Sigma(S^4, S) \\ {\scriptstyle p}\downarrow & & \downarrow{\scriptstyle p} \\ S^4 & \xrightarrow{\ \ \sigma\ \ } & S^4 \end{array}$$

Furthermore, $\tilde{\sigma}^2 = \tau$, and there are exactly two such lifts, namely $\tilde{\sigma}$ and $\tilde{\sigma}^3$.

Proof Let $N(S)$ be an equivariant tubular neighborhood of S and $E = S^4 \backslash N(S)$ be the surface exterior. Denote by $\tilde{E}$ the double cover of E corresponding to the kernel G of $\pi_1(E) \to H_1(E; \mathbb{Z}/2\mathbb{Z})$. We also choose a basepoint $s \in E$ and lifts $\tilde{s}, \tilde{t} \in \tilde{E}$ with $p(\tilde{s}) = s$ and $p(\tilde{t}) = \sigma(s)$.

Since G is the unique index-2 subgroup of $\pi_1(E, s)$, it is a characteristic subgroup. Hence G is also the image of $\pi_1(\sigma \circ p)\colon \pi_1(\tilde{E}, \tilde{t}) \to \pi_1(E, s)$. Then by the covering space lifting property, since $\mathrm{im}(\pi_1(\sigma \circ p)) \subseteq \mathrm{im}(\pi_1(p)\colon \pi_1(\tilde{E}, \tilde{s}) \to \pi_1(E, s))$, there is a unique map $\tilde{\sigma}\colon (\tilde{E}, \tilde{t}) \to (\tilde{E}, \tilde{s})$ such that $p \circ \tilde{\sigma} = \sigma \circ p$. By the equivariant tubular neighborhood theorem [3, Theorem VI.2.2], ∂E can be identified with the unit normal bundle of S, where σ preserves S^1 fibers. Lifting this bundle structure to $\partial \tilde{E}$, p gives a bijection between the set of fibers of ∂E and the set of fibers of $\partial \tilde{E}$ (p restricts to a two-to-one covering on each fiber). In particular, $\tilde{\sigma}$ preserves the set of S^1 fibers on the S^1–bundle boundary of $\tilde{E}$. By extending this action over each D^2 fiber, we can (smoothly) extend $\tilde{\sigma}$ to the tubular neighborhood $p^{-1}(N(S)) \subset \Sigma(S^4, S)$ such that $p \circ \tilde{\sigma} = \sigma \circ p$.

Finally, $p \circ \tilde{\sigma} = \sigma \circ p$ implies that $p \circ \tilde{\sigma}^2 = \sigma^2 \circ p = p$, so that $\tilde{\sigma}^2$ is either the identity map, or else the nontrivial deck transformation τ on $\Sigma(B^4, S)$. Note that, in either case, $\tilde{\sigma}^4$ is the identity map. However, σ acts by π–rotation on an equivariant meridian α of a fixed point of σ. Indeed, if σ acted by reflection or identity on α, then there would be fixed points disjoint from S. In the branched cover we then have that $\tilde{\sigma}$ acts by $\frac{\pi}{2}$–rotation on $p^{-1}(\alpha)$. Thus $\tilde{\sigma}$ has order 4 and $\tilde{\sigma}^2 = \tau$, as desired. Finally, we note that there are exactly two lifts, $\tilde{\sigma}$ and $\tau \circ \tilde{\sigma} = \tilde{\sigma}^3$, one for each choice of $\tilde{t}$. $\qquad\square$

Corollary 2.2 *Let (K, σ) be a strongly negative amphichiral knot with double branched cover $\Sigma(S^3, K)$. Let $S \subset B^4$ be a smooth properly embedded surface with boundary K which is invariant under an extension of σ to B^4 (which we again call σ). Then there is a lift $\tilde{\sigma}\colon \Sigma(B^4, S) \to \Sigma(B^4, S)$ such that $\tilde{\sigma}^2 = \tau$ (and hence $\tilde{\sigma}^4 = \mathrm{Id}$) and $p \circ \tilde{\sigma} = \sigma \circ p$. In fact, there are exactly two such lifts, namely $\tilde{\sigma}$ and $\tilde{\sigma}^3$.*

Proof Take the double of $\Sigma(B^4, S)$ to obtain a closed connected surface in S^4, then apply Proposition 2.1 and restrict to $\Sigma(B^4, S)$. $\qquad\square$

Proposition 2.3 *Every strongly negative amphichiral knot (K, σ) bounds a smooth properly embedded surface $S \subset B^4$ which is invariant under the cone of σ.*

Proof First we fix a symmetric diagram for (K, σ), from which we will produce an equivariant unknotting sequence. Since each equivariant pair of crossing changes produces an equivariant genus-2 cobordism,

this will imply that (K, σ) is equivariantly cobordant to the unknot. Then we note that the unknot bounds a smooth disk in B^4 (given by the cone of the unknot), which is invariant under the cone of σ.

For the equivariant unknotting sequence, separate K at the two fixed points of σ into two arcs, α and β. Now, for each equivariant pair of crossings between α and β, either α is the overstrand in both crossings, or β is. Hence we can perform equivariant crossing changes so that α is always the overstrand in crossings between α and β. Then we can pull α and β apart to get a knot of the form $J \# -J$, where the symmetry exchanges J and $-J$. Finally, any unknotting sequence for J produces an equivariant unknotting sequence for $J \# -J$, as desired. □

We conclude by lifting σ to the double branched cover of K.

Proposition 2.4 *Let (K, σ) be a strongly negative amphichiral knot. Then there exist exactly two lifts of σ to $\Sigma(S^3, K)$. Moreover, each such lift $\tilde{\sigma}$ has $\tilde{\sigma}^2 = \tau$, where $\tau: \Sigma(S^3, K) \to \Sigma(S^3, K)$ is the nontrivial deck transformation action, and hence $\tilde{\sigma}$ has order 4.*

Proof The proof is essentially the same as that of [Proposition 2.1]. It can also be obtained by restricting the lifts in [Corollary 2.2] to the boundary $\Sigma(S^3, K)$, using the surface guaranteed by [Proposition 2.3]. □

3 A condition on the determinant

It is implicit in the work of Goeritz [10] that the determinant of an amphichiral knot can be written as the sum of two squares (see also [24] for the converse and [9] for a partial generalization). In this section we reprove this theorem for strongly negative amphichiral knots, and show that the same condition must hold on the square root of the determinant if K is equivariantly slice.

Theorem 1.2 *Let (K, σ) be a strongly negative amphichiral knot. Then $\det(K)$ is a sum of two squares. Furthermore, if (K, σ) is equivariantly slice, then $\det(K)$ is the square of a sum of two squares.*

Before we give a proof of the theorem, we need a few lemmas.

Lemma 3.1 *Let A be an abelian group, and let $\Sigma(X, Y)$ be the double cover of a manifold X (possibly with boundary), branched over a properly embedded submanifold $Y \subset X$ with nontrivial deck transformation involution $\tau: \Sigma(X, Y) \to \Sigma(X, Y)$. Suppose that $H_n(X; A) = 0$. Then $\tau_*(x) = -x$ for all $x \in H_n(\Sigma(X, Y); A)$.*

Proof Since $H_n(X; A) = 0$, the image of the transfer homomorphism $T: H_n(X; A) \to H_n(\Sigma(X, Y); A)$ is 0. For any $x \in H_n(\Sigma(X, Y); A)$, we have that $x + \tau_*(x)$ is in the image of T and hence is 0. Thus $\tau_*(x) = -x$. □

Letting $(X, Y) = (S^3, K)$ in [Lemma 3.1], we observe that τ_* fixes only the identity element since $H_1(\Sigma(S^3, K); A)$ has no elements of order 2.

Lemma 3.2 [4, Lemma 3] *Let K be slice with slice disk $D \subset B^4$ and A be a torsion-free abelian group. If the image of $H_1(\Sigma(S^3, K); A)$ in $H_1(\Sigma(B^4, D); A)$ has order m, then $|H_1(\Sigma(S^3, K); A)| = m^2$.*

Proof The proof is as in [4, Lemma 3], noting that since A is torsion free the universal coefficient theorem does not introduce any unwanted Tor terms. □

Lemma 3.3 *Suppose (K, σ) has an equivariant slice disk D. Then the kernel of the map*

$$i_*: H_1(\Sigma(S^3, K); A) \to H_1(\Sigma(B^4, D); A),$$

induced by inclusion, is invariant under the induced action of any lift $\tilde{\sigma}: \Sigma(S^3, K) \to \Sigma(S^3, K)$ of σ on homology.

Proof Let $x \in \ker(i_*)$ so that x is a boundary in $\Sigma(B^4, D)$. By Corollary 2.2, there is an extension of the lift $\tilde{\sigma}$ to $\Sigma(B^4, D)$. Hence $\tilde{\sigma}_*(x)$ is also a boundary, and hence contained in $\ker(i_*)$. □

Proof of Theorem 1.2 By Proposition 2.4, σ lifts to an order-4 action $\tilde{\sigma}$ on $\Sigma(S^3, K)$ with $\tilde{\sigma}^2 = \tau$. In particular, Lemma 3.1 implies that all orbits of $\tilde{\sigma}_*: H_1(\Sigma(S^3, K); A) \to H_1(\Sigma(S^3, K); A)$ have order 4, except the orbit consisting of the identity element. Taking coefficients A as the p–adic integers $\mathbb{Z}_p$ for some prime p, we have

$$|H_1(\Sigma(S^3, K); \mathbb{Z}_p)| \equiv 1 \pmod 4.$$

For $p \equiv 3 \pmod 4$, this implies that $|H_1(\Sigma(S^3, K); \mathbb{Z}_p)|$ is an even power of p. However, by the universal coefficient theorem, $H_1(\Sigma(S^3, K); \mathbb{Z}_p) \cong H_1(\Sigma(S^3, K); \mathbb{Z}) \otimes \mathbb{Z}_p$ and hence the prime decomposition of $|H_1(\Sigma(S^3, K); \mathbb{Z})| = \det(K)$ contains an even power of p. By the sum of two squares theorem, we then have that $\det(K)$ is the sum of two squares.

Now suppose that (K, σ) has an equivariant slice disk $D \subset B^4$. By Lemma 3.2 with p–adic coefficients, the kernel of $H_1(\Sigma(S^3, K); \mathbb{Z}_p) \to H_1(\Sigma(B^4, D); \mathbb{Z}_p)$ is a square-root order subgroup of $H_1(\Sigma(S^3, K); \mathbb{Z}_p)$, and by Lemma 3.3, this subgroup is invariant under the action of $\tilde{\sigma}_*$. In particular this subgroup must consist of the identity plus a (finite) collection of order-4 orbits, so that

$$\sqrt{|H_1(\Sigma(S^3, K); \mathbb{Z}_p)|} \equiv 1 \pmod 4.$$

As above, we then have that $\sqrt{\det(K)}$ can be written as the sum of two squares. □

4 An obstruction on Spin^c–structures

In this section we prove Theorem 1.3, giving an obstruction to an alternating strongly negative amphichiral knot bounding an equivariant slice disk D in B^4. We do so by considering Spin^c–structures on the double branched cover and applying Donaldson's theorem. This obstruction is based on the following observation:

Proposition 4.1 *Let $\rho\colon Y \to Y$ be a diffeomorphism of a closed 3–manifold Y. If ρ extends to a diffeomorphism $\rho'\colon X \to X$ of a 4–manifold X with $\partial X = Y$, then*

$$\rho^*(\mathrm{Spin}^c(X)|_Y) = \mathrm{Spin}^c(X)|_Y,$$

where $\rho^*\colon \mathrm{Spin}^c(Y) \to \mathrm{Spin}^c(Y)$ is the induced map on the Spin^c–structures on the boundary.

Proof Since ρ' is a diffeomorphism, $\rho^*(\mathrm{Spin}^c(X)|_Y) = (\rho')^*(\mathrm{Spin}^c(X))|_Y = \mathrm{Spin}^c(X)|_Y$. $\square$

In order to use this proposition, take $Y = \Sigma(S^3, K)$, $X = \Sigma(B^4, D)$ and $\rho = \tilde{\sigma}\colon \Sigma(B^4, D) \to \Sigma(B^4, D)$ a lift of the strongly negative amphichiral symmetry from Corollary 2.2. In order to rule out that $\tilde{\sigma}_*(\mathrm{Spin}^c(X)|_Y) = \mathrm{Spin}^c(X)|_Y$, we will need to compute $\tilde{\sigma}^*\colon \mathrm{Spin}^c(Y) \to \mathrm{Spin}^c(Y)$ and also restrict the possible subsets $\mathrm{Spin}^c(X)|_Y \subset \mathrm{Spin}^c(Y)$ using Donaldson's theorem. Propositions 4.5 and 4.7 combined allow us to compute $\tilde{\sigma}^*\colon \mathrm{Spin}^c(Y) \to \mathrm{Spin}^c(Y)$, and Proposition 4.2 gives restrictions on $\mathrm{Spin}^c(X)|_Y \subset \mathrm{Spin}^c(Y)$. See Section 5 for an example.

We recall the following characterization of Spin^c–structures in terms of characteristic covectors which we will use throughout this section. Let X be a smooth 4–manifold which is either closed with no 2–torsion in $H_1(X)$, or constructed by attaching 2–handles to the 4–ball with ∂X a rational homology sphere. Let Q be the intersection form on X and $\mathrm{Spin}^c(X)$ be the set of Spin^c–structures of X. Then the first Chern class gives a bijection between the Spin^c–structures on X and the characteristic covectors of $H_2(X)$; see [11, Proposition 2.4.16]. More precisely,

$$\mathrm{Spin}^c(X) \cong \mathrm{Char}(H_2(X)) := \{u \in H_2(X)^* \mid u(x) \equiv Q(x, x) \pmod 2 \text{ for all } x \in H_2(X)\}.$$

In the case that $\partial X \neq \varnothing$ this identification induces a bijection

$$\mathrm{Spin}^c(\partial X) \cong \mathrm{Char}(H_2(X))/2i(H_2(X)),$$

where $i\colon H_2(X) \to H_2(X)^*$ is given by $x \mapsto Q(x, -)$ (see for example [21, Section 2.3]).

The following proposition gives restrictions on the set of Spin^c–structures on a 3–manifold which extend over a $\mathbb{Z}/2\mathbb{Z}$–homology 4–ball which it bounds. Analogous statements are discussed in [13, Section 2] and [7, Theorem 5.1].

Proposition 4.2 *Let X be a positive-definite smooth 4–manifold obtained by attaching 2–handles to the 4–ball and with ∂X a rational homology sphere Y. Suppose that Y also bounds a $\mathbb{Z}/2\mathbb{Z}$–homology 4–ball W. The inclusion map $X \to X \cup_Y -W$ induces an embedding $\iota_*\colon (H_2(X), Q) \to (\mathbb{Z}^n, \mathrm{Id})$, where Q is the intersection form of X. Choosing a basis for $H_2(X)$, ι_* is given by an $n \times n$ matrix A, and the Spin^c–structures on Y which extend over W are those of the form*

$$A^{\top}(v) \pmod{2Q} \in \mathrm{Spin}^c(Y) = \mathrm{Char}(H_2(X))/\mathrm{im}(2Q),$$

where $v \in \mathbb{Z}^n$ is any vector with all odd entries, and where elements of $\mathrm{Char}(H_2(X)) \subset \mathrm{Hom}(H_2(X), \mathbb{Z})$ are written in the dual basis.

Proof Let $Z = X \cup_Y -W$, and note that Z is positive definite (see eg [16, Proposition 7]). Hence, by Donaldson's theorem, there is an isomorphism of intersection forms $(H_2(Z)/\operatorname{Tor}, Q_Z) \cong (\mathbb{Z}^n, \operatorname{Id})$, where $n = b_2(X)$. We then have a map $\iota_* \colon (H_2(X), Q) \to (\mathbb{Z}^n, \operatorname{Id})$ induced by the inclusion $\iota \colon X \hookrightarrow Z$. Note that we may identify $\operatorname{Char}(H_2(Z))$ with $\operatorname{Spin}^c(Z)$ (since $H_1(Z)$ has no 2–torsion), and similarly $\operatorname{Char}(H_2(X))$ with $\operatorname{Spin}^c(X)$; see the discussion preceding Proposition 4.2. Applying $\operatorname{Hom}(-, \mathbb{Z})$ gives the map $\iota^* \colon H^2(Z)/\operatorname{Tor} \to H^2(X)$, which induces a map $\iota^* \colon \operatorname{Char}(H_2(Z)) \to \operatorname{Char}(H_2(X))$ on Spin^c–structures. Recall as well that the restriction $r \colon \operatorname{Spin}^c(X) \to \operatorname{Spin}^c(Y)$ is given by the quotient map

$$r \colon \operatorname{Char}(H_2(X)) \to \operatorname{Char}(H_2(X))/2i(H_2(X)),$$

where $i \colon H_2(X) \to \operatorname{Hom}(H_2(X), \mathbb{Z})$ is given by $x \mapsto Q(x, -)$. Hence the restriction map from $\operatorname{Spin}^c(Z)$ to $\operatorname{Spin}^c(Y)$ is given by $r \circ \iota^*$. We then claim that the image of $r \circ \iota^*$ is precisely the Spin^c–structures on Y which extend over W. Indeed r is surjective, so all Spin^c–structures on Y extend over X, and hence a Spin^c–structure on Y extends over W if and only if it extends over all of Z.

Combinatorially, we can compute this restriction as follows. Choose a basis for $H_2(X)$, and the dual basis for $\operatorname{Hom}(H_2(X), \mathbb{Z})$. Then ι_* is given by a matrix A, and ι^* is given by A^T. The characteristic covectors of $H_2(Z)$ are given by vectors v in $\mathbb{Z}^n$ with all odd entries. Then the image of ι^* consists of elements of all vectors of the form

$$A^\mathsf{T} v \in \operatorname{Char}(H_2(X)) = \operatorname{Spin}^c(X),$$

written in the dual basis for $\operatorname{Hom}(H_2(X), \mathbb{Z}) \supset \operatorname{Char}(H_2(X))$. The image of $r \circ \iota^*$ then consists of these vectors modulo the column space of $2Q$. □

We now turn to computing $\tilde{\sigma}^* \colon \operatorname{Spin}^c(\Sigma(S^3, K)) \to \operatorname{Spin}^c(\Sigma(S^3, K))$. To do so, begin with a strongly negative amphichiral alternating diagram for K, and let F_+ and F_- be the pair of checkerboard surfaces with F_+ and F_- positive and negative definite, respectively. Note that F_+ and F_- are exchanged by the strongly negative amphichiral symmetry.

Definition 4.3 Take S^4 as the unit sphere in $\mathbb{R}^5$. Define $\sigma_{\operatorname{swap}} \colon S^4 \to S^4$ as the involution

$$(x_1, x_2, x_3, x_4, x_5) \mapsto (x_1, -x_2, -x_3, -x_4, -x_5).$$

On the equatorial $S^3 = \{(x_1, x_2, x_3, x_4, 0) \mid x_1^2 + x_2^2 + x_3^2 + x_4^2 = 1\}$, $\sigma_{\operatorname{swap}}$ restricts to the (unique[2]) amphichiral symmetry σ with two fixed points $(\pm 1, 0, 0, 0, 0)$. Finally, note that $\sigma_{\operatorname{swap}}$ is orientation-preserving and exchanges the two hemispheres of S^4.

With respect to this involution $\sigma_{\operatorname{swap}}$, we can push F_+ and F_- equivariantly into distinct hemispheres of S^4. By Proposition 2.1 there are two lifts, $\tilde{\sigma}_{\operatorname{swap}}$ and $\tilde{\sigma}'_{\operatorname{swap}}$, of $\sigma_{\operatorname{swap}}$ to an order-4 symmetry of

---

[2]Livesay [19] proved that up to conjugation there is a unique involution on S^3 with exactly two fixed points.

$\Sigma(S^4, F_+ \cup F_-)$. We have that $\tilde{\sigma}_{\text{swap}} = \tilde{\sigma}'_{\text{swap}} \circ \tau$, where τ is the nontrivial deck transformation involution $\tau \colon \Sigma(S^4, F_+ \cup F_-) \to \Sigma(S^4, F_+ \cup F_-)$. Using Lemma 3.1, this implies that

$$-(\tilde{\sigma}_{\text{swap}})_* = (\tilde{\sigma}'_{\text{swap}})_* \colon H_2(\Sigma(S^4, F_+ \cup F_-)) \to H_2(\Sigma(S^4, F_+ \cup F_-)).$$

This immediately implies the following proposition:

Proposition 4.4 *Let $\tilde{\sigma}_{\text{swap}}$ and $\tilde{\sigma}'_{\text{swap}}$ be the two lifts of σ_{swap} to $\Sigma(S^4, F_+ \cup F_-)$. These lifts induce maps $H_2(\Sigma(B^4, F_+)) \to H_2(-\Sigma(B^4, F_-))$ which are equal to $\pm\sigma_* \colon H_1(F_+) \to H_1(F_-)$ under the identification of $H_2(\Sigma(B^4, F_\pm))$ with $H_1(F_\pm)$ from* [12, Theorem 3].

We now use $\tilde{\sigma}_{\text{swap}}$ to help us understand the action of $\tilde{\sigma}$ on Spin^c–structures.

Proposition 4.5 *Let (K, σ) be an alternating strongly negative amphichiral knot with checkerboard surfaces F_+ and F_-, and fix a lift $\tilde{\sigma} \colon \Sigma(S^3, K) \to \Sigma(S^3, K)$; see Proposition 2.4. The induced action $\tilde{\sigma}^* \colon \text{Spin}^c(\Sigma(S^3, K)) \to \text{Spin}^c(\Sigma(S^3, K))$ can be computed as follows. Let $\mathfrak{s} \in \text{Spin}^c(\Sigma(S^3, K))$, let r, r_- and r_+ be the obvious restriction maps in the noncommutative diagram*

$$\text{Spin}^c(\Sigma(S^3, K)) \overset{r}{\longleftarrow} \text{Spin}^c(\Sigma(B^4, F_+)) \overset{r_+}{\longleftarrow} \text{Spin}^c(\Sigma(S^4, F_+ \cup F_-))$$
$$\qquad\qquad (\tilde{\sigma}_{\text{swap}}^{\text{res}})^* \Big\uparrow \qquad\qquad \swarrow\ r_-$$
$$\text{Spin}^c(-\Sigma(B^4, F_-))$$

and let $\bar{\mathfrak{s}} \in \text{Spin}^c(\Sigma(S^4, F_+ \cup F_-))$ be such that $r \circ r_+(\bar{\mathfrak{s}}) = \mathfrak{s}$. Then $\tilde{\sigma}^*(\mathfrak{s}) = r \circ (\tilde{\sigma}_{\text{swap}}^{\text{res}})^* \circ r_-(\bar{\mathfrak{s}})$, where $\tilde{\sigma}_{\text{swap}}^{\text{res}} \colon \text{Spin}^c(-\Sigma(B^4, F_-)) \to \text{Spin}^c(\Sigma(B^4, F_+))$ is the map obtained by restricting $\tilde{\sigma}_{\text{swap}}$, and the lift $\tilde{\sigma}_{\text{swap}}$ is chosen to agree with $\tilde{\sigma}$ on $\Sigma(S^3, K)$.

Proof By construction, $(\tilde{\sigma}_{\text{swap}}|_{\Sigma(S^3, K)})^* = \tilde{\sigma}^*$. Hence the map

$$(\tilde{\sigma}_{\text{swap}})^* \colon \text{Spin}^c(\Sigma(S^4, F_+ \cup F_-)) \to \text{Spin}^c(\Sigma(S^4, F_+ \cup F_-))$$

restricts to $\tilde{\sigma}^* \colon \text{Spin}^c(\Sigma(S^3, K)) \to \text{Spin}^c(\Sigma(S^3, K))$. We then compute

$$\tilde{\sigma}^*(\mathfrak{s}) = \tilde{\sigma}^* \circ r \circ r_+(\bar{\mathfrak{s}}) = r \circ r_+ \circ (\tilde{\sigma}_{\text{swap}})^*(\bar{\mathfrak{s}}) = r \circ (\tilde{\sigma}_{\text{swap}}^{\text{res}})^* \circ r_-(\bar{\mathfrak{s}}),$$

where the final equality holds since $\tilde{\sigma}_{\text{swap}}$ exchanges $\Sigma(B^4, F_+)$ and $\Sigma(B^4, F_-)$ in $\Sigma(S^4, F_+ \cup F_-)$. $\square$

We now consider the complementary checkerboard graph $\mathcal{G}^c(F_+)$, which has a vertex v_i corresponding to each planar region of the knot diagram complementary to F_+ and an edge corresponding to each crossing in the knot diagram. Let γ_i be the simple loop in F_+ running once counterclockwise around the region corresponding to v_i. Applying the isomorphism $H_1(F_+) \cong H_2(\Sigma(B^4, F_+))$ from [12, Theorem 3], we get an element $v_i \in H_2(\Sigma(B^4, F_+))$. We call $\{v_i\}$ the *vertex generating set* of $H_2(\Sigma(B^4, F_+))$, and we declare the vertex generating set of $H_2(-\Sigma(B^4, F_+))$ to be $\{-v_i\}$.

Definition 4.6 Fix a strongly negative amphichiral alternating knot diagram, let $F_\pm$ be the positive and negative definite checkerboard surfaces and let $\mathcal{G}^c(F_\pm)$ be the corresponding complementary checkerboard
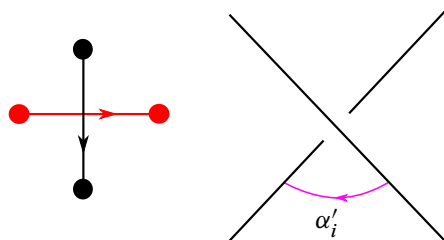
Figure 2: An oriented edge of $\mathcal{G}^c(F_+)$ in black intersecting an edge of $\mathcal{G}^c(F_-)$ in red (left). The orientation on the red edge is induced by the right-hand rule. On the right is the oriented arc α_i' induced from the oriented edge of $\mathcal{G}^c(F_+)$ in black.

graphs, embedded as dual planar graphs. The graphs $\mathcal{G}^c(F_\pm)$ are *compatibly oriented* if their edges are oriented so that intersecting dual edges satisfy the right-hand rule, as in the left of Figure 2.

Suppose $\mathcal{G}^c(F_\pm)$ are compatibly oriented, order the vertices of each of $\mathcal{G}^c(F_\pm)$ so that the strongly negative amphichiral symmetry respects the orderings and enumerate the edges of each graph so that intersecting edges have the same index; see Figure 6 for an example. We call the oriented incidence matrices $J_\pm$ for $\mathcal{G}^c(F_\pm)$ *compatible*. We use the notation J_+^* (resp. J_-^*) to denote the matrix J_+ (resp. J_-) with the last row removed. Recall that, in an oriented incidence matrix A,

$$A_{i,j} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ edge begins at the } i^{\text{th}} \text{ vertex,} \\ -1 & \text{if the } j^{\text{th}} \text{ edge terminates at the } i^{\text{th}} \text{ vertex,} \\ 0 & \text{otherwise.} \end{cases}$$

The following proposition can be used to combinatorially compute the maps r_+ and r_- from Proposition 4.5 in terms of oriented incidence matrices; see Remark 4.9.

Proposition 4.7 *Let D be an alternating knot diagram with positive and negative definite checkerboard surfaces F_+ and F_-, respectively, and let $\mathcal{G}^c(F_\pm)$ be compatibly oriented complementary checkerboard graphs (see Definition 4.6). Then there is an orthonormal basis $\{e_i\}$ of $H_2(\Sigma(S^4, F_+ \cup F_-))$ in bijection with the crossings of D for which the maps $H_2(\pm\Sigma(B^4, F_\pm)) \to H_2(\Sigma(S^4, F_+ \cup F_-))$, induced by inclusion, are given by the transposes $(J_\pm)^{\mathsf{T}}$ of the oriented incidence matrices of $\mathcal{G}^c(F_\pm)$ with respect to the vertex generating sets for $H_2(\pm\Sigma(B^4, F_\pm))$.*

Remark 4.8 The checkerboard surfaces F_+ and F_- are always nonorientable, because they are homeomorphic and at most one checkerboard surface in any diagram can be orientable.

Proof Following [12, proof of Theorem 3], $\Sigma(B^4, F_+)$ (and similarly $\Sigma(B^4, F_-)$) can be constructed as follows. Let D_1 denote the manifold obtained by cutting open B^4 along the trace of an isotopy which pushes $\text{int}(F_+)$ into $\text{int}(B^4)$. The manifold D_1 is homeomorphic to B^4 and the part exposed by the cut is given by a tubular neighborhood N_+ of F_+ in $S^3 \cong \partial D_1$. Let D_2 be another copy of D_1, and let $\iota: N_+ \to N_+$ be the involution given by reflecting each fiber. Then

$$\Sigma(B^4, F_+) = (D_1 \cup -D_2)/(x \in N_+ \subset D_1 \sim \iota(x) \in N_+ \subset D_2).$$
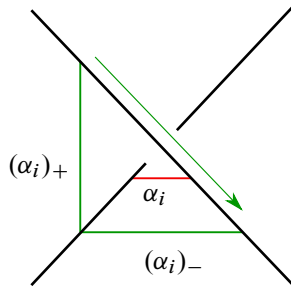
Figure 3: The arcs $(\alpha_i)_+$ and $(\alpha_i)_-$ are contained in the horizontal and vertical checkerboard surfaces, respectively. The green arrow indicates an isotopy between them in S^3. Lifting this to $\Sigma(S^3, K)$, we see that the self pairing of the sphere e_i is 1.

There is an isomorphism $\phi\colon (H_1(F_+), Q_{F_+}) \to (H_2(\Sigma(B^4, F_+)), Q_+)$, where Q_{F_+} is the Gordon–Litherland form and Q_+ is the intersection form, which is given as follows. Letting a be a 1–cycle in F_+,

$$\phi([a]) = [(\text{cone on } a \text{ in } D_1) - (\text{cone on } \iota(a) \text{ in } D_2)].$$

In their interiors, the surfaces F_+ and F_- in S^3 intersect in a collection of k arcs $\alpha_1, \ldots, \alpha_k$, one for each crossing of D. The I–subbundle of N_+ over α_i is a disk $D^2_+(\alpha_i) \subset D_1$ with boundary $\tilde{\alpha}_i$, the preimage of α_i in $\Sigma(S^3, K)$. (The disk $D^2_+(\alpha_i)$ is also the trace of α_i under the isotopy pushing $\text{int}(F_+)$ into $\text{int}(B^4)$.) Note that $D^2_+(\alpha_i)$ is properly embedded in $\Sigma(B^4, F_+)$. Similarly, there is a disk $D^2_-(\alpha_i)$ properly embedded in $\Sigma(B^4, F_-)$, and gluing these disks along $\tilde{\alpha}_i$ gives a sphere e_i in $\Sigma(S^4, F_+ \cup F_-)$.

Note that $e_1, \ldots, e_k$ are in correspondence with the edges of $\mathcal{G}^c(F_+)$ (and $\mathcal{G}^c(F_-)$). Furthermore, the orientation on an edge E_i in $\mathcal{G}^c(F_+)$ induces an orientation on the corresponding e_i as follows. First, orient the arc α_i going into the page of the knot diagram (away from the reader). Next, push the interior of α_i into the region corresponding to the terminal vertex of E_i and then out of the page of the diagram (toward the reader) so that it is disjoint from $F_+ \cup F_-$. Call the resulting arc α'_i; see Figure 2. Recall that $\Sigma(B^4, F_+) = D_1 \cup -D_2$ as an oriented manifold. Then the orientation of $\alpha'_i \subset D_1$ determines an orientation on the union of $\alpha'_i \subset D_1$ with $-\alpha'_i \subset -D_2$, which is locally isotopic within $\Sigma(S^3, K)$ to $\tilde{\alpha}_i$. This orientation on $\tilde{\alpha}_i$ then determines an orientation on $D^2_+(\alpha_i)$ as its oriented boundary, and this orientation on $D^2_+(\alpha_i)$ extends to an orientation on $e_i = D^2_+(\alpha_i) \cup D^2_-(\alpha_i)$.

We now show that $\{e_1, \ldots, e_k\}$ is an orthonormal basis for $H_2(\Sigma(S^4, F_+ \cup F_-))$. Note that

$$b_2(\Sigma(S^4, F_+ \cup F_-)) = b_2(\Sigma(B^4, F_+)) + b_2(\Sigma(B^4, F_-)),$$

since $\Sigma(S^3, K)$ is a rational homology sphere. However, $b_2(\Sigma(B^4, F_\pm)) = n_\pm - 1$, where $n_\pm$ is the number of vertices of $\mathcal{G}^c(F_\pm)$. From the Euler characteristic of the sphere of the knot diagram, we get $2 = n_+ - k + n_-$ since $\mathcal{G}^c(F_+)$ and $\mathcal{G}^c(F_-)$ are dual graphs. Hence $b_2(\Sigma(S^4, F_+ \cup F_-)) = k$. Thus it suffices to show that $e_1, \ldots, e_k$ are orthonormal. Observe that e_i and e_j are disjoint for $i \neq j$, so it is enough to show that $e_i \cdot e_i = 1$. Consider the arcs $(\alpha_i)_\pm$ shown in Figure 3, where $(\alpha_i)_\pm \subset F_\pm$ and $(\alpha_i)_+$ intersects $(\alpha_i)_-$
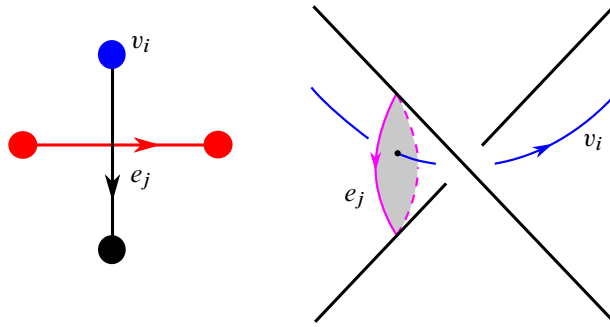
Figure 4: If $v_i \in \mathcal{G}^c(F_+)$ is the starting endpoint of an edge corresponding to e_j, then $e_j \cdot v_i = 1$. The magenta loop is the boundary of the gray disk $e_j \cap N_+$, and is oriented so that the arc coming out of the page is isotopic (keeping the endpoints on K) to α'_j (see Figure 2) in the complement of $F_+ \cup F_-$.

at a single point. Observe that the preimages $(\tilde\alpha_i)_\pm \subset \Sigma(B^4, F_\pm)$ of $(\alpha_i)_\pm$ bound disks $D^2_\pm(\alpha_i)'$ parallel to $D^2_\pm(\alpha_i)$ in $\Sigma(B^4, F_\pm)$. There is an isotopy in S^3 between $(\alpha_i)_+$ and $(\alpha_i)_-$ intersecting α_i in a single point, which induces an isotopy between $(\tilde\alpha_i)_+$ and $(\tilde\alpha_i)_-$. Gluing $D^2_+(\alpha_i)'$ to $D^2_-(\alpha_i)'$ along the (image of the) isotopy in $\Sigma(S^3, K)$ defines a push-off of e_i which has a single positive transverse intersection with e_i.

Recall that an element $v_i \in H_2(\Sigma(B^4, F_+))$ of the vertex generating set is represented by a sphere which intersects $N_+ \subset \Sigma(B^4, F_+)$ in a loop $\gamma_i \subset F_+$. By construction, $e_j \cap \Sigma(B^4, F_+)$ is the disk $D^2_+(\alpha_j)$ contained in $N_+ \subset \Sigma(B^4, F_+)$. Hence $v_i \cdot e_j$ can be computed locally in N_+. Diagrammatically (see Figure 4), we draw $\partial D_1 = S^3$ and think of N_+ as a neighborhood of $F_+ \subset S^3$. Specifically, $v_i \cdot e_j = 0$ if the edge corresponding to e_j and v_i are not incident, $v_i \cdot e_j = 1$ if the edge corresponding to e_j begins at v_i, and $v_i \cdot e_j = -1$ if the edge corresponding to e_j terminates at v_i. A similar argument applies to the vertex generating set of $H_2(-\Sigma(B^4, F_-))$. □

Remark 4.9 Proposition 4.7 combinatorially determines the maps

$$r_\pm \colon \mathrm{Spin}^c(\Sigma(S^4, F_+ \cup F_-)) \to \mathrm{Spin}^c(\pm\Sigma(B^4, F_\pm))$$

from Proposition 4.5. Specifically, the maps $r_\pm$ are given by taking the duals of

$$H_2(\pm\Sigma(B^4, F_\pm)) \to H_2(\Sigma(S^4, F_+ \cup F_-)),$$

then restricting to characteristic vectors.

We conclude the section with a proof of Theorem 1.3 from the introduction:

Proof of Theorem 1.3 Let $Y = \Sigma(S^3, K)$ and $X_\pm = \Sigma(B^4, F_\pm)$. We identify each of $H_2(X_\pm)$ with the $\mathbb{Z}$–span of $\mathrm{Vert}(\mathcal{G}^c(F_\pm))\backslash\{v_\pm\}$, where $\{v_+, v_-\}$ is the pair of σ–invariant vertices removed when defining $J^*_\pm$. Note that $X_\pm$ can be constructed by attaching 2–handles to the 4–ball (see for example the proof of Lemma 3.6 in [22]). Hence, using the dual basis for $H_2(X_\pm)^*$, we may identify

$$\mathrm{Spin}^c(X_\pm) \cong \mathrm{Char}(\mathbb{Z}^n, A_\pm) \quad \text{and} \quad \mathrm{Spin}^c(Y) \cong \mathrm{Char}(\mathbb{Z}^n, A_+)/\mathrm{im}(2A_+);$$

see the discussion before Proposition 4.2. With respect to these choices of dual bases, we may choose a lift $\tilde{\sigma}$ of σ to Y so that $\tilde{\sigma}^*_{\mathrm{swap}}\colon H_2(-X_-)^* \to H_2(X_+)^*$ is the identity matrix by Proposition 4.4; this determines the map on Spin^c–structures. Since Y is a rational homology sphere, $b_2(\Sigma(S^4, F_+ \cup F_-)) = b_2(\Sigma(B^4, F_+)) + b_2(\Sigma(B^4, F_-)) = n + n$. Using the orthonormal basis for $H_2(\Sigma(S^4, F_+ \cup F_-)) \cong \mathbb{Z}^{2n}$ from Proposition 4.7, we may identify

$$\mathrm{Spin}^c(\Sigma(S^4, F_+ \cup F_-)) \cong \{v \in \mathbb{Z}^{2n} \mid v \equiv (1, 1, \ldots, 1)^\top \pmod 2\}.$$

By Proposition 4.7 (see also Remark 4.9), the maps $r_\pm$ in Proposition 4.5 are given by $J^*_\pm$. Proposition 4.5 then shows that the map $\tilde{\sigma}^*\colon \mathrm{Spin}^c(Y) \to \mathrm{Spin}^c(Y)$ is determined by

$$\tilde{\sigma}^*[J^*_+ v] = [J^*_- v] \quad \text{for all } v \in \mathbb{Z}^{2n} \text{ with } v \equiv (1, 1, \ldots, 1)^\top \pmod 2.$$

Finally, let D be an equivariant slice disk for K. By Proposition 4.2, the set of Spin^c–structures of Y which extend over $\Sigma(B^4, D)$ is given by

$$S = \{[u] \in \mathrm{Spin}^c(Y) \mid u = A^\top v \text{ for some } v \in \mathbb{Z}^n \text{ with } v \equiv (1, 1, \ldots, 1)^\top \pmod 2\},$$

and by Corollary 2.2 there is a lift $\Sigma(B^4, D) \to \Sigma(B^4, D)$ which restricts to the lift $\tilde{\sigma}$ on Y. Hence, by Proposition 4.1, S is $\tilde{\sigma}^*$–invariant.   □

5 An alternating slice strongly negative amphichiral example

In this section we give an example of a strongly negative amphichiral knot which Theorem 1.3 shows is not equivariantly slice.

Example 5.1 Consider the slice knot $K = 12a_{1105}$ along with the strongly negative amphichiral alternating diagram shown in Figure 5. Theorem 1.3 obstructs K from being equivariantly slice. Note that Theorem 1.2 does not provide an obstruction since $\det(K) = 17^2$. Let F_+ (resp. F_-) be the positive (resp. negative) definite checkerboard surface for the knot diagram in Figure 5. In Figure 6 we draw corresponding compatibly oriented complementary checkerboard graphs $\mathcal{G}^c(F_\pm)$. The edges in each
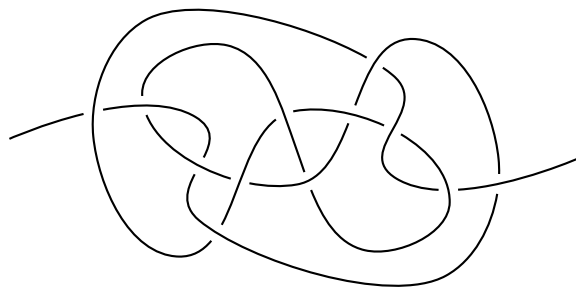
Figure 5: A strongly negative amphichiral symmetry on $12a_{1105}$. The symmetry is π–rotation within the plane of the diagram followed by a reflection across the plane of the diagram.
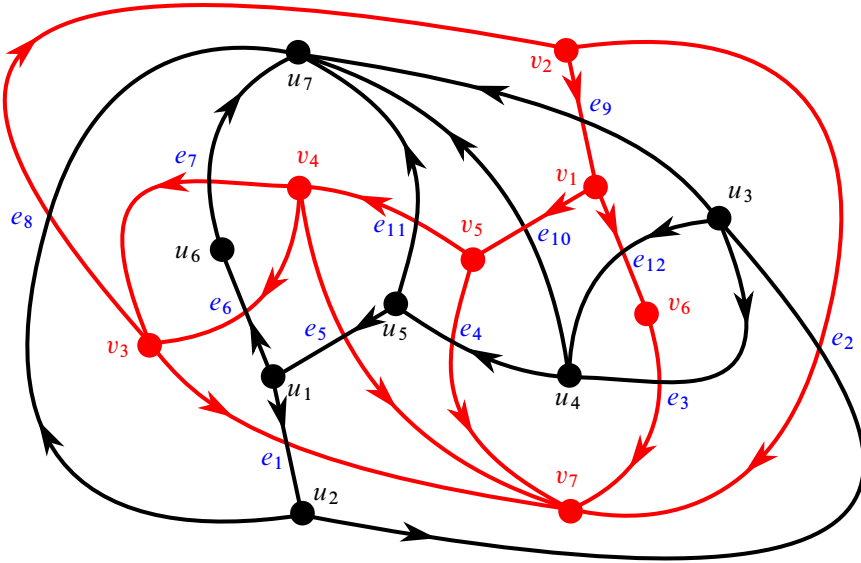
Figure 6: The pair of complementary checkerboard graphs of the alternating diagram for $12a_{1105}$ in Figure 5. They are exchanged by the strongly negative amphichiral symmetry. $\mathcal{G}^c(F_+)$ is black and $\mathcal{G}^c(F_-)$ is red. The $\{e_i\}$ correspond to crossings in the knot diagram.

graph are enumerated by the crossings e_i shown in Figure 6. Using u_7 and v_7 for the last row of the oriented incidence matrices $J_\pm$ (which we remove), we have

$$
J_+^* =
\begin{bmatrix}
1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix},
$$

$$
J_-^* =
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1
\end{bmatrix}.
$$

From these we can compute the Goeritz matrix for F_+:

$$
A_+ = J_+^*(J_+^*)^\mathsf{T} =
\begin{bmatrix}
3 & -1 & 0 & 0 & -1 & -1 \\
-1 & 3 & -1 & 0 & 0 & 0 \\
0 & -1 & 4 & -2 & 0 & 0 \\
0 & 0 & -2 & 4 & -1 & 0 \\
-1 & 0 & 0 & -1 & 3 & 0 \\
-1 & 0 & 0 & 0 & 0 & 2
\end{bmatrix}.
$$

We now combinatorially enumerate all possible lattice embeddings $A: (\mathbb{Z}^6, A_+) \to (\mathbb{Z}^6, \mathrm{Id})$, up to automorphisms of $\mathbb{Z}^6$. Using a computer program[3] we enumerate integer matrices A satisfying $A^\mathsf{T} A = A_+$, up to permutations and sign changes of the rows of A. We find two possibilities for A, which we denote by A_1 and A_2; their transposes are

$$
A_1^\mathsf{T} = \begin{bmatrix} -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & -1 & -1 \\ -1 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad A_2^\mathsf{T} = \begin{bmatrix} -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 1 & 0 \\ -1 & 0 & -1 & -1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.
$$

Neither matrix satisfies the $\tilde{\sigma}^*$–invariance condition in Theorem 1.3. We will show this for the matrix A_1; the computation for A_2 is similar. For A_1, we compute that the set

$$
S = \{[u] \in \mathrm{Spin}^c(Y) \mid u = A_1^\mathsf{T} v \text{ for some } v \in \mathbb{Z}^n \text{ with } v \equiv (1, 1, \ldots, 1)^\mathsf{T} \pmod 2\}
$$

consists of the 17 classes represented by the following vectors:

$$
\begin{bmatrix} 1 \\ 1 \\ -2 \\ -2 \\ 1 \\ -2 \end{bmatrix}, \begin{bmatrix} 3 \\ -3 \\ 2 \\ 0 \\ -1 \\ -2 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \\ 0 \\ -2 \\ -1 \\ -2 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ -4 \\ 4 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -3 \\ 4 \\ -2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -2 \\ 4 \\ -3 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ -1 \\ -2 \\ 2 \\ -3 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ -3 \\ 6 \\ -4 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 0 \\ 2 \\ -3 \\ 2 \end{bmatrix},
$$

$$
\begin{bmatrix} 1 \\ -1 \\ 0 \\ -2 \\ 3 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \\ -6 \\ 4 \\ 1 \\ -2 \end{bmatrix}, \begin{bmatrix} -3 \\ 1 \\ 2 \\ -2 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 2 \\ -4 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \\ -4 \\ 2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 4 \\ -4 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -3 \\ 3 \\ -2 \\ 0 \\ 1 \\ 2 \end{bmatrix}.
$$

We will show that this collection S of Spin^c–structures on $\Sigma(S^3, K)$ is not $\tilde{\sigma}^*$–invariant. Specifically, we will show that the Spin^c–structure represented by the second vector $\mathfrak{s} = (3, -3, 2, 0, -1, -2)^\mathsf{T}$ is mapped by $\tilde{\sigma}^*$ to a Spin^c–structure not contained in S.

Consider the vector

$$
\tilde{\mathfrak{s}} = (7, 3, 3, 3, 1, -3, -5, 1, 1, 1, 1, 1)^\mathsf{T} \in \mathbb{Z}^{12}.
$$

Multiplying, we see that $J_+^*(\tilde{\mathfrak{s}}) = \mathfrak{s}$ and $J_-^*(\tilde{\mathfrak{s}}) = (1, 3, 16, -8, 3, 2)^\mathsf{T}$. A straightforward linear algebra computation shows that $(1, 3, 16, -8, 3, 2)^\mathsf{T}$ is not equivalent modulo $2A_+$ to any of the 17 vectors in S. Hence $\tilde{\sigma}^*[J_+^*(\tilde{\mathfrak{s}})] = [J_-^*(\tilde{\mathfrak{s}})]$ is not in S. Along with a similar computation for A_2, this implies that K is not equivariantly slice, by Theorem 1.3.

---

[3]The equation $A^\mathsf{T} A = A_+$ implies that each column of A has bounded norm, so there are finitely many possibilities to check for A.

6 Heegaard Floer correction terms

In this section we give a necessary condition on the Heegaard Floer correction terms $d(\Sigma(S^3, K), \mathfrak{s})$, also known as d–invariants, for a knot to be strongly negative amphichiral. In the case of periodic knots, a similar type of condition was proved by Jabuka and Naik in [17]. As in the case of periodic knots, we first need invariance of the d–invariants.

Lemma 6.1 *Let Y be a rational homology 3–sphere with $\mathfrak{s} \in \mathrm{Spin}^c(Y)$ and $\sigma : Y \to Y$ an orientation-reversing diffeomorphism. Then*

$$d(Y, \sigma^*(\mathfrak{s})) = -d(Y, \mathfrak{s}).$$

Proof This follows directly from the diffeomorphism invariance of Heegaard Floer homology. □

Along with the following lemma, this implies our final theorem below.

Lemma 6.2 *For any knot $K \subset S^3$, the deck transformation involution τ of the double branched cover $\Sigma(S^3, K)$ acts on the set of Spin^c–structures by conjugation.*

Proof The first Chern class $c_1 : \mathrm{Spin}^c(\Sigma(S^3, K)) \to H^2(\Sigma(S^3, K); \mathbb{Z})$ is an isomorphism, since $\Sigma(S^3, K)$ is a $\mathbb{Z}/2\mathbb{Z}$ homology sphere, and by Poincaré duality we also have an isomorphism

$$H^2(\Sigma(S^3, K); \mathbb{Z}) \cong H_1(\Sigma(S^3, K); \mathbb{Z}).$$

By Lemma 3.1, τ acts as the negative of the identity on $H_1(\Sigma(S^3, K); \mathbb{Z})$, which then induces conjugation on the set of Spin^c–structures under these natural isomorphisms. □

Theorem 1.4 *Let (K, σ) be a strongly negative amphichiral knot and let $\tilde{\sigma}$ be a lift of σ to $\Sigma(S^3, K)$ (see Proposition 2.1). Then the orbits of the d–invariants of $\Sigma(S^3, K)$ under the action of $\tilde{\sigma}$ satisfy:*

- *There is exactly one orbit $\{\mathfrak{s}_0\}$ of order 1. Moreover, $d(\Sigma(S^3, K), \mathfrak{s}_0) = 0$.*
- *Other orbits $\{\mathfrak{s}, \tilde{\sigma}(\mathfrak{s}), \tilde{\sigma}^2(\mathfrak{s}), \tilde{\sigma}^3(\mathfrak{s})\}$ have order 4, and $d(\Sigma(S^3, K), \tilde{\sigma}^i(\mathfrak{s})) = (-1)^i r$ for some $r \in \mathbb{Q}$.*

Proof Since $\tilde{\sigma}$ has order 4, the $\tilde{\sigma}^*$–orbits of the Spin^c–structures will have order 1, 2 or 4. Let $\tau = \tilde{\sigma}^2$ be the deck transformation action on $\Sigma(S^3, K)$, and note that τ^* acts on the set of Spin^c–structures by conjugation by Lemma 6.2. Hence, if a Spin^c–structure is not fixed by conjugation, then it will have a $\tilde{\sigma}^*$–orbit of length 4. On the other hand, since $\Sigma(S^3, K)$ is a $\mathbb{Z}/2\mathbb{Z}$–homology sphere, there is a unique Spin^c–structure $\mathfrak{s}_0$ fixed by conjugation. Furthermore, since $|H_1(\Sigma(S^3, K))|$ is odd there are an odd number of Spin^c–structures, and hence $\mathfrak{s}_0$ has a $\tilde{\sigma}^*$–orbit of length 1. □

Example 6.3 The d–invariants of $\Sigma(S^3, 6_1)$, appropriately oriented, are

$$-\tfrac{4}{9}, -\tfrac{4}{9}, 0, 0, 0, \tfrac{2}{9}, \tfrac{2}{9}, \tfrac{8}{9}, \tfrac{8}{9}.$$

Since these are not of the form required by Theorem 1.4, 6_1 is not strongly negative amphichiral. We compare this to the strongly negative amphichiral knot 6_3, for which $\Sigma(S^3, 6_3)$ has d–invariants
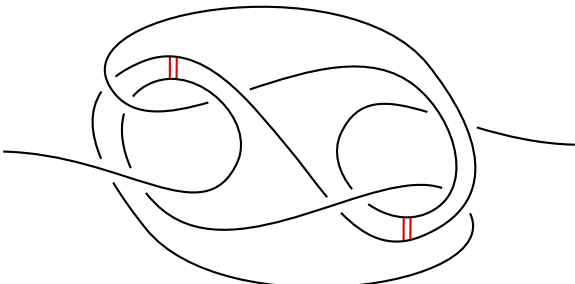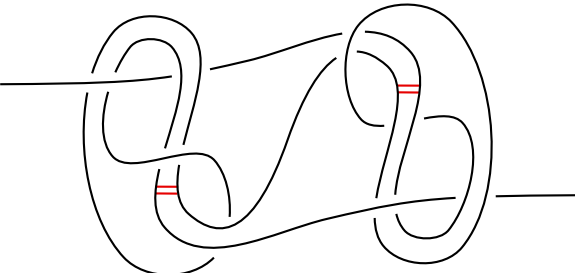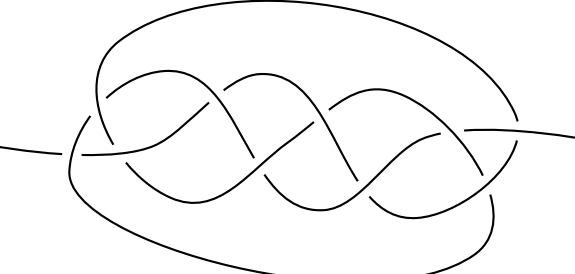
$$0, \tfrac{8}{13}, -\tfrac{8}{13}, \tfrac{8}{13}, -\tfrac{8}{13}, \tfrac{6}{13}, -\tfrac{6}{13}, \tfrac{6}{13}, -\tfrac{6}{13}, \tfrac{2}{13}, -\tfrac{2}{13}, \tfrac{2}{13}, -\tfrac{2}{13}.$$
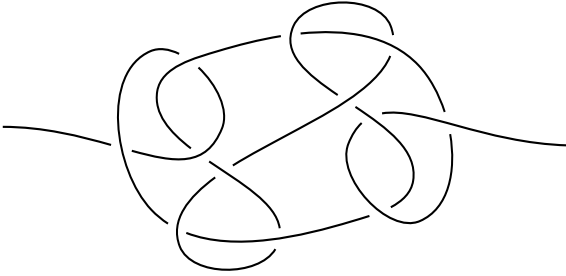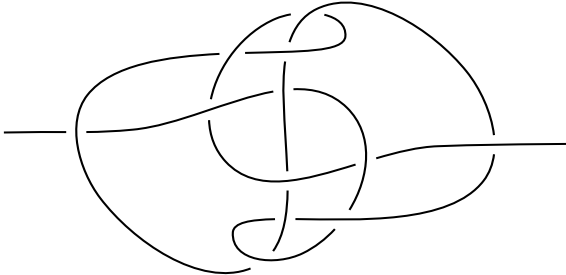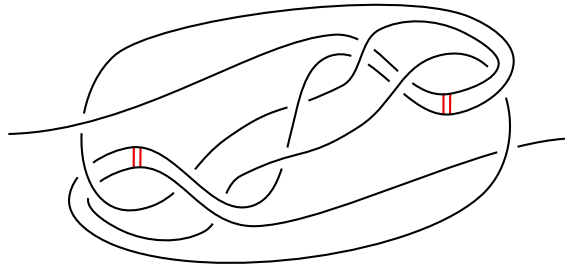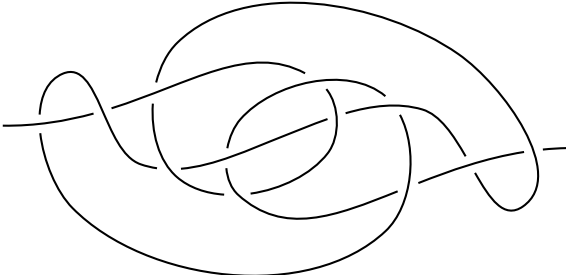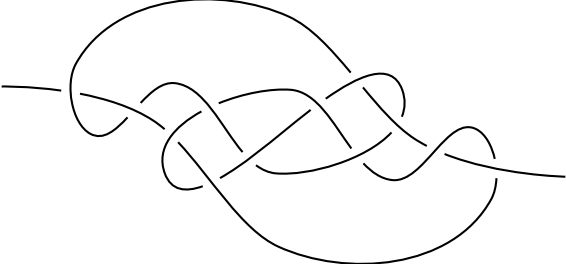
7 A table of slice strongly negative amphichiral prime knots with 12 or fewer crossings
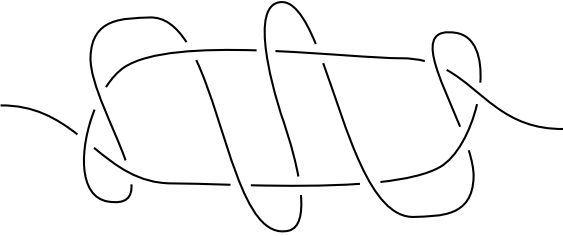
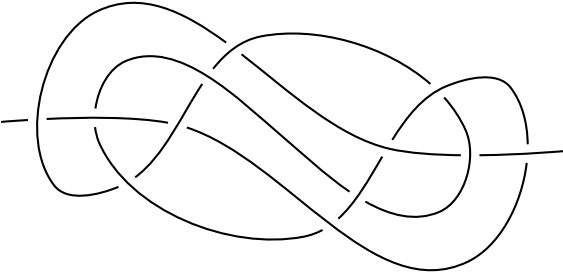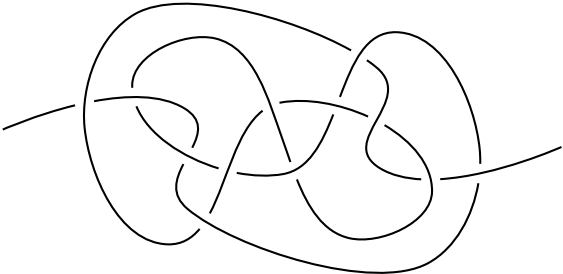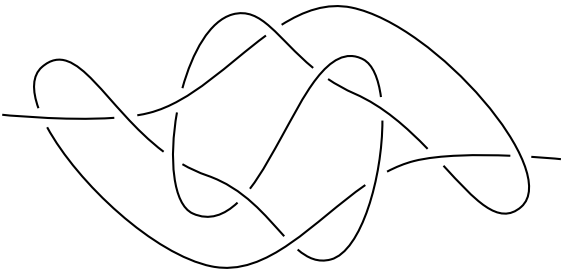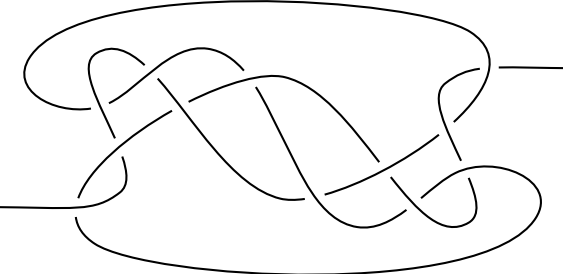We conclude with a table of all slice strongly negative amphichiral prime knots with 12 or fewer crossings. These are categorized as follows:
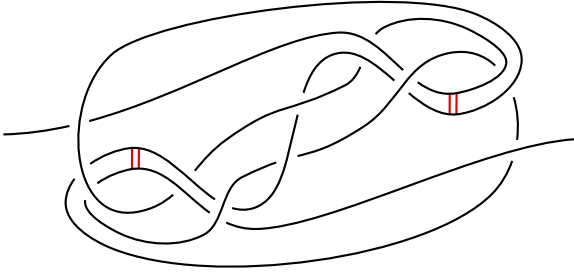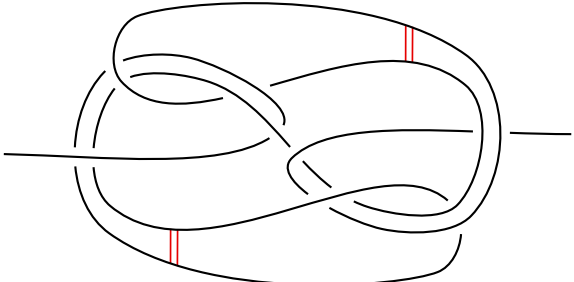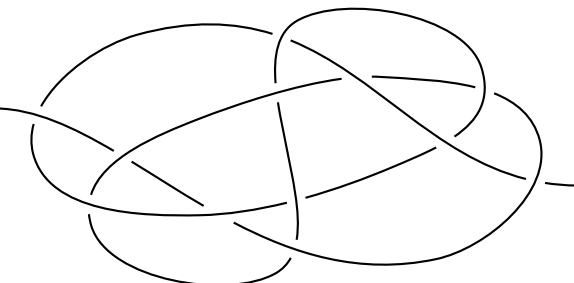
(Rib)   Knots for which we have found an equivariant ribbon diagram. We indicate this with a pair of equivariant bands (in red), which reduce the knot to a 3–component unlink.

(Det)   Knots for which Theorem 1.2 obstructs an equivariant slice disk.

($Spin^c$)   Knots for which the obstruction from Theorem 1.2 fails, but Theorem 1.3 obstructs an equivariant slice disk.

(Unk)   Knots for which we were unable to find or obstruct an equivariant slice disk.

We also include the knot determinant and whether the knot is equivariantly slice.

name	diagram	eq. slice	category	det
8_9		yes	(Rib)	5^2
10_{99}		yes	(Rib)	9^2
10_{123}		no	(Det)	11^2

name	diagram	eq. slice	category	det
$12a_{435}$		no	(Det)	15^2
$12a_{458}$		unknown	(Unk)	17^2
$12a_{819}$		yes	(Rib)	13^2
$12a_{887}$		unknown	(Unk)	17^2
$12a_{990}$		no	(Det)	15^2

name	diagram	eq. slice	category	det
$12a_{477}$		unknown	(Unk)	13^2
$12a_{1019}$		no	(Det)	19^2
$12a_{1105}$		no	(Spin^c)	17^2
$12a_{1202}$		no	(Spin^c)	13^2
$12a_{1225}$		no	(Det)	15^2

name	diagram	eq. slice	category	det
$12a_{1269}$		yes	(Rib)	13^2
$12n_{462}$		yes	(Rib)	5^2
$12n_{706}$		no	(Det)	7^2

References

[1] **F Bonahon**, *Difféotopies des espaces lenticulaires*, Topology 22 (1983) 305–314 MR Zbl

[2] **K Boyle**, **A Issa**, *Equivariant 4–genera of strongly invertible and periodic knots*, J. Topol. 15 (2022) 1635–1674 MR Zbl

[3] **G E Bredon**, *Introduction to compact transformation groups*, Pure and Applied Math. 46, Academic, New York (1972) MR Zbl

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

[5] **J C Cha**, **K H Ko**, *On equivariant slice knots*, Proc. Amer. Math. Soc. 127 (1999) 2175–2182 MR Zbl

[6] **J F Davis**, **S Naik**, *Alexander polynomials of equivariant slice and ribbon knots in S^3*, Trans. Amer. Math. Soc. 358 (2006) 2949–2964 MR Zbl

[7] **A Donald**, *Embedding Seifert manifolds in S^4*, Trans. Amer. Math. Soc. 367 (2015) 559–595 MR Zbl

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

[9]   **S Friedl**, **A N Miller**, **M Powell**, *Linking forms of amphichiral knots*, preprint (2017)   arXiv 1706.07940

[10] **L Goeritz**, *Knoten und quadratische Formen*, Math. Z. 36 (1933) 647–654   MR   Zbl

[11] **R E Gompf**, **A I Stipsicz**, 4–*Manifolds and Kirby calculus*, Graduate Studies in Math. 20, Amer. Math. Soc., Providence, RI (1999)   MR   Zbl

[12] **C M Gordon**, **R A Litherland**, *On the signature of a link*, Invent. Math. 47 (1978) 53–69   MR   Zbl

[13] **J Greene**, **S Jabuka**, *The slice-ribbon conjecture for 3–stranded pretzel knots*, Amer. J. Math. 133 (2011) 555–580   MR   Zbl

[14] **T Grove**, **S Jabuka**, *On the periodic non-orientable 4–genus a knot*, J. Knot Theory Ramifications 30 (2021) art. id. 150061   MR   Zbl

[15] **C Hodgson**, **J H Rubinstein**, *Involutions and isotopies of lens spaces*, from "Knot theory and manifolds" (D Rolfsen, editor), Lecture Notes in Math. 1144, Springer (1985) 60–96   MR   Zbl

[16] **A Issa**, **D McCoy**, *On Seifert fibered spaces bounding definite manifolds*, Pacific J. Math. 304 (2020) 463–480   MR   Zbl

[17] **S Jabuka**, **S Naik**, *Periodic knots and Heegaard Floer correction terms*, J. Eur. Math. Soc. 18 (2016) 1651–1674   MR   Zbl

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

[19] **G R Livesay**, *Involutions with two fixed points on the three-sphere*, Ann. of Math. 78 (1963) 582–593   MR   Zbl

[20] **A N Miller**, *Amphichiral knots with large 4–genus*, Bull. Lond. Math. Soc. 54 (2022) 624–634   MR   Zbl

[21] **A Némethi**, *On the Ozsváth–Szabó invariant of negative definite plumbed 3–manifolds*, Geom. Topol. 9 (2005) 991–1042   MR   Zbl

[22] **P Ozsváth**, **Z Szabó**, *On the Heegaard Floer homology of branched double-covers*, Adv. Math. 194 (2005) 1–33   MR   Zbl

[23] **M Sakuma**, *On strongly invertible knots*, from "Algebraic and topological theories" (M Nagata, S Araki, A Hattori, N Iwahori, editors), Kinokuniya, Tokyo (1986) 176–196   MR   Zbl

[24] **A Stoimenow**, *Square numbers, spanning trees and invariants of achiral knots*, Comm. Anal. Geom. 13 (2005) 591–631   MR   Zbl

Department of Mathematics, University of British Columbia
Vancouver, BC, Canada

Department of Mathematics, University of British Columbia
Vancouver, BC, Canada

kboyle@math.ubc.ca,   aissa@math.ubc.ca

Computing the Morava K–theory of real Grassmannians using chromatic fixed point theory

NICHOLAS J KUHN

CHRISTOPHER J R LLOYD

We study $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$, the 2–local Morava K–theories of the real Grassmannians, about which very little has been previously computed. We conjecture that the Atiyah–Hirzebruch spectral sequences computing these all collapse after the first possible nonzero differential $d_{2^{n+1}-1}$, and give much evidence that this is the case.

We use a novel method to show that higher differentials can't occur: we get a lower bound on the size of $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$ by constructing a C_4–action on our Grassmannians and then applying the chromatic fixed point theory of the authors' previous paper. In essence, we bound the size of $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$ by computing $K(n-1)^*(\mathrm{Gr}_d(\mathbb{R}^m)^{C_4})$.

Meanwhile, the size of $E_{2^{n+1}}$ is given by Q_n–homology, where Q_n is Milnor's n^{th} primitive mod 2 cohomology operation. Whenever we are able to calculate this Q_n–homology, we have found that the size of $E_{2^{n+1}}$ agrees with our lower bound for the size of $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$. We have two general families where we prove this: $m \leq 2^{n+1}$ and all d, and $d = 2$ and all m and n. Computer calculations have allowed us to check many other examples with larger values of d.

1 Introduction

Let $\mathrm{Gr}_d(\mathbb{R}^m)$ be the real Grassmannian of k–planes in $\mathbb{R}^m$, a much studied compact manifold of dimension $d(m-d)$ admitting the structure of a CW complex with $\binom{m}{d}$ "Schubert cells".

Much is known about the ordinary cohomology of these spaces:

(1) $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$ is generated by Stiefel–Whitney classes satisfying standard relations. It has total dimension $\binom{m}{d}$.

(2) $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Q})$ is generated by Pontryagin classes, along with, in some cases, an odd-dimensional class. For fixed d, and $\epsilon = 0$ or 1, the total dimension of $H^*(\mathrm{Gr}_d(\mathbb{R}^{2-\epsilon+2l}); \mathbb{Q})$ is polynomial of degree $\lfloor d/2 \rfloor$ as a function of $l \geq 0$.

(3) If m is even, then $\mathrm{Gr}_d(\mathbb{R}^m)$ is oriented. Furthermore, the inclusion $\mathrm{Gr}_d(\mathbb{R}^{m-1}) \hookrightarrow \mathrm{Gr}_d(\mathbb{R}^m)$ induces an epimorphism in rational cohomology.

(4) Nontrivial torsion in $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z})$ has order 2. The mod 2 Bockstein spectral sequence (BSS) collapses after the first differential. Equivalently, the mod 2 Adams spectral sequence (ASS) converging to $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z})$ collapses at E_2.

Much less is known about other cohomology theories applied to these Grassmannians. In this paper, we study $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$ for $n \geq 1$. Here $K(n)^*(X)$ denotes the 2–local n^{th} Morava K–theory of a space X, a graded vector space over the graded field $K(n)_* = \mathbb{Z}/2[v_n^{\pm}]$ with $|v_n| = 2^{n+1} - 2$. We let $k(n)$ denote the connective cover of $K(n)$: $k(n)_* = \mathbb{Z}/2[v_n]$.

Viewing $H\mathbb{Q}$ as $K(0)$ and $H\mathbb{Z}$ as $k(0)$, our discovery is that analogues of statements (2)–(4) above appear to hold for all n, with the Atiyah–Hirzebruch spectral sequence (AHSS) replacing the Bockstein spectral sequence in statement (4). Furthermore, the analogue of statement (1) holds through a much bigger range than one would expect from dimension considerations.

In the next two subsections, we describe our main results.

1.1 Results proved using chromatic fixed point theory

Given a finite complex X and $n \geq 0$, we let $k_n(X) = \dim_{K(n)_*} K(n)^*(X)$.

Theorem 1.1 *If $m \leq 2^{n+1}$, then $k_n(\mathrm{Gr}_d(\mathbb{R}^m)) = \binom{m}{d}$. Thus, in this range, the AHSS converging to $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$ collapses at E_2.*

We note that this collapsing range is surprisingly large, as dimension considerations just imply collapsing if $d(m-d) < 2^{n+1}$.

For larger m, we have the following lower bound.

Theorem 1.2 *Let $m = 2^{n+1} - \epsilon + 2l$ with $\epsilon = 0$ or 1, and $l \geq 0$. Then*

$$k_n(\mathrm{Gr}_d(\mathbb{R}^m)) \geq \sum_{i=0}^{\lfloor d/2 \rfloor} \binom{2^{n+1} - \epsilon}{d - 2i}\binom{l}{i}.$$

Conjecture 1.3 *Equality always holds in this last theorem.*

The biggest novelty of this paper is our method for proving Theorems 1.1 and 1.2: we make use of chromatic fixed point theory to prove these nonequivariant results.

The blue shift theorem of Barthel, Hausmann, Naumann, Nikolaus, Noel and Stapleton [2] says that if C is a finite cyclic p–group and X is a finite C–CW complex, then

$$\widetilde{K}(n)^*(X) = 0 \implies \widetilde{K}(n-1)^*(X^C) = 0;$$

see also Balderrama and the first author [1]. In [8], we upgraded this. Specialized to cyclic groups, [8, Theorem 2.17] says the following.

Theorem 1.4 *If C is a finite cyclic p–group, and X is a finite C–CW complex, then*

$$k_n(X) \geq k_{n-1}(X^C).$$

Note that, in these statements, $K(n)_*$ means Morava K–theory at the prime p.

As $\binom{m}{d}$ is an evident upper bound for $k_n(\mathrm{Gr}_d(\mathbb{R}^m))$, to prove Theorem 1.1, it suffices to show that $k_n(\mathrm{Gr}_d(\mathbb{R}^m)) \geq \binom{m}{d}$ in the stated range. Using Theorem 1.4, we will show this by induction on n using a C_2–action on $\mathrm{Gr}_d(\mathbb{R}^m)$ induced by an m–dimensional real representation of $C = C_2$.

We will similarly prove Theorem 1.2 for $n \geq 1$ by using a C_4–action on $\mathrm{Gr}_d(\mathbb{R}^m)$ induced by an m–dimensional real representation of $C = C_4$.

In both cases, it will be quite easy to compute $k_{n-1}(\mathrm{Gr}_d(\mathbb{R}^m)^C)$.

Details of this will be in Section 2.

1.2 Results about the Q_n–homology of the Grassmannians

Conjecture 1.3 follows from a conjectural calculation that only involves $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$, viewed as a module over the Steenrod algebra.

Let Q_n for $n = 0, 1, 2, \ldots$ be the Milnor primitives: the elements in the mod 2 Steenrod algebra recursively defined by $Q_0 = \mathrm{Sq}^1$, and $Q_n = [Q_{n-1}, \mathrm{Sq}^{2^n}]$. These satisfy $Q_n^2 = 0$, and we let $k_{Q_n}(X)$ denote the total dimension of the Q_n–homology of X,

$$H^*(X; Q_n) = \frac{Z^*(X; Q_n)}{B^*(X; Q_n)},$$

where

$$Z^*(X; Q_n) = \ker\{Q_n \colon H^*(X; \mathbb{Z}/2) \to H^{*+2^{n+1}-1}(X; \mathbb{Z}/2)\},$$

$$B^*(X; Q_n) = \mathrm{im}\{Q_n \colon H^{*-2^{n+1}+1}(X; \mathbb{Z}/2) \to H^*(X; \mathbb{Z}/2)\}.$$

As will be reviewed in Section 3.1, the first differential in the AHSS converging to $K(n)^*(X)$ is $d_{2^{n+1}-1}$, with formula

$$d_{2^{n+1}-1}(x) = Q_n(x)v_n$$

for all $x \in E_2^{*,0}(X) = H^*(X; \mathbb{Z}/2)$. This makes it not hard to check the next lemma.

Lemma 1.5 *If X is a finite complex, $k_{Q_n}(X) \geq k_n(X)$ is always true, and the following are equivalent:*

(a) $k_{Q_n}(X) = k_n(X)$;

(b) *the AHSS, when $n \geq 1$, or the BSS, when $n = 0$, computing $K(n)^*(X)$ collapses at $E_{2^{n+1}}$;*

(c) *the ASS computing $k(n)^*(X)$ collapses at E_2.*

We apply this to our situation. First, Theorem 1.1 has the following nontrivial algebraic consequence.

Corollary 1.6 *If $m \leq 2^{n+1}$, then Q_n acts trivially on $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$.*

For an algebraic proof of this result using the methods of Section 3.5, see the second author's thesis [10, page 75].

For $m > 2^{n+1}$, we believe the following is true.

Conjecture 1.7 *Let $m = 2^{n+1} - \epsilon + 2l$ with $\epsilon = 0$ or 1, and $l \geq 0$. Then*

$$k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^m)) = \sum_{i=0}^{\lfloor d/2 \rfloor} \binom{2^{n+1}-\epsilon}{d-2i}\binom{l}{i}.$$

Comparison with Theorem 1.2 shows that when Conjecture 1.7 is true, one can conclude

- $k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^m)) = k_n(\mathrm{Gr}_d(\mathbb{R}^m))$, and Conjecture 1.3 is true;
- the AHSS computing $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$ collapses at $E_{2^{n+1}}$;
- the ASS computing $k(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$ collapses at E_2;
- $k_n(\mathrm{Gr}_d(\mathbb{R}^{2^{n+1}-\epsilon+2l}))$ is polynomial of degree $\lfloor d/2 \rfloor$ as a function of l.

Known rational calculations imply that the conjecture is true when $n = 0$. It is also easy to show that the conjecture is true when $d = 1$, and one calculates

$$k_n(\mathrm{Gr}_1(\mathbb{R}^m)) = \begin{cases} m & \text{if } 1 \leq m \leq 2^{n+1}, \\ 2^{n+1} - \epsilon & \text{if } m = 2^{n+1} - \epsilon + 2l. \end{cases}$$

With much more work we prove the following.

Theorem 1.8 *Conjecture 1.7 is true when $d = 2$. Thus the Atiyah–Hirzebruch spectral sequence computing $K(n)^*(\mathrm{Gr}_2(\mathbb{R}^m))$ collapses at $E_{2^{n+1}}$, the Adams spectral sequence computing $k(n)^*(\mathrm{Gr}_2(\mathbb{R}^m))$ collapses at E_2, and we have the calculation*

$$k_n(\mathrm{Gr}_2(\mathbb{R}^m)) = \begin{cases} \binom{m}{2} & \text{if } 2 \leq m \leq 2^{n+1}, \\ \binom{2^{n+1}-\epsilon}{2} + l & \text{if } m = 2^{n+1} - \epsilon + 2l. \end{cases}$$

We are firm believers in our conjectures. For more evidence, the second author has made extensive computer calculations verifying Conjecture 1.7 in hundreds more cases with larger values of d; see the tables in the appendix.

For $d \geq 2$, computing the size of $H^*(\mathrm{Gr}_d(\mathbb{R}^m); Q_n)$ seems tricky. We have organized our efforts by studying how these numbers change as m is increased as follows.

Let $C_d(\mathbb{R}^m)$ denote the cofiber of the inclusion $\mathrm{Gr}_d(\mathbb{R}^{m-1}) \to \mathrm{Gr}_d(\mathbb{R}^m)$, so there is a cofiber sequence $\mathrm{Gr}_d(\mathbb{R}^{m-1}) \xrightarrow{i} \mathrm{Gr}_d(\mathbb{R}^m) \xrightarrow{p} C_d(\mathbb{R}^m)$. In Section 3.3, $C_d(\mathbb{R}^m)$ is identified as the Thom space of the canonical normal bundle over $\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})$, and in Section 3.4 we study the Q_n–module

$\tilde{H}^*(C_d(\mathbb{R}^m); \mathbb{Z}/2)$, viewed as $H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}); \mathbb{Z}/2)$ equipped with an explicit twisted Q_n–action. One has an induced short exact sequence of modules over the Steenrod algebra

$$0 \to \tilde{H}^*(C_d(\mathbb{R}^m); \mathbb{Z}/2) \xrightarrow{p^*} H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2) \xrightarrow{i^*} H^*(\mathrm{Gr}_d(\mathbb{R}^{m-1}); \mathbb{Z}/2) \to 0,$$

inducing a long exact sequence on Q_n–homology.

When m is even, we see much orderly behavior.

Theorem 1.9 *Let m be even.*

(a) $H^{d(m-d)}(\mathrm{Gr}_d(\mathbb{R}^m); Q_n) \simeq \mathbb{Z}/2$, *ie the nonzero top-dimensional cohomology class is not in the image of Q_n for all n.*

(b) *The chain complex $\left(\tilde{H}^*(C_d(\mathbb{R}^m); Q_n)\right)$ is dual to the chain complex*

$$\left(H^{d(m-d)-*}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}); Q_n)\right).$$

(c) *If [Conjecture 1.7](#) is true for $(n, d, m-1)$ and $(n, d-1, m-1)$, then it is true for (n, d, m). Furthermore, $\mathrm{Gr}_d(\mathbb{R}^m)$ will then be $k(n)$–oriented, and the cofiber sequence above will induce short exact sequences*

$$0 \to \tilde{H}^*(C_d(\mathbb{R}^m); Q_n) \xrightarrow{p^*} H^*(\mathrm{Gr}_d(\mathbb{R}^m); Q_n) \xrightarrow{i^*} H^*(\mathrm{Gr}_d(\mathbb{R}^{m-1}); Q_n) \to 0,$$
$$0 \to \tilde{K}(n)^*(C_d(\mathbb{R}^m)) \xrightarrow{p^*} K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m)) \xrightarrow{i^*} K(n)^*(\mathrm{Gr}_d(\mathbb{R}^{m-1})) \to 0.$$

We prove [Theorem 1.9](#) in [Section 4](#). We make use of the additive basis $\{s_\lambda\}$ dual to the classical Schubert cells. Here λ runs through partitions having at most d parts, each no bigger than $m-d$. In [9], Cristian Lennart gave a combinatorial formula for $Q_n(s_\lambda)$, and we use this to prove (a). Duality statement (b) follows quite formally from (a), and (c) follows easily from (b).

When m is odd, the analogues of statements (a) and (b) are false, and, for $d \geq 3$, the full behavior of the connecting map in the Q_n–homology long exact sequence,

$$\delta: H^*(\mathrm{Gr}_d(\mathbb{R}^{m-1}); Q_n) \to \tilde{H}^{*+2^{n+1}-1}(C_d(\mathbb{R}^m); Q_n),$$

is as yet unclear to the authors. In [Section 6](#), we will prove analogues of Theorems [1.1](#) and [1.2](#) for $C_d(\mathbb{R}^m)$, and then speculate on behavior of δ that would be compatible with all of our computations.

However, when $d = 2$, we have the following result.

Theorem 1.10 *Let $m > 2^{n+1}$ be odd. Then $k_{Q_n}(C_2(\mathbb{R}^m)) = 2^{n+1} - 2$ and the map*

$$\tilde{H}^*(C_2(\mathbb{R}^m); Q_n) \xrightarrow{p^*} H^*(\mathrm{Gr}_2(\mathbb{R}^m); Q_n)$$

is zero, so there is a short exact sequence

$$0 \to H^*(\mathrm{Gr}_2(\mathbb{R}^m); Q_n) \xrightarrow{i^*} H^*(\mathrm{Gr}_2(\mathbb{R}^{m-1}); Q_n) \xrightarrow{\delta} \tilde{H}^{*+2^{n+1}-1}(C_2(\mathbb{R}^m); Q_n) \to 0.$$

From this, Theorem 1.8 quickly follows and one can deduce that, in this case, there is a short exact sequence

$$0 \to K(n)^*(\mathrm{Gr}_2(\mathbb{R}^m)) \xrightarrow{i^*} K(n)^*(\mathrm{Gr}_2(\mathbb{R}^{m-1})) \xrightarrow{\delta} \widetilde{K}(n)^{*+1}(C_2(\mathbb{R}^m)) \to 0.$$

We prove Theorem 1.10 in Section 5. The tools we use are very different from those used in proving Theorem 1.9: we work with the classical presentation of $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$ as a ring of Stiefel–Whitney classes.

1.3 Comparison with other work

When comparing our work to what has come before, the first thing to say is that the outcome of our calculations — though not the methods — are in line with the classical calculations first made by C Ehresmann in 1937 [3]. He determined the additive structure of both $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$ and $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Q})$. He also showed that all the torsion in $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z})$ was of order 2; in modern terms this is equivalent to showing that the Bockstein spectral sequence computing $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z})$ collapses after the first nonzero differential given by $Q_0 = \mathrm{Sq}^1 = \beta$.

Calculating the Morava K–theories of $\mathrm{Gr}_d(\mathbb{R}^\infty) = BO(d)$ was done first by Kono and Yagita [7], and then, with a simpler proof, by Kitchloo and Wilson [6]. Again, the AHSS computing $K(n)^*(BO(d))$ collapses after the first nonzero differential, but the collapsing is for an elementary reason: $H^*(BO(d); Q_n)$ is concentrated in even degrees. Indeed, one quickly learns that the complexification map $BO(d) \to BU(d)$ induces an epimorphism $K(n)^*(BU(d)) \to K(n)^*(BO(d))$, so $K(n)^*(BO(d))$ is generated by Chern classes $c_1, \dots, c_d$.

An equivalent statement is that $H^*(BO(d); Q_n)$ is generated by the classes $w_1^2, \dots, w_d^2$. These will still be permanent classes in the AHSS converging to $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$, but now we have odd-dimensional classes as well, with the number of these seemingly growing as d and m grow.

Finally, we point out that we do not attempt to describe $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$ as a $K(n)^*$–algebra. Our results do tell us something about this, however. In the situation of Theorem 1.1, the known algebra $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2) \otimes K(n)^*$ will be an associated graded. Similarly, whenever our conjecture is valid, $H^*(\mathrm{Gr}_d(\mathbb{R}^m); Q_n) \otimes K(n)^*$ would be an associated graded of the $K(n)^*$–algebra $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$. What is still needed, and might be necessary to prove our conjectural collapsing in general, are sensible constructions of classes in odd degrees.

Acknowledgements

2 The proofs of Theorems 1.1 and 1.2

In this section we prove Theorems 1.1 and 1.2 by using our chromatic fixed point theorem Theorem 1.4.

2.1 A fixed point formula

Let G be a finite group, and let V be an m–dimensional real representation of G. Then $\mathrm{Gr}_d(V)$, the space of d–planes in V, is a model for $\mathrm{Gr}_d(\mathbb{R}^m)$ with an evident G–action. Here we describe $\mathrm{Gr}_d(V)^G$, its space of G–fixed points.

To state this, we need some notation. Let $V_1, \dots, V_k$ be the irreducible real representations of G, let $r_i = \dim_{\mathbb{R}} V_i$, and let $\mathbb{D}_i = \mathrm{End}_{\mathbb{R}[G]}(V_i, V_i)$. Each of the endomorphism algebras $\mathbb{D}_i$ will be a finite-dimensional real division algebra, and thus isomorphic to $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$, and $\dim_{\mathbb{R}} \mathbb{D}_i$ will divide r_i.

Proposition 2.1 *If $V = V_1^{m_1} \oplus \cdots \oplus V_k^{m_k}$, then there is a homeomorphism*

$$\mathrm{Gr}_d(V)^G = \bigsqcup_{j_1 r_1 + \cdots + j_k r_k = d} \mathrm{Gr}_{j_1}(\mathbb{D}_1^{m_1}) \times \cdots \times \mathrm{Gr}_{j_k}(\mathbb{D}_k^{m_k}).$$

Proof The fixed point space $\mathrm{Gr}_d(V)^G$ will be the space of sub–G–modules $W < V$ of real dimension d. Such a G–module W will decompose canonically as $W = W_1 \oplus \cdots \oplus W_k$, with $W_i < V_i^{m_i}$. If $d_i = \dim_{\mathbb{R}} W_i$, then $d_1 + \cdots + d_k = d$. Thus we have a decomposition

$$\mathrm{Gr}_d(V)^G = \bigsqcup_{d_1 + \cdots + d_k = d} \mathrm{Gr}_{d_1}(V_1^{m_1})^G \times \mathrm{Gr}_{d_2}(V_2^{m_2})^G \times \cdots \times \mathrm{Gr}_{d_k}(V_k^{m_k}).$$

A submodule W_i of $V_i^{m_i}$ must be isomorphic to V_i^j for some j; thus $\mathrm{Gr}_{d_i}(V_i^{m_1})^G$ will be empty unless $d_i = j_i r_i$ for some j_i.

Finally, using that $\mathrm{Hom}_{\mathbb{R}[G]}(V_i^{j_i}, V_i^{m_i}) = \mathrm{Hom}_{\mathbb{D}}(\mathbb{D}^{j_i}, \mathbb{D}^{m_i})$, one deduces that the submodules of $V_i^{m_i}$ isomorphic to $V_i^{j_i}$ correspond to the $\mathbb{D}$–subspaces of $\mathbb{D}^{m_i}$ of dimension j_i over $\mathbb{D}$. Thus there is a homeomorphism

$$\mathrm{Gr}_{j_i r_i}(V_i^{m_i})^G = \mathrm{Gr}_{j_i}(\mathbb{D}_i^{m_i}). \qquad \square$$

Corollary 2.2 *If $V = V_1^{m_1} \oplus \cdots \oplus V_k^{m_k}$, then, for any n,*

$$k_n(\mathrm{Gr}_d(V)^G) = \sum_{j_1 r_1 + \cdots + j_k r_k = d} k_n(\mathrm{Gr}_{j_1}(\mathbb{D}_1^{m_1})) \cdots k_n(\mathrm{Gr}_{j_k}(\mathbb{D}_k^{m_k})).$$

Proof A consequence of the Künneth theorem for $K(n)_*$ is that $k_n(X \times Y) = k_n(X)k_n(Y)$. Thus the corollary follows from the proposition. $\qquad \square$

Remark 2.3 If $\mathbb{D} = \mathbb{C}$ or $\mathbb{H}$, then $\mathrm{Gr}_d(\mathbb{D}^m)$ has a CW structure with $\binom{m}{d}$ cells that are all even-dimensional, and thus $k_n(\mathrm{Gr}_d(\mathbb{D}^m)) = \binom{m}{d}$ for all n.

2.2 Proof of Theorem 1.1

Theorem 1.1 says that if $m \leq 2^{n+1}$ then $k_n(\mathrm{Gr}_d(\mathbb{R}^m)) = \binom{m}{d}$. Using Theorem 1.4 and Proposition 2.1, we prove this by induction on n.

The $n = 0$ case of the theorem is easy to check, as

$$\mathrm{Gr}_d(\mathbb{R}^0) = \begin{cases} * & \text{if } d = 0, \\ \varnothing & \text{otherwise,} \end{cases} \quad \text{and} \quad \mathrm{Gr}_d(\mathbb{R}^1) = \begin{cases} * & \text{if } d = 0, 1, \\ \varnothing & \text{otherwise.} \end{cases}$$

For the inductive step, assume that if $p \leq 2^n$ then $k_{n-1}(\mathrm{Gr}_d(\mathbb{R}^p)) = \binom{p}{d}$.

Let $m \leq 2^{n+1}$. As it is clear that $k_n(\mathrm{Gr}_d(\mathbb{R}^m)) \leq \binom{m}{d}$, our goal is to show that $k_n(\mathrm{Gr}_d(\mathbb{R}^m)) \geq \binom{m}{d}$.

Let C_2 be the cyclic group of order 2. To get our needed lower bound, our strategy will be to make $\mathbb{R}^m$ into a C_2–module, and then apply Theorem 1.4.

The group C_2 has two irreducible 1–dimensional real representations; call them L_1 and L_2. Since $m \leq 2^{n+1}$, we can write m as $m = p + q$ with both $p \leq 2^n$ and $q \leq 2^n$. Now let $V = L_1^p \oplus L_2^q$, an m–dimensional real representation of C_2.

Applying Proposition 2.1, we see that

$$\mathrm{Gr}_d(V)^{C_2} = \bigsqcup_{i+j=d} \mathrm{Gr}_i(\mathbb{R}^p) \times \mathrm{Gr}_j(\mathbb{R}^q).$$

Applying Theorem 1.4 to this, we learn that

$$\begin{aligned} k_n(\mathrm{Gr}_d(\mathbb{R}^m)) &\geq \sum_{i+j=d} k_{n-1}(\mathrm{Gr}_i(\mathbb{R}^p)) k_{n-1}(\mathrm{Gr}_j(\mathbb{R}^q)) \\ &= \sum_{i+j=d} \binom{p}{i}\binom{q}{j} \quad \text{(by inductive hypothesis)} \\ &= \binom{m}{d}. \end{aligned}$$

Remark 2.4 The same inductive proof can be used to prove the classical result that

$$\dim_{\mathbb{Z}/2} H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2) = \binom{m}{d}$$

for all m and d, with our chromatic fixed point theorem Theorem 1.4 replaced by the classical theorem of Ed Floyd [4, Theorem 4.4]: if the cyclic group C_p acts on a finite CW complex X, then $\dim_{\mathbb{Z}/p} H^*(X; \mathbb{Z}/p) \geq \dim_{\mathbb{Z}/p} H^*(X^{C_p}; \mathbb{Z}/p)$. It would be interesting to know if this argument was known to Floyd, or others, like Bob Stong, who regularly worked with these sorts of group actions.

2.3 Proof of Theorem 1.2

The strategy of the proof of Theorem 1.2 is the same as the proof in the last subsection: we get a lower bound on $k_n(\mathrm{Gr}_d(\mathbb{R}^m))$ by letting a cyclic 2–group act on $\mathbb{R}^m$ and applying Theorem 1.4.

In this case, the representation theory of C_2 is not rich enough to give us a big enough lower bound, but a well chosen real representation of the group C_4 of order 4 works better. Curiously, in our calculation of k_{n-1} of the resulting fixed point space, we are able to use our already proven Theorem 1.1, so the proof is not by induction, but more direct.

The group C_4 has three irreducible real representations: L_1 and L_2 of dimension 1, and R of real dimension 2. Note that $\mathrm{End}_{\mathbb{R}[C_4]}(R) \simeq \mathbb{C}$.

Now let $m = 2^{n+1} - \epsilon + 2l$ with $\epsilon = 0$ or 1, and $l \geq 0$. We define an m–dimensional real representation V of C_4 by $V = L_1^{2^n} \oplus L_2^{2^n - \epsilon} \oplus R^l$.

Applying Proposition 2.1, we see that

$$\mathrm{Gr}_d(V)^{C_4} = \bigsqcup_{j+k+2i=d} \mathrm{Gr}_j(\mathbb{R}^{2^n}) \times \mathrm{Gr}_k(\mathbb{R}^{2^n - \epsilon}) \times \mathrm{Gr}_i(\mathbb{C}^l).$$

Applying Theorem 1.4 to this, we learn that

$$\begin{aligned}
k_n(\mathrm{Gr}_d(\mathbb{R}^m)) &\geq \sum_{j+k+2i=d} k_{n-1}(\mathrm{Gr}_j(\mathbb{R}^{2^n}))k_{n-1}(\mathrm{Gr}_k(\mathbb{R}^{2^n-\epsilon}))k_{n-1}(\mathrm{Gr}_i(\mathbb{C}^l)) \\
&= \sum_{j+k+2i=d} \binom{2^n}{j}\binom{2^n-\epsilon}{k}\binom{l}{i} \quad \text{(using Theorem 1.1)} \\
&= \sum_i \left[\sum_{j+k=d-2i} \binom{2^n}{j}\binom{2^n-\epsilon}{k} \right] \binom{l}{i} \\
&= \sum_i \binom{2^{n+1}-\epsilon}{d-2i}\binom{l}{i}.
\end{aligned}$$

3 The Q_n homology of $\mathrm{Gr}_d(\mathbb{R}^m)$: background material

3.1 The AHSS and the ASS for Morava K–theory

Let $n \geq 1$. We recall the structure of the AHSS converging to $K(n)^*(X)$ (as always, in this paper, with $p = 2$). It is a spectral sequence of graded $K(n)^* = \mathbb{Z}/2[v_n^\pm]$ algebras with

$$E_2^{*,\star}(X) = H^*(X; K(n)^\star) = H^*(X; \mathbb{Z}/2)[v_n^\pm].$$

Here v_n has cohomological degree $2 - 2^{n+1}$.

Sparseness of the rows implies that the differential d_r will be zero unless $r = s(2^{n+1} - 2) + 1$ for some s. The first possible nonzero differential, $d_{2^{n+1}-1}$, satisfies the following formula [15]: given $x \in E_2^{*,0}(X) = H^*(X; \mathbb{Z}/2)$,

$$d_{2^{n+1}-1}(x) = Q_n(x)v_n.$$

It follows that $E_{2^{n+1}}(X) \simeq H^*(X; Q_n)[v_n^{\pm}]$, and so the dimension of $E_{2^{n+1}}(X)$ as a $K(n)^*$–vector space will equal $k_{Q_n}(X)$, the dimension of the Q_n–homology of X. One immediately deduces part of Lemma 1.5: $k_{Q_n}(X) = k_n(X)$ if and only if the AHSS converging to $K(n)^*(X)$ collapses at $E_{2^{n+1}}(X)$.

To continue with the proof of Lemma 1.5, let $cE_r^{*,\star}(X)$ denote the terms of the AHSS computing $k(n)^*(X)$, a 4th quadrant spectral sequence. Note that $cE_2^{*,\star} = H^*(X; \mathbb{Z}/2)[v_n]$ embeds in $E_2^{*,\star}(X) = H^*(X; \mathbb{Z}/2)[v_n^{\pm}]$, and equals it for $\star \leq 0$, and that the latter spectral sequence is obtained from the former by inverting v_n.

It follows that $cE_{2^{n+1}}^{*,\star}(X) = E_{2^{n+1}}^{*,\star}(X)$ for $\star < 0$, with the map on the 0–line between the spectral sequences corresponding to the epimorphism $Z^*(X; Q_n) \twoheadrightarrow H^*(X; Q_n)$. From this, one sees that any higher differential in the $k(n)^*(X)$ AHSS would be detected in the $K(n)^*(X)$ AHSS. Since this second spectral sequence is the localization of the first, we can conclude that the $K(n)^*(X)$ AHSS collapses at $E_{2^{n+1}}(X)$ if and only if the $k(n)^*(X)$ AHSS collapses at $cE_{2^{n+1}}(X)$.

Next we note that the AHSS spectral sequence $cE_r^{*,\star}(X)$ identifies with the ASS computing $k(n)^*(X)$ with suitable reindexing, with $cE_{2^{n+1}}^{*,\star}(X)$ corresponding to the Adams E_2 term. Firstly, a result of C R F Maunder [11] implies that the AHSS converging to $[X, k(n)]_*$ can be constructed by taking the Postnikov filtration of the spectrum $k(n)$. But the Postnikov tower for $k(n)$ is also an Adams tower: as described in the survey paper [14, Section 5], there is a cofibration sequence

$$\Sigma^{2^{n+1}-2}k(n) \xrightarrow{v_n} k(n) \xrightarrow{\pi} H\mathbb{Z}/2 \xrightarrow{\bar{Q}_n} \Sigma^{2^{n+1}-1}k(n)$$

such that $\Sigma^{2^{n+1}-1}\pi \circ \bar{Q}_n = Q_n$ and π induces the epimorphism $A \to A/AQ_n$ on mod 2 cohomology.

Finally, we note that, when $n = 0$, one still has the cofibration sequence as above, now with $v_0 = 2$, so that the ASS for $k(0) = H\mathbb{Z}$ is similarly related to the Bockstein spectral sequence.

3.2 The description of $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$ via Stiefel–Whitney classes

We recall classical results that are either explicitly in [12] or can easily be deduced from the material there.

Let $w_1, \ldots, w_d$ denote the Stiefel–Whitney classes of the canonical d–dimensional bundle γ_d over $\mathrm{Gr}_d(\mathbb{R}^\infty)$. One has

$$H^*(\mathrm{Gr}_d(\mathbb{R}^\infty); \mathbb{Z}/2) = \mathbb{Z}/2[w_1, \ldots, w_d].$$

Dual classes $\bar{w}_1, \bar{w}_2, \ldots$ are defined by the equation

$$(1 + w_1 + \cdots + w_d)(1 + \bar{w}_1 + \bar{w}_2 + \cdots) = 1,$$

and this allows one to write the classes $\bar{w}_k$ as polynomials in $w_1, \ldots, w_d$.

The inclusion $\mathrm{Gr}_d(\mathbb{R}^m) \hookrightarrow \mathrm{Gr}_d(\mathbb{R}^\infty)$ then induces a surjective ring homomorphism

$$H^*(\mathrm{Gr}_d(\mathbb{R}^\infty); \mathbb{Z}/2) \to H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$$

with kernel $J(d, m - d) = (\bar{w}_k \mid k > m - d)$. Now $\bar{w}_k$ can be interpreted as $w_k(\gamma_d^\perp)$, where $\gamma_d^\perp$ is the $(m - d)$–dimensional bundle complementary to γ_d.

We record some useful consequences. To state these, it is useful to let

$$i : \mathrm{Gr}_d(\mathbb{R}^{m-1}) \hookrightarrow \mathrm{Gr}_d(\mathbb{R}^m)$$

be the inclusion induced by the inclusion $\mathbb{R}^{m-1} \hookrightarrow \mathbb{R}^m$, and to let

$$j : \mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}) \hookrightarrow \mathrm{Gr}_d(\mathbb{R}^m)$$

be the inclusion sending $V \subset \mathbb{R}^{m-1}$ to $V \oplus \mathbb{R} \subset \mathbb{R}^m$.

Lemma 3.1 (a) *The ideal $J(d, m - d)$ is generated by the d classes $\bar{w}_{m-d+1}, \bar{w}_{m-d+2}, \ldots, \bar{w}_m$.*

(b) *In $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$, $w_d \bar{w}_{m-d} = 0$.*

(c) $\ker\{i^*\} = (\bar{w}_{m-d}) \subset H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$.

(d) $\ker\{j^*\} = (w_d) \subset H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$.

Proof Statement (a) follows from the recursive relations among the $\bar{w}_k$. Statement (b) follows from the equation

$$(1 + w_1 + \cdots + w_d)(1 + \bar{w}_1 + \cdots + \bar{w}_{m-d}) = 1,$$

which holds in $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$. Statement (c) follows from the fact that

$$J(d, m - 1 - d) = J(d, m) + (\bar{w}_{m-d}),$$

and (d) follows from (c), noting that j can be written as the composite

$$\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}) \simeq \mathrm{Gr}_{m-d}(\mathbb{R}^{m-1}) \xrightarrow{i} \mathrm{Gr}_{m-d}(\mathbb{R}^m) \simeq \mathrm{Gr}_d(\mathbb{R}^m),$$

where the indicated homeomorphisms are given by taking complementary subspaces (and, in cohomology, these maps swap each w_i with a $\bar{w}_j$). $\qquad\square$

We end this subsection with a couple more facts about $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$.

An additive basis for $H^q(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$ is given by the monomials

$$\left\{ w_1^{r_1} w_2^{r_2} \cdots w_d^{r_d} \,\Big|\, \sum_{i=1}^{d} r_i \leq m - d \right\},$$

so the top-dimensional class is w_d^{m-d} in degree $d(m-d)$; see [5].

The Wu formulae [12, page 94] are closed formulae for $\mathrm{Sq}^i w_j$, and, in theory, formulae for $Q_n(w_j)$ follow.

3.3 A description of the cofiber $C_d(\mathbb{R}^m)$ and its cohomology

Recall that $C_d(\mathbb{R}^m)$ is defined as the cofiber of the inclusion $\mathrm{Gr}_d(\mathbb{R}^{m-1}) \xrightarrow{i} \mathrm{Gr}_d(\mathbb{R}^m)$. This cofiber can be identified as a Thom space as follows.

Proposition 3.2 *Let $S(\gamma_{d-1}^{\perp})$ and $D(\gamma_{d-1}^{\perp})$ be the sphere and disk bundles associated to*

$$\gamma_{d-1}^{\perp} \to \mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}).$$

There is a pushout

$$
\begin{array}{ccc}
S(\gamma_{d-1}^{\perp}) & \xrightarrow{\;f\;} & \mathrm{Gr}_d(\mathbb{R}^{m-1}) \\
\downarrow & & \downarrow \\
D(\gamma_{d-1}^{\perp}) & \xrightarrow{\;f\;} & \mathrm{Gr}_d(\mathbb{R}^m),
\end{array}
$$

inducing a homeomorphism $f: \mathrm{Th}(\gamma_{d-1}^{\perp}) \xrightarrow{\sim} C_d(\mathbb{R}^m)$, such that the composite

$$\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}) \xrightarrow[\sim]{0-section} D(\gamma_{d-1}^{\perp}) \xrightarrow{\;f\;} \mathrm{Gr}_d(\mathbb{R}^m)$$

is the map j of Lemma 3.1.

Proof Recall that

$$D(\gamma_{d-1}^{\perp}) = \{(V, v) \mid V \in \mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}), v \in V^{\perp}, |v| \leq 1\},$$
$$S(\gamma_{d-1}^{\perp}) = \{(V, v) \mid V \in \mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}), v \in V^{\perp}, |v| = 1\}.$$

We define $f: D(\gamma_{d-1}^{\perp}) \to \mathrm{Gr}_d(\mathbb{R}^m)$ by the formula

$$f(V, v) = V + \langle v + \sqrt{1 - |v|^2} e_m \rangle,$$

where e_m is the mth standard basis vector in $\mathbb{R}^m$. We claim this f has the needed properties.

First, note that $f(V, \mathbb{0}) = V + \langle e_m \rangle = V \oplus \mathbb{R} = j(V)$.

Second, $(V, v) \in S(\gamma_{d-1}^{\perp})$ if and only if $f(V, v) = V + \langle v \rangle$, and so is an element of $\mathrm{Gr}_d(\mathbb{R}^{m-1})$. Furthermore, $f: S(\gamma_{d-1}^{\perp}) \to \mathrm{Gr}_d(\mathbb{R}^{m-1})$ is surjective: given any $W \in \mathrm{Gr}_d(\mathbb{R}^{m-1})$, if we choose any $(d-1)$–dimensional subspace V of W, and a unit length vector $v \in W$ in the 1–dimensional orthogonal complement, then $f(V, v) = W$.

Finally, we need to check that f is bijective on $\mathring{D}(\gamma_{d-1}^{\perp}) = D(\gamma_{d-1}^{\perp}) - S(\gamma_{d-1}^{\perp})$. To check this, let $W \in \mathrm{Gr}_d(\mathbb{R}^m)$ be a d–dimensional subspace of $\mathbb{R}^m$ not contained in $\mathbb{R}^{m-1}$, so that

$$V = W \cap \mathbb{R}^{m-1} \in \mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}).$$

Let $V^{\perp}$ be the complement of V in of $\mathbb{R}^m$ so that $W \cap V^{\perp}$ is one-dimensional, and let v be the unique unit vector $v \in W \cap V^{\perp}$ such that v has positive m^{th} coordinate. Let $\pi \colon \mathbb{R}^m \to \mathbb{R}^{m-1}$ be the standard projection. We claim that $f(V, \pi(v)) = W$ and $(V, \pi(v))$ is the unique point in $\mathring{D}(\gamma_{d-1}^{\perp})$ with this property: since $|v| = 1$ the mth component of v is $\sqrt{1 - |\pi(v)|^2}$; thus $v = \pi(v) + \sqrt{1 - |\pi(v)|^2} e_m$ and so $f(V, \pi(v)) = V + \langle v \rangle = W$. $\qquad\square$

Let $u_{\gamma_{d-1}^{\perp}} \in \widetilde{H}^{m-d}(\mathrm{Th}(\gamma_{d-1}^{\perp}))$ be the Thom class of $\gamma_{d-1}^{\perp} \to \mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})$. Then $\widetilde{H}^{m-d}(\mathrm{Th}(\gamma_{*}^{\perp}))$ is a free rank 1 $H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}))$–module on $u_{\gamma_{d-1}^{\perp}}$. Meanwhile, $H^*(\mathrm{Gr}_d(\mathbb{R}^m))$ is a $H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}))$–module via j^*, and the ideal $(\bar{w}_{m-d}) = \widetilde{H}^*(C_d(\mathbb{R}^m))$ is a submodule. The proposition thus implies the following.

Corollary 3.3 *The map* $f^* \colon \widetilde{H}^*(C_d(\mathbb{R}^m)) \xrightarrow{\sim} \widetilde{H}^*(\mathrm{Th}(\gamma_{d-1}^{\perp}))$ *is an isomorphism of free rank 1* $H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}))$*–modules, and* $f^*(\bar{w}_{m-d}) = u_{\gamma_{d-1}^{\perp}}$.

3.4 The characteristic class associated to Q_n and a twisted Q_n–module

Let $\alpha_n \in H^{2^{n+1}-1}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}))$ be defined as the element satisfying

$$Q_n(\bar{w}_{m-d}) = \alpha_n \bar{w}_{m-d} \in \widetilde{H}^*(C_d(\mathbb{R}^m)).$$

Then define

$$\widehat{Q}_n \colon H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})) \to H^{*+2^{n+1}-1}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}))$$

by the formula

$$\widehat{Q}_n(x) = Q_n(x) + x\alpha_n.$$

Proposition 3.4 $\widehat{Q}_n^2 = 0$, *and the chain complex* $\left(H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})), \widehat{Q}_n \right)$ *is isomorphic to the chain complex* $\left(\widetilde{H}^{*+m-d}(C_d(\mathbb{R}^m)), Q_n \right)$.

Proof Let $\Theta \colon H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})) \to \widetilde{H}^{*+m-d}(C_d(\mathbb{R}^m))$ be the isomorphism established in the last subsection, $\Theta(x) = x\bar{w}_{m-d}$. The proposition follows once we check that $\Theta(\widehat{Q}_n(x)) = Q_n(\Theta(x))$;

$$\begin{aligned}
\Theta(\widehat{Q}_n(x)) &= \widehat{Q}_n(x)\bar{w}_{m-d} \\
&= (Q_n(x) + x\alpha_n)\bar{w}_{m-d} \\
&= Q_n(x)\bar{w}_{m-d} + x(\alpha_n \bar{w}_{m-d}) \\
&= Q_n(x)\bar{w}_{m-d} + x Q_n(\bar{w}_{m-d}) \\
&= Q_n(x\bar{w}_{m-d}) \\
&= Q_n(\Theta(x)). \qquad\square
\end{aligned}$$

It is useful to put the class α_n in context. Given any element a in the Steenrod algebra $\mathcal{A}$, one gets a characteristic class $w_a(\xi) \in H^{|a|}(B; \mathbb{Z}/2)$ associated to any real vector bundle $\xi \to B$; $w_a(\xi)$ is defined as the element satisfying $a(u_\xi) = w_a(\xi)u_\xi \in \tilde{H}^{\dim \xi + |a|}(\mathrm{Th}(\xi); \mathbb{Z}/2)$, where u_ξ is the Thom class of ξ. So, for example, $w_{\mathrm{Sq}^n}(\xi) = w_n(\xi)$, and, relevant for us, our class α_n equals $w_{Q_n}(\xi)$ when $\xi = \gamma_{d-1}^\perp \to \mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})$.

We have the following characterization of w_{Q_n}.

Proposition 3.5 w_{Q_n} *is the unique characteristic class satisfying the following two properties*:

(a) $w_{Q_n}(\xi \oplus \nu) = w_{Q_n}(\xi) + w_{Q_n}(\nu)$;

(b) *if* $\gamma \to B$ *is one-dimensional, then* $w_{Q_n}(\gamma) = w_1(\gamma)^{2^{n+1}-1}$.

Proof Property (a) follows from the fact that Q_n is primitive in $\mathcal{A}$ (or, equivalently, that Q_n acts a derivation). To see property (b), one first calculates that $Q_n(t) = t^{2^n+1} \in \mathbb{Z}/2[t] = H^*(\mathbb{R}P^\infty; \mathbb{Z}/2)$, recalling that $Q_0 = \mathrm{Sq}^1$, and $Q_n = \mathrm{Sq}^{2^n} Q_{n-1} + Q_{n-1}\mathrm{Sq}^{2^n}$. Then property (b) follows, since if γ is the universal line bundle over $\mathbb{R}P^\infty$, then $u_\gamma = t$. Uniqueness follows from the splitting principle. $\square$

Remark 3.6 Thus $w_{Q_n}(\xi)$ agrees with the "s–class" $s_{2^{n+1}-1}(\xi)$, analogous to the class of the same name for complex vector bundles as defined in [12, Section 16]. (These s_I are *not* the same as the s_λ of the next subsection; these are two conflicting and standard usages.)

Since $\gamma_{d-1}^\perp \oplus \gamma_{d-1}$ is trivial, property (b) has the following consequence.

Corollary 3.7 $$\alpha_n = w_{Q_n}(\gamma_{d-1}) \in H^{2^{n+1}-1}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}); \mathbb{Z}/2).$$

3.5 The description of $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$ via Schubert cells, and Lenart's formula

For the purposes of proving Theorem 1.9, we use an alternative description of $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$.

We recall the cell structure of $\mathrm{Gr}_d(\mathbb{R}^{d+c})$ as described in [12, Section 6]. A *Schubert symbol* $\lambda = (\lambda_1, \ldots, \lambda_d)$ of $\mathrm{Gr}_d(\mathbb{R}^m)$ is a sequence of integers

$$m - d \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0.$$

The *weight* of λ is defined to be $\sum_i \lambda_i$ and is denoted $|\lambda|$. Such a λ is a partition contained inside of a $d \times (m-d)$ grid when depicted as Young diagrams — diagrams with λ_i boxes in the ith row.

To each such partition is associated a Schubert cell $e(\lambda)$ of dimension λ in $\mathrm{Gr}_d(\mathbb{R}^m)$ defined by

$$e(\lambda) = \{V \in \mathrm{Gr}_d(\mathbb{R}^m) \mid \dim(V \cap \mathbb{R}^{i+\lambda_{d+1-i}}) \geq i \text{ for } 1 \leq i \leq d\}.$$
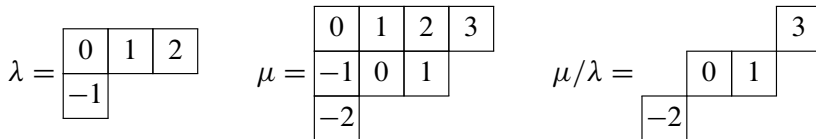
This cell decomposition of the Grassmannian leads to the dual Schubert cell basis for $H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$ with basis elements $s_\lambda \in H^{|\lambda|}(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$.

With this notation, one has that $w_i = s_{(1^i)}$ and $\bar{w}_j = s_{(j)}$. Although we don't use this here, it is worth noting that the cohomology ring structure in this basis is described by the Littlewood–Richardson rule of symmetric function theory.

To state Lenart's formula for calculating Q_n on a Schubert basis element [9], we need some combinatorial definitions. Given a Young diagram λ that includes into another Young diagram μ, one can form the complement μ/λ. For example,

$$\lambda = \qquad \mu = \qquad \mu/\lambda =$$
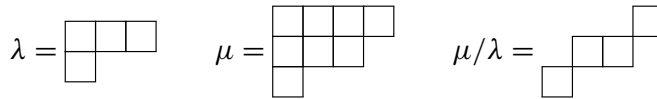
The *content* of a box b of μ in row i and column j is defined to be $c(b) = j - i$. For a box b in the skew shape μ/λ, we define its content to be the content of b embedded in μ. Here we fill in the contents of the diagrams from above:

$$\lambda = \boxed{\begin{matrix} 0 & 1 & 2 \\ -1 \end{matrix}} \qquad \mu = \boxed{\begin{matrix} 0 & 1 & 2 & 3 \\ -1 & 0 & 1 \\ -2 \end{matrix}} \qquad \mu/\lambda = \boxed{\begin{matrix} & & & 3 \\ 0 & 1 \\ -2 \end{matrix}}$$

A skew-shape is said to be *connected* when each pair of boxes in the diagram is connected by a sequence of boxes that each share an edge. A shape λ is called a *border strip*, if it is connected and does not contain a 2×2 block of boxes. A shape satisfying just the second condition is called a *broken border* strip, and in particular, a border strip is an example of a broken border strip with just one connected component. If λ is a broken border strip, then we denote by $cc(\lambda)$ the number of connected components of λ. If λ is not a broken border strip, then we define $cc(\lambda) = \infty$. For example, in the next diagram, λ_1 is a border strip, λ_2 is a broken border strip that is not a border strip, and λ_3 is an example of a shape that is neither:

$$\lambda_1 = \qquad \lambda_2 = \qquad \lambda_3 =$$
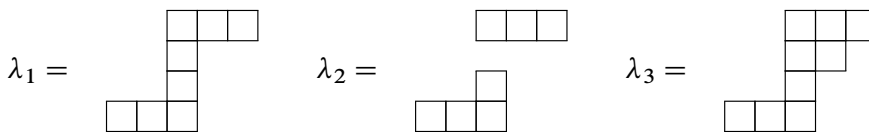
A *sharp corner* of a broken border strip is a box with no north, no west and no northwest neighbors. A *dull corner* is a box with both north and west neighbors, but no northwest neighbor. Let $C(\mu/\lambda)$ denote the set of sharp and dull corners of μ/λ. For example, in the following diagram the sharp corners have been labeled S and the dull corners have been labeled D:

$$\boxed{\begin{matrix} & & S & & \\ & & & & \\ S & & D & & \end{matrix}}$$
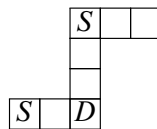
We are now ready to state Lenart's formula from [9]:

$$(3\text{-}1) \qquad Q_n(s_\lambda) = \sum_{\substack{\mu \supset \lambda : |\mu| - |\lambda| = 2^{n+1} - 1 \\ cc(\mu/\lambda) \leq 2}} d_{\lambda\mu} s_\mu,$$

where μ/λ must be a broken border strip and

$$(3\text{-}2) \qquad d_{\lambda\mu} = \begin{cases} \sum_{b \in C(\mu/\lambda)} c(b) & \text{if } \mu/\lambda \text{ is connected,} \\ 1 & \text{if } \mu/\lambda \text{ is disconnected.} \end{cases}$$

Example 3.8 As an example we compute Q_1 on $w_1 = s_\square$ in the Schubert basis in $\mathrm{Gr}_2(\mathbb{R}^6)$. There are three basis elements in degree four,

$$\mu_1 = \square\square\square\square, \quad \mu_2 = \begin{array}{c}\square\square\\\square\end{array}, \quad \mu_3 = \boxplus.$$

To compute $Q_n(s_\square)$ using (3-1) we must consider each complement. Let $\lambda = \square$. For μ_1,

$$\mu_1/\lambda = \boxed{\begin{array}{c|c|c|c} 0 & 1 & 2 & 3 \end{array}} \Big/ \boxed{\ } = \boxed{\begin{array}{c|c|c} 1 & 2 & 3 \end{array}}$$

The complement is a border strip and there is just one sharp corner (the left most corner) and no dull corners. The content of the sharp corner is 1 modulo two; hence $d_{\lambda\mu_1} = 1$, and so s_{μ_1} is in the expansion of $Q_1(s_\lambda)$. Next we consider

$$\mu_2/\lambda = \boxed{\begin{array}{c|c|c} 0 & 1 & 2 \\\hline -1 \end{array}} \Big/ \boxed{\ } = \boxed{\begin{array}{c|c} & 1 & 2 \\\hline -1 \end{array}}$$

This is a disconnected broken border strip; hence $d_{\lambda\mu_2} = 1$, and so s_{μ_2} is in the expansion. Finally,

$$\mu_3/\lambda = \boxed{\begin{array}{c|c} 0 & 1 \\\hline -1 & -2 \end{array}} \Big/ \boxed{\ } = \boxed{\begin{array}{c|c} & 1 \\\hline -1 & -2 \end{array}}$$

There are two sharp corners, one of content -1 and the other of content 1. There is also one dull corner of content -2. This means $d_{\lambda\mu_3} = (-1) + 1 + 2 \equiv 0$, and so s_{μ_3} is not in the expansion. Hence,

$$Q_1(s_\square) = s_{\square\square\square\square} + s_{\begin{array}{c}\square\square\\\square\end{array}}.$$

4 Results about $H^*(\mathrm{Gr}_d(\mathbb{R}^m); Q_n)$ when m is even

Proof of Theorem 1.9(a) We are going to show that $Q_n(s_\lambda) = 0$ for each Schubert basis element s_λ in degree $d(m-d) - 2^{n+1} + 1$. Since $s_{(d^{(m-d)})}$ is the only class in degree $d(m-d)$,

$$Q_n(s_\lambda) = d_{\lambda(d^{(m-d)})} s_{(d^{(m-d)})},$$

where $d_{\lambda(d^{(m-d)})}$ is given by (3-2). We must only consider λ such that $(d^{(m-d)})/\lambda$ is a broken border strip. As $(d^{(m-d)})$ is a $d \times (m-d)$ grid the complement $(d^{(m-d)})/\lambda$ is always connected and so if $(d^{(m-d)})/\lambda$ is a broken border strip it must be, in particular, a border strip. If $(d^{(m-d)})/\lambda$ is a border strip, then it must be one of three types:

(1) $(d^{(m-d)})/\lambda$ is the last row of $(d^{(m-d)})$,

(2) $(d^{(m-d)})/\lambda$ is the last column of $(d^{(m-d)})$,

(3) $(d^{(m-d)})/\lambda$ is the union of the last row and last column of $(d^{(m-d)})$.

We will show that $d_{\lambda(d^{(m-d)})} = 0$ in each of these cases. As m was assumed to be even, the content of the right most bottom box of $(d^{(m-d)})$ is also even.

(1) For the first case, there is just one sharp corner, namely the left most box, and there are no dull corners. Since the strip is of odd length, namely, $2^{n+1} - 1$, the leftmost box and the rightmost box have the same content modulo two. Hence, the content of this sharp corner is zero modulo two, and so $d_{\lambda(d^{(m-d)})} = 0$.

(2) For the second case, the argument is exactly the same, but with the sharp corner on the top.

(3) For the third case, the content of the sharp corner on the bottom left and the content of the sharp corner on the top right agree modulo two, because the border strip is of odd length. There is one dull corner in the bottom right and it is zero modulo two. Thus, the two sharp corners cancel and the dull corner contributes nothing.

Thus, in all cases $Q_n(s_\lambda) = 0$ for s_λ in degree $d(m - d) - 2^{n+1} + 1$. This completes the proof that the top class is not in the image of Q_n for even m. □

Proof of Theorem 1.9(b) We wish to prove that, when m is even, the chain complexes $\big(\tilde{H}^*(C_d(\mathbb{R}^m)); Q_n\big)$ and $\big(H^{d(m-d)-*}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})); Q_n\big)$ are dual.

By Proposition 3.4, $\big(\tilde{H}^{*+m-d}(C_d(\mathbb{R}^m)); Q_n\big)$ is isomorphic to the chain complex

$$\big(H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})); \hat{Q}_n\big),$$

where we recall that $\hat{Q}_n(y) = Q_n(y) + y\alpha_n$, and that $\alpha_n \bar{w}_{m-d} = Q_n(\bar{w}_{m-d}) \in H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$.

So we need to check that the chain complexes

$$\big(H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})); \hat{Q}_n\big) \quad \text{and} \quad \big(H^{(d-1)(m-d)-*}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})); Q_n\big)$$

are dual. This means we need to show that, if $x, y \in H^*(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}); \mathbb{Z}/2)$ satisfy

$$|x| + |y| + |Q_n| = (d - 1)(m - 1),$$

then

$$Q_n(x)y = x\hat{Q}_n(y).$$

By Theorem 1.9(a), we know that

$$Q_n(xy\bar{w}_{m-d}) = 0 \in H^{d(m-d)}(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2).$$

Thus, in $H^{d(m-d)}(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$,

$$
\begin{aligned}
0 &= Q_n(xy\bar{w}_{m-d}) \\
&= Q_n(x)y\bar{w}_{m-d} + xQ_n(y)\bar{w}_{m-d} + xyQ_n(\bar{w}_{m-d}) \\
&= Q_n(x)y\bar{w}_{m-d} + xQ_n(y)\bar{w}_{m-d} + xy\alpha_n\bar{w}_{m-d} \\
&= (Q_n(x)y + xQ_n(y) + xy\alpha_n)\bar{w}_{m-d} \\
&= (Q_n(x)y + x\hat{Q}_n(y))\bar{w}_{m-d},
\end{aligned}
$$

and we conclude that $0 = Q_n(x)y + x\hat{Q}_n(y) \in H^{(d-1)(m-d)}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}); \mathbb{Z}/2)$. $\qquad\square$

Proof of Theorem 1.9(c) Recall that $k_{Q_n}(X)$ denotes the rank of the Q_n–homology $H^*(X; Q_n)$. Similarly, let $\bar{k}_{Q_n}(X)$ denote the rank of $\tilde{H}^*(X; Q_n)$.

Let $m = 2^{n+1} - \epsilon + 2l$ with $\epsilon = 0$ or 1, and $l \geq 0$. Let

$$
k_n^G(d, m) = \sum_{i=0}^{\lfloor d/2 \rfloor} \binom{2^{n+1} - \epsilon}{d - 2i} \binom{l}{i}.
$$

We start with the first part of Theorem 1.9(c). This asserts that, when m is even, if we assume that

$$
k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^{m-1})) = k_n^G(d, m-1) \quad \text{and} \quad k_{Q_n}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})) = k_n^G(d-1, m-1),
$$

then we can conclude that $k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^m)) = k_n^G(d, m)$.

Theorem 1.2 tells us that $k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^m)) \geq k_n^G(d, m)$.

Since we have a short exact sequence

$$
0 \to \tilde{H}^*(C_d(\mathbb{R}^m)) \to H^*(\mathrm{Gr}_d(\mathbb{R}^m)) \to H^*(\mathrm{Gr}_d(\mathbb{R}^{m-1})) \to 0,
$$

we see that

$$
k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^{m-1})) + \bar{k}_{Q_n}(C_d(\mathbb{R}^m)) \geq k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^m)),
$$

with equality if and only if the associated long exact Q_n–homology sequence is still short exact.

Since m is even, Theorem 1.9(b) applies, and tells us that $\bar{k}_{Q_n}(C_d(\mathbb{R}^m)) = k_{Q_n}(\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1}))$.

Putting this all together, under our assumptions,

$$
k_n^G(d, m-1) + k_n^G(d-1, m-1) \geq k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^m)) \geq k_n^G(d, m).
$$

That these would be, in fact, equalities, follows from the next lemma.

Lemma 4.1 *If $m = 2^{n+1} + 2l$ with $l \geq 0$, then*

$$
k_n^G(d, m-1) + k_n^G(d-1, m-1) = k_n^G(d, m).
$$

Proof We compute

$$k_n^G(d, m-1) + k_n^G(d-1, m-1) = \sum_i \left[\binom{2^{n+1}-1}{d-2i} + \binom{2^{n+1}-1}{d-2i-1} \right] \binom{l}{i}$$

$$= \sum_i \binom{2^{n+1}}{d-2i} \binom{l}{i}$$

$$= k_n^G(d, m). \qquad \square$$

When all of this happens, we then see that the Q_n–homology long exact sequence really is still short exact, and also that the $K(n)$–AHSS must collapse for these three spaces. Thus there is also a short exact sequence

$$0 \to \widetilde{K}(n)^*(C_d(\mathbb{R}^m)) \xrightarrow{p^*} K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m)) \xrightarrow{i^*} K(n)^*(\mathrm{Gr}_d(\mathbb{R}^{m-1})) \to 0.$$

Finally, the top cohomology class in $H^{d(m-d)}(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Z}/2)$ will be a permanent cycle in the AHSS computing $K(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$ and thus also in the AHSS computing $k(n)^*(\mathrm{Gr}_d(\mathbb{R}^m))$, and this is equivalent to saying that $\mathrm{Gr}_d(\mathbb{R}^m)$ is $k(n)$–oriented. $\qquad \square$

5 Results about $H^*(\mathrm{Gr}_d(\mathbb{R}^m); Q_n)$ when $d = 2$

In this section we present our results about the Q_n–homology of $\mathrm{Gr}_2(\mathbb{R}^m)$, with the focus on understanding the case when m has the form $2^{n+1} - 1 + 2l$.

To begin with, we know that

- $H^*(\mathrm{Gr}_2(\mathbb{R}^m); \mathbb{Z}/2) = \mathbb{Z}/2[w_1, w_2]/(\bar{w}_{m-1}, \bar{w}_m)$;

- in $H^*(\mathrm{Gr}_2(\mathbb{R}^m); \mathbb{Z}/2)$, the ideal $\widetilde{H}^*(C_2(\mathbb{R}^m); \mathbb{Z}/2)$ has an additive basis $\{w_1^i \bar{w}_{m-2} \mid 0 \le i \le m-2\}$.

Now we collect results that hold in $H^*(\mathrm{Gr}_2(\mathbb{R}^\infty); \mathbb{Z}/2)$.

Lemma 5.1 *In* $H^*(\mathrm{Gr}_2(\mathbb{R}^\infty); \mathbb{Z}/2)$,

(a) $\bar{w}_0 = 1$, $\bar{w}_1 = w_1$, *and, recursively,* $\bar{w}_k = w_1 \bar{w}_{k-1} + w_2 \bar{w}_{k-2}$;

(b) $w_2^j \bar{w}_k = \sum_i \binom{j}{i} w_1^{j-i} \bar{w}_{k+j+i}$;

(c) $\bar{w}_k = \sum_j \binom{k-j}{j} w_1^{k-2j} w_2^j$;

(d) $\bar{w}_{2^b-1} = w_1^{2^b-1}$ *for all* $b \ge 0$;

(e) $\bar{w}_{2^b-2} = \sum_{c=0}^{b-1} w_1^{2^b-2^{c+1}} w_2^{2^c-1}$ *for all* $b \ge 1$.

Proof The homogeneous components of the equation $0 = (1 + w_1 + w_2)(1 + \bar{w}_1 + \bar{w}_2 + \cdots)$ give statement (a).

Statement (b) is proved by induction on j. The case when $j = 0$ is trivial, and statement (a) rewrites as $w_2 \bar{w}_k = w_1 \bar{w}_{k+1} + \bar{w}_{k+2}$, which is the case when $j = 1$. One then computes

$$
\begin{aligned}
w_2^j \bar{w}_k &= w_2(w_2^{j-1} \bar{w}_k) \\
&= \sum_i \binom{j-1}{i} w_1^{j-1-i} w_2 \bar{w}_{k+j-1+i} \\
&= \sum_i \binom{j-1}{i} [w_1^{j-i} \bar{w}_{k+j+i} + w_1^{j-1-i} \bar{w}_{k+j+i+1}] \\
&= \sum_i \left[\binom{j-1}{i} + \binom{j-1}{i-1} \right] w_1^{j-i} \bar{w}_{k+j+i} \\
&= \sum_i \binom{j}{i} w_1^{j-i} \bar{w}_{k+j+i}.
\end{aligned}
$$

For (c), note that $\bar{w}_k$ is the homogeneous component of degree k in

$$
\bar{w} = (1 + w_1 + w_2)^{-1} = \sum_{t=0}^{\infty} (w_1 + w_2)^t.
$$

Statement (d) follows from (c):

$$
\bar{w}_{2^b-1} = \sum_j \binom{2^b-1-j}{j} w_1^{k-2j} w_2^j = w_1^{2^b-1},
$$

using that $\binom{2^b-1-j}{j} \equiv 1 \bmod 2$ only if $j = 0$.

Similarly, statement (e) follows from (c):

$$
\bar{w}_{2^b-2} = \sum_j \binom{2^b-2-j}{j} w_1^{k-2j} w_2^j = \sum_{c=0}^{b-1} w_1^{2^b-2^{c+1}} w_2^{2^c-1},
$$

using that $\binom{2^b-2-j}{j} \equiv 1 \bmod 2$ if and only if $j = 2^c - 1$ with $0 \le j \le b - 1$. $\qquad\square$

Now we determine the action of Q_n on various classes.

Lemma 5.2 *In* $H^*(\mathrm{Gr}_2(\mathbb{R}^\infty); \mathbb{Z}/2)$, $Q_n(w_1) = w_1^{2^{n+1}} = w_1 \bar{w}_{2^{n+1}-1}$.

Proof The first equality here was already noted in the proof of Proposition 3.5, and the second follows from Lemma 5.1(d). $\qquad\square$

Lemma 5.3 *In* $H^*(\mathrm{Gr}_2(\mathbb{R}^\infty); \mathbb{Z}/2)$,

$$
Q_n(w_2) = \sum_{c=0}^{n} w_1^{2^{n+1}-2^{c+1}+1} w_2^{2^c} = w_1 w_2 \bar{w}_{2^{n+1}-2}.
$$

Proof The second equality here follows from Lemma 5.1(e), so we just need to check the first. We do this by induction on n, where the $n = 0$ case is the easily checked: $Q_0(w_2) = \mathrm{Sq}^1(w_2) = w_1 w_2$.

Before proceeding with the inductive step, we make two observations.

The first is that for $n \geq 1$, $Q_n(w_2) = \mathrm{Sq}^{2^n} Q_{n-1}(w_2)$ because the other term, $Q_{n-1}\mathrm{Sq}^{2^n}(w_2)$, will be zero. This is clear if $n \geq 2$ as then $\mathrm{Sq}^{2^n}(w_2) = 0$, and when $n = 1$, we observe that $Q_0\mathrm{Sq}^2(w_2) = \mathrm{Sq}^1(w_2^2) = 0$.

The second observation is that $\mathrm{Sq}(w_2) = w_2(1 + w_1 + w_2)$, so $\mathrm{Sq}(w_2^{2^c}) = w_2^{2^c}(1 + w_1^{2^c} + w_2^{2^c})$, and thus

$$
\mathrm{Sq}^j(w_2^{2^c}) = \begin{cases} w_2^{2^c} & \text{if } j = 0, \\ w_1^{2^c} w_2^{2^c} & \text{if } j = 2^c, \\ w_2^{2^{c+1}} & \text{if } j = 2^{c+1}, \\ 0 & \text{otherwise.} \end{cases}
$$

Now we check the inductive step of our proof.

$$
\begin{aligned}
Q_n(w_2) &= \mathrm{Sq}^{2^n} Q_{n-1}(w_2) && \text{(by our first observation)} \\
&= \sum_{c=0}^{n-1} \mathrm{Sq}^{2^n}(w_1^{2^n - 2^{c+1} + 1} w_2^{2^c}) && \text{(by inductive hypothesis)} \\
&= \sum_{c=0}^{n-1} \sum_j \mathrm{Sq}^{2^n - j}(w_1^{2^n - 2^{c+1} + 1}) \mathrm{Sq}^j(w_2^{2^c}).
\end{aligned}
$$

By our second observation, the only possible nonzero terms in this double sum are when $j = 0, 2^c, 2^{c+1}$. The terms with $j = 0$ are all zero, as $\mathrm{Sq}^{2^n}(w_1^{2^n - 2^{c+1} + 1}) = 0$ by the unstable condition. Similarly, the only nonzero term with $j = 2^c$ is the term $w_1^{2^{n+1} - 1} w_2$, when $c = 0$. Finally, one gets $w_1^{2^{n+1} - 2^{c+2} + 1} w_2^{2^{c+1}}$ when $j = 2^{c+1}$ for all $0 \leq c \leq n - 1$. One is left with

$$
Q_n(w_2) = \sum_{c=0}^{n} w_1^{2^{n+1} - 2^{c+1} + 1} w_2^{2^c},
$$

completing our induction. $\qquad\square$

Remark 5.4 The referee has pointed out that the first equality in the last lemma appears in [13, page 508].[1]

We now turn our attention to the behavior of Q_n on $\widetilde{H}^*(C_2(\mathbb{R}^m); \mathbb{Z}/2)$.

Lemma 5.5 In $\widetilde{H}^*(C_2(\mathbb{R}^m); \mathbb{Z}/2)$, $Q_n(\bar{w}_{m-2}) = w_1^{2^{n+1} - 1} \bar{w}_{m-2}$.

Proof By Corollary 3.7, $Q_n(\bar{w}_{m-2}) = w_{Q_n}(\gamma_1) \bar{w}_{m-2}$, where $\gamma_1 \to \mathrm{Gr}_1(\mathbb{R}^{m-1})$ is the canonical line bundle, and Proposition 3.5 tells us that $w_{Q_n}(\gamma_1) = w_1^{2^{n+1} - 1}$. $\qquad\square$

---

[1]There is a slight misprint, and a proof is just hinted at.

Remark 5.6 This lemma also admits a proof using the Schubert cell perspective; see [10, Lemma 4.9.13].

As Q_n is a derivation, the lemma, together with the calculation $Q_n(w_1) = w_1^{2^{n+1}}$, allows one to easily compute the Q_n–homology of $C_2(\mathbb{R}^m)$. What results is the following.

Proposition 5.7   (a)   In $\widetilde{H}^*(C_2(\mathbb{R}^m); \mathbb{Z}/2)$,

$$Q_n(w_1^i \bar{w}_{m-2}) = \begin{cases} w_1^{2^{n+1}-1+i} \bar{w}_{m-2} & \text{if } i \text{ is even,} \\ 0 & \text{if } i \text{ is odd.} \end{cases}$$

(b)   If $m \leq 2^{n+1}$, then Q_n acts as zero on $\widetilde{H}^*(C_2(\mathbb{R}^m); \mathbb{Z}/2)$. Thus $\tilde{k}_{Q_n}(C_2(\mathbb{R}^m)) = m - 1$.

(c)   If $m > 2^{n+1}$ and is even, then the classes

$$\{w_1^{2j-1} \bar{w}_{m-2} \mid 1 \leq j \leq 2^n - 1\} \quad \text{and} \quad \{w_1^{m-2j} \bar{w}_{m-2} \mid 1 \leq j \leq 2^n\}$$

represent the Q_n–homology classes. Thus $\tilde{k}_{Q_n}(C_2(\mathbb{R}^m)) = 2^{n+1} - 1$.

(d)   If $m > 2^{n+1}$ and is odd, then the classes

$$\{w_1^{2j-1} \bar{w}_{m-2} \mid 1 \leq j \leq 2^n - 1\} \quad \text{and} \quad \{w_1^{m-1-2j} \bar{w}_{m-2} \mid 1 \leq j \leq 2^n - 1\}$$

represent the Q_n–homology classes. Thus $\tilde{k}_{Q_n}(C_2(\mathbb{R}^m)) = 2^{n+1} - 2$.

Proof of Theorem 1.10   Let $m = 2^{n+1} + 1 + 2l$. We need to prove that the map

$$\widetilde{H}^*(C_2(\mathbb{R}^m); Q_n) \xrightarrow{p^*} H^*(\mathrm{Gr}_2(\mathbb{R}^m); Q_n)$$

is zero; ie we need to show that representatives of the Q_n–homology classes in $\widetilde{H}^*(C_2(\mathbb{R}^m); \mathbb{Z}/2)$ are in the image of Q_n when regarded in $H^*(\mathrm{Gr}_2(\mathbb{R}^m); \mathbb{Z}/2)$.

By Proposition 5.7(d), these representatives are in two families,

$$w_1^{1+2j} \bar{w}_{2^{n+1}-1+2l} \quad \text{and} \quad w_1^{2l+2+2j} \bar{w}_{2^{n+1}-1+2l},$$

both with $0 \leq j \leq 2^n - 2$.

If we can find $a, b \in H^*(\mathrm{Gr}_2(\mathbb{R}^m); \mathbb{Z}/2)$ such that

$$Q_n(a) = w_1 \bar{w}_{2^{n+1}-1+2l} \quad \text{and} \quad Q_n(b) = w_1^{2l+2} \bar{w}_{2^{n+1}-1+2l},$$

we will be done, as then

$$Q_n(w_1^{2j} a) = w_1^{1+2j} \bar{w}_{2^{n+1}-1+2l} \quad \text{and} \quad Q_n(w_1^{2j} b) = w_1^{2l+2+2j} \bar{w}_{2^{n+1}-1+2l}.$$

Thus the next two propositions finish the proof.   □

Proposition 5.8   In $H^*(\mathrm{Gr}_2(\mathbb{R}^\infty); \mathbb{Z}/2)$,

$$Q_n(w_1 \bar{w}_{2l}) = w_1 \bar{w}_{2^{n+1}-1+2l}.$$

Proposition 5.9 In $H^*(\mathrm{Gr}_2(\mathbb{R}^{2^{n+1}+1+2l}); \mathbb{Z}/2)$,

$$Q_n(w_2^{2l+1}) = w_1^{2l+2}\bar{w}_{2^{n+1}-1+2l}.$$

Before proving these, we first run through how Theorem 1.10 leads to the proof of Theorem 1.8.

Proof of Theorem 1.8 Our goal is to show that if $m = 2^{n+1} - \epsilon + 2l$ with $\epsilon = 0$ or 1, and $l \geq 0$, then $k_{Q_n}(\mathrm{Gr}_2(\mathbb{R}^m)) = \binom{2^{n+1}-\epsilon}{2} + l$, the lower bound coming from Theorem 1.2.

We prove this by induction on m, with the two cases when $l = 0$ already covered by Theorem 1.1. The case when m is even is covered by Theorem 1.9(c), as we know our calculations are right for $(n, 1, m-1)$, and by induction we can assume the theorem for $(n, 2, m-1)$.

Suppose m is odd, so $\epsilon = 1$ and $m - 1 = 2^{n+1} + 2(l - 1)$. By induction, we can assume that $k_{Q_n}(\mathrm{Gr}_2(\mathbb{R}^{m-1})) = \binom{2^{n+1}}{2} + (l - 1)$. Then

$$\begin{aligned}
k_{Q_n}(\mathrm{Gr}_2(\mathbb{R}^m)) &= k_{Q_n}(\mathrm{Gr}_2(\mathbb{R}^{m-1})) - \bar{k}_{Q_n}(C_2(\mathbb{R}^m)) \quad \text{(by Theorem 1.10)} \\
&= \binom{2^{n+1}}{2} + (l - 1) - (2^{n+1} - 2) \quad \text{(by Proposition 5.7(d))} \\
&= \binom{2^{n+1}-1}{2} + l. \qquad \square
\end{aligned}$$

It remains to prove Propositions 5.8 and 5.9.

Proof of Proposition 5.8 We prove by induction on l that

$$Q_n(w_1\bar{w}_{2l}) = w_1\bar{w}_{2^{n+1}-1+2l}$$

holds in $H^*(\mathrm{Gr}_2(\mathbb{R}^\infty); \mathbb{Z}/2)$.

We start the induction by checking both the $l = 0$ and $l = 1$ cases.

When $l = 0$, this reads $Q_n(w_1) = w_1\bar{w}_{2^{n+1}}$, which was proved in Lemma 5.2.

We check the $l = 1$ case using both Lemmas 5.2 and 5.3,

$$\begin{aligned}
Q_n(w_1\bar{w}_2) &= Q_n(w_1(w_2 + w_1^2)) \\
&= Q_n(w_1 w_2 + w_1^3) \\
&= Q_n(w_1)w_2 + w_1 Q_n(w_2) + w_1^2 Q_n(w_1) \\
&= w_1 w_2 \bar{w}_{2^{n+1}-1} + w_1^2 w_2 \bar{w}_{2^{n+1}-2} + w_1^3 \bar{w}_{2^{n+1}-1} \\
&= w_1[w_2\bar{w}_{2^{n+1}-1} + w_1(w_2\bar{w}_{2^{n+1}-2} + w_1\bar{w}_{2^{n+1}-1})] \\
&= w_1[w_2\bar{w}_{2^{n+1}-1} + w_1\bar{w}_{2^{n+1}}] \\
&= w_1\bar{w}_{2^{n+1}+1}.
\end{aligned}$$

For the inductive case, we use the identity $\bar{w}_k = w_2^2 \bar{w}_{k-4} + w_1^2 \bar{w}_{k-2}$ which holds for all $k \geq 4$. Then

$$
\begin{aligned}
Q_n(w_1 \bar{w}_{2l}) &= Q_n(w_1 w_2^2 \bar{w}_{2(l-2)} + w_1^3 \bar{w}_{2(l-1)}) \\
&= w_2^2 Q_n(w_1 \bar{w}_{2(l-2)}) + w_1^2 Q_n(w_1 \bar{w}_{2(l-1)}) \\
&= w_1 w_2^2 \bar{w}_{2^{n+1}+2(l-2)-1} + w_1^3 \bar{w}_{2^{n+1}+2(l-1)-1} \\
&= w_1 [w_2^2 \bar{w}_{2^{n+1}+2(l-2)-1} + w_1^2 \bar{w}_{2^{n+1}+2(l-1)-1}] \\
&= w_1 \bar{w}_{2^{n+1}+2l-1}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square
\end{aligned}
$$

Proof of Proposition 5.9 We wish to prove that

$$
Q_n(w_2^{2l+1}) = w_1^{2l+2} \bar{w}_{2^{n+1}-1+2l}
$$

holds in $H^*(\mathrm{Gr}_2(\mathbb{R}^{2^{n+1}+1+2l}); \mathbb{Z}/2)$.

We begin with a calculation in $H^*(\mathrm{Gr}_2(\mathbb{R}^\infty); \mathbb{Z}/2)$,

$$
\begin{aligned}
Q_n(w_2^{2l+1}) &= w_2^{2l} Q_n(w_2) \\
&= w_1 w_2^{2l+1} \bar{w}_{2^{n+1}-2} && \text{(using Lemma 5.3)} \\
&= \sum_i \binom{2l+1}{i} w_1^{2l+2-i} \bar{w}_{2^{n+1}+2l-1+i} && \text{(using Lemma 5.1(b)).}
\end{aligned}
$$

When we project this sum onto

$$
H^*(\mathrm{Gr}_2(\mathbb{R}^{2^{n+1}+1+2l}); \mathbb{Z}/2) = \mathbb{Z}/2[w_1, w_2]/(\bar{w}_k \mid k \geq 2^{n+1} + 2l),
$$

only the term with $i = 0$ is not zero. In other words

$$
Q_n(w_2^{2l+1}) = w_1^{2l+2} \bar{w}_{2^{n+1}-1+2l}
$$

holds in $H^*(\mathrm{Gr}_2(\mathbb{R}^{2^{n+1}+1+2l}); \mathbb{Z}/2)$. $\square$

6 Towards the conjectures

As organized in this paper, we are trying to calculate $H^*(\mathrm{Gr}_d(\mathbb{R}^m); Q_n)$ by induction on m (and d) with two steps:

- calculate $\widetilde{H}^*(C_d(\mathbb{R}^m); Q_n)$, recalling that $C_d(\mathbb{R}^m)$ is the Thom space of a bundle over $\mathrm{Gr}_{d-1}(\mathbb{R}^{m-1})$;

- calculate $\delta \colon H^*(\mathrm{Gr}_d(\mathbb{R}^{m-1}); Q_n) \to \widetilde{H}^{*+2^{n+1}-1}(C_d(\mathbb{R}^m); Q_n)$.

When m is even, Theorem 1.9 says we can carry through with this plan. In this section we speculate about how things might go when m is odd.

Firstly, we have the analogues of Theorems 1.1 and 1.2 for

$$
\bar{k}_n(C_d(\mathbb{R}^m)) = \dim_{K(n)_*} \widetilde{K}(n)^*(C_d(\mathbb{R}^m)).
$$

Theorem 6.1 *If $m \leq 2^{n+1}$, then $\bar{k}_n(C_d(\mathbb{R}^m)) = \binom{m-1}{d-1}$.*

Proof Theorem 1.1 implies that, if $m \leq 2^{n+1}$, the inclusion $\mathrm{Gr}_d(\mathbb{R}^{m-1}) \to \mathrm{Gr}_d(\mathbb{R}^m)$ induces an inclusion $K(n)_*(\mathrm{Gr}_d(\mathbb{R}^{m-1})) \to K(n)_*(\mathrm{Gr}_d(\mathbb{R}^m))$, as this is true in mod p homology. Thus

$$\bar{k}_n(C_d(\mathbb{R}^m)) = k_n(\mathrm{Gr}_d(\mathbb{R}^m)) - k_n(\mathrm{Gr}_d(\mathbb{R}^{m-1})) = \binom{m}{d} - \binom{m-1}{d} = \binom{m-1}{d-1}. \qquad \square$$

Theorem 6.2 *Let $m = 2^{n+1} - \epsilon + 2l$ with $\epsilon = 0$ or 1, and $l \geq 0$. Then*

$$k_n(C_d(\mathbb{R}^m)) \geq \sum_{i=0}^{\lfloor d/2 \rfloor} \binom{2^{n+1}-1-\epsilon}{d-1-2i}\binom{l}{i}.$$

Proof The proof is similar to the proof of Theorem 1.2, with a little tweak.

If V is a real representation of C_4, and W is a subrepresentation, let $C_d(V, W)$ denote the cofiber of the inclusion $\mathrm{Gr}_d(W) \hookrightarrow \mathrm{Gr}_d(V)$; this is a based C_4 space.

If $\dim V = m$ and $\dim W = m - 1$, then $C_d(\mathbb{R}^m) = C_d(V, W)$ and thus $\bar{k}_n(C_d(\mathbb{R}^m)) \geq \bar{k}_n(C_d(V, W)^{C_4})$, by our chromatic fixed point theorem, Theorem 1.4. Furthermore, $C_d(V, W)^{C_4}$ will be the cofiber of the inclusion $\mathrm{Gr}_d(W)^{C_4} \hookrightarrow \mathrm{Gr}_d(V)^{C_4}$.

Now we choose V and W. Recall that L_1 and L_2 were the one-dimensional real representations of C_4 and R was the two-dimensional irreducible. We let $V = L_1^{2^n} \oplus L_2^{2^n-\epsilon} \oplus R^l$ and $W = L_1^{2^n-1} \oplus L_2^{2^n-\epsilon} \oplus R^l$.

Proposition 2.1 tells us that

$$\mathrm{Gr}_d(V)^{C_4} = \bigsqcup_{j+k+2i=d} \mathrm{Gr}_j(\mathbb{R}^{2^n}) \times \mathrm{Gr}_k(\mathbb{R}^{2^n-\epsilon}) \times \mathrm{Gr}_i(\mathbb{C}^l)$$

and

$$\mathrm{Gr}_d(W)^{C_4} = \bigsqcup_{j+k+2i=d} \mathrm{Gr}_j(\mathbb{R}^{2^n-1}) \times \mathrm{Gr}_k(\mathbb{R}^{2^n-\epsilon}) \times \mathrm{Gr}_i(\mathbb{C}^l),$$

so

$$C_d(V, W)^{C_4} = \bigvee_{j+k+2i=d} C_j(\mathbb{R}^{2^n}) \wedge \mathrm{Gr}_k(\mathbb{R}^{2^n-\epsilon})_+ \wedge \mathrm{Gr}_i(\mathbb{C}^l)_+.$$

Thus,

$$\bar{k}_n(C_d(\mathbb{R}^m)) \geq \sum_{j+k+2i=d} \bar{k}_{n-1}(C_j(\mathbb{R}^{2^n})) k_{n-1}(\mathrm{Gr}_k(\mathbb{R}^{2^n-\epsilon})) k_{n-1}(\mathrm{Gr}_i(\mathbb{C}^l))$$

$$= \sum_{j+k+2i=d} \binom{2^n-1}{j-1}\binom{2^n-\epsilon}{k}\binom{l}{i} \quad \text{(using Theorems 6.1 and 1.1)}$$

$$= \sum_i \left[\sum_{j+k=d-2i} \binom{2^n-1}{j-1}\binom{2^n-\epsilon}{k} \right]\binom{l}{i}$$

$$= \sum_i \binom{2^{n+1}-1-\epsilon}{d-1-2i}\binom{l}{i}. \qquad \square$$

Conjecture 6.3 *Equality holds in Theorem 6.2.*

As before, this would be implied by a conjectural calculation of the Q_n homology of $C_d(\mathbb{R}_n)$.

Conjecture 6.4 *Let $m = 2^{n+1} - \epsilon + 2l$ with $\epsilon = 0$ or 1, and $l \geq 0$. Then*

$$\bar{k}_{Q_n}(C_d(\mathbb{R}^m)) = \sum_i \binom{2^{n+1}-1-\epsilon}{d-1-2i}\binom{l}{i}.$$

Our various conjectures imply a conjecture about the behavior of the boundary map

$$\delta \colon H^*(\mathrm{Gr}_d(\mathbb{R}^{m-1}); Q_n) \to \tilde{H}^{*+2^{n+1}-1}(C_d(\mathbb{R}^m); Q_n)$$

when $m = 2^{n+1} - \epsilon + 2l$. Let $k_n^\delta(d, m)$ denote the dimension of the image of this map.

Conjecture 1.7 says that $k_{Q_n}(\mathrm{Gr}_d(\mathbb{R}^m)) = k_n^G(d, m)$, where

$$k_n^G(d, m) = \sum_i \binom{2^{n+1}-\epsilon}{d-2i}\binom{l}{i}.$$

Conjecture 6.4 similarly says that $\bar{k}_{Q_n}(C_d(\mathbb{R}^m)) = \bar{k}_n^C(d, m)$, where

$$\bar{k}_n^C(d, m) = \sum_i \binom{2^{n+1}-1-\epsilon}{d-1-2i}\binom{l}{i}.$$

If these conjectures are true, then the exactness of the Q_n–homology long exact sequence would imply that

$$k_n^G(d, m) + 2k_n^\delta(d, m) = k_n^G(d, m-1) + \bar{k}_n^C(d, m),$$

so that

$$k_n^\delta(d, m) = \tfrac{1}{2}\big[k_n^G(d, m-1) + \bar{k}_n^C(d, m) - k_n^G(d, m)\big].$$

As expected, the right hand side here is zero if m is even, ie $\epsilon = 0$.

When m is odd, so $\epsilon = 1$, the right hand side is not zero, but can be rearranged as in the following lemma.

Lemma 6.5 *If $m = 2^{n+1} - 1 + 2l$ and $l > 0$, then*

$$\tfrac{1}{2}\big[k_n^G(d, m-1) + \bar{k}_n^C(d, m) - k_n^G(d, m)\big] = \sum_i \binom{2^{n+1}-2}{d-1-2i}\binom{l-1}{i}.$$

Proof We expand $k_n^G(d, m-1)$:

$$k_n^G(d, m-1) = \sum_i \binom{2^{n+1}}{d-2i}\binom{l-1}{i}$$

$$= \sum_i \Big[\binom{2^{n+1}-2}{d-2i} + 2\binom{2^{n+1}-2}{d-1-2i} + \binom{2^{n+1}-2}{d-2-2i}\Big]\binom{l-1}{i}.$$

We rewrite $k_n^G(d,m) - \bar{k}_n^C(d,m)$:

$$
k_n^G(d,m) - \bar{k}_n^C(d,m) = \sum_i \left[\binom{2^{n+1}-1}{d-2i} - \binom{2^{n+1}-2}{d-1-2i} \right] \binom{l}{i}
$$

$$
= \sum_i \binom{2^{n+1}-2}{d-2i} \binom{l}{i}
$$

$$
= \sum_i \binom{2^{n+1}-2}{d-2i} \left[\binom{l-1}{i} + \binom{l-1}{i-1} \right]
$$

$$
= \sum_i \left[\binom{2^{n+1}-2}{d-2i} + \binom{2^{n+1}-2}{d-2-2i} \right] \binom{l-1}{i}.
$$

Subtracting our second expression from the first, and dividing by two, proves the lemma. $\qquad\square$

Thus we can add the following to our conjectures.

Conjecture 6.6 *If $m = 2^{n+1} - 1 + 2l$ and $l > 0$, then*

$$
k_n^\delta(d,m) = \sum_i \binom{2^{n+1}-2}{d-1-2i} \binom{l-1}{i}.
$$

Example 6.7 Suppose that $n = 0$, so $m = 2l + 1$. Conjecture 6.4 predicts that

$$
\dim_{\mathbb{Q}} H^*(C_d(\mathbb{R}^{2l+1}); \mathbb{Q}) = \begin{cases} 0 & \text{if } d \text{ is even,} \\ \binom{l}{c} & \text{if } d = 2c + 1. \end{cases}
$$

Similarly, Conjecture 6.6 predicts that

$$
k_0^\delta(d, 2l+1) = \begin{cases} 0 & \text{if } d \text{ is even,} \\ \binom{l-1}{c} & \text{if } d = 2c + 1. \end{cases}
$$

Noting that $k_0^\delta(d, 2l+1)$ can be viewed as the dimension of the cokernel of the map

$$
i^* \colon H^*(\mathrm{Gr}_d(\mathbb{R}^{2l+1}); \mathbb{Q}) \to H^*(\mathrm{Gr}_d(\mathbb{R}^{2l}); \mathbb{Q}),
$$

one can check that our conjectures do correspond to the known behavior of i^* — it takes Pontryagin classes to Pontryagin classes — together with the computations

$$
\dim_{\mathbb{Q}} H^*(\mathrm{Gr}_d(\mathbb{R}^m); \mathbb{Q}) = \begin{cases} \binom{l}{c} & \text{if } m = 2l + 1 \text{ and } d = 2c \text{ or } 2c + 1, \\ 2\binom{l-1}{c} & \text{if } m = 2l \text{ and } d = 2c + 1. \end{cases}
$$

Appendix   Tables

We present some tables of calculations made by the second author that support Conjecture 1.7. Calculational algorithms used are documented in [10, Appendix B]. For larger Grassmannians the authors used the University of Virginia Rivanna high-performance computing system. The white cells are the conjectured values which have not been checked due to computational limitations. The tables are necessarily symmetric in c and d.

c / d	1	2	3	4	5	6	7	8	9	10	11
1	2	3	4	3	4	3	4	3	4	3	4
2	3	6	4	7	5	8	6	9	7	10	8
3	4	4	8	7	12	10	16	13	20	16	24
4	3	7	7	14	12	22	18	31	25	41	33
5	4	5	12	12	24	22	40	35	60	51	84
6	3	8	10	22	22	44	40	75	65	116	98
7	4	6	16	18	40	40	80	75	140	126	224
8	3	9	13	31	35	75	75	150	140	266	238
9	4	7	20	25	60	65	140	140	280	266	504
10	3	10	16	41	51	116	126	266	266	532	504
11	4	8	24	33	84	98	224	238	504	504	1008
12	3	11	19	52	70	168	196	434	462	966	966
13	4	9	28	42	112	140	336	378	840	882	1848
14	3	12	22	64	92	232	288	666	750	1632	1716
15	4	10	32	52	144	192	480	570	1320	1452	3168
16	3	13	25	77	117	309	405	975	1155	2607	2871
17	4	11	36	63	180	255	660	825	1980	2277	5148
18	3	14	28	91	145	400	550	1375	1705	3982	4576
19	4	12	40	75	220	330	880	1155	2860	3432	8008
20	3	15	31	106	176	506	726	1881	2431	5863	7007
21	4	13	44	88	264	418	1144	1573	4004	5005	12012
22	3	16	34	122	210	628	936	2509	3367	8372	10374
23	4	14	48	102	312	520	1456	2093	5460	7098	17472
24	3	17	37	139	247	767	1183	3276	4550	11648	14924
25	4	15	52	117	364	637	1820	2730	7280	9828	24752
26	3	18	40	157	287	924	1470	4200	6020	15848	20944
27	4	16	56	133	420	770	2240	3500	9520	13328	34272
28	3	19	43	176	330	1100	1800	5300	7820	21148	28764
29	4	17	60	150	480	920	2720	4420	12240	17748	46512
30	3	20	46	196	376	1296	2176	6596	9996	27744	38760
31	4	18	64	168	544	1088	3264	5508	15504	23256	62016
32	3	21	49	217	425	1513	2601	8109	12597	35853	51357
33	4	19	68	187	612	1275	3876	6783	19380	30039	81396
34	3	22	52	239	477	1752	3078	9861	15675	45714	67032
35	4	20	72	207	684	1482	4560	8265	23940	38304	105336
36	3	23	55	262	532	2014	3610	11875	19285	57589	86317
37	4	21	76	228	760	1710	5320	9975	29260	48279	134596
38	3	24	58	286	590	2300	4200	14175	23485	71764	109802
39	4	22	80	250	840	1960	6160	11935	35420	60214	170016
40	3	25	61	311	651	2611	4851	16786	28336	88550	138138
41	4	23	84	273	924	2233	7084	14168	42504	74382	212520
42	3	26	64	337	715	2948	5566	19734	33902	108284	172040
43	4	24	88	297	1012	2530	8096	16698	50600	91080	263120
44	3	27	67	364	782	3312	6348	23046	40250	131330	212290
45	4	25	92	322	1104	2852	9200	19550	59800	110630	322920
46	3	28	70	392	852	3704	7200	26750	47450	158080	259740
47	4	26	96	348	1200	3200	10400	22750	70200	133380	393120
48	3	29	73	421	925	4125	8125	30875	55575	188955	315315
49	4	27	100	375	1300	3575	11700	26325	81900	159705	475020
50	3	30	76	451	1001	4576	9126	35451	64701	224406	380016

▨ $cd \leq 2^{n+1} - 1$	▨ projective spaces	▨ Theorem 1.8
▨ conjecture verified	▨ conjecture	

Table 1: $k_1(\mathrm{Gr}_d(\mathbb{R}^{d+c}))$.

c \ d	1	2	3	4	5	6	7	8	9	10	11
1	2	3	4	5	6	7	8	7	8	7	8
2	3	6	10	15	21	28	22	29	23	30	24
3	4	10	20	35	56	42	64	49	72	56	80
4	5	15	35	70	56	98	78	127	101	157	125
5	6	21	56	56	112	98	176	147	248	203	328
6	7	28	42	98	98	196	176	323	277	480	402
7	8	22	64	78	176	176	352	323	600	526	928
8	7	29	49	127	147	323	323	646	600	1126	1002
9	8	23	72	101	248	277	600	600	1200	1126	2128
10	7	30	56	157	203	480	526	1126	1126	2252	2128
11	8	24	80	125	328	402	928	1002	2128	2128	4256
12	7	31	63	188	266	668	792	1794	1918	4046	4046
13	8	25	88	150	416	552	1344	1554	3472	3682	7728
14	7	32	70	220	336	888	1128	2682	3046	6728	7092
15	8	26	96	176	512	728	1856	2282	5328	5964	13056
16	7	33	77	253	413	1141	1541	3823	4587	10551	11679
17	8	27	104	203	616	931	2472	3213	7800	9177	20856
18	7	34	84	287	497	1428	2038	5251	6625	15802	18304
19	8	28	112	231	728	1162	3200	4375	11000	13552	31856
20	7	35	91	322	588	1750	2626	7001	9251	22803	27555
21	8	29	120	260	848	1422	4048	5797	15048	19349	46904
22	7	36	98	358	686	2108	3312	9109	12563	31912	40118
23	8	30	128	290	976	1712	5024	7509	20072	26858	66976
24	7	37	105	395	791	2503	4103	11612	16666	43524	56784
25	8	31	136	321	1112	2033	6136	9542	26208	36400	93184
26	7	38	112	433	903	2936	5006	14548	21672	58072	78456
27	8	32	144	353	1256	2386	7392	11928	33600	48328	126784
28	7	39	119	472	1022	3408	6028	17956	27700	76028	106156
29	8	33	152	386	1408	2772	8800	14700	42400	63028	169184
30	7	40	126	512	1148	3920	7176	21876	34876	97904	141032
31	8	34	160	420	1568	3192	10368	17892	52768	80920	221952
32	7	41	133	553	1281	4473	8457	26349	43333	124253	184365
33	8	35	168	455	1736	3647	12104	21539	64872	102459	286824
34	7	42	140	595	1421	5068	9878	31417	53211	155670	237576
35	8	36	176	491	1912	4138	14016	25677	78888	128136	365712
36	7	43	147	638	1568	5706	11446	37123	64657	192793	302233
37	8	37	184	528	2096	4666	16112	30343	95000	158479	460712
38	7	44	154	682	1722	6388	13168	43511	77825	236304	380058
39	8	38	192	566	2288	5232	18400	35575	113400	194054	574112
40	7	45	161	727	1883	7115	15051	50626	92876	286930	472934
41	8	39	200	605	2488	5837	20888	41412	134288	235466	708400
42	7	46	168	773	2051	7888	17102	58514	109978	345444	582912
43	8	40	208	645	2696	6482	23584	47894	157872	283360	866272
44	7	47	175	820	2226	8708	19328	67222	129306	412666	712218
45	8	41	216	686	2912	7168	26496	55062	184368	338422	1050640
46	7	48	182	868	2408	9576	21736	76798	151042	489464	863260
47	8	42	224	728	3136	7896	29632	62958	214000	401380	1264640
48	7	49	189	917	2597	10493	24333	87291	175375	576755	1038635
49	8	43	232	771	3368	8667	33000	71625	247000	473005	1511640
50	7	50	196	967	2793	11460	27126	98751	202501	675506	1241136

☐ $cd \leq 2^{n+1} - 1$ ▨ projective spaces   ☐ Theorem 1.8
☐ conjecture verified   ☐ conjecture   ☐ Theorem 1.1

Table 2: $k_2(\mathrm{Gr}_d(\mathbb{R}^{d+c}))$.

c \ d	1	2	3	4	5	6	7	8	9	10	11
1	2	3	4	5	6	7	8	9	10	11	12
2	3	6	10	15	21	28	36	45	55	66	78
3	4	10	20	35	56	84	120	165	220	286	364
4	5	15	35	70	126	210	330	495	715	1001	1365
5	6	21	56	126	252	462	792	1287	2002	3003	4368
6	7	28	84	210	462	924	1716	3003	5005	8008	6370
7	8	36	120	330	792	1716	3432	6435	11440	9438	15808
8	9	45	165	495	1287	3003	6435	12870	11440	20878	17810
9	10	55	220	715	2002	5005	11440	11440	22880	20878	38688
10	11	66	286	1001	3003	8008	9438	20878	20878	41756	38688
11	12	78	364	1365	4368	6370	15808	17810	38688	38688	77376
12	13	91	455	1820	3458	9828	12896	30706	33774	72462	72462
13	14	105	560	1470	4928	7840	20736	25650	59424	64338	136800
14	15	120	470	1940	3928	11768	16824	42474	50598	114936	123060
15	16	106	576	1576	5504	9416	26240	35066	85664	99404	222464
16	15	121	485	2061	4413	13829	21237	56303	71835	171239	194895
17	16	107	592	1683	6096	11099	32336	46165	118000	145569	340464
18	15	122	500	2183	4913	16012	26150	72315	97985	243554	292880
19	16	108	608	1791	6704	12890	39040	59055	157040	204624	497504
20	15	123	515	2306	5428	18318	31578	90633	129563	334187	422443
21	16	109	624	1900	7328	14790	46368	73845	203408	278469	700912
22	15	124	530	2430	5958	20748	37536	111381	167099	445568	589542
23	16	110	640	2010	7968	16800	54336	90645	257744	369114	958656
24	15	125	545	2555	6503	23303	44039	134684	211138	580252	800680
25	16	111	656	2121	8624	18921	62960	109566	320704	478680	1279360
26	15	126	560	2681	7063	25984	51102	160668	262240	740920	1062920
27	16	112	672	2233	9296	21154	72256	130720	392960	609400	1672320
28	15	127	575	2808	7638	28792	58740	189460	320980	930380	1383900
29	16	113	688	2346	9984	23500	82240	154220	475200	763620	2147520
30	15	128	590	2936	8228	31728	66968	221188	387948	1151568	1771848
31	16	114	704	2460	10688	25960	92928	180180	568128	943800	2715648
32	15	129	605	3065	8833	34793	75801	255981	463749	1407549	2235597
33	16	115	720	2575	11408	28535	104336	208715	672464	1152515	3388112
34	15	130	620	3195	9453	37988	85254	293969	549003	1701518	2784600
35	16	116	736	2691	12144	31226	116480	239941	788944	1392456	4177056
36	15	131	635	3326	10088	41314	95342	335283	644345	2036801	3428945
37	16	117	752	2808	12896	34034	129376	273975	918320	1666431	5095376
38	15	132	650	3458	10738	44772	106080	380055	750425	2416856	4179370
39	16	118	768	2926	13664	36960	143040	310935	1061360	1977366	6156736
40	15	133	665	3591	11403	48363	117483	428418	867908	2845274	5047278
41	16	119	784	3045	14448	40005	157488	350940	1218848	2328306	7375584
42	15	134	680	3725	12083	52088	129566	480506	997474	3325780	6044752
43	16	120	800	3165	15248	43170	172736	394110	1391584	2722416	8767168
44	15	135	695	3860	12778	55948	142344	536454	1139818	3862234	7184570
45	16	121	816	3286	16064	46456	188800	440566	1580384	3162982	10347552
46	15	136	710	3996	13488	59944	155832	596398	1295650	4458632	8480220
47	16	122	832	3408	16896	49864	205696	490430	1786080	3653412	12133632
48	15	137	725	4133	14213	64077	170045	660475	1465695	5119107	9945915
49	16	123	848	3531	17744	53395	223440	543825	2009520	4197237	14143152
50	15	138	740	4271	14953	68348	184998	728823	1650693	5847930	11596608

Legend:
- $cd \leq 2^{n+1} - 1$
- conjecture verified
- projective spaces
- conjecture
- Theorem 1.8
- Theorem 1.1

Table 3: $k_3(\mathrm{Gr}_d(\mathbb{R}^{d+c}))$.

d \ c	1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10	11
2	3	6	10	15	21	28	36	45	55	66
3	4	10	20	35	56	84	120	165	220	286
4	5	15	35	70	126	210	330	495	715	1001
5	6	21	56	126	252	462	792	1287	2002	3003
6	7	28	84	210	462	924	1716	3003	5005	8008
7	8	36	120	330	792	1716	3432	6435	11440	19448
8	9	45	165	495	1287	3003	6435	12870	24310	43758
9	10	55	220	715	2002	5005	11440	24310	48620	92378
10	11	66	286	1001	3003	8008	19448	43758	92378	184756
11	12	78	364	1365	4368	12376	31824	75582	167960	352716
12	13	91	455	1820	6188	18564	50388	125970	293930	646646
13	14	105	560	2380	8568	27132	77520	203490	497420	1144066
14	15	120	680	3060	11628	38760	116280	319770	817190	1961256
15	16	136	816	3876	15504	54264	170544	490314	1307504	3268760
16	17	153	969	4845	20349	74613	245157	735471	2042975	5311735
17	18	171	1140	5985	26334	100947	346104	1081575	3124550	8436285
18	19	190	1330	7315	33649	134596	480700	1562275	4686825	13123110
19	20	210	1540	8855	42504	177100	657800	2220075	6906900	20030010
20	21	231	1771	10626	53130	230230	888030	3108105	10015005	30045015
21	22	253	2024	12650	65780	296010	1184040	4292145	14307150	44352165
22	23	276	2300	14950	80730	376740	1560780	5852925	20160075	64512240
23	24	300	2600	17550	98280	475020	2035800	7888725	28048800	52240890
24	25	325	2925	20475	118755	593775	2629575	10518300	22789650	75030540
25	26	351	3276	23751	142506	736281	3365856	8625006	31414656	60865896
26	27	378	3654	27405	169911	906192	2799486	11424492	25589136	86455032
27	28	406	4060	31465	201376	767746	3567232	9392752	34981888	70258648
28	29	435	4495	35960	174406	942152	2973892	12366644	28563028	98821676
29	30	465	4960	31930	206336	799676	3773568	10192428	38755456	80451076
30	31	496	4526	36456	178932	978608	3152824	13345252	31715852	112166928
31	32	466	4992	32396	211328	832072	3984896	11024500	42740352	91475576
32	31	497	4557	36953	183489	1015561	3336313	14360813	35052165	126527741
33	32	467	5024	32863	216352	864935	4201248	11889435	46941600	103365011
34	31	498	4588	37451	188077	1053012	3524390	15413825	38576555	141941566
35	32	468	5056	33331	221408	898266	4422656	12787701	51364256	116152712
36	31	499	4619	37950	192696	1090962	3717086	16504787	42293641	158446353
37	32	469	5088	33800	226496	932066	4649152	13719767	56013408	129872479
38	31	500	4650	38450	197346	1129412	3914432	17634199	46208073	176080552
39	32	470	5120	34270	231616	966336	4880768	14686103	60894176	144558582
40	31	501	4681	38951	202027	1168363	4116459	18802562	50324532	194883114
41	32	471	5152	34741	236768	1001077	5117536	15687180	66011712	160245762
42	31	502	4712	39453	206739	1207816	4323198	20010378	54647730	214893492
43	32	472	5184	35213	241952	1036290	5359488	16723470	71371200	176969232
44	31	503	4743	39956	211482	1247772	4534680	21258150	59182410	236151642
45	32	473	5216	35686	247168	1071976	5606656	17795446	76977856	194764678
46	31	504	4774	40460	216256	1288232	4750936	22546382	63933346	258698024
47	32	474	5248	36160	252416	1108136	5859072	18903582	82836928	213668260
48	31	505	4805	40965	221061	1329197	4971997	23875579	68905343	282573603
49	32	475	5280	36635	257696	1144771	6116768	20048353	88953696	233716613
50	31	506	4836	41471	225897	1370668	5197894	25246247	74103237	307819850

$cd \leq 2^{n+1} - 1$		projective spaces		Theorem 1.8
conjecture verified		conjecture		Theorem 1.1

Table 4: $k_4(\mathrm{Gr}_d(\mathbb{R}^{d+c}))$.

References

[1]   **W Balderrama**, **N J Kuhn**, *An elementary proof of the chromatic Smith fixed point theorem*, Homology Homotopy Appl. 26 (2024) 131–140   MR Zbl

[2]   **T Barthel**, **M Hausmann**, **N Naumann**, **T Nikolaus**, **J Noel**, **N Stapleton**, *The Balmer spectrum of the equivariant homotopy category of a finite abelian group*, Invent. Math. 216 (2019) 215–240   MR Zbl

[3]   **C Ehresmann**, *Sur la topologie de certaines variétés algébriques réelles*, J. Math. Pures Appl. 16 (1937) 69–100   Zbl

[4]   **E E Floyd**, *On periodic maps and the Euler characteristics of associated spaces*, Trans. Amer. Math. Soc. 72 (1952) 138–147   MR Zbl

[5]   **J Jaworowski**, *An additive basis for the cohomology of real Grassmannians*, from "Algebraic topology Poznań 1989" (S Jackowski, B Oliver, K Pawałowski, editors), Lecture Notes in Math. 1474, Springer (1991) 231–234   MR Zbl

[6]   **N Kitchloo**, **W S Wilson**, *The Morava K–theory of $BO(q)$ and $MO(q)$*, Algebr. Geom. Topol. 15 (2015) 3049–3058   MR Zbl

[7]   **A Kono**, **N Yagita**, *Brown–Peterson and ordinary cohomology theories of classifying spaces for compact Lie groups*, Trans. Amer. Math. Soc. 339 (1993) 781–798   MR Zbl

[8]   **N J Kuhn**, **C J R Lloyd**, *Chromatic fixed point theory and the Balmer spectrum for extraspecial 2–groups*, preprint (2020)   arXiv 2008.00330   To appear in Amer. J. Math.

[9]   **C Lenart**, *The combinatorics of Steenrod operations on the cohomology of Grassmannians*, Adv. Math. 136 (1998) 251–283   MR Zbl

[10] **C J R Lloyd**, *Applications of chromatic fixed point theory*, PhD thesis, University of Virginia (2021) Available at https://libraetd.lib.virginia.edu/public_view/h702q715z

[11] **C R F Maunder**, *The spectral sequence of an extraordinary cohomology theory*, Proc. Cambridge Philos. Soc. 59 (1963) 567–574   MR Zbl

[12] **J W Milnor**, **J D Stasheff**, *Characteristic classes*, Annals of Mathematics Studies No. 76, Princeton Univ. Press (1974)   MR Zbl

[13] **B Schuster**, *Morava K–theory of groups of order 32*, Algebr. Geom. Topol. 11 (2011) 503–521   MR Zbl

[14] **U Würgler**, *Morava K–theories: a survey*, from "Algebraic topology" (S Jackowski, B Oliver, K Pawałowski, editors), Lecture Notes in Math. 1474, Springer (1991) 111–138   MR Zbl

[15] **N Yagita**, *On the Steenrod algebra of Morava K–theory*, J. London Math. Soc. 22 (1980) 423–438   MR Zbl

Department of Mathematics, University of Virginia
Charlottesville, VA, United States
Arlington, VA, United States

njk4x@virginia.edu,   cjl8zf@virginia.edu

Slope gap distributions of Veech surfaces

Luis Kumanduri

Anthony Sanchez

Jane Wang

The slope gap distribution of a translation surface is a measure of how random the directions of the saddle connections on the surface are. It is known that Veech surfaces, a highly symmetric type of translation surface, have gap distributions that are piecewise real analytic. Beyond that, however, very little is currently known about the general behavior of the slope gap distribution, including the number of points of nonanalyticity or the tail.

We show that the limiting gap distribution of slopes of saddle connections on a Veech translation surface is always piecewise real analytic with *finitely* many points of nonanalyticity. We do so by taking an explicit parametrization of a Poincaré section to the horocycle flow on $\mathrm{SL}(2, \mathbb{R})/\mathrm{SL}(X, \omega)$ associated to an arbitrary Veech surface (X, ω), and establishing a key finiteness result for the first return map under this flow. We use the finiteness result to show that the tail of the slope gap distribution of a Veech surface always has quadratic decay.

1 Introduction

We will study the slope gap distributions of Veech surfaces, a highly symmetric type of translation surface. *Translation surfaces* can be defined geometrically as finite collections of polygons with sides identified in parallel opposite pairs. If we place these polygons in the complex plane $\mathbb{C}$, the surface inherits a Riemann surface structure from $\mathbb{C}$, and the one-form dz gives rise to a well-defined holomorphic one-form on the surface. This leads to a second equivalent definition of a translation surfaces as a pair (X, ω) where X is a Riemann surface and ω is a holomorphic one-form on the surface. Every translation surface locally has the structure of $(\mathbb{C}, dz)$, except for at finitely many points that have total angle around them $2\pi k$ for some integer $k \geq 2$. These points are called *cone points* and correspond to the zeros of the one-form ω. A zero of order k gives rise to a cone point of angle $2\pi(k + 1)$.

A translation surface inherits a flat metric from $\mathbb{C}$. *Saddle connections* are then straight-line geodesics connecting two cone points with no cone points in the interior. The *holonomy vector* of a saddle connection γ is then the vector describing how far and in what direction the saddle connection travels:

$$\boldsymbol{v}_\gamma = \int_\gamma \omega.$$

We will be interested in the distribution of directions of these vectors for various translation surfaces.
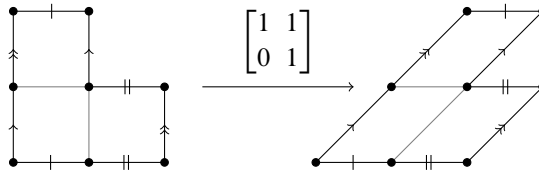
Figure 1: A matrix in $\mathrm{SL}(2, \mathbb{R})$ acting on a translation surface.

There is a natural $\mathrm{SL}(2, \mathbb{R})$ action on translation surfaces coming from the linear action of matrices on $\mathbb{R}^2$, as can be seen in Figure 1.

Sometimes this action produces a symmetry of the surface (X, ω). That is, after acting on the surface by the matrix, it is possible to cut and paste the new surface so that it looks like the original surface again. The collection of these symmetries is the stabilizer under the $\mathrm{SL}(2, \mathbb{R})$ action and is called the *Veech group* of the surface. It will be denoted by $\mathrm{SL}(X, \omega)$ and is a subgroup of $\mathrm{SL}(2, \mathbb{R})$. When the Veech group $\mathrm{SL}(X, \omega)$ of a translation surface has finite covolume in $\mathrm{SL}(2, \mathbb{R})$, the surface (X, ω) is called a *Veech surface*. Sometimes such surfaces are also called lattice surfaces since $\mathrm{SL}(X, \omega)$ is a lattice in $\mathrm{SL}(X, \mathbb{R})$. Veech surfaces have many nice properties, such as satisfying the *Veech dichotomy*: in any direction, every infinite trajectory on the surface is periodic or every infinite trajectory is equidistributed. For more information about translation and Veech surfaces see Hubert and Schmidt [6] and Zorich [13].

From work of Vorobets [12], it is known that, for almost every translation surface (X, ω) with respect to the Masur–Veech volume on any strata of translation surfaces (for details about Masur–Veech volume and strata, please see [13]), the angles of the saddle connections equidistribute in S^1. That is, if we let

$$\Lambda(X, \omega) := \{\text{holonomy vectors of saddle connections of } (X, \omega)\}$$

and normalize the circle to have total length 1, then for any interval $I \subset S^1$, as we let $R \to \infty$, the proportion of vectors in $\Lambda(X, \omega)$ of length $\leq R$ that have direction in the interval I converges to the length of I.

A finer measure of the randomness of the saddle connection directions of a surface is its *gap distribution*, which we will now define. The idea of the gap distribution is that it records the limiting distribution of the spacings between the set of angles (or in our case, slopes) of the saddle connection directions of length up to a certain length R. We will be working with slope gap distributions rather than angle gap distributions because the slope gap distribution has deep ties to the horocycle flow on strata of translation surfaces. Thus, dynamical tools relating to the horocycle flow can be more easily applied to analyze the slope gap distribution.

Let us restrict our attention to the first quadrant and to slopes of at most 1, and define

$$\mathbb{S}(X, \omega) := \{\text{slope}(\boldsymbol{v}) \mid \boldsymbol{v} \in \Lambda(X, \omega), \ 0 < \mathrm{Re}(\boldsymbol{v}) \text{ and } 0 \leq \mathrm{Im}(\boldsymbol{v}) \leq \mathrm{Re}(\boldsymbol{v})\}.$$

We also allow ourselves to restrict to slopes of saddle connections of at most some length R in the ℓ_∞ metric, and define

$$\mathbb{S}_R(X, \omega) := \{\text{slope}(\boldsymbol{v}) \mid \boldsymbol{v} \in \Lambda(X, \omega), \ 0 < \mathrm{Re}(\boldsymbol{v}) \text{ and } 0 \leq \mathrm{Im}(\boldsymbol{v}) \leq \mathrm{Re}(\boldsymbol{v}) \leq R\}.$$

We let $N(R)$ denote the number of unique slopes $N(R) := |\mathbb{S}_R(X, \omega)|$. By results of Masur [7; 8], the growth of the number of saddle connections of length at most R in any translation surface is quadratic in R. We can order the slopes:

$$\mathbb{S}_R(X, \omega) = \{0 \leq s_0^R < s_1^R < \cdots < s_{N(R)-1}^R\}.$$

Since $N(R)$ grows quadratically in R, we now define the *renormalized slope gaps* of (X, ω) to be

$$\mathbb{G}_R(X, \omega) := \{R^2(s_i^R - s_{i-1}^R) \mid 1 \leq i \leq N(R) - 1 \text{ and } s_i \in \mathbb{S}_R(X, \omega)\}.$$

If there exists a limiting probability distribution function $f : [0, \infty) \to [0, \infty)$ for the renormalized slope gaps

$$\lim_{R \to \infty} \frac{|\mathbb{G}_R(X, \omega) \cap (a, b)|}{N(R)} = \int_a^b f(x)\, dx,$$

then f is called the *slope gap distribution* of the translation surface (X, ω). If the sequence of slopes of holonomy vectors of increasing length of a translation surface were independent and identically distributed uniform $[0, 1]$ random variables, then a probability computation shows that the gap distribution would be a Poisson process of intensity 1. In all computed examples of slope gap distributions, however, this is not the case.

We give a brief overview of the literature on gap distributions of translation surfaces. In [2], Athreya and Chaika analyzed the gap distributions for typical surfaces and showed that, for almost every translation surface (with respect to the Masur–Veech volume), the gap distribution exists. They also showed that a translation surface is a Veech surface if and only if it has *no small gaps*, that is, if $\liminf_{R \to \infty}(\min(\mathbb{G}_R(X, \omega))) > 0$. In a later work [3], Athreya, Chaika and Lelièvre explicitly computed the gap distribution of the golden L, and in [1] Athreya gives an overview of results and techniques about gap distributions. Another relevant work is a paper by Taha [10] studying cross sections to the horocycle and geodesic flows on quotients of $\mathrm{SL}(2, \mathbb{R})$ by Hecke triangle groups. The computation of slope gap distributions involved understanding the first return map of the horocycle flow to a particular transversal of a quotient of $\mathrm{SL}(2, \mathbb{R})$.

In [11], Uyanik and Work computed the gap distribution of the octagon, and also showed that the gap distribution of any Veech surface exists and is piecewise real analytic. In [9], Sanchez went on to study the gap distributions of doubled slit tori. Up until then, all known slope gap distributions were for Veech surfaces. The above articles focus on gap distributions of *specific* translation surfaces and their $\mathrm{SL}(2, \mathbb{R})$ orbits. This work applies to any Veech surface and gives insight to the general behavior of the graph of the slope gap distribution of Veech surfaces. In fact, outside of [2], where it is shown that there are no small gaps, there is no other work in this direction with this level of generality.

Uyanik and Work gave an algorithm to compute the gap distribution of any Veech surface and showed that the gap distribution was piecewise analytic. However, their algorithm does not necessarily terminate in finite time and can make it seem like the gap distribution can have infinitely many points of nonanalyticity,

as we will see in Section 2.3. We improve upon their algorithm to guarantee termination in finite time and show as a result that every Veech translation surface has a gap distribution with finitely many points of nonanalyticity. Uyanik and Work's algorithm starts by taking a transversal to the horocycle flow which a priori may break up into infinitely many components under the return map. Our key observation is that, by carefully choosing this transversal using the geometry of our surface, it will only break up into finitely many pieces, which will give the following theorem:

Theorem 1 *The slope gap distribution of any Veech surface has finitely many points of nonanalyticity.*

In addition, we show that the tail of the gap distribution of any Veech surface has a quadratic decay. Let $f(t) \sim g(t)$ mean that the ratio is bounded above and below by two positive constants as $t \to \infty$.

Theorem 2 *The slope gap distribution of any Veech surface has quadratic tail decay. That is, if f denotes the density function of the slope gap distribution, then*

$$\int_t^\infty f(x)\,dx \sim t^{-2}.$$

Thus, our results and the "no small gaps" result of [2] give a good understanding of the graph of the slope gap distribution of Veech surfaces: for some time the graph is identically zero before becoming positive. Afterward the graph has finitely many pieces where it is real analytic and may fluctuate up and down before it begins permanently decaying quadratically.

Organization   In Section 2.1 we will go over background information on slope gap distributions, including how to relate the gap distribution to return times to a Poincaré section of the horocycle flow. In Section 2.2, we will outline the algorithm of Uyanik and Work, and observe some possible modifications. In Section 2.3, we will see how a couple steps of Uyanik and Work's algorithm apply to a specific Veech surface. A priori, the first return map to the Poincaré section breaks the section into infinitely many pieces, but after making some modifications to the parametrization we will see that there are in fact finitely many pieces. In Section 3 we will give a proof of Theorem 1. The strategy of the proof is to apply a compactness argument to show finiteness under our modified parametrization of the Poincaré section. We will show that, on a compact set that includes the Poincaré section, every point has a neighborhood that can contribute at most finitely many points of nonanalyticity to the gap distribution. This will give us that the slope gap distribution has finitely many points of nonanalyticity overall. In Section 4, as an application of Theorem 1, we prove quadratic decay of the slope gap distribution of Veech surfaces. Finally, in Section 5 we discuss a few further questions regarding slope gap distributions of translation surfaces.

The authors are grateful to the referee for a careful reading of the manuscript and many useful suggestions.

2 Background

2.1 A Poincaré section for the horocycle flow

In this section, we review a general strategy for computing the gap distribution of a translation surface by relating slope gap distributions to the horocycle flow. For more background and proofs of the statements given here, see [4] or [3].

Suppose that we wish to compute the slope gap distribution of a translation surface (X, ω). We let $\Lambda(X, \omega)$, sometimes shortened to just Λ, be the set of holonomy vectors of the surface. We may start by considering all of the holonomy vectors of (X, ω) in the first quadrant, with ℓ_∞ norm $\leq R$. If we act on (X, ω) by the matrix

$$g_{-2\log(R)} = \begin{bmatrix} 1/R & 0 \\ 0 & R \end{bmatrix},$$

the slopes of the holonomy vectors of $g_{-2\log(R)}(X, \omega)$ in $[0, 1] \times [0, R^2]$ are the same as R^2 times the slopes of the holonomy vectors of (X, ω) in $[0, R] \times [0, R]$, as we can see in Figure 2.

Another important observation is that the horocycle flow

$$h_s = \begin{bmatrix} 1 & 0 \\ -s & 1 \end{bmatrix}$$
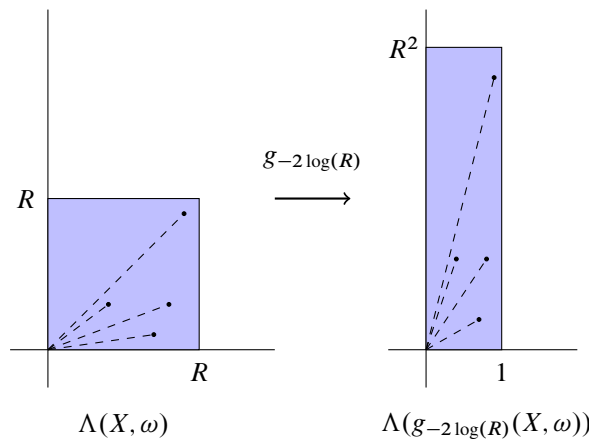
Figure 2: Upon renormalizing a surface (X, ω) by applying $g_{-2\log(R)}$, the slopes of the saddle connections of (X, ω) scale by R^2.

changes slopes of holonomy vectors by s. That is,

$$\text{slope}(h_s(z)) = \text{slope}(z) - s$$

for $z \in \mathbb{R}^2$. As a result, slope differences are preserved by the flow h_s.

Now we let the Veech group of the surface be $\text{SL}(X, \omega)$ and we define a *Poincaré section* or transversal for horocycle flow on $\text{SL}(2, \mathbb{R})/\text{SL}(X, \omega)$. By transversal we mean a subset such that almost every orbit under horocycle flow intersects that subset in a nonempty countable discrete set of times. Two key related notions are given by the *return time* of a point in the transversal, which records how long it takes to return, and the *return map*, which outputs what the point has returned to in the transversal after flowing by the return time. Each of these are explicit in our situation and will be described below.

We consider the transversal given by the surfaces in the $\text{SL}(2, \mathbb{R})$ orbit of (X, ω) with a short horizontal saddle connection of length ≤ 1. That is,

$$\Omega(X, \omega) = \{g \, \text{SL}(X, \omega) \mid g\Lambda \cap ((0, 1] \times \{0\}) \neq \varnothing\}.$$

By [1, Lemma 2.1], $\Omega(X, \omega)$ indeed is transversal for horocycle flow.

Then the slope gaps of (X, ω) for holonomy vectors of ℓ_∞ length $\leq R$ are exactly $1/R^2$ times the set of $N(R) - 1$ first return times to $\Omega(X, \omega)$ of the surface $g_{-2\log(R)}(X, \omega)$ under the horocycle flow h_s for $s \in [0, R^2]$. Here we are thinking of return times as the amount of time between each two successive times that the horocycle flow returns to the Poincaré section. In this way, the slope gaps of (X, ω) are related to the return times of the horocycle flow to the Poincaré section. Summarizing, since $\mathbb{G}_R(X, \omega)$ is the set of slope gaps renormalized by R^2, we have that

$$\mathbb{G}_R(X, \omega) = \{\text{first } N(R) - 1 \text{ return times of } g_{-2\log(R)}(X, \omega) \text{ to } \Omega(X, \omega) \text{ under } h_s\}.$$

For a point z in the Poincaré section $\Omega(X, \omega)$, we denote by $R_h(z)$ the return time of z to $\Omega(X, \omega)$ under the horocycle flow. Then as one lets $R \to \infty$, this renormalization procedure gives us that

$$\lim_{R \to \infty} \frac{|\mathbb{G}_R(X, \omega) \cap (a, b)|}{N(R)} = \mu\{z \in \Omega(X, \omega) \mid R_h(z) \in (a, b)\},$$

where μ is the unique ergodic probability measure on $\Omega(X, \omega)$ for which the first return map under h_s is not supported on a periodic orbit. Computing the slope gap distribution then reduces to finding a convenient parametrization of the Poincaré section for the horocycle flow on $\text{SL}(2, \mathbb{R})/\text{SL}(X, \omega)$, a suitable measure on this parametrization, and the first return time function to this the Poincaré section.

We note that this last point also makes it clear that every surface in the $\text{SL}(2, \mathbb{R})$ orbit of a Veech surface has the same slope gap distribution. We also note that scaling the surface by c scales the gap distribution from $f(x)$ to $(1/c^4)f(x/c^2)$; see [11] for a proof of this latter fact.

2.2 Computing gap distributions for Veech surfaces

In [11], Uyanik and Work developed a general algorithm for computing the slope gap distribution for Veech surfaces. In particular, their algorithm finds a parametrization for the Poincaré section of any Veech surface and calculates the gap distribution by examining the first return time of the horocycle flow to this Poincaré section. In this section, we'll go over the basics of this algorithm. For more details about this algorithm as well as a proof of why it works, please see Uyanik and Work's original paper.

We start by supposing that (X, ω) is a Veech surface with $n < \infty$ cusps. Then we let $\Gamma_1, \ldots, \Gamma_n$ be representatives of the conjugacy classes of maximal parabolic subgroups of $\mathrm{SL}(X, \omega)$. We are going to find a piece of the Poincaré section for each parabolic subgroup Γ_i. The idea here is that the set of shortest holonomy vectors of (X, ω) in each direction breaks up into $\bigcup_{i=1}^{n} \mathrm{SL}(X, \omega) \boldsymbol{v}_i$, where the $\boldsymbol{v}_i$ vectors are in the eigendirections of the generators of each Γ_i.

The Poincaré section is given by those elements $g \in \mathrm{SL}(X, \mathbb{R})/\mathrm{SL}(X, \omega)$ such that $g(X, \omega)$ has a short (length ≤ 1) horizontal holonomy vector:

$$\Omega(X, \omega) = \{g \, \mathrm{SL}(X, \omega) \mid g\Lambda \cap ((0, 1] \times \{0\}) \neq \varnothing\}.$$

Here Λ is the set of holonomy vectors of (X, ω). Up to the action of $\mathrm{SL}(X, \omega)$, these short horizontal holonomy vectors are then just $g\boldsymbol{v}_i$ for a unique $\boldsymbol{v}_i$.

So $\Omega(X, \omega)$ then breaks up into a piece for each Γ_i, which we can parametrize as follows, depending on whether $-I \in \mathrm{SL}(X, \omega)$.

Case 1 $(-I \in \mathrm{SL}(X, \omega))$ In this case, $\Gamma_i \cong \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ and we can choose a generator P_i for the infinite cyclic factor of Γ_i that has eigenvalue 1. Up to possibly replacing P_i with its inverse, there exists a $C_i \in \mathrm{SL}(2, \mathbb{R})$ such that

$$S_i = C_i P_i C_i^{-1} = \begin{bmatrix} 1 & \alpha_i \\ 0 & 1 \end{bmatrix}$$

for some $\alpha_i > 0$ and that $C_i(X, \omega)$ has a shortest horizontal holonomy vector of $(1, 0)$. The piece of the Poincaré section associated to Γ_i is then parametrized by all matrices $\boldsymbol{M}_{a,b}$ that take the saddle connection of $C_i(X, \omega)$ with holonomy vector $(1, 0)$ to a short horizontal with holonomy vector $(|a|, 0)$ of $\boldsymbol{M}_{a,b} C_i(X, \omega)$ with $-1 \leq a < 0$ or $0 < a \leq 1$. With some linear algebra, we can see that this is given by matrices

$$\boldsymbol{M}_{a,b} = \begin{bmatrix} a & b \\ 0 & 1/a \end{bmatrix}$$

with $-1 \leq a < 0$ or $0 < a \leq 1$.

Since S_i and $-I$ are in the Veech group of $C_i(X, \omega)$, this set of $\boldsymbol{M}_{a,b}$ has some redundancies. Quotienting out by $-I$ gives us the set of $\boldsymbol{M}_{a,b}$ with $0 < a \leq 1$ and arbitrary b. If we further quotient out by S_i, we
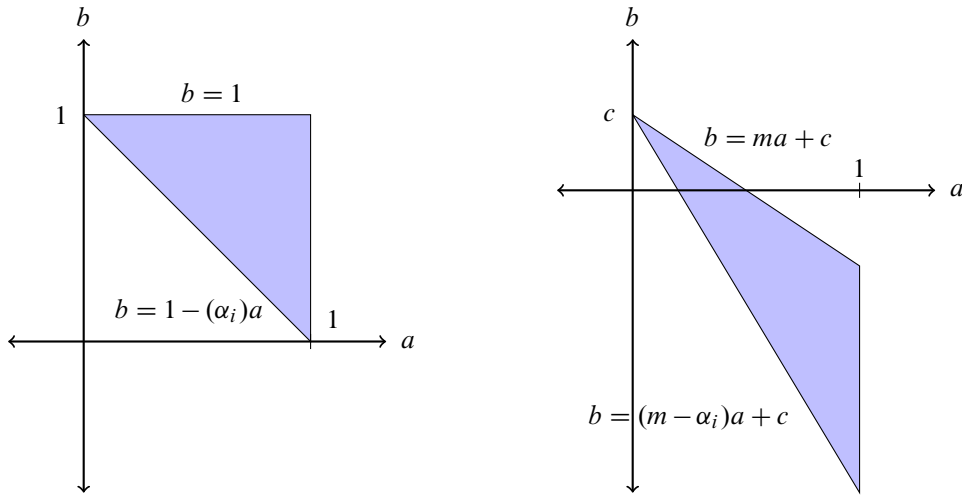
Figure 3: Two possible Poincaré section pieces Ω_i.

see that $M_{a,b}$ is identified with $M_{a,b+n(\alpha_i)a}$ for every $n \in \mathbb{Z}$. The result is that a Poincaré section piece associated to Γ_i can be parametrized by

$$\Omega_i = \{(a,b) \in \mathbb{R}^2 \mid 0 < a \le 1 \text{ and } 1 - (\alpha_i)a < b \le 1\},$$

where each $(a,b) \in \Omega_i$ corresponds to $g \, \mathrm{SL}(X,\omega)$ for $g = M_{a,b}C_i$.

Remark 3 While Ω_i is defined in this specific way in Uyanik and Work's paper, there is actually a lot more freedom in defining Ω_i. We just need to choose a fundamental domain for the $M_{a,b}$ matrices under the action of $\langle S_i, -I \rangle$. To do this, we again let $0 < a \le 1$, but for each a we choose a set of b values of length $(\alpha_i)a$ to account for the quotienting out by S_i. For any $m, c \in \mathbb{R}$, another such fundamental domain is

$$\Omega_i = \{(a,b) \in \mathbb{R}^2 \mid 0 < a \le 1 \text{ and } ma + c - (\alpha_i)a < b \le ma + c\}.$$

That is, instead of choosing Ω_i to be a triangle whose top line is $b = 1$ for $0 < a \le 1$, we choose Ω_i to be a triangle whose top line is $b = ma + c$ for some slope m and b–intercept c. We see the distinction between these two Poincaré section pieces in Figure 3.

Furthermore, we can make similar modifications to Ω_i in Case 2 below. In this case, there will be another triangle with $a < 0$, and we have the freedom to choose the top line of the triangles with $a > 0$ and $a < 0$ independently. These modifications will be integral in our finiteness proofs.

Case 2 $(-I \notin \mathrm{SL}(X,\omega))$ This case breaks up into two subcases, depending on whether the generator P_i of $\Gamma_i \cong \mathbb{Z}$ has eigenvalue 1 or -1.

If P_i has eigenvalue 1, then we again can find

$$S_i = C_i P_i C_i^{-1} = \begin{bmatrix} 1 & \alpha_i \\ 0 & 1 \end{bmatrix}$$

for some $\alpha_i > 0$ and that $C_i(X, \omega)$ has a shortest horizontal holonomy vector of $(1, 0)$. We again have that the matrices $M_{a,b}$ parametrize the Poincaré section piece, but now we only can quotient out by the subgroup generated by S_i. The result is that the Poincaré section piece associated to Γ_i can be parametrized by

$$\Omega_i = \{(a, b) \in \mathbb{R}^2 \mid 0 < a \leq 1 \text{ and } 1 - (\alpha_i)a < b \leq 1\} \cup \{(a, b) \in \mathbb{R}^2 \mid -1 \leq a < 0 \text{ and } 1 + (\alpha_i)a < b \leq 1\},$$

where each $(a, b) \in \Omega_i$ corresponds to $g \operatorname{SL}(X, \omega)$ for $g = M_{a,b}C_i$.

When P_i has eigenvalue -1, we can only find $C_i \in \operatorname{SL}(2, \mathbb{R})$ such that

$$S_i = C_i P_i C_i^{-1} = \begin{bmatrix} -1 & \alpha_i \\ 0 & -1 \end{bmatrix},$$

where $\alpha_i > 0$ and $C_i(X, \omega)$ has a shortest horizontal holonomy vector of $(1, 0)$. We again quotient out our set of $M_{a,b}$ matrices by the subgroup generated by S_i. The resulting Poincaré section piece associated to Γ_i can be parametrized by

$$\Omega_i = \{(a, b) \in \mathbb{R}^2 \mid 0 < a \leq 1 \text{ and } 1 - (2\alpha_i)a < b \leq 1\},$$

where each $(a, b) \in \Omega_i$ corresponds to $g \operatorname{SL}(X, \omega)$ for $g = M_{a,b}C_i$.

Having established what each piece of the Poincaré section associated to each Γ_i looks like, we also need to find the measure on the whole Poincaré section. The measure on the Poincaré section is the unique ergodic measure μ on $\Omega(X, \omega)$, which is a scaled copy of the Lebesgue measure on each of these pieces Ω_i of $\mathbb{R}^2$. The scaling factor is the total area of all the pieces of the transversal.

Upon finding the Poincaré section pieces, the return time function of the horocycle flow at a point $M_{a,b}C_i(X, \omega)$ is the smallest positive slope of a holonomy vector of $M_{a,b}C_i(X, \omega)$ which has short horizontal component. This is because of the way the horocycle flow acts on slopes. More precisely, if $v = (x, y)$ is the holonomy vector of $C_i(X, \omega)$ such that $M_{a,b}(x, y)$ is the holonomy vector on $M_{a,b}(x, y)$ with the smallest positive slope among all holonomy vectors with a horizontal component of length ≤ 1, then the return time function at that point $(a, b) \in \Omega_i$ in the Poincaré section is given by the slope of $M_{a,b}(x, y)$, which is

$$\frac{y}{a(ax + by)}.$$

We call such a vector $v = (x, y)$ a *winner* or *winning saddle connection*. We note that while technically v is the holonomy vector of a saddle connection, we will often use the terms holonomy vector and saddle connection interchangeably. Our proof that the slope gap distribution of a Veech surface has finitely many points of nonanalyticity will rely on us showing that each piece Ω_i of the Poincaré section has finitely many winners.

Each such v would then be a winner on a convex polygonal piece of Ω_i, an example of which is given in Figure 6. Furthermore, the cumulative distribution function of the slope gap distribution would then be given by areas between the hyperbolic return time function level curves (see Figure 16 for an example
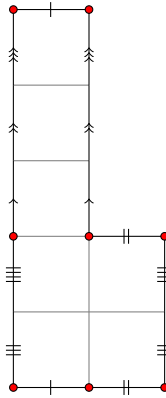
Figure 4: The surface $\mathscr{L}$ with cone point in red.

picture) and the sides of these polygons, and would therefore be piecewise real analytic with finitely many points of nonanalyticity. For more details about this process and a worked example of a computation of a slope gap distribution, please see [3].

2.3 Examples and difficulties

In this section, we will give an example of difficulties that arise from the choice of parametrization of the Poincaré section. In particular, it is possible for there to be infinitely many winning saddle connections under certain parametrizations, but only finitely many different winners under a different parametrization. For full computations of a gap distribution we refer to [3; 11].

We will take the surface $\mathscr{L}$ in Figure 4 and analyze the winning saddle connection on the component Ω_1 of the Poincaré section corresponding to the parabolic subgroup of $\mathrm{SL}(\mathscr{L})$ generated by $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. $\mathscr{L}$ is a 7–square square-tiled surface with a single cone point.

Since $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is in the Veech group and $\mathscr{L}$ has a length-1 horizontal saddle connection, the corresponding piece of the Poincaré section Ω_1 can be parametrized by matrices

$$M_{a,b} = \begin{bmatrix} a & b \\ 0 & a^{-1} \end{bmatrix}$$

with $0 < a \le 1$ and $1 - a < b \le 1$. Notice that $\mathscr{L}$ has all saddle connections with coordinates $(n, 2)$ and $(n, 3)$ for $n \in \mathbb{Z}$, and no saddle connection with y–coordinate 1.

Proposition 4 *In a neighborhood of the point $(0, 1)$ on Ω_1, the winning saddle connection always has y–coordinate 2.*

Proof Take a saddle connection $v = (n, k)$ with $k > 0$ such that $M_{a,b}v$ has horizontal component ≤ 1. We will show that if $k > 2$ and $a < \frac{1}{3}$, there is a saddle connection $w = (m, 2)$ such that the slope of $M_{a,b}w$ is less than the slope of $M_{a,b}v$, and $M_{a,b}w$ has short horizontal component. Since there are no saddle connections with $k = 1$, this implies that the winning saddle connection must have y–coordinate 2.
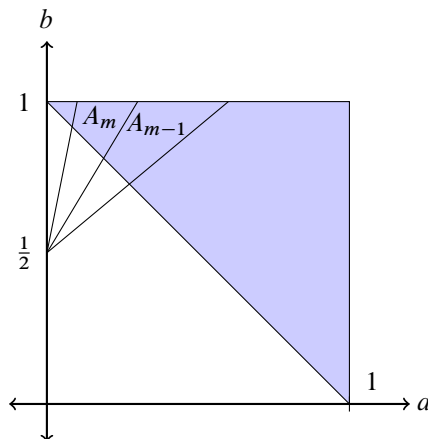
Figure 5: Regions A_m where $(-m, 2)$ is a winner.

The x–coordinate of $M_{a,b}w$ is $ma + 2b$. Since $\mathcal{L}$ has all $(m, 2)$ saddle connections, we may choose an m so that $1 - a < ma + 2b \leq 1$. The condition that $\mathrm{slope}(M_{a,b}w) < \mathrm{slope}(M_{a,b}v)$ rearranges to

$$na + kb < \tfrac{1}{2}k(ma + 2b).$$

If $k > 2$, then since $\mathcal{L}$ is square-tiled $k \geq 3$, and when $a < \tfrac{1}{3}$ we have that $ma + 2b > 1 - a \geq \tfrac{2}{3}$; thus $\tfrac{1}{2}k(ma + 2b) > 1$. Since $M_{a,b}v$ has a short horizontal component $na + kb \leq 1$, so the above inequality is always true. $\qquad\square$

Let A_m be the region where the saddle connection $(-m, 2)$ is the winning saddle connection. By Proposition 4, in the top left corner of Ω_1, A_m is the region where $M_{a,b}(-m, 2) = (2b - ma, 2a^{-1})$ has smallest slope among all saddle connections with y–coordinate 2 and short horizontal component. The slope is $2a^{-1}/(2b - ma)$, so minimizing the slope is equivalent to maximizing $2b - ma$ with the constraint that $2b - ma \leq 1$, or in other words, $-m = \lfloor (1 - 2b)/a \rfloor$. But as $a \to 0$ inside the region Ω_1, $b \to 1$, so $-m \sim -1/a \to -\infty$. This implies that there infinitely many saddle connections that occur as winners in the top left corner of the Poincaré section.

By Remark 3 in Section 2.2, we can change the parametrization of the Poincaré section. One problem in our previous parametrization was that there were infinitely many winners in the upper left-hand corner $(0, 1)$ of our Poincaré section. To fix this, we will change our parametrization so that the upper left corner is at $\left(0, \tfrac{1}{2}\right)$ and the slope of the top line of our Poincaré section triangle is nicely compatible with the $(1, 2)$ holonomy vector. This will ensure that there are finitely many winners in the top left corner, and will result in finitely many winners across the entire Poincaré section. We will prove that we can always do this for arbitrary Veech surfaces in Section 3.

We will use the parametrization $0 < a \leq 1$ and $\tfrac{1}{2} - \tfrac{3}{2}a < b \leq \tfrac{1}{2} - \tfrac{1}{2}a$. This parametrization is chosen to ensure that the saddle connection $(1, 2)$ of $\mathcal{L}$ wins in a neighborhood of the top line segment, which prevents the problem that arises in the previous parametrization.
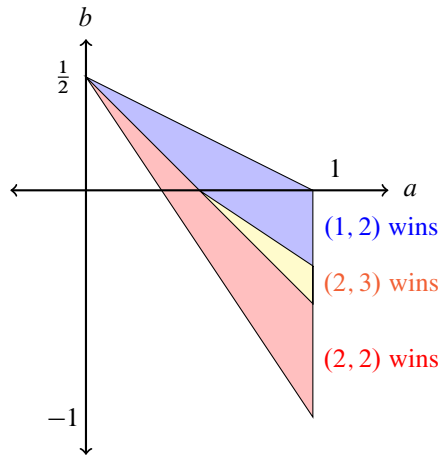
Figure 6: The new Poincaré section breaks up into three pieces, with saddle connection (1,2) winning in the blue region, (2,3) in the yellow region and (2,2) in the red region.

In this case, the only winners are the $(1, 2)$, $(2, 3)$ and $(2, 2)$ saddle connections on $\mathscr{L}$:

(1) $(1, 2)$ wins in the region

$$\{(a, b) \mid 0 < a \leq 1, \ \tfrac{1}{2} - a < b \leq \tfrac{1}{2} - \tfrac{1}{2}a \text{ and } \tfrac{1}{3} - \tfrac{2}{3}a < b\}.$$

(2) $(2, 3)$ wins in the region

$$\{(a, b) \mid \tfrac{1}{2} < a \leq 1 \text{ and } \tfrac{1}{2} - a < b \leq \tfrac{1}{3} - \tfrac{2}{3}a\}.$$

(3) $(2, 2)$ wins in the region

$$\{(a, b) \mid 0 < a \leq 1 \text{ and } \tfrac{1}{2} - \tfrac{3}{2}a < b \leq \tfrac{1}{2} - a\}.$$

To see this, notice that the saddle connection (x, y) is the winner at (a, b) if $M_{a,b}(x, y)$ has smallest positive slope amongst all saddle connections with short horizontal component. $M_{a,b}(x, y)$ has short horizontal component in the region with $0 < a \leq 1$ and $(-x/y)a < b \leq 1/y - (x/y)a$. Minimizing the slope at (a, b) is equivalent to maximizing x/y over all saddle connections with a short horizontal component.

Working out the exact winners then comes down to casework. In this case, $M_{a,b}(m, 2)$ never has a short horizontal component for $m > 2$ and (a, b) in the Poincaré section, and simple casework shows where $(1, 2)$ and $(2, 2)$ are the winners. For saddle connections with y–coordinate greater than 2, we need to understand those with $x/y > \tfrac{1}{2}$ which can potentially win against $(1, 2)$ or $(2, 2)$. In the yellow region $(2, 3)$ wins, as $(2, 2)$ does not have a short horizontal component for (a, b) in that region. All other saddle connections with $y = 3$ and $x \geq 3$ do not have short horizontal component in the Poincaré section. For $y \geq 4$, a similar analysis shows that none of the saddle connections can appear as winners, giving the result.

3 Main theorem

In [Section 2.3](#), we examined the 7–square square-tiled surface $\mathscr{L}$ and saw that, in one parametrization, it looked like the Poincaré section would admit infinitely many winning saddle connections and therefore give the possibility of infinitely many points of nonanalyticity in the slope gap distribution. However, when we strategically chose a different parametrization of this piece of the Poincaré section, there were only finitely many winners. Thus this piece of the Poincaré section could only contribute finitely many points of nonanalyticity to the slope gap distribution.

It is interesting to note that this implies that many of the potential points of nonanalyticity arising from the Uyanik–Work parametrization must cancel each other out and not result in points of nonanalyticity in the slope gap distribution. Choosing a strategic parametrization of the Poincaré section is one of the key ideas of the main theorem of this paper:

Theorem 1 *The slope gap distribution of any Veech surface has finitely many points of nonanalyticity.*

This section is devoted to the proof of this theorem. We will begin by giving an outline of the proof, and then will dive into the details of each step.

3.1 Outline

The idea is that after choosing strategic parametrizations of each piece of the Poincaré section of a Veech translation surface (X, ω), we will use compactness arguments to show that there are finitely many winners on each piece.

(1) We begin with a Veech translation surface (X, ω) and focus on a piece of its Poincaré section corresponding to one maximal parabolic subgroup in $\mathrm{SL}(X, \omega)$. Up to multiplication by an element of $\mathrm{GL}(2, \mathbb{R})$, we will assume that the generator of the parabolic subgroup has a horizontal eigenvector and (X, ω) has a horizontal saddle connection of length 1. Based on properties of the saddle connection set of (X, ω), we strategically choose a parametrization T_X of this Poincaré section piece. T_X will be some triangle in the plane.

(2) For any saddle connection $\boldsymbol{v}$ of (X, ω), we will define a strip $S_\Omega(\boldsymbol{v})$ that gives a set of points $(a, b) \in \mathbb{R}_{>0} \times \mathbb{R}$ where $\boldsymbol{v}$ is a potential winning saddle connection on the surface $\boldsymbol{M}_{a,b}(X, \omega) \in T_X$. We will start by showing various properties of these strips that we will make use of later on in the proof.

(3) We will then show that every point $(a, b) \in T_X$ in the interior of a strip has an open neighborhood with finitely many winning saddle connections.

(4) We show that every point (a, b) on the top edge of T_X has an open neighborhood with finitely many winning saddle connections.

(5) We then move on to show that points $(a, b) \in T_X$ that are either in the interior of T_X or on the bottom edge not including the right vertex with $a = 1$ have an open neighborhood with finitely many winning saddle connections.

(6)   Next we show that on the boundary $a = 1$ of T_X there are finitely many winning saddle connections.

(7)   Using the finiteness on the right boundary, we show that any point $(a, b) \in T_X$ with $a = 1$ has an open neighborhood with finitely many winning saddle connections.

(8)   By compactness of T_X, there is a finite cover of T_X with the open neighborhoods of points $(a, b) \in T_X$ that we found in our previous steps. Since each of the these open neighborhoods had finitely many winners, we find that there are finitely many winning saddle connections across all of T_X.

(9)   Finally, we show that finitely many winners on each piece of the Poincaré section implies finitely many points of nonanalyticity of the slope gap distribution.

3.2   Proof

Using the method of [11] outlined in Section 2.2, it will suffice to show that every piece of the Poincaré section can be chosen so that there are only finitely many winning saddle connections. For most of the arguments in this section we will fix a piece of the Poincaré section and will work exclusively with it.

We recall that there is a piece of the Poincaré section for each conjugacy class of a maximal parabolic subgroup in $SL(X, \omega)$. We will now fix such a maximal parabolic subgroup Γ_i and work with the corresponding component of the Poincaré section. Without loss of generality we may assume that (X, ω) has a horizontal saddle connection with x–component 1 and that Γ_i is generated by

$$P_i = \begin{bmatrix} 1 & \alpha_i \\ 0 & 1 \end{bmatrix}.$$

Using the notation of Section 2.2, this is essentially replacing (X, ω) with $C_i(X, \omega)$.

Since (X, ω) is a Veech surface with a horizontal saddle connection it has a horizontal cylinder decomposition [6], and therefore, for all $a \in \mathbb{R}$, there are only finitely many heights $0 \leq h \leq a$ such that (X, ω) has a saddle connection with y–component h. Let $y_0 > 0$ be the shortest vertical component of a saddle connection on (X, ω), and let $x_0 > 0$ be the shortest horizontal component of a saddle connection at height y_0. Our first step is to use this saddle connection to give a parametrization of the Poincaré section that is adapted to the geometry of (X, ω).

By Remark 3, we can choose the following parametrization of this piece of the Poincaré section, as pictured in Figure 7:

$$T_X = \left\{ (a, b) \; \middle| \; 0 < a \leq 1 \text{ and } \frac{1 - x_0 a}{y_0} - na \leq b \leq \frac{1 - x_0 a}{y_0} \right\}.$$

Here n is either α_i or $2\alpha_i$ depending on which one is needed to fully parametrize this piece of the Poincaré section, as described in Section 2.2. In the case where $-I \notin SL(X, \omega)$ and P_1 had eigenvalue 1, the Poincaré section has an additional triangle with $a < 0$. In particular, we can choose this triangle so that it consists of points $(-a, -b)$ for $(a, b) \in T_X$. But if $\boldsymbol{v}$ were the winning saddle connection for $M_{a,b}(X, \omega)$, then $-\boldsymbol{v}$ would be the winning saddle connection for $M_{-a,-b}(X, \omega)$, and hence when proving that there are only finitely many winners, it suffices to consider only the portion with $a > 0$.
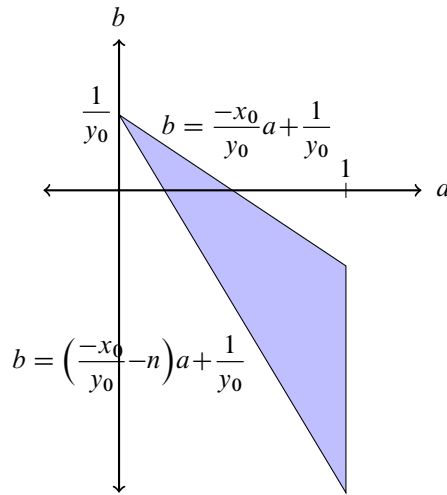
Figure 7: A Poincaré section piece for (X, ω) with $y_0 > 0$ the shortest vertical component and $x_0 > 0$ the shortest corresponding horizontal component of a saddle connection on (X, ω) with vertical component y_0.

Our goal now is to prove that the return time function is piecewise real analytic with finitely many pieces. We will do so by proving that there are finitely many winning saddle connections $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \Lambda(X, \omega)$ such that each point $(a, b) \in T_X$ has a winner $\boldsymbol{M}_{a,b}\boldsymbol{v}_i$ for some $1 \leq i \leq n$. We will repeat this for every T_X corresponding to each maximal parabolic subgroup.

To achieve this goal, we will first define an auxiliary set that will help us understand for what points $(a, b) \in T_X$ a particular $\boldsymbol{v} \in \Lambda(X, \omega)$ is a candidate winner. By a candidate winner, we mean that $\boldsymbol{M}_{a,b}\boldsymbol{v}$ has a positive x–coordinate at most 1 and a positive y–coordinate. If $\boldsymbol{v} = (x, y)$, the x–coordinate condition is the condition that $0 < ax + by \leq 1$. We also note that for $\boldsymbol{M}_{a,b}(x, y)$ to be a winner, we need that $a^{-1}y > 0$. Since $a > 0$ on T_X, this condition reduces to saying that $y > 0$.

Definition 5 Given a saddle connection $\boldsymbol{v} = (x, y)$ with $y > 0$, we define $\mathcal{S}_\Omega(\boldsymbol{v})$ as the strip of points $(a, b) \in \mathbb{R}_{>0} \times \mathbb{R}$ such that $0 < ax + by \leq 1$. This corresponds to the set of surfaces $\boldsymbol{M}_{a,b}(X, \omega)$ for which $\boldsymbol{M}_{a,b}\boldsymbol{v}$ is a potential winning saddle connection.

Let us note some properties of these strips $\mathcal{S}_\Omega(\boldsymbol{v})$ that we will use repeatedly in our proofs. We recall that we are assuming without loss of generality that (X, ω) has a short horizontal saddle connection of length 1. Considering the particular piece T_X of the Poincaré section, we recall that T_X is parametrized by matrices

$$\boldsymbol{M}_{a,b} = \begin{bmatrix} a & b \\ 0 & a^{-1} \end{bmatrix}$$

so that $\boldsymbol{M}_{a,b}(X, \omega)$ has a horizontal saddle connection of length ≤ 1.
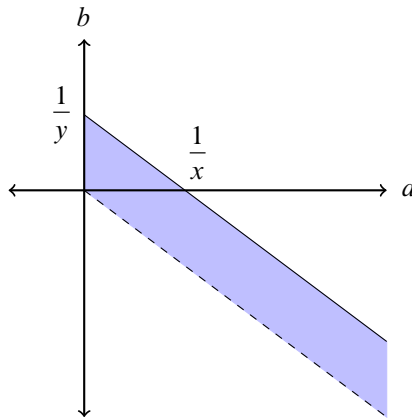
Figure 8: A strip $\mathcal{S}_\Omega(v)$ for $v = (x, y)$. Here $y > 0$. The slope of the upper and lower lines of the strip is $-x/y$.

Then, because (X, ω) is a Veech surface, it breaks up into horizontal cylinders, and therefore there exists a $y_0 > 0$ such that there is a saddle connection with height y_0 and furthermore that every saddle connection with positive height has height $\geq y_0$.

With these assumptions in place, we note the following useful properties of the strips $\mathcal{S}_\Omega(v)$ that are used implicitly throughout the proof:

(1) The strip $\mathcal{S}_\Omega(v)$ for $v = (x, y)$ is sandwiched between a solid line that intersects the b–axis at $1/y$ and a dotted line that intersects the b–axis at 0. Both lines have slope $-x/y$. We also know that $y \geq y_0$, so $1/y \leq 1/y_0$.

(2) Fixing any $c > 0$, there are only finitely many y–coordinates of saddle connections v such that $S_\Omega(v)$ intersects the y–axis at any point $\geq c$.

This is because (X, ω) being a Veech surface and having a horizontal saddle connection implies that the surface breaks up into finitely many horizontal cylinders of heights $h_1, \ldots, h_n$ and every saddle connection with positive y–component must have a y–component that is a nonnegative linear combination of these h_i. Since there are finitely many such y values $\leq 1/c$, there are finitely many strips that intersect the y–axis at points $\geq c$.

(3) At a particular point $(a, b) \in T_X$, the winner is the saddle connection $v = (x, y) \in \Lambda(X, \omega)$ such that $M_{a,b}v = (ax + by, a^{-1}y)$ has the least slope among those saddle connections satisfying $0 < ax + by \leq 1$ and $a^{-1}y > 0$. Since $a > 0$ for any point in Ω_i and the reciprocal of the slope of $M_{a,b}v$ is $a^2x/y + ab$, this corresponds to the saddle connection with the greatest reciprocal slope, which corresponds to having the greatest x/y with $y > 0$.

In terms of our strips, we're fixing the point (a, b) and looking for the strip $\mathcal{S}_\Omega(v)$ that contains (a, b) and has the least slope, since each strip has slope $-x/y$. We further note that, due to our choice of Poincaré

section, no saddle connection $v = (x, y)$ with $x < 0$ can ever be the winner at a point $(a, b) \in T_X$, since either (a, b) will not be in the strip $\mathcal{S}_\Omega(v)$ or the saddle connection (x_0, y_0) that defined T_X would win over (x, y) at (a, b). Because of this, from now on we will only consider saddle connections with $x \geq 0$ (and $y > 0$) when looking for potential winners.

(4) For any given $y > 0$, there are only finitely many saddle connection vectors $v = (x, y)$ of (X, ω) with $x \geq 0$ such that $S_\Omega(v)$ intersects T_X.

This is because $S_\Omega(v)$ does not intersect T_X for x/y larger than some constant C that depends on T_X and y. Specifically, we can let $C = x_0/y_0 + n$, the negative of the slope of the bottom line that defines the triangle T_X. Since the saddle connection set is discrete, there are finitely many $x \geq 0$ for a given y such that $x/y \leq C$.

With these facts established, let us first prove a lemma that shows that winning saddle connections exist and that will be useful in proving Lemma 12.

Lemma 6 *Let $(a, b) \in T_X$ be such that $(a, 0)$ is a short horizontal saddle connection of $M_{a,b}(X, \omega)$. Then $M_{a,b}(X, \omega)$ either has a saddle connection $v = (x, y)$ with $0 < x < a$ and $y > 0$, or there exist two saddle connections $v_1 = (a, y)$ and $v_2 = (0, y)$ with $y > 0$. This implies that every point in T_X has a winning saddle connection, or equivalently that every point in T_X is in some strip $S_\Omega(v)$ for some $v = (x, y)$ with $y > 0$.*

Proof Let us take a horizontal saddle connection on our surface $M_{a,b}(X, \omega)$ with holonomy vector $(a, 0)$, connecting two (possibly identical) cone points p and q. Then we will consider developing a width-a vertical strip on our surface extending upward with the open horizontal segment from p to q as its base. Since our surface is of finite area, this vertical strip must eventually hit a cone point r or come back to overlap our original open segment from p to q. Now we're going to define our vectors v, or v_1 and v_2, in each case.

In the former case when the top edge of our vertical strip hits a cone point r in the interior of the edge, the straight segment from p to r cutting through our vertical strip gives us v.

The latter case when the top edge of our vertical strip comes back to overlap our original open segment breaks up into two cases. If we have an incomplete overlap, then the top edge contains the cone point
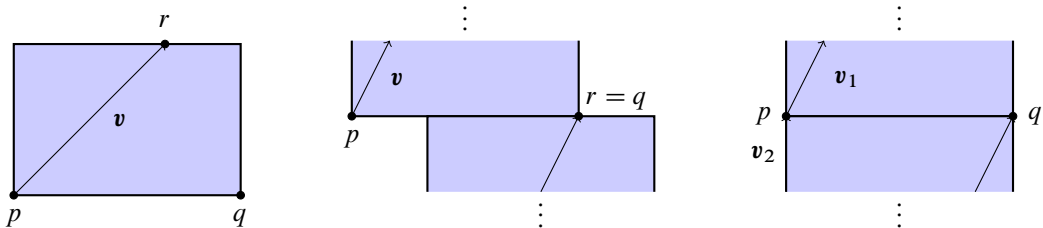
Figure 9: The vectors v or v_1 and v_2 in the three different cases of vertical strip.
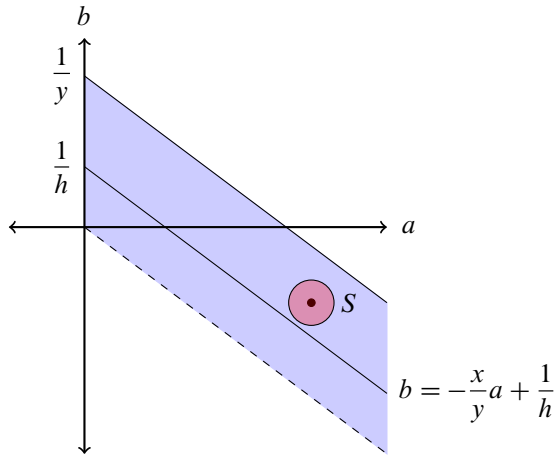
Figure 10: A choice of $1/h$ for a particular S.

$r = p$ or $r = q$, and the saddle connection from p on the bottom edge to r on the top edge gives us $\boldsymbol{v}$. If we have a complete overlap, then the saddle connection from p on the bottom edge to q on the top edge gives us our vector $\boldsymbol{v}_1$ and the saddle connection from p on the bottom edge to p on the top edge gives us our vector $\boldsymbol{v}_2$.

In any of these cases, letting $\boldsymbol{v}' = M_{a,b}^{-1}(\boldsymbol{v})$ or $M_{a,b}^{-1}(\boldsymbol{v}_1)$ gives us that $S_\Omega(\boldsymbol{v}')$ contains our initial point (a, b) and $\boldsymbol{v}'$ is a possible winning saddle connection. $\square$

The following lemma will help us show that there are finitely many winning saddle connections on certain sets in T_X:

Lemma 7 *Let S be a closed set that is a subset of $S_\Omega(\boldsymbol{v})$ for $\boldsymbol{v} = (x, y)$ with $y > 0$. Then there are finitely many winning saddle connections on S.*

Proof Let S be a closed set contained in $S_\Omega(\boldsymbol{v})$ for a saddle connection $\boldsymbol{v} = (x, y)$ of (X, ω) with $y > 0$. By definition, $\boldsymbol{v}$ is a potential winning saddle connection on all of S. That is, for any point $(a, b) \in S$, $M_{a,b}\boldsymbol{v}$ has positive y component and positive and short (≤ 1) x component.

We recall that for a point $(a', b') \subset S$ to have winner $\boldsymbol{v}' = (x', y') \neq \boldsymbol{v} = (x, y)$, we need that $\boldsymbol{v}'$ is a saddle connection of (X, ω), $x'/y' > x/y$, and that $(a', b') \subset S_\Omega(\boldsymbol{v}')$.

This corresponds to the strip $S_\Omega(\boldsymbol{v}')$ having a smaller slope than $S_\Omega(\boldsymbol{v})$ and still intersecting S. Given that S is closed and the bottom boundary of $S_\Omega(\boldsymbol{v})$ is open, there exists an $h > 0$ such that the line S is completely on or above the line $b = -(x/y)a + 1/h$. Furthermore, since the left boundary of $S_\Omega(\boldsymbol{v})$ is open, S is a positive distance away from the y–axis.

Then for $S_\Omega(\boldsymbol{v}')$ to intersect S and for $x'/y' > x/y$, we need that $y' < h$, since otherwise the strip $S_\Omega(\boldsymbol{v}')$ would have y–intercepts $1/y' \leq 1/h$ and 0 and would have smaller slope than that of $S_\Omega(\boldsymbol{v})$ and would therefore not intersect S.

But since (X, ω) is a Veech surface with a horizontal saddle connection, it decomposes into finitely many horizontal cylinders. Therefore, the set of possible vertical components y' of saddle connections are a discrete subset of $\mathbb{R}$, and thus there are finitely many vertical components of saddle connections that satisfy $y' < h$. Since there are finitely many saddle connections in the vertical strip $(0, 1] \times (0, \infty)$ with vertical component less than h, there are finitely many possible winning saddle connections on S. □

We recall that our goal is to show that every point $(a, b) \in T_X$ has a neighborhood on which there are finitely many winners. This will allow us to use a compactness argument to prove that there are finitely many winners on all of T_X. Building off of the previous lemma, we show in the next lemma that certain points $(a, b) \in T_X$ have an open neighborhood on which there are finitely many winners:

Lemma 8 *Let* (a, b) *be in the interior of some strip* $S_\Omega(\boldsymbol{v})$. *Then there exists a neighborhood of* (a, b) *with finitely many winning saddle connections.*

Proof Let (a, b) be in the interior of the strip $S_\Omega(\boldsymbol{v})$ for $\boldsymbol{v} = (x, y)$ with $y > 0$ and $x \geq 0$. Then we can find an $\epsilon > 0$ such that the closed ball of radius ϵ around (a, b) remains in the interior of the strip. That is, we choose an $\epsilon > 0$ such that

$$\overline{B_\epsilon((a, b))} \subset S_\Omega(\boldsymbol{v}).$$

We can then use Lemma 7 to conclude that there are finitely many winning saddle connections on $\overline{B_\epsilon((a, b))}$, and therefore on $B_\epsilon((a, b))$. □

We now look at points $(a, b) \in T_X$ that lie on the top edge of T_X and show that these points have a neighborhood with finitely many winners.

Lemma 9 *For any* (a, b) *that lies on the top edge of* T_X, *including the point* $(0, 1/y_0)$, *there exists a neighborhood* $B_\epsilon((a, b))$ *such that there are finitely many winning saddle connections on* $B_\epsilon((a, b)) \cap T_X$.

Proof We recall that T_X is a triangle bounded by the lines $b = (-x_0/y_0)a + 1/y_0$ on top, the line $a = 1$ on the right and the line $b = (-x_0/y_0 - n)a + 1/y_0$ on the bottom.

We break up the proof of this lemma into cases, depending on the location of $(a, b) \in T_X \cup \{(0, 1/y_0)\}$:

(1) $b = (-x_0/y_0)a + 1/y_0$ These points are on the top line of T_X. We recall that y_0 was chosen to be the least $y > 0$ for which X has a saddle connection (x, y_0). Then x_0 was the least $x > 0$ for which (x, y_0) was a saddle connection of X.

Let (a, b) be any point on the top line of T_X and let $\boldsymbol{v} = (x_0, y_0)$. Then (a, b) is on the top line of the strip $S_\Omega((x_0, y_0))$. We can find an $\epsilon > 0$ such that $\overline{B_\epsilon((a, b))} \cap S_\Omega((x_0, y_0))$ is a closed subset of $S_\Omega((x_0, y_0))$. By Lemma 7, there are then finitely many winners on $B_\epsilon((a, b)) \cap S_\Omega((x_0, y_0))$.

(2) $(a, b) = (0, 1/y_0)$ This point is not in T_X but is the top left corner of the triangle that makes up T_X. We can find a $y_1 > y_0$ such that every saddle connection (x, y) of X with $y > y_0$ must satisfy that $y \geq y_1$. Thus, we can choose an $\epsilon > 0$ such that $B_\epsilon((0, 1/y_0)) \cap T_X \subset S_\Omega((x_0, y_0))$ and no strip $S_\Omega((x, y))$, for
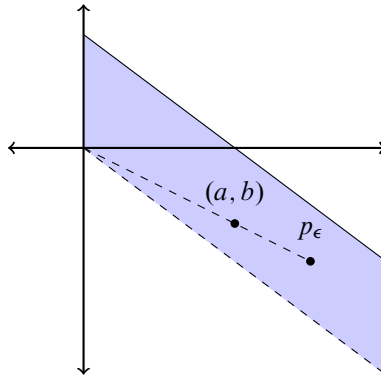
Figure 11: The strip $\mathcal{S}_\Omega(\boldsymbol{w}_\epsilon)$.

a saddle connection with $y > y_0$ and $x \geq 0$, intersects $B_\epsilon((0, 1/y_0))$. This would imply that the only possible winning saddle connections on $B_\epsilon((0, 1/y_0))$ are of the form (x, y_0) for $x \geq x_0$.

But if we fix $y = y_0$, since the set of saddle connections (x, y_0) is discrete and T_X is bounded below by the line $b = (-x_0/y_0 - n)a + 1/y_0$, there are only finitely many saddle connections $\boldsymbol{v} = (x, y_0)$ of (X, ω) whose strip $\mathcal{S}_\Omega(\boldsymbol{v})$ intersects $B_\epsilon((0, 1/y_0))$ — exactly those x such that $x_0 \leq x \leq x_0 + n y_0$. We have shown then that only finitely many strips $\mathcal{S}_\Omega(\boldsymbol{v})$, for holonomy vectors $\boldsymbol{v}$ that could win over (x_0, y_0), intersect $B_\epsilon((0, 1/y_0))$, and therefore there are only finitely many winners on this neighborhood. $\qquad\square$

Having established that points $(a, b) \in T_X$ on the top edge of T_X have neighborhoods with finitely many winners, we now turn to points $(a, b) \in T_X$ that lie in the interior of T_X or on the bottom edge of T_X.

Lemma 10 *For any (a, b) that lies in the interior of T_X or on the bottom edge of T_X, excluding the vertex with $a = 1$, there exists a neighborhood $B_\epsilon((a, b))$ such that there are finitely many winning saddle connections on $B_\epsilon((a, b)) \cap T_X$.*

Proof By Lemma 8, it suffices to show that (a, b) lies on the interior of a strip $\mathcal{S}_\Omega(\boldsymbol{v})$ for some saddle connection $\boldsymbol{v}$.

Because (a, b) is in T_X, it must lie in some strip $\mathcal{S}_\Omega(\boldsymbol{v})$. If (a, b) is in the interior of $\mathcal{S}_\Omega(\boldsymbol{v})$, then we are done. Otherwise, if (a, b) is on the boundary of $\mathcal{S}_\Omega(\boldsymbol{v})$, we consider the points $p_\epsilon = ((1 + \epsilon)a, (1 + \epsilon)b)$, with winner $\boldsymbol{w}_\epsilon$. Since (a, b) lies in the interior of T_X or on the bottom edge of T_X, for $\epsilon > 0$ sufficiently small p_ϵ also lies in T_X. Moreover, notice that p_ϵ and (a, b) lie on the same line through the origin. This immediately implies that (a, b) lies in the interior of $\mathcal{S}_\Omega(\boldsymbol{w}_\epsilon)$, as seen in Figure 11.

Indeed, by the definition of $\mathcal{S}_\Omega(\boldsymbol{v})$ for any holonomy vector $\boldsymbol{v}$ as a half-open strip with the open bottom boundary passing through the origin, for all points $p \in \mathcal{S}_\Omega(\boldsymbol{v})$ the points tp for $0 < t < 1$ lie in the interior, which gives the desired result. $\qquad\square$

The combination of our previous lemmas shows that, for all $(a, b) \in T_X$ away from the right vertical boundary, there are only finitely many winners in a neighborhood of (a, b). We also want to show that,

for each $(1, b)$ on the right vertical boundary, there are only finitely many winners in a neighborhood. We will do this in two steps. First we will show that there are finitely many winning saddle connections along the right boundary of T_X. We will then use this result to prove that every point $(1, b)$ on the right boundary of T_X has a neighborhood with finitely many winning saddle connections.

For our first result, we will need the following definition:

Definition 11 Given $(a, b) \in \mathbb{R}^2$, define the set $\mathcal{S}_\Lambda(a, b)$ as the strip of vectors $\boldsymbol{v} = (x, y) \in \mathbb{R}^2$ such that $0 < ax + by \le 1$ and $y > 0$. This corresponds to the set of vectors that are potential winners on the surface $\boldsymbol{M}_{a,b}(X, \omega)$.

We think of this definition as a sort of dual to Definition 5, where instead of thinking of the surfaces corresponding to a particular winning saddle connection, we think about the set of possible coordinates of winning saddle connections for a particular surface.

Lemma 12 *There are only finitely many winning saddle connections along the right vertical boundary* $a = 1$ *of* T_X.

Proof By Lemma 6, we know that every point $(1, b)$ on the right boundary of T_X has a winning saddle connection. The set of $b \in \mathbb{R}$ such that $(1, b) \in T_X$ is some interval $[c, d]$. We note that since $\left[\begin{smallmatrix} 1 & \alpha \\ 0 & 1 \end{smallmatrix}\right]$ is in the Veech group of our surface for some $\alpha > 0$, it suffices to show that there are finitely many winners for $b \in [c + n\alpha, d + n\alpha]$ for any $n \in \mathbb{Z}$. This is because (x, y) is the winner for b' if and only if $(x - n\alpha y, y)$ is the winner for $b' + n\alpha$. For convenience, we will prove that there are finitely many winners for $b \in [M, N] = [c + n\alpha, d + n\alpha]$ for an n such that $M, N > 0$.

For each such b, we let $\boldsymbol{v}_b$ be its corresponding winning saddle connection. We wish to show that the set of vectors $\boldsymbol{v}_b$ is finite. We suppose that $\{\boldsymbol{v}_b\}$ is infinite. Then, since $b \in [M, N]$, we must be able to find a convergent subsequence of $b_i \in \mathbb{R}$ with corresponding winning saddle connections (x_i, y_i) such that $b_i \to b'$ and $b', b_i \in [M, N]$ for all i. In particular, $b' > 0$.

We claim now that $\mathcal{S}_\Lambda(1, b')$ cannot have a winning saddle connection, which would contradict Lemma 6. This corresponds to a saddle connection (x, y) in the strip $\mathcal{S}_\Lambda(1, b')$ that maximizes x/y. The strip $\mathcal{S}_\Lambda(1, b')$ satisfies that $y > 0$ and $0 < x + b'y \le 1$, or alternatively that $-(1/b')x < y \le -(1/b')x + 1/b'$. We recall that $b' > 0$. Figure 12 shows a depiction of this strip.

We suppose that the winning saddle connection (x', y') for b' lies in the interior of $\mathcal{S}_\Lambda(1, b')$. If $x'/y' > x_i/y_i$ and $(x', y') \in \mathcal{S}_\Lambda(1, b_i)$, then (x_i, y_i) could not be the winner for $(1, b_i)$ because (x', y') beats it and is still in the strip $\mathcal{S}_\Lambda(1, b_i)$.

We let $\mathcal{C}_{b'}$ be the cone given by the intersection of $y < (y'/x')x$ and $y > (y'/(x'-1))x - y'/(x'-1)$. We notice that if $(x_i, y_i) \in \mathcal{C}_{b'}$, then it follows that $(x', y') \in \mathcal{S}_\Lambda(1, b_i)$. One can see this algebraically or visually by noting that if (x_i, y_i) is in the cone $\mathcal{C}_{b'}$ as depicted in Figure 13, then $\mathcal{S}_\Lambda(1, b_i)$ contains (x_i, y_i) and is bounded by two lines with x–intercepts 0 and 1 and therefore must contain the point (x', y'). Furthermore, the first inequality defining the cone gives us that $x/y > x'/y'$.
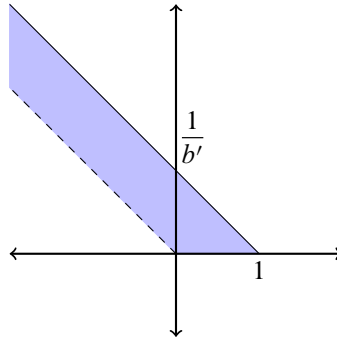
Figure 12: The strip $\mathcal{S}_\Lambda(1, b')$.

Therefore, if (x_i, y_i) is a winning saddle connection for some $(1, b_i)$, it cannot be in the open cone $\mathcal{C}_{b'}$ as defined above. Since $b_i \to b'$, this implies that for any $\epsilon > 0$ we can find an n large enough that, for all $i \geq n$, the strips $S_\Lambda(1, b_i)$ all lie in a region $\mathcal{S}_\epsilon$ that is the region where $(-1/b' + \epsilon)x \leq y \leq (-1/b' - \epsilon)x + (1/b' + \epsilon)$ and $y > 0$. Specifically, we will choose an ϵ such that the slopes of the two bounding lines of $\mathcal{S}_\epsilon$ are wedged between the slopes of the bounding lines of $\mathcal{C}_{b'}$. That is, we will choose $\epsilon > 0$ such that $(-1/b' - \epsilon) > y'/x'$ and $(-1/b' + \epsilon) < y'/(x' - 1)$. We call this latter region $\mathcal{S}_\epsilon$. Figure 13 illustrates these regions.

Given these conditions, we notice that $\mathcal{S}_\epsilon \setminus \mathcal{C}_{b'}$ is a compact set. With the possible exception of one point that equals (x', y'), the winning saddle connections (x_i, y_i) for $i \geq n$ must all be in this region. But the set of holonomy vectors of saddle connections of (X, ω), of which $\{(x_i, y_i)\}$ is a subset, is a discrete subset of $\mathbb{R}^2$ with no accumulation points, and so there are only finitely many $(x_i, y_i) \in \mathcal{S}_\epsilon \setminus \mathcal{C}_{b'}$. This is a
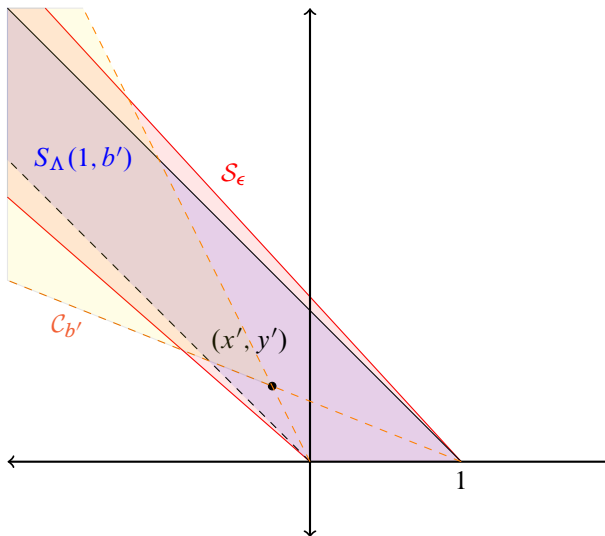
Figure 13: The strip $\mathcal{S}_\Lambda(1, b')$ with its winner (x', y') and cone $\mathcal{C}_{b'}$, along with the region $\mathcal{S}_\epsilon$ containing the winners (x_i, y_i) for $i \geq n$.
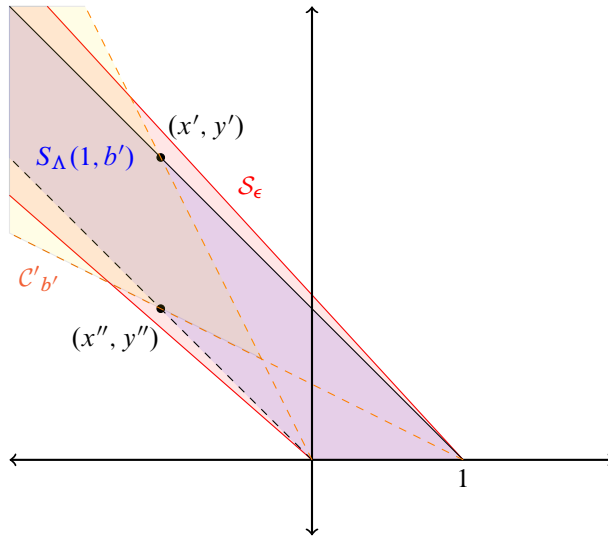
Figure 14: The strip $S_\Lambda(1, b')$ with its winner (x', y'), the vector (x'', y'') on its open boundary and its cone $\mathcal{C}'_{b'}$, along with the region $\mathcal{S}_\epsilon$ containing the winners (x_i, y_i) for $i \geq n$.

contradiction, since the set $\{(x_i, y_i)\}$ is infinite. Hence, if $S_\Lambda((1, b'))$ contained a point (x', y'), it could not be in the interior of the strip.

We also consider the case when (x', y') is in on the boundary of $S_\Lambda(1, b')$. That is, we suppose that (x', y') is on the line $y = -(1/b')x + 1/b'$. If there exists a saddle connection in the interior of $S_\Lambda(1, b')$, we can appeal to the reasoning in the previous case to find a contradiction. Else, after potentially applying a shear to our surface, Lemma 6 guarantees that there is also a holonomy vector (x'', y'') on the open boundary $y = -(1/b')x$ of $S_\Lambda(1, b')$.

We now consider the cone $\mathcal{C}'_{b'}$ given by the intersection of the regions

$$y < \frac{y'}{x'}x \quad \text{and} \quad y > \frac{y''}{x'' - 1}x - \frac{y''}{x'' - 1}.$$

Similar to the previous case, we can find n large enough that the strips $S_\Lambda(1, b_i)$ all lie in a region $\mathcal{S}_\epsilon$ that is defined by $(-1/b' + \epsilon)x \leq y \leq (-1/b' - \epsilon)x + (1/b' + \epsilon)$ and $y > 0$. Here we again choose $\epsilon > 0$ such that the slopes of the two bounding lines of $\mathcal{S}_\epsilon$ are wedged between the slopes of the bounding lines of $\mathcal{C}'_{b'}$. That is, we will choose $\epsilon > 0$ such that $(-1/b' - \epsilon) > y'/x'$ and $(-1/b' + \epsilon) < y''/(x'' - 1)$. We call this latter region $\mathcal{S}_\epsilon$. Figure 14 illustrates these regions.

Since the set $\{(x_i, y_i)\}$ has no accumulation points and $\mathcal{S}_\epsilon \backslash \mathcal{C}'_{b'}$ is compact, all but finitely many of the (x_i, y_i) for $i \geq n$ must lie in the cone $\mathcal{C}'_{b'}$ and not be equal to (x', y') or (x'', y''). Let us consider one of these (x_i, y_i). The corresponding strip $\mathcal{S}_\Lambda(1, b_i)$ is the region between two parallel lines that intersect the x–axis at 1 and 0, including the line through 1 but not including the line through 0. Therefore $S_\Lambda(1, b_i)$ must either contain (x', y') or (x'', y''), depending on if $b_i \leq b'$ or $b_i > b'$, respectively. If it contains (x', y'), then by similar reasoning as in the previous case (x', y') beats (x_i, y_i), and so (x_i, y_i) could not

have been the winner for $(1, b_i)$. If it contains (x'', y''), then either (x'', y'') beats (x_i, y_i), which means that (x_i, y_i) was not the winner, or (x_i, y_i) was in the interior or $\mathcal{S}_\Lambda((1, b'))$, which contradicts that the interior of $S_\Lambda(1, b')$ did not contain any saddle connections. In either case, we have a contradiction.

Since we found a contradiction in both the cases when there was saddle connection in the interior and on the boundary of $S_\Lambda(1, b')$, we see that there must have been only finitely many winners on the right vertical boundary of T_X. □

We can now use the previous lemma to show that points on the right boundary of T_X have a neighborhood with finitely many winners.

Lemma 13 *Given any point $(a, b) \in T_X$ with $a = 1$, there exists a neighborhood $B_\epsilon((a, b))$ such that there are finitely many winning saddle connections on $B_\epsilon((a, b)) \cap T_X$.*

Proof Suppose that we have a point $(a, b) \in T_X$ with $a = 1$ and $b = b'$. Then Lemma 6 guarantees that $(1, b')$ is in some strip $S_\Omega(v)$. If $(1, b')$ is in the interior of $S_\Omega(v)$, then Lemma 8 shows that there is a neighborhood of $(1, b')$ in T_X with finitely many potential winners.

We now consider the case where $(1, b')$ is not in the interior of any strip. This means that $(1, b')$ is on the top boundary of some strip $S_\Omega(v)$. We will first deal with the case where $(1, b')$ is not on the top boundary of T_X: Every point $(1, b' + c)$ for $c > 0$ small enough must also be in some winning strip. Since Lemma 12 tells us that there are finitely many winning saddle connections on the right boundary of T_X where $a = 1$, this then implies that $(1, b')$ is on the bottom boundary of some other strip $S_\Omega(w)$, where w is the winning saddle connection for all $(1, b' + c)$ for $c > 0$ small enough.

Because there are finitely many winning saddle connections on the $a = 1$ line of T_X by Lemma 12, we can now choose an $\epsilon > 0$ small enough that w is the winning saddle connection for $(1, b' + c)$ and v is the winning saddle connection for $(1, b' - c)$ for any $0 < c \le \epsilon$.

We claim now that there are finitely many winning saddle connections on $B_\epsilon((1, b'))$. We recall that for a point $(a, b) \in B_\epsilon((1, b')) \cap T_X$ to have a winning saddle connection other than v or w, there must be a strip $S_\Omega(u)$ for a saddle connection u that is steeper (has more negative slope) than $S_\Omega(v)$ or $S_\Omega(w)$ (whichever is the winner at (a, b)) and that contains (a, b).

Shrinking ϵ if necessary, $B_\epsilon((1, b'))$ lies above the line $b = -(x/y)a + 1/h$ for some $h > 0$ and $(x, y) = v$. Then, as in the proof of Lemma 7, we can show that there are finitely many strips of saddle connections u of (X, ω) with strips $S_\Omega(u)$ intersecting $B_\epsilon((1, b'))$ and that are at least as steep as $S_\Omega(v)$.

If $S_\Omega(u)$ is at most as steep as $S_\Omega(w)$, then it cannot win for any point in $B_\epsilon((1, b')) \cap T_X$ since w or v would win instead.

If $S_\Omega(u)$ has steepness strictly between that of w and v, then for u to be a winner for some point $(a, b) \in B_\epsilon((1, b')) \cap T_X$ we must have that $(a, b) \in S_\Omega(u) \cap (S_\Omega(w) \backslash S_\Omega(v))$. But then, by slope considerations, $S_\Omega(u)$ must also intersect the $a = 1$ boundary of T_X in $B_\epsilon((1, b'))$ above the point $(1, b')$. But this contradicts that w and v were the only winners on the right boundary of T_X in $B_\epsilon((1, b'))$.
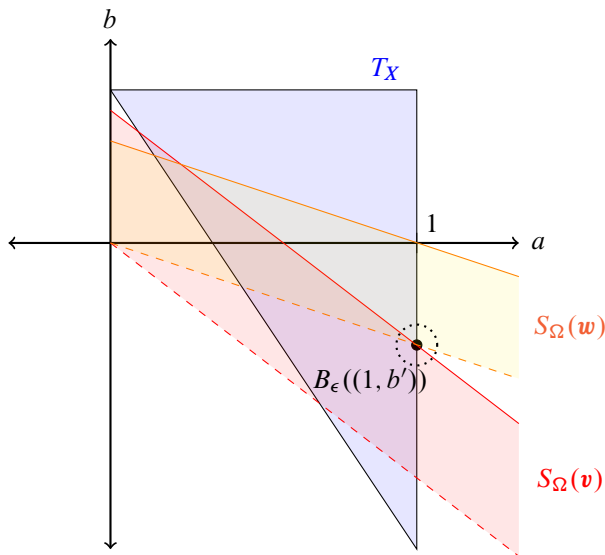
Figure 15: The winning strips $S_\Omega(v)$ and $S_\Omega(w)$ near $(1, b')$ on the right boundary of T_X.

Hence, only the finitely many saddle connections u with strips that intersect $B_\epsilon((1, b'))$ and have slope steeper than that of $S_\Omega(v)$ can be winners on $B_\epsilon((1, b')) \cap T_X$. □

Combining these lemmas shows that for all points in T_X there are finitely many winners in a neighborhood, and hence by compactness there are finitely many winners on T_X.

Proof of Theorem 1 We will consider $\overline{T}_X = T_X \cup \{(0, 1/y_0)\}$. This is a compact set. We showed in Lemmas 9, 10 and 13 that, for any point $(a, b) \in \overline{T}_X$, we can find a neighborhood $B_\epsilon((a, b))$ such that there are finitely many possible winning saddle connections on $B_\epsilon((a, b)) \cap T_X$. Since $\overline{T}_X$ is compact, it is covered by finitely many of these neighborhoods. Since a finite union of finite sets is finite, the set of possible winners on T_X is finite.

Each winning saddle connection v_i would then be a winner on a convex (see the remark below) polygonal piece of T_X. The cumulative distribution function of the slope gap distribution would then be given by the sums of areas between the level curves of the hyperbolic return time functions $y/(a(ax + by))$, as described in Section 2.2, and the sides of these polygons. Since there are finitely many polygonal pieces, the cumulative distribution function and therefore also the slope gap distribution would be piecewise real analytic with finitely many points of nonanalyticity. □

Remark 14 While it is not necessary for the proof of Theorem 1, we can see that each v is a winner on a convex polygonal piece of Ω_i. The convexity arises because the region where v wins is the intersection of finitely many convex regions: the strip $S_\Omega(v)$, the triangular region T_X, and finitely many half-planes that are the upper piece of the complement of $S_\Omega(v')$ for other vectors v' that win on some region of $S_\Omega(v) \cap T_X$.

4 Quadratic tail decay

As an application of the finiteness result (Theorem 1), we prove:

Theorem 2 *The slope gap distribution of any Veech surface has quadratic tail decay. That is, if f denotes the density function of the slope gap distribution, then*

$$\int_t^\infty f(x)\,dx \sim t^{-2}.$$

Proof We find the decay of the tail on a piece of the Poincaré section given by the triangle T_X. Doing this for all the pieces gives the decay of the tail.

The proof of Theorem 1 shows that there exists a minimal finite set of saddle connections $F \subset \Lambda(X,\omega)$ such that, for any point in the triangle T_X, there is some $\boldsymbol{v} \in F$ with $\boldsymbol{M}_{a,b}\boldsymbol{v}$ being the winning saddle connection. Let $S_\Omega(\boldsymbol{v}) \subset T_X$ denote the strip where $\boldsymbol{M}_{a,b}\boldsymbol{v}$ could win and $W_\Omega(\boldsymbol{v}) \subset S_\Omega(\boldsymbol{v})$ denote where $\boldsymbol{M}_{a,b}\boldsymbol{v}$ does win.

Fix $\boldsymbol{v} = (x,y) \in F$. Then the tail on the piece $W_\Omega(\boldsymbol{v})$ is proportional to the area of the set of points (a,b) in $W_\Omega(\boldsymbol{v})$ with

$$\mathrm{slope}(\boldsymbol{M}_{a,b}\boldsymbol{v}) = \frac{y}{a^2 x + aby} > t \iff \frac{1}{at} - \frac{x}{y}a > b.$$

Let $m = x/y$. By adding the contribution that $W_\Omega(\boldsymbol{v})$ gives on the tail for each $\boldsymbol{v} \in F$, we get the full contribution to the tail. In what follows we work on one such winning saddle connection $\boldsymbol{v}$. Hence, it suffices to understand the portion of $W_\Omega(\boldsymbol{v})$ below the hyperbola $b = 1/(at) - ma$. Notice that this hyperbola approaches the line $b = -ma$ from above. Moreover, notice that the line $b = -ma$ is the bottom boundary of the strip $S_\Omega(\boldsymbol{v})$.

We have three situations, depending on how the line $b = -ma$ intersects T_X, as shown in Figure 16.

(1) Suppose $b = -ma$ doesn't intersect T_X. This means that the line $b = -ma$ avoids the bottom edge of T_X for $a \in [0,1]$. In this case we will only find contribution to the tail when the vertical of $\boldsymbol{v}$ is y_0, since otherwise we can choose large enough t so that the hyperbola misses $W_\Omega(\boldsymbol{v})$.
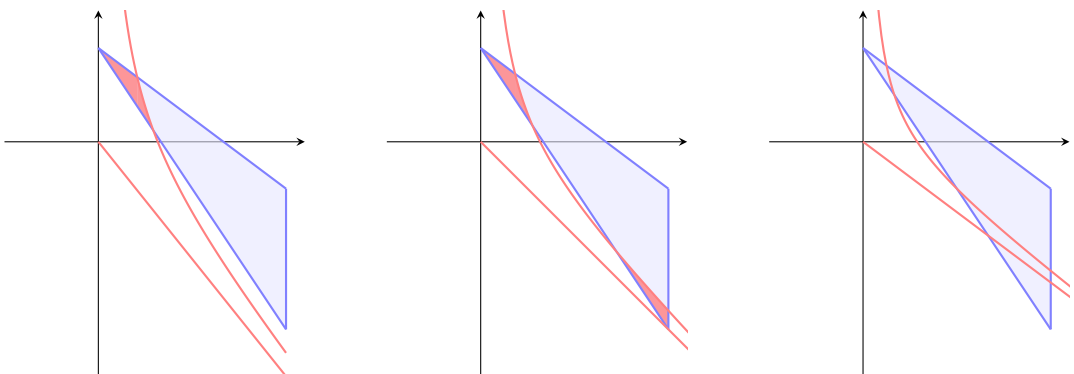
Figure 16: An illustration of cases (1)–(3) in the proof of Theorem 2 (from left to right).

An upper bound for the contribution of $W_\Omega(v)$ is just the part underneath the hyperbola and inside T_X. For t large, the hyperbola $b = 1/(at) - ma$ intersects the triangle twice. First it intersects at the top through the boundary line $b = 1/y_0 - (x_0/y_0)a$ at the point

$$a_{\text{top}}^+ = \frac{-1 + \sqrt{1 + 4y_0(my_0 - x_0)/t}}{2(my_0 - x_0)},$$

and then leaves through the bottom boundary line $b = 1/y_0 - (x_0/y_0 + n)a$ at the point

$$a_{\text{bot}}^+ = \frac{-1 + \sqrt{1 + 4y_0(my_0 - (x_0 + ny_0))/t}}{2(my_0 - (x_0 + ny_0))}.$$

Thus, the contribution is given by

$$\int_{a=0}^{a_{\text{top}}^+} \int_{b=1/y_0 - (x_0/y_0 + n)a}^{1/y_0 - (x_0/y_0)a} 1 \, db \, da + \int_{a=a_{\text{top}}^+}^{a_{\text{bot}}^+} \int_{b=1/y_0 - (x_0/y_0 + n)a}^{1/(at) - ma} 1 \, db \, da.$$

The first integral evaluates to $\frac{1}{2}n(a_{\text{top}}^+)^2$ and, by using a Taylor series on the square root, can be shown to decay like t^{-2}.

The second integral evaluates to

$$\left(\frac{1}{t} \log(a) + \frac{1}{2}\left(\frac{x_0}{y_0} + n - m \right)a^2 - \frac{1}{y_0}a \right)\Bigg|_{a=a_{\text{top}}^+}^{a_{\text{bot}}^+}.$$

By performing a Taylor series approximation on a_{top}^+ and a_{bot}^+, we get that the second integral decays like t^{-3}.

Thus, the total decay on the integral is like t^{-2}.

(2) Now consider the case when $b = -ma$ intersects T_X at the bottom vertex of T_X. In this case $m = x_0/y_0 + n - 1/y_0$. If the vertical of y is the same as y_0, then we get a contribution to the tail at the top of T_X as in case (1). In fact, this is the only way we can get contribution at the top of T_X.

Now we find the contribution on the bottom of T_X. Thus, we are interested in the intersection of the hyperbola $b = 1/(at) - ma$ with the bottom boundary line of T_X given by $1/y_0 - (x_0/y_0 + n)a$. This is the point

$$a_{\text{bot}}^- = \frac{-1 - \sqrt{1 + 4y_0(my_0 - (x_0 + ny_0))/t}}{2(my_0 - (x_0 + ny_0))}.$$

In fact, using that the line $b = -ma$ intersects the bottom of T_X, we get that $m = x_0/y_0 + n - 1/y_0$ and so we can see

$$a_{\text{bot}}^- = \frac{1}{2}\left(1 + \sqrt{1 - \frac{4y_0}{t}} \right).$$

The contribution is then given by

$$\int_{a=a_{\text{bot}}^-}^{1} \int_{b=1/y_0 - (x_0/y_0 + n)a}^{1/(at) - ma} 1 \, db \, da.$$

This integral evaluates to

$$\frac{1}{2}\left(\frac{x_0}{y_0} + n - m\right) - \frac{1}{y_0} - \frac{1}{t}\log(a_{\text{bot}}^-) - \frac{1}{2}\left(\frac{x_0}{y_0} + n - m\right)(a_{\text{bot}}^-)^2 + \frac{a_{\text{bot}}^-}{y_0}.$$

By doing a Taylor series approximation on a_{bot}^- we can show that the decay is like t^{-2}.

(3) Now suppose that the line $b = -ma$ does intersect T_X and this intersection is above the bottom vertex of T_X, ie above $b = 1/y_0 - (x_0/y_0 + n)$. We have two regions to consider: the top of the triangle and the region above between the hyperbola $b = 1/(at) - ma$ and $b = -ma$. The behavior at the top of the triangle is identical to cases (1) and (2), and only occurs when the vertical y is the same as y_0. Thus we have quadratic decay there. We now focus on the second region and observe that each point on the bottom edge of $S_\Omega(v)$ must be in some other winning strip $S_\Omega(v')$. There are finitely many such v', and we number them $v_1, \dots, v_n$. Thus $W_\Omega(v) \subset (S_\Omega(v) - \bigcup_{i=1}^n S_\Omega(v_i))$, which is some polygonal region whose closure is completely above the bottom boundary of $S_\Omega(v)$, $b = -ma$. Since the hyperbola $b = 1/(at) - ma$ approaches $b = -ma$ as $t \to \infty$, for all t large enough the hyperbola is completely below $W_\Omega(v)$ and therefore $W_\Omega(v)$ has no contribution to the tail.

Adding up the contribution of every $v \in F$, we see that there is a constant $C > 0$ such that

$$\int_t^\infty f(x)\, dx \leq \frac{C}{t^2}.$$

Now we compute a lower bound. Let $v_0 = (x_0, y_0)$ be the saddle connection used to define T_X, $S_\Omega(v_0)$ denote the associated strip, and $b = 1/(at) - (x_0/y_0)a$ be the associated hyperbola. We will use this specific saddle connection to find a lower bound to $\int_t^\infty f(x)\, dx$, essentially by using the argument from case (1) of the upper bound. That is, by analyzing the behavior at the top of the triangle. Either v_0 is the winning saddle connection for every point on $S_\Omega(v_0)$ or there is some other saddle connection v for which it is the winning saddle connection on $S_\Omega(v_0) \cap S_\Omega(v)$. We deal with both cases.

(i) If v_0 is the winning saddle connection for every point on $S_\Omega(v_0)$, then a lower bound to $\int_t^\infty f(x)\, dx$ comes from the part underneath the hyperbola $b = 1/(at) - (x_0/y_0)a$ and inside $S_\Omega(v_0)$. We can choose t large enough that the hyperbola intersects $S_\Omega(v_0)$ only once, at the point

$$a_{\text{top}}^+ = \frac{-1 + \sqrt{1 + 4y_0(my_0 - x_0)/t}}{2(my_0 - x_0)}$$

with contribution given by

$$\int_{a=0}^{a_{\text{top}}^+} \int_{b=1/y_0-((x_0/y_0)+n)a}^{1/y_0-x_0/y_0 a} 1\, db\, da.$$

Earlier we showed this decays like t^{-2}.

(ii) In the case that there is some other saddle connection v that is the winning saddle connection on $S_\Omega(v_0) \cap S_\Omega(v)$ we have two subcases, depending on whether v has the same vertical as v_0 or not. In the latter case we can choose t large enough that the contribution is the same as case (1). We now focus on when the vertical of v and v_0 is the same. Furthermore, since we are looking for any lower bound, it

suffices to assume that v has the least negative slope among all vectors that win in the intersection of $S_\Omega(v_0) \cap S_\Omega(v)$. The contribution is given by

$$\int_{a=0}^{a_{\text{top}}^+} \int_{b=1/y_0-(x/y)a}^{1/y_0-(x_0/y_0)a} 1 \, db \, da.$$

The integral evaluates to $\frac{1}{2}(x/y - x_0/y_0)(a_{\text{top}}^+)^2$ and, by using a Taylor series on the square root, can be shown to decay like t^{-2}. $\qquad\square$

5 Further questions

We end with a few questions for further exploration:

(1) Are there bounds on the number of points of nonanalyticity of the slope gap distribution of a Veech surface?

In [5], linear upper and lower bounds in terms of n on the number of points were found for the translation surface given by gluing opposite sides of the $2n$–gon. These surfaces each have two cusps and have genus that grows linearly in n. This shows that bounds on the number of points of nonanalyticity based on the number of cusps is impossible. However, we can ask if there are bounds based on the genus of the surface.

(2) What can be said about the gap distributions of non-Veech surfaces?

In [2] it was shown that the limiting slope gap distribution exists for almost every translation surface, and in [9] the slope gap distributions for a special family of non-Veech surfaces were shown to be piecewise real analytic. We can ask if the limiting slope gap distributions are always piecewise real analytic, and if so, are there always finitely many points of nonanalyticity?

(3) Where do the points of nonanalyticity lie?

Beyond just understanding the number of points of nonanalyticity, we can ask about number-theoretic properties of the points themselves. In every example known to the authors of a limiting slope gap distribution, after rescaling, the points of nonanalyticity lie in the trace field of the Veech group. Given that the gap distribution is computed by integrating areas between hyperbolas in regions related to the geometry of the surface, it is natural to conjecture that points of nonanalyticity lie in quadratic extensions of the trace field.

References

[1]   **J S Athreya**, *Gap distributions and homogeneous dynamics*, from "Geometry, topology, and dynamics in negative curvature" (C S Aravinda, F T Farrell, J-F Lafont, editors), Lond. Math. Soc. Lect. Note Ser. 425, Cambridge Univ. Press (2016) 1–31   MR Zbl

[2] **J S Athreya**, **J Chaika**, *The distribution of gaps for saddle connection directions*, Geom. Funct. Anal. 22 (2012) 1491–1516 MR Zbl

[3] **J S Athreya**, **J Chaika**, **S Lelièvre**, *The gap distribution of slopes on the golden L*, from "Recent trends in ergodic theory and dynamical systems" (S Bhattacharya, T Das, A Ghosh, R Shah, editors), Contemp. Math. 631, Amer. Math. Soc., Providence, RI (2015) 47–62 MR Zbl

[4] **J S Athreya**, **Y Cheung**, *A Poincaré section for the horocycle flow on the space of lattices*, Int. Math. Res. Not. 2014 (2014) 2643–2690 MR Zbl

[5] **J Berman**, **T McAdam**, **A Miller-Murthy**, **C Uyanik**, **H Wan**, *Slope gap distribution of saddle connections on the 2n–gon*, Discrete Contin. Dyn. Syst. 43 (2023) 1–56 MR Zbl

[6] **P Hubert**, **T A Schmidt**, *An introduction to Veech surfaces*, from "Handbook of dynamical systems, 1B" (B Hasselblatt, A Katok, editors), Elsevier, Amsterdam (2006) 501–526 MR Zbl

[7] **H Masur**, *Lower bounds for the number of saddle connections and closed trajectories of a quadratic differential*, from "Holomorphic functions and moduli, I" (D Drasin, C J Earle, F W Gehring, I Kra, A Marden, editors), Math. Sci. Res. Inst. Publ. 10, Springer (1988) 215–228 MR Zbl

[8] **H Masur**, *The growth rate of trajectories of a quadratic differential*, Ergodic Theory Dynam. Systems 10 (1990) 151–176 MR Zbl

[9] **A Sanchez**, *Gaps of saddle connection directions for some branched covers of tori*, Ergodic Theory Dynam. Systems 42 (2022) 3191–3245 MR Zbl

[10] **D Taha**, *On cross sections to the geodesic and horocycle flows on quotients of* $SL(2, \mathbb{R})$ *by Hecke triangle groups* G_q, preprint (2019) arXiv 1906.07250

[11] **C Uyanik**, **G Work**, *The distribution of gaps for saddle connections on the octagon*, Int. Math. Res. Not. 2016 (2016) 5569–5602 MR Zbl

[12] **Y Vorobets**, *Periodic geodesics on generic translation surfaces*, from "Algebraic and topological dynamics" (S Kolyada, Y Manin, T Ward, editors), Contemp. Math. 385, Amer. Math. Soc., Providence, RI (2005) 205–258 MR Zbl

[13] **A Zorich**, *Flat surfaces*, from "Frontiers in number theory, physics, and geometry, I" (P Cartier, B Julia, P Moussa, P Vanhove, editors), Springer (2006) 437–583 MR Zbl

Department of Mathematics, Massachusetts Institute of Technology
Cambridge, MA, United States

Department of Mathematics, University of California San Diego
La Jolla, CA, United States

Department of Mathematics and Statistics, University of Maine
Orono, ME, United States

luisk@mit.edu,   ans032@ucsd.edu,   jane.wang@maine.edu

Embedding calculus for surfaces

Manuel Krannich

Alexander Kupers

We prove convergence of the Goodwillie–Weiss embedding calculus for spaces of embeddings into a manifold of dimension at most two, so in particular for diffeomorphisms between surfaces. We also relate the Johnson filtration of the mapping class group of a surface to a certain filtration arising from embedding calculus.

1 Introduction

For smooth manifolds M and N, and an embedding $e_\partial\colon \partial M \hookrightarrow \partial N$, we write $\mathrm{Emb}_\partial(M, N)$ for the space of embeddings that agree with e_∂ on ∂M, equipped with the smooth topology. Embedding calculus à la Goodwillie and Weiss provides a space $T_\infty\mathrm{Emb}_\partial(M, N)$ and a map

$$(1) \qquad \mathrm{Emb}_\partial(M, N) \to T_\infty\mathrm{Emb}_\partial(M, N),$$

which approximates the space of embeddings through restrictions to subsets diffeomorphic to a finite collection of open discs and a collar. The space $T_\infty\mathrm{Emb}_\partial(M, N)$ arises as a homotopy limit of a tower of maps whose homotopy fibres have an explicit description in terms of the configuration spaces of M and N — see Weiss [27] — so its homotopy type is sometimes easier to study than that of $\mathrm{Emb}_\partial(M, N)$. The main result in this context is due to Goodwillie, Klein and Weiss [11; 12] and says that if the difference of the dimension of N and the relative handle dimension of the boundary inclusion $\partial M \subset M$ is at least three, then embedding calculus *converges* in the sense that (1) is a weak homotopy equivalence. If this assumption is not met, little is known about for which choices of M and N embedding calculus converges (but see Remark 1.1(ii) and (vi) below).

1.1 Convergence in low dimensions

In the first part of this work, we study (1) when the target N has dimension at most two. Our main result shows that embedding calculus always converges under this assumption, even though the assumption on the handle codimension is not satisfied.

Theorem A *For compact manifolds M and N with $\dim(N) \leq 2$, the map*

$$\mathrm{Emb}_\partial(M, N) \to T_\infty \mathrm{Emb}_\partial(M, N)$$

is a weak homotopy equivalence for any embedding $e_\partial : \partial M \hookrightarrow \partial N$.

Perhaps the most interesting (hence eponymous) instance of Theorem A is when $M = N$ is a surface Σ and $e_\partial = \mathrm{id}_{\partial \Sigma}$. In this case Theorem A specialises to the following:

Corollary B *For a compact surface Σ, possibly with boundary and nonorientable, the map*

$$\mathrm{Diff}_\partial(\Sigma) \to T_\infty \mathrm{Emb}_\partial(\Sigma, \Sigma)$$

is a weak homotopy equivalence.

Remark 1.1   (i)   We prove Theorem A as a special case of a more general result that also treats embedding spaces of *triads* (see Theorem 3.1).

  (ii)   Theorem A is special to dimension at most 2: in [18], we show that this results fails for $N = D^3$ and for most high-dimensional compact manifolds N. In the language of that paper, Theorem A implies that the smooth Disc–structure space $\mathcal{S}_\partial^{\mathrm{Disc}}(N)$ is contractible if $\dim(N) \leq 2$.

  (iii)   The proof of Theorem A does not rely on Goodwillie, Klein and Weiss's convergence results.

  (iv)   Theorem A is stronger than Corollary B, even if $\dim(M) = \dim(N) = 2$. It implies that $T_\infty \mathrm{Emb}_\partial(\Sigma, \Sigma') = \varnothing$ if Σ and Σ' are connected compact surfaces that are not diffeomorphic.

  (v)   Composition induces an E_1–structure on $T_\infty \mathrm{Emb}_\partial(M, M)$ with respect to which the map

$$\mathrm{Emb}_\partial(M, M) \to T_\infty \mathrm{Emb}_\partial(M, M)$$

  is an E_1–map. For a compact manifold M, the E_1–space $\mathrm{Emb}_\partial(M, M) = \mathrm{Diff}_\partial(M)$ is grouplike, but it is not known whether the same holds for $T_\infty \mathrm{Emb}_\partial(M, M)$. Theorem A implies that this is the case if $\dim(M) \leq 2$.

  (vi)   Theorem A provides a class of examples for which the map $\mathrm{Emb}_\partial(M, N) \to T_\infty \mathrm{Emb}(M, N)$ is a weak equivalence in handle codimension less than three. A few examples of this form were known before; see Knudsen and Kupers [17, Theorem C, Section 6.2.4]. In contrast, there are some cases for which it is known that embedding calculus does not converge, such as for $M = D^1$ and $N = D^3$ by an argument due to Goodwillie.

1.2   Embedding calculus and the Johnson filtration

The *Johnson filtration*

$$\pi_0 \mathrm{Diff}_\partial(\Sigma) = \mathcal{J}(0) \supset \mathcal{J}(1) \supset \mathcal{J}(2) \supset \cdots$$

of the mapping class group $\pi_0 \mathrm{Diff}_\partial(\Sigma)$ of an orientable surface Σ of genus g with one boundary component is the filtration by the kernels of the action of $\pi_0 \mathrm{Diff}_\partial(\Sigma)$ on the quotients of the fundamental

group $\pi_1(\Sigma, *)$ based at the point in the boundary, by the constituents of its lower central series. By work of Moriyama [21], this filtration can be recovered from the action of $\pi_0\mathrm{Diff}_\partial(\Sigma)$ on the compactly supported cohomology of the configuration spaces of the punctured surface $\Sigma \setminus \{*\}$. It is reasonable to expect a relationship between the Johnson filtration and embedding calculus, as the latter may be viewed as the study of embeddings via their induced maps between the homotopy types of configuration spaces of thickened points in source and target.

The second part of this work serves to establish one such a relationship: we introduce a filtration

$$(2) \qquad \pi_0\mathrm{Diff}_\partial(\Sigma) = T\mathscr{I}_{\partial/2}^{H\mathbb{Z}}(0) \supset T\mathscr{I}_{\partial/2}^{H\mathbb{Z}}(1) \supset T\mathscr{I}_{\partial/2}^{H\mathbb{Z}}(2) \supset \cdots$$

arising from the cardinality filtration of embedding calculus in $H\mathbb{Z}$–modules applied to the space of self-embeddings fixed on an interval in the boundary (see Section 4 for precise definitions), and we use [21] to show that this filtration contained in the Johnson filtration

$$T\mathscr{I}_{\partial/2}^{H\mathbb{Z}}(k) \subset \mathscr{I}(k) \quad \text{for } k \geq 0.$$

Acknowledgements

2 Generalities on spaces of embeddings and embedding calculus

We begin by fixing some conventions on spaces of embeddings, followed by recalling various known properties of embedding calculus and complementing them with some new properties such as a lemma for lifting embeddings along covering spaces in the context of embedding calculus.

2.1 Spaces of embeddings and maps

All our manifolds will be smooth and may be noncompact, disconnected, or nonorientable. A *manifold triad* is a manifold M together with a decomposition of its boundary $\partial M = \partial_0 M \cup \partial_1 M$ into two codimension-zero submanifolds that intersect at a set $\partial(\partial_0 M) = \partial(\partial_1 M)$ of corners. Any of these sets may be empty or disconnected. If this decomposition is not specified, we implicitly take $\partial_0 M = \partial M$ and $\partial_1 M = \varnothing$.

When studying embeddings between manifolds triads M and N, we always fix a *boundary condition*, ie an embedding $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$, and only consider embeddings $e \colon M \hookrightarrow N$ that restrict to e_{∂_0} on $\partial_0 M$ and have near $\partial_0 M$ the form $e_{\partial_0} \times \mathrm{id}_{[0,1)} \colon \partial_0 M \times [0,1) \hookrightarrow \partial_0 N \times [0,1)$ with respect to collars of $\partial_0 M$ and $\partial_0 N$. We denote the space of such embeddings in the weak $\mathscr{C}^\infty$–topology by $\mathrm{Emb}_{\partial_0}(M,N)$. We replace the subscript ∂_0 by ∂ to indicate that $\partial_0 M = \partial M$, and drop the subscript if we want to emphasise that $\partial_0 M = \varnothing$ holds. As a final piece of notation, given manifold triads M and L, we consider $M \sqcup L$ as a manifold triad via $\partial_0(M \sqcup L) = \partial_0 M \sqcup \partial_0 L$.

Similarly, we also consider the space of bundle maps $\mathrm{Bun}_{\partial_0}(TM, TN)$. By this we mean the space of fibrewise injective linear maps $TM \to TN$ that restrict to the derivative $d(e_{\partial_0})$ on $T\partial_0 M$, in the compact-open topology. Taking derivatives induces a map $\mathrm{Emb}_{\partial_0}(M,N) \to \mathrm{Bun}_{\partial_0}(TM, TN)$ which we may postcompose with the forgetful map $\mathrm{Bun}_{\partial_0}(TM, TN) \to \mathrm{Map}_{\partial_0}(M,N)$ to the space of continuous maps extending e_{∂_0}, equipped with the compact-open topology.

2.2 Manifold calculus

Given manifold triads M and N and a boundary condition $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ as above, Goodwillie and Weiss's *embedding calculus* [12; 27] gives a space $T_\infty \mathrm{Emb}_{\partial_0}(M,N)$ (or rather, a homotopy type) together with a map

(3) $$\mathrm{Emb}_{\partial_0}(M,N) \to T_\infty \mathrm{Emb}_{\partial_0}(M,N).$$

Embedding calculus *converges* if the map (3) is a weak homotopy equivalence (shortened to *weak equivalence* throughout this work). This fits into the more general context of *manifold calculus*, and we shall need this generalisation at several places.

2.2.1 Manifold calculus in terms of presheaves Among the various models for the map (3) and manifold calculus in general, that of Boavida de Brito and Weiss in terms of presheaves [1] is most convenient for our purposes. We refer to Section 8 of their work for a proof of the equivalence between this model and the classical model of [27].

To recall their model (in a slightly more general setting; see Remark 2.5), we fix a $(d-1)$–manifold K possibly with boundary, thought of as $\partial_0 M$ for manifold triads M. We write Disc_K for the topologically enriched category whose objects are smooth d–dimensional manifold triads that are diffeomorphic (as triads) to $K \times [0,1) \sqcup T \times \mathbb{R}^d$ for a finite set T with $\partial_0(K \times [0,1) \sqcup T \times \mathbb{R}^d) = K \times \{0\}$, and whose morphisms are given by spaces of embeddings of triads as described in Section 2.1. If K is clear from the context, we abbreviate Disc_K by $\mathrm{Disc}_{\partial_0}$.

We write $\mathrm{PSh}(\mathrm{Disc}_{\partial_0})$ for the topologically enriched category of space-valued enriched presheaves on $\mathrm{Disc}_{\partial_0}$, and we consider it as a category with weak equivalences by declaring a morphism of presheaves to be a weak equivalence if it is a weak equivalence on all its values. Localising at these weak equivalences

(for instance as described in [7]) gives rise to a topologically enriched category $\mathsf{PSh}(\mathsf{Disc}_{\partial_0})^{\mathrm{loc}}$ together with an enriched functor

(4) $$\mathsf{PSh}(\mathsf{Disc}_{\partial_0}) \to \mathsf{PSh}(\mathsf{Disc}_{\partial_0})^{\mathrm{loc}}.$$

Denoting by $\mathsf{Man}_{\partial_0}$ the topologically enriched category with objects all manifold triads M with an identification $\partial_0 M \cong K$ and morphism spaces the spaces of embeddings of triads, a presheaf $F \in \mathsf{PSh}(\mathsf{Disc}_{\partial_0})$ induces a new presheaf $T_\infty F \in \mathsf{PSh}(\mathsf{Man}_{\partial_0})$ by setting

$$T_\infty F(M) := \mathsf{Map}_{\mathsf{PSh}(\mathsf{Disc}_{\partial_0})^{\mathrm{loc}}}(\mathsf{Emb}_{\partial_0}(-, M), F).$$

If F is the restriction of a presheaf $F \in \mathsf{PSh}(\mathsf{Man}_{\partial_0})$, then we have a composition of maps of presheaves

(5) $$F(M) \xrightarrow{\cong} \mathsf{Map}_{\mathsf{PSh}(\mathsf{Man}_{\partial_0})}(\mathsf{Emb}_{\partial_0}(-, M), F) \to T_\infty F(M)$$

on $\mathsf{Man}_{\partial_0}$ where the first map is given by the enriched Yoneda lemma and the second is induced by the restriction along $\mathsf{Disc}_{\partial_0} \subset \mathsf{Man}_{\partial_0}$ and the functor (4). Note that this is a weak equivalence whenever $M \in \mathsf{Disc}_{\partial_0}$, that is, manifold calculus *converges* on manifolds diffeomorphic to the disjoint union of a collar on $\partial_0 M$ and a finite number of open discs.

Example 2.1 (embedding calculus) For triads M and N and a boundary condition $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$, we have a presheaf $\mathsf{Emb}_{\partial_0}(-, N)$ of embeddings of triads extending e_{∂_0}. Choosing $K = \partial_0 M$, the map (5) gives rise to a model for the embedding calculus map (3),

(6) $$\mathsf{Emb}_{\partial_0}(M, N) \to \mathsf{Map}_{\mathsf{PSh}(\mathsf{Disc}_{\partial_0})^{\mathrm{loc}}}(\mathsf{Emb}_{\partial_0}(-, M), \mathsf{Emb}_{\partial_0}(-, N)) = T_\infty \mathsf{Emb}_{\partial_0}(M, N).$$

Remark 2.2 There are several alternative points of view on the maps (5) and (6), for instance in terms of modules over variants of the little discs operad; see [1, Section 6] or [26].

2.2.2 A smaller model In some situations, it is convenient to replace $\mathsf{Disc}_{\partial_0}$ by a smaller equivalent category. There is a chain of enriched functors

(7) $$\mathsf{Disc}_{\partial_0}^\bullet \to \mathsf{Disc}_{\partial_0}^{\mathrm{sk}} \to \mathsf{Disc}_{\partial_0}.$$

The right arrow is the inclusion of the full subcategory $\mathsf{Disc}_{\partial_0}^{\mathrm{sk}} \subset \mathsf{Disc}_{\partial_0}$ on the objects $\partial_0 M \times [0, 1) \sqcup \underline{n} \times \mathbb{R}^d$ for $\underline{n} = \{1, \ldots, n\}$ with $n \geq 0$. The category $\mathsf{Disc}_{\partial_0}^\bullet$ has the same objects as $\mathsf{Disc}_{\partial_0}^{\mathrm{sk}}$ and space of morphisms pairs (s, e) of a parameter $s \in (0, 1]$ and an embedding of triads

$$e \colon \partial_0 M \times [0, 1) \sqcup \underline{n} \times \mathbb{R}^d \to \partial_0 M \times [0, 1) \sqcup \underline{m} \times \mathbb{R}^d$$

with $e|_{\partial_0 M \times [0,1)} = \mathrm{id}_{\partial_0 M} \times s \cdot (-)$, where $s \cdot (-) \colon [0, 1) \to [0, 1)$ is multiplication by s. Composition is given by composing embeddings and multiplying parameters, and the functor to $\mathsf{Disc}_{\partial_0}^{\mathrm{sk}}$ forgets the parameters. Both functors in (7) are Dwyer–Kan equivalences, the first by a variant of the proof of the contractibility of the space of collars and the second by definition, so we may equivalently define $T_\infty F(-)$ using any of the three categories (7).

2.2.3 Two properties of manifold calculus

The following two properties of the functor

$$\mathsf{PSh}(\mathsf{Disc}_{\partial_0}) \ni F \mapsto T_\infty F \in \mathsf{PSh}(\mathsf{Man}_{\partial_0}) \tag{8}$$

will be of use:

(a) **Homotopy limits** The mapping spaces resulting from the localisation (4) can be viewed equivalently as the derived mapping spaces formed with respect to the projective model structure on $\mathsf{PSh}(\mathsf{Disc}_{\partial_0})$; see [1, Section 3.1]. That is, the functor (8) models the homotopy right Kan-extension along the inclusion $\mathsf{Disc}_{\partial_0} \subset \mathsf{Man}_{\partial_0}$ [1, Section 4.2]. The functor (8) thus preserves homotopy limits in the projective model structures, which are computed objectwise.

(b) $\mathcal{J}_\infty$–**covers and descent** If F is the restriction of a presheaf $F \in \mathsf{PSh}(\mathsf{Man}_{\partial_0})$ then $T_\infty F$ can be seen alternatively as the *homotopy $\mathcal{J}_\infty$–sheafification* of F: for $1 \le k \le \infty$ (we will only use the cases $k = 1, \infty$), a nonempty open cover $\mathcal{U}$ of a triad M is called a *Weiss k–cover* if every $U \in \mathcal{U}$ contains an open collar on $\partial_0 M$ and every finite subset of cardinality $\le k$ of $\mathrm{int}(M)$ is contained in some element of $\mathcal{U}$. An enriched presheaf on $\mathsf{Man}_{\partial_0}$ is a *homotopy $\mathcal{J}_k$–sheaf* if it satisfies descent for Weiss k–covers in sense of [1, Definition 2.2]. Note that a homotopy $\mathcal{J}_1$–sheaf is a homotopy sheaf in the usual sense, and a homotopy $\mathcal{J}_k$–sheaf is also a homotopy $\mathcal{J}_{k'}$–sheaf for any $k' \ge k$. By [1, Theorem 1.2], the functor

$$\mathsf{PSh}(\mathsf{Man}_{\partial_0}) \ni F \mapsto T_\infty F \in \mathsf{PSh}(\mathsf{Man}_{\partial_0})$$

together with the natural transformation $\mathrm{id}_{\mathsf{PSh}(\mathsf{Man}_{\partial_0})} \Rightarrow T_\infty$ is a model for the homotopy $\mathcal{J}_\infty$–sheafification. In particular, if F is already a $\mathcal{J}_\infty$–sheaf, then $F \to T_\infty F$ is a weak equivalence, so any map $F \to G$ in $\mathsf{PSh}(\mathsf{Man}_{\partial_0})$ with G a homotopy $\mathcal{J}_k$–sheaf for some $1 \le k \le \infty$ factors over $F \to T_\infty F$ up to weak equivalence.

It is often convenient to use a stronger version of descent, namely with respect to *complete* Weiss ∞–covers $\mathcal{U}$, which are Weiss ∞–covers that contain a Weiss ∞–cover of any finite intersection of elements in $\mathcal{U}$. Regarding $\mathcal{U}$ as a poset ordered by inclusion, the map induced by restriction

$$T_\infty F(M) \to \operatorname*{holim}_{U \in \mathcal{U}} T_\infty F(U)$$

is a weak equivalence by [17, Lemma 6.7].

Remark 2.3 At several points in the remainder of this work, we will construct maps between spaces of the form $T_\infty \mathsf{Emb}_{\partial_0}(M, N)$ by using the descent property from Section 2.2.3(b). Strictly speaking, these will only be *weak maps*, ie zigzags of maps whose wrong-way maps are weak equivalences. This will be good enough for all purposes. More formally, a weak map $X \to Y$ gives an actual morphism from X and Y in the localisation of the category of spaces at the weak equivalences, and all our statements involving weak maps can be viewed as taking place in this localisation. In particular, when we say that a square involving weak maps *commutes up to canonical homotopy* then we mean that the square can be enhanced in a preferred way to a homotopy commutative square in this localisation.

2.3 Properties of embedding calculus

We explain various features of embedding calculus which illustrate that $T_\infty\mathrm{Emb}_{\partial_0}(M, N)$ has formally similar properties to $\mathrm{Emb}_{\partial_0}(M, N)$ even in situations where embedding calculus need not converge.

(a) **Postcomposition with embeddings** Given triads M, N, and K, with boundary conditions

$$e_{\partial_0 M}: \partial_0 M \hookrightarrow \partial_0 N \quad \text{and} \quad e_{\partial_0 N}: \partial_0 N \hookrightarrow \partial_0 K,$$

there is a map

$$T_\infty\mathrm{Emb}_{\partial_0}(M, N) \times \mathrm{Emb}_{\partial_0}(N, K) \to T_\infty\mathrm{Emb}_{\partial_0}(M, K)$$

that is associative in the evident sense and compatible with the composition maps for embeddings spaces, both up to higher coherent homotopy.

In the model of Section 2.2.1, these maps are given by applying the map

$$(9) \qquad \mathrm{Emb}_{\partial_0}(N, K) \to \mathrm{Map}_{\mathsf{PSh}(\mathrm{Disc}_{\partial_0 M})^{\mathrm{loc}}}(\mathrm{Emb}_{\partial_0}(-, N), \mathrm{Emb}_{\partial_0}(-, K))$$

induced by postcomposition in the second factor, followed by composition in $\mathsf{PSh}(\mathrm{Disc}_{\partial_0 M})^{\mathrm{loc}}$. Note that the codomain of (9) does in general *not* agree with $T_\infty\mathrm{Emb}_{\partial_0}(N, K)$.

(b) **Naturality and isotopy invariance** In the situation of (a), if we assume $\dim(M) = \dim(N)$, then there are composition maps

$$(10) \qquad T_\infty\mathrm{Emb}_{\partial_0}(M, N) \times T_\infty\mathrm{Emb}_{\partial_0}(N, K) \to T_\infty\mathrm{Emb}_{\partial_0}(M, K)$$

that are associative in the evident sense and compatible with (9) and the composition for embeddings, up to higher coherent homotopy. Combining this with (a), we see that like spaces of embeddings, $T_\infty\mathrm{Emb}_{\partial_0}(-, -)$ is isotopy-invariant in source and target: if $M \subset M'$ is a subtriad with $\partial_0 M \subset \partial_0 M'$ such that there is an embedding of triads $M' \hookrightarrow M$ which is inverse to the inclusion up to isotopy of triads, then the maps

$$T_\infty\mathrm{Emb}_{\partial_0}(M', N) \to T_\infty\mathrm{Emb}_{\partial_0}(M, N) \quad \text{and} \quad T_\infty\mathrm{Emb}_{\partial_0}(L, M) \to T_\infty\mathrm{Emb}_{\partial_0}(L, M')$$

induced by restriction and inclusion are weak equivalences. Here L is any other triad with a boundary condition $e_{\partial_0}: \partial_0 L \hookrightarrow \partial_0 M$.

In the model described in Section 2.2.1, the composition map (10) can implemented as follows: the codimension-0 embedding $e_{\partial_0 M}: \partial_0 M \hookrightarrow \partial_0 N$ induces enriched functors

$$(e_{\partial_0 M})_*: \mathrm{Disc}^\bullet_{\partial_0 M} \to \mathrm{Disc}^\bullet_{\partial_0 N} \quad \text{and} \quad (e_{\partial_0 M})^*: \mathsf{PSh}(\mathrm{Disc}^\bullet_{\partial_0 N}) \to \mathsf{PSh}(\mathrm{Disc}^\bullet_{\partial_0 M}).$$

Writing $d := \dim(M) = \dim(N)$, $(e_{\partial_0 M})_*$ sends objects $\partial_0 M \times [0, 1) \sqcup \underline{n} \times \mathbb{R}^d$ to $\partial_0 N \times [0, 1) \sqcup \underline{n} \times \mathbb{R}^d$. For morphisms, $(e_{\partial_0 M})_*$ keeps the parameter s fixed and sends an embedding e to the embedding given by $\mathrm{id}_{\partial_0 N} \times (s \cdot (-))$ on $\partial_0 M \times [0, 1)$ and by $(e_{\partial_0 M} \times [0, 1) \sqcup \mathrm{id}_{\underline{n} \times \mathbb{R}^d}) \circ e|_{\underline{n} \times \mathbb{R}^d}$ on $\underline{n} \times \mathbb{R}^d$. The functor $(e_{\partial_0 M})^*$ is given by precomposition with $(e_{\partial_0 M})_*$. The restriction maps

$$\mathrm{Emb}_{\partial_0 N}(\partial_0 N \times [0, 1) \sqcup \underline{n} \times \mathbb{R}^d, N) \to \mathrm{Emb}_{\partial_0 M}(\partial_0 M \times [0, 1) \sqcup \underline{n} \times \mathbb{R}^d, N)$$

are weak equivalences by the contractibility of spaces of collars, and similarly for $\mathrm{Emb}_{\partial_0}(-, K)$, so we have weak equivalences in $\mathsf{PSh}(\mathsf{Disc}^\bullet_{\partial_0 M})$,

$$(e_{\partial M})^* \mathrm{Emb}_{\partial_0 N}(-, N) \xrightarrow{\simeq} \mathrm{Emb}_{\partial_0 M}(-, N), \quad (e_{\partial M})^* \mathrm{Emb}_{\partial_0 N}(-, K) \xrightarrow{\simeq} \mathrm{Emb}_{\partial_0 M}(-, K).$$

Using the model

$$T_\infty \mathrm{Emb}_{\partial_0}(M, N) \simeq \mathrm{Map}_{\mathsf{PSh}(\mathsf{Disc}^\bullet_{\partial_0 M})^{\mathrm{loc}}}(\mathrm{Emb}_{\partial_0}(-, M), \mathrm{Emb}_{\partial_0}(-, N)),$$

the composition (10) is given by applying $(e_{\partial M})^*$ to the second factor, composition in the category $\mathsf{PSh}(\mathsf{Disc}^\bullet_{\partial_0 M})^{\mathrm{loc}}$, and using the weak equivalences of presheaves above.

(c) **Convergence on disjoint unions of discs** Embedding calculus converges if the domain M is diffeomorphic (as a triad) to $\partial_0 M \times [0, 1) \sqcup T \times \mathbb{R}^d$ for a finite set T, where

$$\partial_0(\partial_0 M \times [0, 1) \sqcup T \times \mathbb{R}^d) = \partial_0 M \times \{0\}.$$

This follows from the corresponding fact for manifold calculus (see Section 2.2.1). By isotopy invariance, it remains true with $T \times \mathbb{R}^d$ replaced by $T_1 \times \mathbb{R}^d \sqcup T_2 \times D^d$ for finite sets T_i.

(d) **Comparison to bundle maps** The derivative map $\mathrm{Emb}_{\partial_0}(M, N) \to \mathrm{Bun}_{\partial_0}(TM, TN)$ fits into a natural commutative diagram (up canonical homotopy) of the form

(11)
$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(M, N) & \longrightarrow \mathrm{Bun}_{\partial_0}(TM, TN) \longrightarrow \mathrm{Map}_{\partial_0}(M, N) \\
\downarrow & \nearrow \\
T_\infty \mathrm{Emb}_{\partial_0}(M, N)
\end{array}
$$

which is compatible with composition maps from (10) up to higher coherent homotopy. This follows from Section 2.2.3(b) by observing that the target in the natural transformation $\mathrm{Emb}_{\partial_0}(-, N) \to \mathrm{Bun}_{\partial_0}(-, TN)$ is a homotopy $\mathcal{J}_1$–sheaf, so the map $\mathrm{Bun}_{\partial_0}(-, TN) \to T_\infty \mathrm{Bun}_{\partial_0}(-, TN)$ is a weak equivalence of presheaves.

(e) **Extension by the identity** Suppose that we have another triad Q with an identification of $\partial_0 Q$ with a codimension-zero submanifold of $\partial_0 M$. Then we can form, up to smoothing corners, the triad $M \cup Q = M \cup_{\partial_0 Q} Q$ with $\partial_0(M \cup Q) = (\partial_0 M \setminus \mathrm{int}(\partial_0 Q)) \cup \partial_1 Q$. If M and N are of the same dimension and we are further given a boundary condition $e_{\partial_0} : \partial_0 M \hookrightarrow \partial_0 N$, we can form $N \cup Q$ in the same manner. Extending embeddings by the identity gives a map $\mathrm{Emb}_{\partial_0}(M, N) \to \mathrm{Emb}_{\partial_0}(M \cup Q, N \cup Q)$ (strictly speaking this requires the addition of collars to the definitions to guarantee the glued map is smooth but we forego the addition of this contractible space of data), which can be shown to fit into a diagram

(12)
$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(M, N) & \longrightarrow & \mathrm{Emb}_{\partial_0}(M \cup Q, N \cup Q) \\
\downarrow & & \downarrow \\
T_\infty \mathrm{Emb}_{\partial_0}(M, N) & \dashrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(M \cup Q, N \cup Q)
\end{array}
$$

commutative up to preferred homotopy. The existence of the dashed map in (12) is proved by noting that $T_\infty \mathrm{Emb}_{\partial_0}(- \cup Q, N \cup Q)$ is a homotopy $\mathcal{J}_\infty$–sheaf on $\mathrm{Disc}_{\partial_0 M}$; see Section 2.2.3(b).

(f) **Isotopy extension**   Suppose that the triads M and N are both d–dimensional, and $e_{\partial_0 M} : \partial_0 M \hookrightarrow \partial_0 N$ is a boundary condition. Fix a compact d–dimensional submanifold triad $P \subset M$ (so, in particular, $\partial_0 P = \partial_0 M \cap \partial P$) and consider the induced boundary condition $e_{\partial_0} : \partial_0 M \supset \partial_0 P \hookrightarrow \partial_0 N$. Suppose that embedding calculus converges for triad embeddings of triads of the form $P \sqcup T \times \mathbb{R}^d \hookrightarrow N$ for finite sets T in the sense that the map

$$\mathrm{Emb}_{\partial_0}(P \sqcup T \times \mathbb{R}^d, N) \to T_\infty \mathrm{Emb}_{\partial_0}(P \sqcup T \times \mathbb{R}^d, N)$$

is a weak equivalence. Then, fixing a triad embedding $e \colon P \hookrightarrow N$ disjoint from $\partial N \setminus e_{\partial_0 M}(\partial_0 P)$, there is a map of fibration sequences

$$
\begin{array}{ccccc}
\mathrm{Emb}_{\partial_0}\big(M \setminus \mathrm{int}(P), N \setminus \mathrm{int}(e(P))\big) & \longrightarrow & \mathrm{Emb}_{\partial_0}(M, N) & \longrightarrow & \mathrm{Emb}_{\partial_0}(P, N) \\
\downarrow & & \downarrow & & \downarrow{\scriptstyle\simeq} \\
T_\infty \mathrm{Emb}_{\partial_0}\big(M \setminus \mathrm{int}(P), N \setminus \mathrm{int}(e(P))\big) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(M, N) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(P, N)
\end{array}
$$

whose right square results from (10) and whose left square is an instance of the diagram (12). The homotopy fibres are taken over the embedding e and its image in $T_\infty \mathrm{Emb}_{\partial_0}(P, N)$, and

$$\partial_0(M \setminus \mathrm{int}(P)) := \partial_1 P \cup \partial_0 M \setminus \mathrm{int}(\partial_0 P)$$

with boundary condition induced by e and $e_{\partial_0 M}$. For the upper row, this is a form of the usual parametrised isotopy extension theorem. For the lower row, this is a mild generalisation of a result of Knudsen and Kupers [17, Theorem 6.1 and Remarks 6.4 and 6.5]. Note that every triad embedding $P \hookrightarrow N$ is disjoint from $\partial N \setminus e_{\partial_0}(\partial_0 P)$ up to isotopy of triad embeddings, so if we would like to draw conclusions about all homotopy fibres of the right horizontal maps, it suffices to restrict to embeddings of this form.

We record the following immediate corollary of properties (c) and (f) which will allow us to restrict to triads with $\partial_0 M \neq \varnothing$ when proving convergence results.

Lemma 2.4   *Let M and N be d–dimensional triads, $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ a boundary condition, and $D^d \subset \mathrm{int}(M)$ an embedded disc. The map*

$$\mathrm{Emb}_{\partial_0}(M, N) \to T_\infty \mathrm{Emb}_{\partial_0}(M, N)$$

is a weak equivalence if and only if for all embeddings $e \colon D^d \hookrightarrow \mathrm{int}(N)$, the map

$$\mathrm{Emb}_{\partial_0}\big(M \setminus \mathrm{int}(D^d), N \setminus \mathrm{int}(e(D^d))\big) \to T_\infty \mathrm{Emb}_{\partial_0}\big(M \setminus \mathrm{int}(D^d), N \setminus \mathrm{int}(e(D^d))\big)$$

is a weak equivalence, where $\partial_0(M \setminus \mathrm{int}(D^d)) = \partial_0 M \cup \partial D^d$ and $\partial_0\big(N \setminus \mathrm{int}(e(D^d))\big) = \partial_0 N \cup \partial e(D^d)$.

Proof This is an instance of the fact that for a commutative square

$$
\begin{array}{ccc}
E & \longrightarrow & B \\
\downarrow & & \downarrow{\scriptstyle\simeq} \\
E' & \longrightarrow & B'
\end{array}
$$

whose right arrow is a weak equivalence, the map $E \to E'$ is a weak equivalence if and only if the map $\mathrm{hofib}(E \to B) \to \mathrm{hofib}(E' \to B')$ is a weak equivalence for all choices of basepoints. We apply this to the commutative square induced by restriction

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(M, N) & \longrightarrow & \mathrm{Emb}(D^d, N) \\
\downarrow & & \downarrow \\
T_\infty\mathrm{Emb}_{\partial_0}(M, N) & \longrightarrow & T_\infty\mathrm{Emb}(D^d, N)
\end{array}
$$

whose right-hand map is a weak equivalence by the convergence on discs (property (c)). By isotopy extension (property (f)), the map on homotopy fibres over an embedding $e\colon D^d \hookrightarrow \mathrm{int}(N)$ agrees with the second map in the statement, so the claim follows. $\qquad\square$

We continue with a pair of remarks about these properties:

Remark 2.5 Boavida de Brito and Weiss [1, Section 9] restrict their attention to the case $\partial_0 M = \partial M$, but this turns out to be no less general: given a manifold triad M, the manifold triad $M \setminus \partial_1 M$ with $\partial_0(M \setminus \partial_1 M) = \mathrm{int}(\partial_0 M) = \partial(M \setminus \partial_1 M)$ is isotopy equivalent to M, so there is a weak equivalence $T_\infty\mathrm{Emb}_{\partial_0}(M, N) \simeq T_\infty\mathrm{Emb}_{\partial}(M \setminus \partial_1 M, N \setminus \partial_1 N)$ by item (b) above.

Remark 2.6 As a consequence of property (d) above, to show that the map of Corollary B on path components $\pi_0\mathrm{Diff}_{\partial}(\Sigma) \to \pi_0 T_\infty\mathrm{Emb}_{\partial}(\Sigma, \Sigma)$ is injective, it suffices to prove that

$$
(13) \qquad\qquad\qquad\qquad \pi_0\mathrm{Diff}_{\partial}(\Sigma) \to \pi_0\mathrm{hAut}_{\partial}(\Sigma)
$$

is injective, which is true for all compact surfaces and can be seen as follows.

First, one reduces to the case of connected surfaces. For this, it suffices to show that closed connected surfaces are homotopy equivalent if and only if they are diffeomorphic, which is a consequence of the fact that closed surfaces are classified by orientability and the Euler characteristic, and both of these are preserved by homotopy equivalences relative to the boundary. In the connected case, the claimed injectivity is proved for instance in [3, Theorem 4.6], with the exception of $\Sigma = S^2$ and $\Sigma = \mathbb{R}P^2$. These two cases can settled using the fibre sequence resulting from restricting to an embedded 2–disc and the fact that the mapping class groups of a disc and a Möbius strip are trivial; see [24, Theorem B; 8, Theorem 3.4].

In fact, the forgetful map (13) is often an isomorphism: for closed orientable surfaces of positive genus this is an instance of the Dehn–Nielsen–Baer theorem [9, Theorem 8.1], but there is also an argument for most surfaces with boundary [3, Theorem 1.1(1)].

The proof of Theorem A relies on some additional properties of embedding calculus which we establish in the ensuing subsections. These properties are not very surprising, but seem to have not appeared in the literature before.

2.4 Thickened embeddings

The first property concerns the behaviour of embedding calculus upon replacing the domain M by a thickening, that is, a vector bundle V over M.

Fix manifold triads M and N and a k–dimensional vector bundle $p : V \to M$. We consider V as a triad via $\partial_0 V := p^{-1}(\partial_0 M)$. Fixing a boundary condition $e_{\partial_0} : \partial_0 V \hookrightarrow \partial_0 N$, we obtain a boundary condition $e'_{\partial_0} : \partial_0 M \hookrightarrow \partial_0 N$ by restriction along the zero-section $M \subset V$. From (11), we obtain the solid arrows in the diagram

(14)
$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(V, N) & \longrightarrow & \mathrm{Emb}_{\partial_0}(M, N) \\
\downarrow & & \downarrow \\
T_\infty \mathrm{Emb}_{\partial_0}(V, N) & \dashrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(M, N) \\
\downarrow & & \downarrow \\
\mathrm{Bun}_{\partial_0}(TV, TN) & \longrightarrow & \mathrm{Bun}_{\partial_0}(TM, TN)
\end{array}
$$

Lemma 2.7 *There exists a dashed map in* (14) *such that the diagram commutes up to preferred homotopy and the two subsquares are homotopy cartesian.*

Proof Let $\mathcal{O}$ be the poset of open subsets $U \subset M$ containing a collar on $\partial_0 M$. Taking derivatives as well as restricting embeddings and bundle maps induces a commutative diagram

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(p^{-1}(-), N) & \longrightarrow & \mathrm{Emb}_{\partial_0}(-, N) \\
\downarrow & & \downarrow \\
& & T_\infty \mathrm{Emb}_{\partial_0}(-, N) \\
& & \downarrow \\
\mathrm{Bun}_{\partial_0}(Tp^{-1}(-), TN) & \longrightarrow & \mathrm{Bun}_{\partial_0}(T-, TN) \xrightarrow{\simeq} T_\infty \mathrm{Bun}_{\partial_0}(T-, TN)
\end{array}
$$

of space-valued presheaves on $\mathcal{O}$, where the bottom equivalence results from the discussion in Section 2.3(d). Since homotopy pullbacks of presheaves are computed objectwise, this is a homotopy-cartesian square of presheaves. We define a new presheaf $F(-)$ on $\mathcal{O}$ as the homotopy pullback

(15)
$$
\begin{array}{ccc}
F(-) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(-, N) \\
\downarrow & & \downarrow \\
\mathrm{Bun}_{\partial_0}(Tp^{-1}(-), TN) & \longrightarrow & T_\infty \mathrm{Bun}_{\partial_0}(T-, TN)
\end{array}
$$

The result will follow by evaluation at $M \in \mathcal{O}$ once we provide an identification

$$
F(M) \simeq T_\infty \mathrm{Emb}_{\partial_0}(p^{-1}(M), N) = T_\infty \mathrm{Emb}_{\partial_0}(V, N)
$$

compatible with the maps to $\mathrm{Bun}_{\partial_0}(TV, TN)$ and from $\mathrm{Emb}_{\partial_0}(V, N)$. It follows from Section 2.2.3(b) and (c) that it suffices to verify that

(a) F satisfies descent for the complete J_∞–cover $\mathcal{U} \subset \mathcal{O}$ given by those open subsets $U \subset M$ equal to a collar on $\partial_0 M$ and a finite collection of open discs, and

(b) the map $\mathrm{Emb}_{\partial_0}(p^{-1}(-), N) \to F(-)$ is a weak equivalence when evaluated on $U \in \mathcal{U}$.

For (a), we observe that all entries but $F(-)$ in the homotopy pullback diagram (15) defining $F(-)$ satisfy descent with respect to $\mathcal{J}_\infty$–covers, so $F(-)$ does as well. For (b), we observe that on $U \in \mathcal{U}$, the right vertical map of (15) is a weak equivalence so it suffices to verify that

$$\mathrm{Emb}_{\partial_0}(p^{-1}(U), N) \to \mathrm{Bun}_{\partial_0}(Tp^{-1}(U), TN)$$

is a weak equivalence. This is indeed the case because $p^{-1}(U)$ is a disjoint union of a collar on $\partial_0 V$ and a finite collection of open discs.   □

We derive from Lemma 2.7 two lemmas that will allow us to interpolate between convergence questions for $\mathrm{Emb}_{\partial_0}(M, N)$ and for $\mathrm{Emb}_{\partial_0}(V, N)$.

Lemma 2.8 *Let M and N be manifold triads, $p \colon V \to M$ be a vector bundle considered as a triad by $\partial_0 V = p^{-1}(\partial_0 M)$, and $e_{\partial_0} \colon \partial_0 V \to \partial_0 N$ be a boundary condition. Then the map*

$$\mathrm{Emb}_{\partial_0}(V, N) \to T_\infty \mathrm{Emb}_{\partial_0}(V, N)$$

is a weak equivalence if the map $\mathrm{Emb}_{\partial_0}(M, N) \to T_\infty \mathrm{Emb}_{\partial_0}(M, N)$ is a weak equivalence with boundary condition obtained by restricting e_{∂_0} to $\partial_0 M \subset \partial_0 V$.

Proof This follows from the upper homotopy cartesian square in (14) provided by Lemma 2.7.   □

Lemma 2.9 *Let M be a d–dimensional manifold triad, N be a $(d+k)$–dimensional manifold triad, and $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ be a boundary condition. Then the map*

$$\mathrm{Emb}_{\partial_0}(M, N) \to T_\infty \mathrm{Emb}_{\partial_0}(M, N)$$

is a weak equivalence if the map $\mathrm{Emb}_{\partial_0}(V, N) \to T_\infty \mathrm{Emb}_{\partial_0}(V, N)$ is a weak equivalence for all k–dimensional vector bundles $V \to M$ and boundary conditions $\partial_0 V \hookrightarrow \partial N$ extending e_{∂_0}.

Proof We write $T_\infty \mathrm{Emb}_{\partial_0}(M, N)_\beta$ for the path component of an element $\beta \in T_\infty \mathrm{Emb}_\partial(M, N)$ and $\mathrm{Emb}_{\partial_0}(M, N)_\beta$ for the union of path components mapping to the component of β. It suffices to prove that $\mathrm{Emb}_{\partial_0}(M, N)_\beta \to T_\infty \mathrm{Emb}_{\partial_0}(M, N)_\beta$ is a weak equivalence for all β.

Writing $\beta' \in \mathrm{Bun}_{\partial_0}(TM, TN)$ for the image of β under $T_\infty \mathrm{Emb}_{\partial_0}(M, N) \to \mathrm{Bun}_{\partial_0}(TM, TN)$ from Section 2.3(d), we choose a metric on TN, let V be the vector bundle over M whose fibre over $m \in M$ is the orthogonal complement to $\beta'(T_m M)$ in $T_{\beta'(m)} N$, and extend the boundary condition $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ to $\partial_0 V$ by exponentiation. Writing $\mathrm{Emb}_{\partial_0}(V, N)_\beta$ and $T_\infty \mathrm{Emb}_{\partial_0}(V, N)_\beta$ for the

unions of the path components mapping to β in (14), Lemma 2.7 yields a homotopy pullback

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(V, N)_\beta & \longrightarrow & \mathrm{Emb}_{\partial_0}(M, N)_\beta \\
\downarrow{\simeq} & & \downarrow \\
T_\infty\mathrm{Emb}_{\partial_0}(V, N)_\beta & \longrightarrow & T_\infty\mathrm{Emb}_{\partial_0}(M, N)_\beta
\end{array}
$$

whose left vertical map a weak equivalence by assumption. By construction, β' lifts to a bundle map in $\mathrm{Bun}_{\partial_0}(TV, TN)$ under the bottom horizontal map in (14), so it follows from Lemma 2.7 that $T_\infty\mathrm{Emb}_{\partial_0}(V, N)_\beta$ is nonempty. As $T_\infty\mathrm{Emb}_{\partial_0}(M, N)_\beta$ is path-connected, this implies that the left vertical map in the homotopy pullback is a weak equivalence. $\square$

2.5 Lifting along covering maps

The second property is concerned with the problem of lifting embeddings of triads $M \hookrightarrow N$ along covering maps $\pi \colon \widetilde{N} \to N$. To state the result, we consider the cover $\widetilde{N}$ as a triad by setting $\partial_0\widetilde{N} := \pi^{-1}(\partial_0 N)$ and $\partial_1\widetilde{N} := \pi^{-1}(\partial_1 N)$, and fix a boundary condition $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ as well as a lift $\tilde{e}_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0\widetilde{N}$. We pick a homotopy class $[\alpha] \in \pi_0\mathrm{Map}_{\partial_0}(M, N)$ such that there exists a lift $[\tilde{\alpha}] \in \pi_0\mathrm{Map}_{\partial_0}(M, \widetilde{N})$. We shall assume that $\partial_0 M \to M$ is 0–connected, so that this lift is unique. We write

$$
\mathrm{Emb}_{\partial_0}(M, N)_\alpha \subset \mathrm{Emb}_{\partial_0}(M, N) \quad \text{and} \quad T_\infty\mathrm{Emb}_{\partial_0}(M, N)_\alpha \subset T_\infty\mathrm{Emb}_{\partial_0}(M, N)
$$

for the unions of the path components that map to $[\alpha] \in \pi_0\mathrm{Map}_{\partial_0}(M, N)$ via the maps in (11). We similarly define subspaces $\mathrm{Emb}_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}} \subset \mathrm{Emb}_{\partial_0}(M, \widetilde{N})$ and $T_\infty\mathrm{Emb}_\partial(M, \widetilde{N})_{\tilde{\alpha}} \subset T_\infty\mathrm{Emb}_\partial(M, \widetilde{N})$.

Lemma 2.10 *In this situation, there exists a dashed map making the diagram*

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(M, N)_\alpha & \longrightarrow & \mathrm{Emb}_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}} \\
\downarrow & & \downarrow \\
T_\infty\mathrm{Emb}_{\partial_0}(M, N)_\alpha & \dashrightarrow & T_\infty\mathrm{Emb}_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}}
\end{array}
$$

commute up to homotopy. Here the top map is given by sending an embedding $\beta \in \mathrm{Emb}_\partial(M, N)_\alpha$ to its unique lift $\tilde{\beta} \in \mathrm{Emb}_\partial(M, \widetilde{N})_{\tilde{\alpha}}$ extending $\tilde{e}_\partial$.

Proof Let $\mathrm{Emb}_{\partial_0}^\pi(-, \widetilde{N}) \subset \mathrm{Emb}_{\partial_0}(-, \widetilde{N})$ be the presheaf on $\mathrm{Disc}_{\partial_0}$ of those embeddings that remain an embedding after composition with π. This fits in a pullback diagram

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}^\pi(-, \widetilde{N}) & \xrightarrow{\ \pi\circ-\ } & \mathrm{Emb}_{\partial_0}(-, N) \\
\downarrow & & \downarrow \\
\mathrm{Map}_{\partial_0}(-, \widetilde{N}) & \xrightarrow{\ \pi\circ-\ } & \mathrm{Map}_{\partial_0}(-, N)
\end{array}
$$

of presheaves on $\mathrm{Disc}_{\partial_0 M}$ whose vertical maps are given by inclusion. This is homotopy cartesian in the projective model structure on $\mathrm{PSh}(\mathrm{Disc}_{\partial_0})$, since $(\pi \circ -) \colon \mathrm{Map}_{\partial_0}(-, \widetilde{N}) \to \mathrm{Map}_{\partial_0}(-, N)$ is a objectwise

fibration by the lifting property of covering maps. Evaluating at M and using that $T_\infty(-)$ preserves homotopy limits by Section 2.2.3(a), we arrive at a commutative cube

$$
\begin{array}{ccc}
\mathrm{Emb}^\pi_{\partial_0}(M, \widetilde{N}) & \longrightarrow & \mathrm{Emb}_{\partial_0}(M, N) \\
& T_\infty\mathrm{Emb}^\pi_{\partial_0}(M, \widetilde{N}) \longrightarrow T_\infty\mathrm{Emb}_{\partial_0}(M, N) & \\
\mathrm{Map}_{\partial_0}(M, \widetilde{N}) & \longrightarrow\!\!\!\twoheadrightarrow & \mathrm{Map}_{\partial_0}(M, N) \\
& T_\infty\mathrm{Map}_{\partial_0}(M, \widetilde{N}) \longrightarrow\!\!\!\twoheadrightarrow T_\infty\mathrm{Map}_{\partial_0}(M, N) &
\end{array}
$$

with front and back faces homotopy cartesian, and bottom diagonal maps weak equivalences since $\mathrm{Map}_{\partial_0}(-, \widetilde{N})$ and $\mathrm{Map}_{\partial_0}(-, N)$ are homotopy $\mathcal{J}_1$–sheaves (see Section 2.2.3(b)). By the uniqueness of lifts (this uses that $\partial_0 M \to M$ is 0–connected), the bottom horizontal maps become weak equivalences when we restrict domain and target to the path components of $[\tilde{\alpha}]$ and $[\alpha]$ respectively. Doing so and using the homotopy pullback property, the top of the cube provides a commutative square

$$
\begin{CD}
\mathrm{Emb}^\pi_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}} @>{\simeq}>> \mathrm{Emb}_{\partial_0}(M, N)_\alpha \\
@VVV @VVV \\
T_\infty\mathrm{Emb}^\pi_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}} @>{\simeq}>> T_\infty\mathrm{Emb}_{\partial_0}(M, N)_\alpha
\end{CD}
$$

with horizontal weak equivalences. The top map is even a homeomorphism, by the uniqueness of lifts. Using the inclusion of presheaves $\mathrm{Emb}^\pi_{\partial_0}(-, \widetilde{N}) \subset \mathrm{Emb}_{\partial_0}(-, \widetilde{N})$, we obtain a commutative diagram

$$
\begin{CD}
\mathrm{Emb}_{\partial_0}(M, N)_\alpha @<{\cong}<< \mathrm{Emb}^\pi_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}} @>>> \mathrm{Emb}_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}} \\
@VVV @VVV @VVV \\
T_\infty\mathrm{Emb}_{\partial_0}(M, N)_\alpha @<{\simeq}<< T_\infty\mathrm{Emb}^\pi_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}} @>>> T_\infty\mathrm{Emb}_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}}
\end{CD}
$$

whose top composition is given by sending an embedding to its unique lift extending $\tilde{e}_\partial$, so we obtain a map $T_\infty\mathrm{Emb}_{\partial_0}(M, N)_\alpha \to T_\infty\mathrm{Emb}_{\partial_0}(M, \widetilde{N})_{\tilde{\alpha}}$, as desired.   □

Remark 2.11 If α has no lift, then there is no component of $\mathrm{Emb}_{\partial_0}(M, \widetilde{N})$ mapping to $[\alpha]$ under composition with π. In this case, the above argument shows that there is also no component of $T_\infty\mathrm{Emb}_{\partial_0}(M, \widetilde{N})$ mapping to $[\alpha]$ under the map of Section 2.3(d) and composition with π.

2.6 Adding a collar to the source

The third property concerns the behaviour of embedding calculus when adding a disjoint collar to the domain.

We fix triads M and N and a boundary condition $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$. Given a compact $(\dim(M)-1)$–manifold K, we replace M by the triad $M \sqcup K \times [0, 1)$ with $\partial_0(M \sqcup K \times [0, 1)) = \partial_0 M \sqcup K \times \{0\}$ and fix an extension $e'_{\partial_0} \colon \partial_0(M \sqcup K \times [0, 1)) \hookrightarrow \partial N$ of e_{∂_0} as boundary condition. By contractibility of the space of collars, the restriction map

$$\mathrm{Emb}_{\partial_0 M \sqcup K \times \{0\}}(M \sqcup K \times [0, 1), N) \to \mathrm{Emb}_{\partial_0}(M, N)$$

is a weak equivalence. Embedding calculus has this property as well:

Lemma 2.12 *In this situation, both horizontal maps in the diagram induced by restriction*

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0 M \sqcup K \times \{0\}}(M \sqcup K \times [0, 1), N) & \xrightarrow{\;\simeq\;} & \mathrm{Emb}_{\partial_0 M}(M, N) \\
\downarrow & & \downarrow \\
T_\infty \mathrm{Emb}_{\partial_0 M \sqcup K \times \{0\}}(M \sqcup K \times [0, 1), N) & \xrightarrow{\;\simeq\;} & T_\infty \mathrm{Emb}_{\partial_0 M}(M, N)
\end{array}
$$

are weak equivalences.

Proof Let $\mathcal{U}$ be the open cover of $M \sqcup K \times [0, 1)$ given by subsets of the form $U = V \sqcup K \times [0, 1)$ where $V \subset M$ is the union of a open subset diffeomorphic to a collar on $\partial_0 M$ and a finite disjoint union of open discs. This is a complete Weiss ∞–cover of $M \sqcup K \times [0, 1)$, and $\mathcal{U}' = \{U \cap M \mid U \in \mathcal{U}\}$ is a complete Weiss ∞–cover of M. Restriction thus induces a commutative diagram

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0 M \sqcup K \times \{0\}}(M \sqcup K \times [0, 1), N) & \longrightarrow & \mathrm{holim}_{U \in \mathcal{U}} \mathrm{Emb}_{\partial_0 M \sqcup K \times \{0\}}(U, N) \\
\downarrow & & \downarrow{\scriptstyle \simeq} \\
T_\infty \mathrm{Emb}_{\partial_0 M \sqcup K \times \{0\}}(M \sqcup K \times [0, 1), N) & \xrightarrow{\;\simeq\;} & \mathrm{holim}_{U \in \mathcal{U}} T_\infty \mathrm{Emb}_{\partial_0 M \sqcup K \times \{0\}}(U, N)
\end{array}
$$

whose bottom horizontal map is a weak equivalences by Section 2.2.3(b) and whose right vertical map is a weak equivalence by Section 2.3(c). Similarly, we have a square

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0 M}(M, N) & \longrightarrow & \mathrm{holim}_{U \in \mathcal{U}} \mathrm{Emb}_{\partial_0 M}(U \cap M, N) \\
\downarrow & & \downarrow{\scriptstyle \simeq} \\
T_\infty \mathrm{Emb}_{\partial_0 M}(M, N) & \xrightarrow{\;\simeq\;} & \mathrm{holim}_{U \in \mathcal{U}} T_\infty \mathrm{Emb}_{\partial_0 M}(U \cap M, N)
\end{array}
$$

which receives a map from the former square by restriction, so it suffices to show that the maps

$$\mathrm{Emb}_{\partial_0 M \sqcup K \times \{0\}}(U, N) \to \mathrm{Emb}_{\partial_0 M}(U \cap M, N)$$

are weak equivalence. This follows from the contractibility of spaces of collars. $\square$

Combined with Lemma 2.4 this yields the following lemma, which is often useful to justify the hypothesis needed to apply isotopy extension for embedding calculus (see Section 2.3(f)).

Lemma 2.13 *Let M and N be d–dimensional triads, and $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ a boundary condition. Then the map*

$$\mathrm{Emb}_{\partial_0}(M \sqcup (T \times \mathbb{R}^d), N) \to T_\infty \mathrm{Emb}_{\partial_0}(M \sqcup (T \times \mathbb{R}^d), N)$$

is a weak equivalence for any finite set T, if the maps $\mathrm{Emb}_{\partial_0}(M, N') \to T_\infty \mathrm{Emb}_{\partial_0}(M, N')$ are weak equivalences for all d–dimensional triads N' and all boundary conditions $e'_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N'$.

Proof By induction over $|T|$ it suffices to prove the case $|T| = 1$. In that case, it suffices by Lemma 2.4 to prove that for all embeddings $e \colon D^d \hookrightarrow \mathrm{int}(N)$ the map

$$\mathrm{Emb}_\partial\big(M \sqcup (\mathbb{R}^d \setminus \mathrm{int}(D^d)), N \setminus \mathrm{int}(e(D^d))\big) \to T_\infty \mathrm{Emb}_\partial\big(M \sqcup (\mathbb{R}^d \setminus \mathrm{int}(D^d)), N \setminus \mathrm{int}(e(D^d))\big)$$

is a weak equivalences. By Lemma 2.12 we may then forget the collars $(\mathbb{R}^d \setminus \mathrm{int}(D^d))$ on ∂D^d from the source, so the result follows. $\qquad\square$

2.7 Taking disjoint unions

The fourth and final general property of embedding calculus we shall discuss concerns taking disjoint unions in source and target. Its full strength is not needed to prove the main results of this paper — only Corollary 2.15 is — but we believe it to be of independent interest.

Let M, M', N, and N' be triads with $\dim(M) = \dim(M')$ and $\dim(N) = \dim(N')$. Given boundary conditions $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ and $e'_{\partial_0} \colon \partial_0 M' \hookrightarrow \partial_0 N'$, we consider the boundary condition

$$e_{\partial_0} \sqcup e'_{\partial_0} \colon \partial_0(M \sqcup M') \hookrightarrow \partial_0(N \sqcup N').$$

Disjoint union of embeddings induces

$$\mathrm{Emb}_{\partial_0}(M, N) \times \mathrm{Emb}_{\partial_0}(M', N') \to \mathrm{Emb}_{\partial_0}(M \sqcup M', N \sqcup N')$$

which is a weak equivalence (in fact, a homeomorphism) if both inclusions $\partial_0 M \hookrightarrow M$ and $\partial_0 M' \hookrightarrow M'$ are 0–connected. Embedding calculus has this property as well:

Lemma 2.14 *In this situation, there is a dashed weak equivalence that makes*

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(M, N) \times \mathrm{Emb}_{\partial_0}(M', N') & \xrightarrow{\ \simeq\ } & \mathrm{Emb}_{\partial_0}(M \sqcup M', N \sqcup N') \\
\downarrow & & \downarrow \\
T_\infty \mathrm{Emb}_{\partial_0}(M, N) \times T_\infty \mathrm{Emb}_{\partial_0}(M', N') & \dashrightarrow^{\simeq} & T_\infty \mathrm{Emb}_{\partial_0}(M \sqcup M', N \sqcup N')
\end{array}
$$

commute up to preferred homotopy.

Proof As in the proof of Lemma 2.12, the property of embedding calculus we shall use is descent for complete Weiss ∞–covers (see Section 2.2.3(b)).

We take $\mathcal{U}_M$ to be the open cover of M given by open subsets $U \subset M$ that are diffeomorphic to a collar on $\partial_0 M$ and a finite disjoint union of open discs, and similarly for $\mathcal{U}_{M'}$. We take $\mathcal{U}_{M \sqcup M'}$ to be the open cover of $M \sqcup M'$ given by unions of an element of $\mathcal{U}_M$ and an element of $\mathcal{U}_{M'}$. The covers $\mathcal{U}_M$, $\mathcal{U}_{M'}$, and $\mathcal{U}_{M \sqcup M'}$ are all complete Weiss ∞–covers.

We consider $\mathcal{U}_{M \sqcup M'}$ as a poset ordered by inclusion and let $\mathrm{Emb}_{\partial_0}^{\sqcup}(-, N \sqcup N')$ be the presheaf on $\mathcal{U}_{M \sqcup M'}$ that sends $U \sqcup U'$ with $U \in \mathcal{U}_M$ and $U' \in \mathcal{U}_{M'}$ to the subspace $\mathrm{Emb}_{\partial_0}^{\sqcup}(U \sqcup U', N \sqcup N') \subset \mathrm{Emb}_{\partial_0}(U \sqcup U', N \sqcup N')$ which maps U into N and U' into N'. Defining $\mathrm{Map}_{\partial_0}^{\sqcup}(-, N \sqcup N')$ similarly, we have a homotopy pullback diagram of presheaves on $\mathcal{U}_{M \sqcup M'}$,

(16)
$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}^{\sqcup}(-, N \sqcup N') & \longrightarrow & \mathrm{Emb}_{\partial_0}(-, N \sqcup N') \\
\downarrow & & \downarrow \\
\mathrm{Map}_{\partial_0}^{\sqcup}(-, N \sqcup N') & \longrightarrow & \mathrm{Map}_{\partial_0}(-, N \sqcup N')
\end{array}
$$

and this remains a homotopy pullback when taking homotopy limits over $\mathcal{U}_{M \sqcup M'}$.

To identify the term

$$
\underset{U \sqcup U' \in \mathcal{U}_{M \sqcup M'}}{\mathrm{holim}} \mathrm{Emb}_{\partial_0}^{\sqcup}(U \sqcup U', N \sqcup N')
$$

we note that there are isomorphisms $\mathcal{U}_{M \sqcup M'} \cong \mathcal{U}_M \times \mathcal{U}_{M'}$ of categories, and

$$
\mathrm{Emb}_{\partial_0}^{\sqcup}(-, N \sqcup N') \cong \mathrm{Emb}_{\partial_0}(-, N) \times \mathrm{Emb}_{\partial_0}(-, N')
$$

of presheaves, so the Fubini theorem for homotopy limits implies that this homotopy limit is given by

$$
\underset{U \in \mathcal{U}_M}{\mathrm{holim}}\, \mathrm{Emb}_{\partial_0}(U, N) \times \underset{U' \in \mathcal{U}_M}{\mathrm{holim}}\, \mathrm{Emb}_{\partial_0}(U', N').
$$

Combining descent with the fact that embedding calculus converges on $U \in \mathcal{U}_M$ and $U' \in \mathcal{U}_{M'}$ by Section 2.3(c), we conclude that

$$
\underset{U \sqcup U' \in \mathcal{U}_{M \sqcup M'}}{\mathrm{holim}} \mathrm{Emb}_{\partial_0}^{\sqcup}(U \sqcup U', N \sqcup N') \simeq T_\infty \mathrm{Emb}_{\partial_0}(M, N) \times T_\infty \mathrm{Emb}_{\partial_0}(M', N').
$$

The same analysis holds for $\mathrm{Map}_{\partial_0}^{\sqcup}(-, M \sqcup M')$ and since this is a homotopy $\mathcal{J}_1$–sheaf (see Section 2.2.3(b)), we conclude that

$$
\underset{U \sqcup U' \in \mathcal{U}_{M \sqcup M'}}{\mathrm{holim}} \mathrm{Map}_{\partial_0}^{\sqcup}(U \sqcup U', N \sqcup N') \simeq \mathrm{Map}_{\partial_0}(M, N) \times \mathrm{Map}_{\partial_0}(M', N').
$$

By the same argument (using descent, convergence on $U \sqcup U' \in \mathcal{U}_{M \sqcup M'}$, and that $\mathrm{Map}_{\partial_0}(-, N \sqcup N')$ is a homotopy $\mathcal{J}_1$–sheaf), we have weak equivalences

$$
\underset{U \sqcup U' \in \mathcal{U}_{M \sqcup M'}}{\mathrm{holim}} \mathrm{Emb}_{\partial_0}(U \sqcup U', N \sqcup N') \simeq T_\infty \mathrm{Emb}_{\partial_0}(M \sqcup M', N \sqcup N'),
$$

$$
\underset{U \sqcup U' \in \mathcal{U}_{M \sqcup M'}}{\mathrm{holim}} \mathrm{Map}_{\partial_0}(U \sqcup U', N \sqcup N') \simeq \mathrm{Map}_{\partial_0}(M \sqcup M', N \sqcup N'),
$$

so altogether we obtain a homotopy pullback diagram of the form

$$
\begin{array}{ccc}
T_\infty \mathrm{Emb}_{\partial_0}(M,N) \times T_\infty \mathrm{Emb}_{\partial_0}(M',N') & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(M \sqcup M', N \sqcup N') \\
\downarrow & & \downarrow \\
\mathrm{Map}_{\partial_0}(M,N) \times \mathrm{Map}_{\partial_0}(M',N') & \longrightarrow & \mathrm{Map}_{\partial_0}(M \sqcup M', N \sqcup N')
\end{array}
$$

The condition that $\partial_0 M \hookrightarrow M$ and $\partial_0 M' \hookrightarrow M'$ are 0–connected implies that the bottom map is a weak equivalence, so the top map is a weak equivalence as well. The proof is finished by tracing through the weak equivalences to see that this makes the square in the statement homotopy commute.   □

Taking $M' = \varnothing$, which is the only case used in this paper, Lemma 2.14 says:

Corollary 2.15   *In this situation, in the diagram induced by the inclusion $N \hookrightarrow N \sqcup N'$,*

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(M,N) & \xrightarrow{\ \simeq\ } & \mathrm{Emb}_{\partial_0}(M, N \sqcup N') \\
\downarrow & & \downarrow \\
T_\infty \mathrm{Emb}_{\partial_0}(M,N) & \xrightarrow{\ \simeq\ } & T_\infty \mathrm{Emb}_{\partial_0}(M, N \sqcup N')
\end{array}
$$

both horizontal maps are weak equivalences.

Remark 2.16   Corollary 2.15 admits an alternative proof along the lines of Lemma 2.10: one observes there is a homotopy pullback diagram of presheaves on $\mathrm{Disc}_{\partial_0 M}$ given by

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}(-,N) & \longrightarrow & \mathrm{Emb}_{\partial_0}(-, N \sqcup N') \\
\downarrow & & \downarrow \\
\mathrm{Map}_{\partial_0}(-,N) & \longrightarrow & \mathrm{Map}_{\partial_0}(-, N \sqcup N')
\end{array}
$$

Taking T_∞ and evaluating at M yields a homotopy pullback diagram of spaces and if $\partial_0 M \to M$ is 0–connected, the map $\mathrm{Map}_{\partial_0}(M,N) \to \mathrm{Map}_{\partial_0}(M, N \sqcup N')$ is a weak equivalence and hence so is the map $T_\infty \mathrm{Emb}_{\partial_0}(M,N) \to T_\infty \mathrm{Emb}_{\partial_0}(M, N \sqcup N')$.

3   Convergence in low dimensions

In this section we make use of the properties of embedding calculus discussed in the previous section to prove the following convergence result. Theorem A is included as the special case $\partial_0 M = \partial M$.

Theorem 3.1   *For compact manifolds triads M and N with $\dim(N) \leq 2$, the map*

$$
\mathrm{Emb}_{\partial_0}(M,N) \to T_\infty \mathrm{Emb}_{\partial_0}(M,N)
$$

is a weak equivalence for any boundary condition $e_{\partial_0} : \partial_0 M \hookrightarrow \partial_0 N$.

Convention 3.2 Throughout this section, we adopt following conventions on triads:

(i) We write $I = [0, 1]$ and call the manifold triads I and $I \times [0, 1]$ with $\partial_0 I = \{0, 1\}$ and $\partial_0(I \times [0, 1]) = \{0, 1\} \times [0, 1]$ the *arc* and the *strip*. We will use the convention and notation from Section 2.1, so embeddings $I \times [0, 1]$ into a triad N will always be assumed to extend a boundary condition $e_{\partial_0} \colon \{0, 1\} \times [0, 1] \hookrightarrow \partial_0 N$ which will either be specified or is clear from the context. We consider I as a submanifold of $I \times [0, 1]$ via the inclusion $\{\frac{1}{2}\} \times [0, 1] \subset I \times [0, 1]$, so a boundary condition e_{∂_0} as above in particular induces a boundary condition $e_{\partial} \colon \{0, 1\} \hookrightarrow N$ for embedding of the form $I \hookrightarrow N$ by restriction.

(ii) We consider the *cylinder* $S^1 \times [0, 1]$ as a manifold triad with $\partial_0(S^1 \times [0, 1]) = \varnothing$. We consider the *circle* S^1 as the submanifold of $S^1 \times [0, 1]$ via the inclusion $S^1 \times \{\frac{1}{2}\} \hookrightarrow S^1 \times [0, 1]$.

(iii) We consider the *Möbius strip* $\mathrm{Mo} = ([0, 1] \times [0, 1])/\sim$, with $\sim$ the equivalence relation generated by $(0, y) \sim (1, 1 - y)$, as a manifold triad with $\partial_0(\mathrm{Mo}) = \varnothing$. We consider S^1 as the submanifold of Mo via the inclusion $S^1 \times \{\frac{1}{2}\} \hookrightarrow \mathrm{Mo}$.

(iv) We write $\Sigma_{g,n}$ for an orientable compact surface of genus g with n boundary components, considered as a manifold triad with $\partial_0 \Sigma_{g,n} = \partial \Sigma_{g,n}$.

The steps

The proof of Theorem 3.1 is divided into the following steps:

(1) $\dim(M) > \dim(N)$ or $\dim(M) = 0$;

(2) $\dim(M) \leq \dim(N) = 2$, with substeps

    (2.1) M an arc or a strip,

    (2.2) M a circle, a cylinder, or a Möbius band,

    (2.3) M a line bundle over a 1–dimensional triad M' with $\partial_0 M' = \partial M$,

    (2.4) M a general 1–dimensional triad,

    (2.5) $M = D^2$ with $\partial_0 M = \partial M$,

    (2.6) M an orientable genus 0 surface with $\partial_0 M = \partial M$,

    (2.7) M a connected 2–dimensional triad with $\partial_0 M = \partial M$,

    (2.8) M a connected 2–dimensional triad with $\partial_0 M \neq \partial M$,

    (2.9) M a general 2–dimensional triad;

(3) $\dim(M) = \dim(N) = 1$.

To avoid being repetitive, we say that *convergence holds for a pair of triads* (M, N) if the map

$$\mathrm{Emb}_{\partial_0}(M, N) \to T_\infty \mathrm{Emb}_{\partial_0}(M, N)$$

is a weak equivalence for all boundary conditions $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$.

Step (1): Convergence holds for (M, N) if $\dim(M) > \dim(N)$ or $\dim(M) = 0$ Convergence for $\dim(M) = 0$ holds as a result of Section 2.3(c). For $M \neq \varnothing$ and $\dim(M) > \dim(N)$, we consider the

composition $\mathrm{Emb}_{\partial_0}(M, N) \to T_\infty \mathrm{Emb}_{\partial_0}(M, N) \to \mathrm{Bun}_{\partial_0}(TM, TN)$ from Section 2.3(d). If $\dim(M) > \dim(N)$ then the final space in this composition is empty, so the same holds for the first and the second space. This implies convergence.

Step (2.1): Convergence holds for (M, N) if M is an arc or a strip, and $\dim(N) = 2$ We divide this step into two substeps: the case where the boundary condition $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ hits two distinct boundary components of N, and the case where the boundary condition hits a single boundary component. The arguments are inspired by Gramain's work [13] and Hatcher's exposition thereof in [14].

Substep: the boundary condition hits two distinct boundary components of N By Lemma 2.9 and isotopy invariance (see Section 2.3(b)), it suffices to consider the case $M = I \times [0, 1]$ of a strip. To do so, we glue a disc D to the boundary component of N hit by $\{1\}$, and consider $L = (I \times [0, 1] \cup D)$. Smoothing corners and an application of isotopy extension justified by the convergence on discs (see Section 2.3(c) and (f)) yields a map of fibre sequence

$$
\begin{array}{ccccc}
\mathrm{Emb}_{\partial_0}(I \times [0, 1], N) & \longrightarrow & \mathrm{Emb}_{I \times \{0\}}(L, N \cup D) & \longrightarrow & \mathrm{Emb}(D, N \cup D) \\
\downarrow & & \downarrow & & \downarrow {\scriptstyle \simeq} \\
T_\infty \mathrm{Emb}_{\partial_0}(I \times [0, 1], N) & \longrightarrow & T_\infty \mathrm{Emb}_{I \times \{0\}}(L, N \cup D) & \longrightarrow & T_\infty \mathrm{Emb}(D, N \cup D)
\end{array}
$$

with fibres taken over the standard inclusion. Since L is isotopy equivalent to $I \times [0, 1)$ relative to $I \times \{0\}$, the middle vertical map is a weak equivalence by isotopy invariance and the convergence on collars (see Section 2.3(b) and (c)), so the left vertical map is a weak equivalence as well.

Substep: the boundary condition hits a single boundary components of N The case of arcs and strips connecting the same boundary component is harder and its proof is the heart of the overall argument. It relies on Lemma 2.10 on lifting embeddings, which we spell out again in the special case we shall use.

This lemma involves a covering map $\widetilde{N} \to N$, a boundary condition $e_\partial \colon \{0, 1\} \hookrightarrow \partial N$, a path α of $\mathrm{Map}_\partial(I, N)$, and a lift $\tilde{\alpha} \colon I \to \widetilde{N}$ of α whose endpoints induce a boundary condition $e_\partial \colon \{0, 1\} \hookrightarrow \partial \widetilde{N}$. Recall that $\mathrm{Emb}_\partial(I, N)_\alpha \subset \mathrm{Emb}_\partial(I, N)$ and $T_\infty \mathrm{Emb}_\partial(I, N)_\alpha \subset T_\infty \mathrm{Emb}_\partial(I, N)$ denote the collections of path components that map to $[\alpha] \in \pi_0 \mathrm{Map}_\partial(I, N)$ via the maps in (11). Lemma 2.10 for the triad $M = I$ with $\partial_0 I = \{0, 1\}$ then gives:

Lemma 3.3 *In this situation, there exists a dashed map making the diagram*

$$
\begin{array}{ccc}
\mathrm{Emb}_\partial(I, N)_\alpha & \longrightarrow & \mathrm{Emb}_\partial(I, \widetilde{N})_{\tilde{\alpha}} \\
\downarrow & & \downarrow \\
T_\infty \mathrm{Emb}_\partial(I, N)_\alpha & \dashrightarrow & T_\infty \mathrm{Emb}_\partial(I, \widetilde{N})_{\tilde{\alpha}}
\end{array}
$$

commute up to homotopy. Here the top map is given by sending an arc $\gamma \in \mathrm{Emb}_\partial(I, N)_\alpha$ to the unique lift $\tilde{\gamma} \in \mathrm{Emb}_\partial(I, \widetilde{N})_{\tilde{\alpha}}$ starting at $\tilde{\alpha}(0) \in \widetilde{N}$.
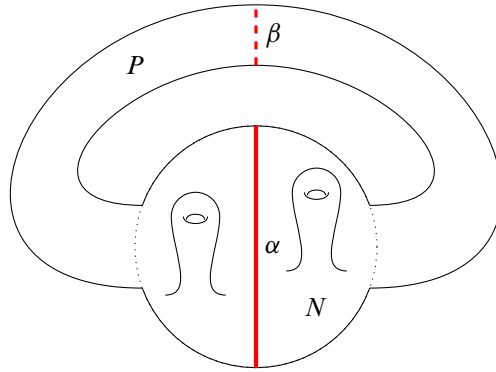
Figure 1: The surface P. The original surface N is the region within the dotted circle.

Using this lemma, we now prove convergence for (M, N) if M is an arc or a strip, $\dim(N) = 2$, and the boundary condition $e_{\partial_0} \colon \partial_0 M \hookrightarrow \partial_0 N$ hits a single boundary components of N.

By Lemma 2.8 and isotopy invariance (see Section 2.3(b)) it suffices to prove the claim for the arc, and by Corollary 2.15, we may assume that the target N is connected. We attach a 1–handle $I \times [0, 1]$ to N to the boundary component hit by $\{0, 1\}$, such that $I \times \{0\}$ and $I \times \{1\}$ are separated on that boundary component by $\{0, 1\}$ and are embedded with opposite orientation, resulting in a new surface P with an additional boundary component; see Figure 1. The composition $\{0, 1\} \hookrightarrow N \subset P$ now hits two distinct boundary components, so the right vertical map in the homotopy-commutative diagram induced by the inclusion $N \subset P$ (see Section 2.3(a))

(17)
$$\begin{array}{ccc}
\operatorname{Emb}_{\partial}(I, N) & \longrightarrow & \operatorname{Emb}_{\partial}(I, P) \\
\downarrow & & \downarrow{\scriptstyle\simeq} \\
T_{\infty}\operatorname{Emb}_{\partial}(I, N) & \longrightarrow & T_{\infty}\operatorname{Emb}_{\partial}(I, P)
\end{array}$$

is a weak equivalence by the previous substep.

We next investigate the set of path components. To do so, we will use that the dashed map in

$$\begin{array}{ccc}
\pi_0\operatorname{Emb}_{\partial}(I, N) & & \\
 & (\pi_0\operatorname{Map}_{\partial}(I, N)) \times_{\pi_0\operatorname{Map}_{\partial}(I, P)} (\pi_0\operatorname{Emb}_{\partial}(I, P)) \longrightarrow \pi_0\operatorname{Emb}_{\partial}(I, P) \\
 & \downarrow & \downarrow \\
 & \pi_0\operatorname{Map}_{\partial}(I, N) \longrightarrow \pi_0\operatorname{Map}_{\partial}(I, P)
\end{array}$$

is surjective: if an embedding $I \hookrightarrow P$ is homotopic to a map $I \to N$, then it is isotopic to an embedding $I \hookrightarrow N$ within the homotopy class of $I \to N$. To see this, use the bigon criterion [9, Sections 1.2.4 and 1.2.7] to isotope $I \hookrightarrow P$ so that its geometric intersection number with the cocore β of the 1–handle is equal to the algebraic intersection number, which is 0 since it is homotopic to a map $I \to N$. With this

in mind, a diagram chase in the factorisation

$$
\begin{array}{ccc}
\pi_0\mathrm{Emb}_\partial(I, N) & \longrightarrow & \pi_0\mathrm{Emb}_\partial(I, P) \\
\downarrow & & \downarrow{\scriptstyle\cong} \\
\pi_0 T_\infty\mathrm{Emb}_\partial(I, N) & \longrightarrow & \pi_0 T_\infty\mathrm{Emb}_\partial(I, P) \\
{\scriptstyle②}\downarrow & & \downarrow \\
\pi_0\mathrm{Map}_\partial(I, N) & \longrightarrow & \pi_0\mathrm{Map}_\partial(I, P)
\end{array}
$$

①

shows that the maps ① and ② have the same image.

Let us now fix a class $[\alpha] \in \pi_0\mathrm{Map}_\partial(I, N)$ in this image. As the map ① is injective because two embedded arcs are isotopic relative to the endpoints if and only if they are homotopic relative to the endpoints (see [10]), there is a unique path component $\mathrm{Emb}_\partial(I, N)_\alpha$ of $\mathrm{Emb}_\partial(I, N)$ mapping to $[\alpha]$. Denoting by $T_\infty\mathrm{Emb}_\partial(I, N)_\alpha \subset T_\infty\mathrm{Emb}_\partial(I, N)$ the union of all path components that map to $[\alpha]$, it suffices to show that the map $\mathrm{Emb}_\partial(I, N)_\alpha \to T_\infty\mathrm{Emb}_\partial(I, N)_\alpha$ is a weak equivalence for all choices of $[\alpha]$. Since $\mathrm{Emb}_\partial(I, N)_\alpha$ is contractible by [13, Théorème 5], the task is to prove that $T_\infty\mathrm{Emb}_\partial(I, N)_\alpha$ is (weakly) contractible as well.

To do so, we will construct a homotopy-commutative diagram

(18)
$$
\begin{array}{ccccc}
\mathrm{Emb}_\partial(I, N)_\alpha & \longrightarrow & \mathrm{Emb}_\partial(I, P)_\alpha & \xrightarrow{(e\circ-)\circ\mathrm{lift}} & \mathrm{Emb}_\partial(I, N)_\alpha \\
\downarrow & & \downarrow{\scriptstyle\simeq} & & \downarrow \\
T_\infty\mathrm{Emb}_\partial(I, N)_\alpha & \longrightarrow & T_\infty\mathrm{Emb}_\partial(I, P)_\alpha & \xrightarrow{(e\circ-)\circ\mathrm{lift}} & T_\infty\mathrm{Emb}_\partial(I, N)_\alpha
\end{array}
$$

whose horizontal compositions are homotopic to the identity. This will finish the proof, since it exhibits $T_\infty\mathrm{Emb}_\partial(I, N)_\alpha$ as a retract of the contractible space $T_\infty\mathrm{Emb}_\partial(I, P)_\alpha \simeq \mathrm{Emb}_\partial(I, P)_\alpha$.

The left square in (18) is obtained by restricting the path components of the homotopy commutative square (17). The right square arises as the composition of two squares

$$
\begin{array}{ccccc}
\mathrm{Emb}_\partial(I, P)_\alpha & \xrightarrow{\mathrm{lift}} & \mathrm{Emb}_\partial(I, \widetilde{P})_{\widetilde{\alpha}} & \xrightarrow{e\circ-} & \mathrm{Emb}_\partial(I, N)_\alpha \\
\downarrow{\scriptstyle\simeq} & & \downarrow & & \downarrow \\
T_\infty\mathrm{Emb}_\partial(I, P)_\alpha & \xrightarrow{\mathrm{lift}} & T_\infty\mathrm{Emb}_\partial(I, \widetilde{P})_{\widetilde{\alpha}} & \xrightarrow{e\circ-} & T_\infty\mathrm{Emb}_\partial(I, N)_\alpha
\end{array}
$$

which we explain now. The surface $\widetilde{P}$ is an appropriate covering space of P: the construction of P gives a decomposition $\pi_1(P) \cong \pi_1(N) * \mathbb{Z}$ and $\widetilde{P}$ is the cover corresponding to the subgroup $\pi_1(N)$. Explicitly, the cover $\widetilde{P}$ can be constructed by cutting P along β to obtain a surface R (see Figure 2) and gluing two copies of the universal cover $\widetilde{R}$ of this surface to the two dashed intervals in the boundary resulting from β. Note that R contains a preferred lift $\widetilde{\alpha}$ of α and hence so does $\widetilde{P}$. We denote the endpoints of α and $\widetilde{\alpha}$ in the various surfaces generically by $\{0, 1\}$. The cover $\widetilde{P}$ has the property that the map $N \to P$ lifts uniquely to $\widetilde{P}$ so that $\{0, 1\}$ is fixed. Moreover, using that the interior of $\widetilde{R}$ is diffeomorphic to $\mathbb{R}^2$, there is an embedding
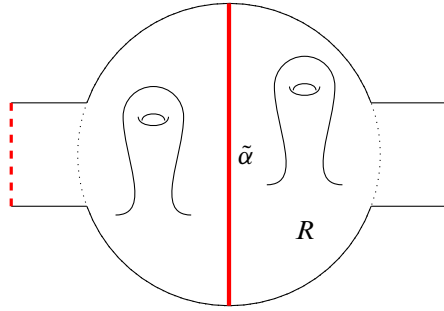
Figure 2: The surface R.

$e\colon \widetilde{P} \hookrightarrow N$ fixing $\{0, 1\}$ such that the composition $N \to \widetilde{P} \to N$ is isotopic to the identity relative to $\{0, 1\}$. Viewing $\widetilde{P}$ as being glued together by three parts — N, the two half-strips resulting from the cut 1–handle, and the two copies of $\tilde{R}$ attached to these two half-strips — this embedding $e\colon \widetilde{P} \hookrightarrow N$ is given by the identity on $N \subset \widetilde{P}$ apart from a neighbourhood of the two arcs in the boundary to which the half-strips are attached, and by pushing the half-strips and the copies of $\tilde{R}$ attached to them into this neighbourhood.

The right square is induced by postcomposition with e, so homotopy commutes in view of Section 2.3(a). The homotopy commutative left square is obtained by invoking the lifting lemma Lemma 3.3 for the covering map $\widetilde{P} \to P$. The top composition in (18) is homotopic to the identity by construction, but it remains to justify this for the bottom composition. Justifying this requires the details of the proof of Lemma 2.10, in particular the presheaf $\mathrm{Emb}_{\partial}^{\pi}(-, \widetilde{P})$ defined there. Viewing N as a submanifold of $\widetilde{P}$ as explained above, the projection $\pi\colon \widetilde{P} \to N$ is isotopic to the identity when restricted to N, so we have a dashed inclusion map of presheaves on $\mathrm{Disc}_{\partial I}$ that makes the triangle in the following diagram commute up to homotopy:

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial}(-, N) & \longhookrightarrow & \mathrm{Emb}_{\partial}(-, P) \\
& \diagdown & \uparrow{\scriptstyle \pi \circ -} \\
\subset & \dashrightarrow & \mathrm{Emb}_{\partial}^{\pi}(-, \widetilde{P}) \longhookrightarrow \mathrm{Emb}_{\partial}(-, \widetilde{P}) \xrightarrow{e \circ (-)} \mathrm{Emb}_{\partial}(-, N)
\end{array}
$$

Moreover, since $N \subset \widetilde{P} \to N$ is isotopic to the identity, the composition $\mathrm{Emb}_{\partial}(-, N) \to \mathrm{Emb}_{\partial}(-, N)$ along the bottom is homotopic to the identity. Applying T_{∞}, evaluating at I, and restricting to path components, we obtain a homotopy commutative diagram

$$
\begin{array}{ccc}
T_{\infty}\mathrm{Emb}_{\partial}(I, N)_{\alpha} & \longrightarrow & T_{\infty}\mathrm{Emb}_{\partial}(I, P)_{\tilde{\alpha}} \\
& \searrow & \uparrow{\scriptstyle \simeq} \\
& & T_{\infty}\mathrm{Emb}_{\partial}^{\pi}(I, \widetilde{P})_{\tilde{\alpha}} \longrightarrow T_{\infty}\mathrm{Emb}_{\partial}(I, \widetilde{P})_{\tilde{\alpha}} \xrightarrow{e \circ (-)} T_{\infty}\mathrm{Emb}_{\partial}(I, N)_{\alpha}
\end{array}
$$

whose composition along the bottom $T_{\infty}\mathrm{Emb}_{\partial}(-, N)_{\alpha} \to T_{\infty}\mathrm{Emb}_{\partial}(-, N)_{\alpha}$ is homotopic to the identity. The composition along the top involving a wrong-way weak equivalence agrees by construction with the bottom composition of (18), so it is homotopic to the identity, as claimed (recall Remark 2.3).
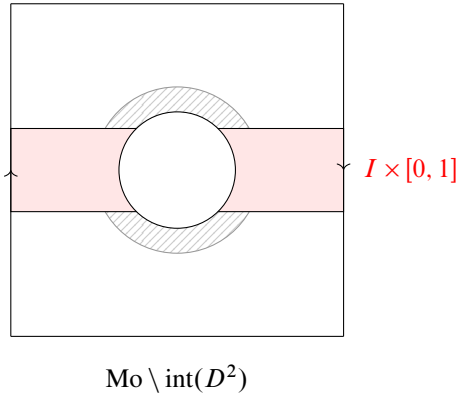
Figure 3: The complement of an open disc in the Möbius strip. The red copy of $I \times [0,1]$ differs up to isotopy equivalence from $\mathrm{Mo} \setminus \mathrm{int}(D^2)$ only in the hatched region which is diffeomorphic to $I \times [0,1) \sqcup I \times [0,1)$.

Step (2.2): Convergence for (M, N) if M is a circle, cylinder, or Möbius strip, and $\dim(N) = 2$
By Lemma 2.9, it suffices to prove the claim for the cylinder and the Möbius strip. We will do so for the Möbius strip $M = \mathrm{Mo}$; the argument for the cylinder is analogous. We pick a disc $D^2 \subset \mathrm{int}(\mathrm{Mo})$. By Lemma 2.4, it suffices to prove that

$$\mathrm{Emb}_{\partial_0}\big(\mathrm{Mo} \setminus \mathrm{int}(D^2), N \setminus \mathrm{int}(e(D^2))\big) \to T_\infty \mathrm{Emb}_{\partial_0}\big(\mathrm{Mo} \setminus \mathrm{int}(D^2), N \setminus \mathrm{int}(e(D^2))\big)$$

is a weak equivalence for all embeddings $e \colon D^2 \hookrightarrow \mathrm{int}(\Sigma)$. To this end, we pick a subtriad $I \times [0,1] \subset \mathrm{Mo} \setminus \mathrm{int}(D^2)$ as in Figure 3 and attempt to show that the vertical restriction maps in the diagram

$$
\begin{array}{ccc}
\mathrm{Emb}_{\partial_0}\big(\mathrm{Mo} \setminus \mathrm{int}(D^2), N \setminus \mathrm{int}(e(D^2))\big) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}\big((\mathrm{Mo} \setminus \mathrm{int}(D^2)), N \setminus \mathrm{int}(e(D^2))\big) \\
\downarrow & & \downarrow \\
\mathrm{Emb}_{\partial_0}\big(I \times [0,1], N \setminus \mathrm{int}(e(D^2))\big) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}\big(I \times [0,1], N \setminus \mathrm{int}(e(D^2))\big)
\end{array}
$$

are weak equivalences. Isotopy extension exhibits the homotopy fibre of the left vertical map up to smoothing corners and isotopy equivalence as $\mathrm{Emb}_\partial\big(I \times [0,1) \sqcup I \times [0,1), N \setminus \mathrm{int}(e(D^2))\big)$ which is contractible by the contractibility of spaces of collars. To see that the right vertical map is an equivalence, one combines this observation with descent with respect to a Weiss ∞–cover of open discs and collars on ∂D^2 similarly to the proof of Lemma 2.12. As the bottom horizontal map is a weak equivalence by step (2.1), the top horizontal map is a weak equivalence as well.

Step (2.3): Convergence for (M, N) if $M = (T_1 \times I \times [0,1]) \sqcup (T_2 \times S^1 \times [0,1]) \sqcup (T_3 \times \mathrm{Mo})$ for (possibly empty) finite sets T_i and $\dim(N) = 2$ The proof is by induction over $t = |T_1| + |T_2| + |T_3|$. The initial case $t = 1$ is provided by steps (2.1) and (2.2). For the induction step, we pick a component of M, say of the form $I \times [0,1]$; the other cases are analogous. We consider $M' := M \setminus I \times [0,1]$. An application of isotopy extension (see Section 2.3(f)) to $P = I \times \big[\frac{1}{4}, \frac{3}{4}\big] \subset I \times [0,1]$, justified by Lemma 2.13

and step (2.1), gives fibre sequences

$$
\begin{array}{ccccc}
\mathrm{Emb}_{\partial_0}\big(M', N \setminus \mathrm{int}(e(P))\big) & \longrightarrow & \mathrm{Emb}_{\partial_0}(M, N) & \longrightarrow & \mathrm{Emb}_{\partial_0}(P, N) \\
\downarrow & & \downarrow & & \downarrow{\scriptstyle\simeq} \\
T_\infty \mathrm{Emb}_{\partial_0}\big(M', N \setminus \mathrm{int}(e(P))\big) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(M, N) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(P, N)
\end{array}
$$

Here we used Lemma 2.12 and isotopy invariance to replace $M' \sqcup (I \times [0, 1] \setminus \mathrm{int}(P))$ in the domain with M'. The left vertical map is a weak equivalence by the induction hypothesis, so the middle vertical map is a weak equivalence too.

Step (2.4): Convergence for (M, N) if $\dim(M) = 1$ and $\dim(N) = 2$ Step (2.3) together with Lemma 2.9 gives the result for those triads of the form $M' = (T_1 \times I) \sqcup (T_2 \times S^1)$ for finite sets T_i and $\partial_0 M' = T_1 \times \{0, 1\}$. The general case, which has

$$
M = (T_1 \times I) \sqcup (T_2 \times S^1) \sqcup (T_3 \times [0, 1]) \sqcup (T_4 \times [0, 1])
$$

for finite sets T_i and $\partial_0 M = (T_1 \times \{0, 1\}) \sqcup (T_3 \times \{0\})$ follows from this by Lemmas 2.12 and 2.13 together with isotopy invariance (see Section 2.3(b)).

Step (2.5): Convergence for (M, N) if $M = D^2$ with $\partial_0 M = \partial M$ and $\dim(N) = 2$ By Corollary 2.15 we may assume that N is connected.

We first prove the case where the target N is *not* diffeomorphic to D^2. In this case $\mathrm{Emb}_\partial(D^2, N) = \varnothing$, so we need to show $T_\infty \mathrm{Emb}_\partial(D^2, N) = \varnothing$. If this were to fail, then the target of the map

$$
T_\infty \mathrm{Emb}_\partial(D^2, N) \to \mathrm{Map}_\partial(D^2, N)
$$

from Section 2.3(d) must be nonempty, so N would be a connected surface with a boundary component whose inclusion is null-homotopic. We claim this is impossible unless $N \cong D^2$. First, if $N = N_1 \natural \cdots \natural N_1$, then $\pi_1(N)$ splits as a free product $\pi_1(N_1) * \cdots * \pi_1(N_n)$ and we may choose this decomposition so that the homotopy class of the boundary inclusion represents the free product of the homotopy classes of boundary inclusions of those components at which we perform the boundary connected sums. By the classification of connected compact surfaces, it then suffices to observe that all boundary inclusions are nontrivial in the fundamental group of the surfaces $\Sigma_{0,2}$, $\Sigma_{1,1}$, and Mo. For $\Sigma_{0,2}$, each inclusion represents a generator of $\pi_1(\Sigma_{0,2}) \cong \mathbb{Z}$, for $\Sigma_{1,1}$ the boundary inclusion represents $xyx^{-1}y^{-1} \in \pi_1(\Sigma_{1,1}) \cong \langle x, y \rangle$, and for the Möbius strip it represents twice a generator in $\pi_1(\mathrm{Mo}) \cong \mathbb{Z}$.

It remains to show that $\mathrm{Emb}_\partial(D^2, D^2) \to T_\infty \mathrm{Emb}_\partial(D^2, D^2)$ a weak equivalence for which we follow the proof of what is sometimes called the *Cerf lemma* [4, Proposition 5]. We consider the triad $H = D^2 \cap \big([-\frac{1}{2}, \infty) \times \mathbb{R}\big)$ with $\partial_0 H = H \cap \partial D^2$ and $\partial_1 H = H \cap \big(\{-\frac{1}{2}\} \times \mathbb{R}\big)$ containing the strip $J = H \cap \big([-\frac{1}{4}, \frac{1}{4}] \times \mathbb{R}\big)$ with $\partial_0 J = J \cap \partial D^2$; see Figure 4. Writing $H_0 = H \setminus \big((-\frac{1}{4}, \frac{1}{4}) \times \mathbb{R}\big) \cap H$ and $D_0^2 = D^2 \setminus \big((-\frac{1}{4}, \frac{1}{4}) \times \mathbb{R}\big) \cap D^2$, an application of isotopy extension (see Section 2.3(f)) justified by step (2.1) in the case $M = J \cong I \times [0, 1]$ and Lemma 2.13 gives a map of fibre sequences with connected
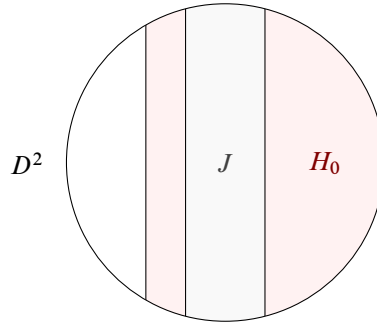
Figure 4: The triads J, $H_0 \subset D^2$. Here H the union of J and H_0, and $H_0' \subset H_0$ is the component to the right of J.

weakly equivalent bases and homotopy fibres over the standard inclusion $J \hookrightarrow D^2$:

$$
\begin{array}{ccccc}
\mathrm{Emb}_{\partial_0}(H_0, D_0^2) & \longrightarrow & \mathrm{Emb}_{\partial_0}(H, D^2) & \longrightarrow & \mathrm{Emb}_{\partial_0}(J, D^2) \\
\downarrow & & \downarrow & & \downarrow \simeq \\
T_\infty \mathrm{Emb}_{\partial_0}(H_0, D_0^2) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(H, D^2) & \longrightarrow & T_\infty \mathrm{Emb}_{\partial_0}(J, D^2)
\end{array}
$$

As H is a closed collar on $\partial_0 H$, the middle vertical map is a weak equivalence by isotopy invariance and the convergence on collars (see Section 2.3(b) and (c)). By Lemma 2.12 we may discard the collar $H_0 \cap ((-\infty, 0] \times \mathbb{R})$ from the source of the left vertical map, and obtain that for $H_0' = H \cap \left(\left[\frac{1}{4}, \infty \right) \times \mathbb{R} \right)$ the map $\mathrm{Emb}_{\partial_0}(H_0', D_0^2) \to T_\infty \mathrm{Emb}_{\partial_0}(H_0', D_0^2)$ is a weak equivalence. Invoking Corollary 2.15 to neglect $D_0^2 \setminus H_0$ from the target and identifying H_0' with a disc upon smoothing corners, we conclude that $\mathrm{Emb}_\partial(D^2, D^2) \to T_\infty \mathrm{Emb}_\partial(D^2, D^2)$ is a weak equivalence.

Step (2.6): Convergence for (M, N) if M is an orientable surface of genus 0 with $n \geq 1$ boundary components and $\partial_0 M = \partial M$ and $\dim(N) = 2$ Note that by gluing $n - 1$ discs to M we obtain a disc D^2. We also glue $n - 1$ discs to the corresponding boundary components of N to obtain a triad N' with a canonical embedding $e : \underline{n-1} \times D^2 \hookrightarrow N'$. Then isotopy extension and the convergence on discs (see Section 2.3(f) and (c)) yields fibre sequences

$$
\begin{array}{ccccc}
\mathrm{Emb}_\partial(M, N) & \longrightarrow & \mathrm{Emb}_\partial(D^2, N') & \longrightarrow & \mathrm{Emb}(\underline{n-1} \times D^2, N') \\
\downarrow & & \downarrow & & \downarrow \simeq \\
T_\infty \mathrm{Emb}_\partial(M, N) & \longrightarrow & T_\infty \mathrm{Emb}(D^2, N') & \longrightarrow & T_\infty \mathrm{Emb}(\underline{n-1} \times D^2, N')
\end{array}
$$

The middle vertical map a weak equivalence by step (2.5), so the left map is one as well.

Step (2.7): Convergence for (M, N) if M is connected, $\partial_0 M = \partial M$, and $\dim(M) = \dim(N) = 2$ As a result of Lemma 2.4, we may assume that $\partial M \neq \varnothing$, so M is a boundary connected sum

$$
\Sigma_{0,n} \natural (\Sigma_{1,1})^{\natural T_1} \natural (\mathbb{R}P^2)^{\natural T_2}
$$
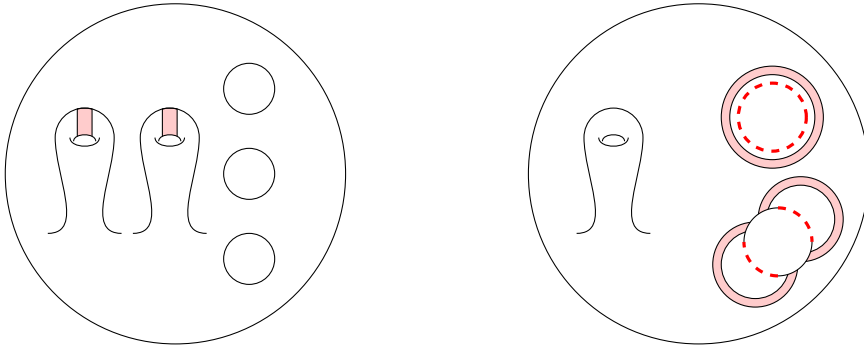
Figure 5: Left: $M = \Sigma_{0,4} \natural (\Sigma_{1,1})^{\natural 2}$ with subtriad $P = \underline{2} \times S^1 \times [0,1] \subset M$ whose complement has genus 0 and 8 boundary components. Right: $M = \Sigma_{0,3} \natural \Sigma_{1,1}$ with $\partial_1 M$ dashed, with subtriad $P = \underline{2} \times I \times [0,1] \sqcup S^1 \times [0,1] \subset M$; the component of $M \setminus \text{int}(P)$ containing $\Sigma_{1,1}$ is M'.

for $n \geq 1$ and possibly empty finite sets T_1 and T_2. Thus we may find an embedding

$$P = (T_1 \times S^1 \times [0,1]) \sqcup (T_2 \times \text{Mo}) \to M$$

such that $M \setminus \text{int}(P) \cong \Sigma_{0,n'}$ with $n' = n + 2|T_1| + |T_2|$; see Figure 5, left, for an example. For any embedding $e \colon M \hookrightarrow N$ extending the boundary condition, an application of isotopy extension (see Section 2.3(f)), justified by step (2.3) and Lemma 2.13, gives a map of fibre sequences

$$
\begin{array}{ccccc}
\text{Emb}_\partial\big(\Sigma_{0,n'}, N \setminus e(\text{int}(P))\big) & \longrightarrow & \text{Emb}_\partial(M, N) & \longrightarrow & \text{Emb}(P, N) \\
\downarrow & & \downarrow & & \downarrow{\scriptstyle \simeq} \\
T_\infty \text{Emb}_\partial\big(\Sigma_{0,n'}, N \setminus e(\text{int}(P))\big) & \longrightarrow & T_\infty \text{Emb}(M, N) & \longrightarrow & T_\infty \text{Emb}(P, N)
\end{array}
$$

whose left vertical map a weak equivalence by step (2.7). Varying the embedding $e \colon M \hookrightarrow N$, we conclude that the middle vertical map is also a weak equivalence.

Step (2.8): Convergence for (M, N) if M is connected, $\partial_0 M \neq \partial M$, and $\dim(M) = \dim(N) = 2$
Choose a triad embedding $P = (T_1 \times I \times [0,1]) \sqcup (T_2 \times S^1 \times [0,1]) \hookrightarrow M$ such that $M \setminus \text{int}(P)$ is the disjoint union of a component M' with $\partial M' = M \cap \partial_0(M \setminus \text{int}(P))$ and collars on components of $\partial_0(M \setminus \text{int}(P))$; see Figure 5, right, for an example. By step (2.3) and Lemma 2.13, we may apply isotopy extension as in step (2.3) to the restriction map $\text{Emb}_{\partial_0}(M, N) \to \text{Emb}_{\partial_0}(P, N)$ and its T_∞-version. From step (2.7) and Lemma 2.12 we see that the map between fibres is a weak equivalence, from which we conclude the claim.

Step (2.9): Convergence for (M, N) if $\dim(M) = \dim(N) = 2$ This is a induction on the number n of components of M. The initial case $n = 1$ is the previous one, and for the induction step we write $M = M' \sqcup M''$ with M' connected. The induction hypothesis applied to M' together with Lemma 2.13 ensures that we may apply isotopy extension (see Section 2.3(f)) to the restriction

$$\text{Emb}_{\partial_0}(M, N) \to \text{Emb}_{\partial_0}(M', N)$$

and its T_∞–version from which the claim follows by noting that the map on fibres is a weak equivalence by applying the induction hypothesis to M''.

Step (3): Convergence for (M, N) if $\dim(M) = \dim(N) = 1$ This can be proved similarly to step (2) but is easier. We outline the argument.

First one proves the case $M = D^1$ with $\partial_0(D^1) = \{-1, 1\}$ by a strategy analogous to step (2.5): one first uses Corollary 2.15 to reduce to $N = D^1$ as in the case for surfaces. Then one takes $H = \left[-\frac{1}{2}, 1\right]$, $J = \left[-\frac{1}{4}, \frac{1}{4}\right]$, and $D_0^1 = D^1 \setminus \mathrm{int}(J)$ and develops a map of fibre sequences

$$
\begin{array}{ccccc}
\mathrm{Emb}_{\partial_0}(H_0, D_0^1) & \longrightarrow & \mathrm{Emb}_{\partial_0}(H, D^1) & \longrightarrow & \mathrm{Emb}_{\partial_0}(J, D^1) \\
\downarrow & & \downarrow{\simeq} & & \downarrow{\simeq} \\
T_\infty\mathrm{Emb}_{\partial_0}(H_0, D_0^1) & \longrightarrow & T_\infty\mathrm{Emb}_{\partial_0}(H, D^1) & \longrightarrow & T_\infty\mathrm{Emb}_{\partial_0}(J, D^1)
\end{array}
$$

similar to step (2.5). Using Lemma 2.12 and Corollary 2.15, the map on fibres agrees with

$$
\mathrm{Emb}_{\partial}(D^1, D^1) \to T_\infty\mathrm{Emb}_{\partial}(D^1, D^1),
$$

so it is a weak equivalence.

Next one shows the case of a general connected triad M: the case $M = S^1$ follows directly by an application of isotopy extension (see Section 2.3(f)) together with the case $M = D^1$ above, and the cases $M = [0, 1]$ with $\partial_0(M) = \{0\}$ or $\partial_0(M) = \varnothing$ hold by Section 2.3(c).

Finally, the case of a possibly disconnected triad M can be settled as in step (2.8).

3.1   Automorphisms of the E_1– and E_2–operad

The above arguments do not rely on the fact that $\mathrm{Diff}_{\partial}(D^d) = \mathrm{Emb}_{\partial}(D^d, D^d)$ is contractible for $d \le 2$ (this is folklore for $d = 1$ and due to Smale for $d = 2$ [24]). Using this fact, we may conclude from Theorem 3.1 that $T_\infty\mathrm{Emb}_{\partial}(D^d, D^d)$ is contractible for $d \le 2$. Combining Theorems 1.2, 1.4, and 6.4 of [2],

$$
T_\infty\mathrm{Emb}_{\partial}(D^d, D^d) \simeq \Omega^{d+1}\mathrm{Aut}^h(E_d)/\mathrm{O}(d)
$$

where $\mathrm{Aut}^h(E_d)/\mathrm{O}(d)$ is the homotopy fibre of the map $B\mathrm{O}(d) \to B\mathrm{Aut}^h(E_d)$ resulting from the standard action of $\mathrm{O}(d)$ on the little discs operad by derived operad automorphisms, so we deduce:

Corollary 3.4 $\Omega^{d+1}\mathrm{Aut}^h(E_d)/\mathrm{O}(d) \simeq *$ for $d \le 2$.

Remark 3.5   Horel [15, Theorem 8.5] proved that $\mathrm{Aut}^h(E_2)/\mathrm{O}(2) \simeq *$ with different methods. His proof crucially uses that the spaces of k–arity operations in the operad E_2 are $K(\pi, 1)$ for all k. This fact can also be used to give an alternative proof of $\Omega^2\mathrm{Aut}^h(E_2) \simeq *$ (and thus of Corollary 3.4): the derived mapping space $\mathrm{Map}^h(O, P)$ between operads O and P can be computed as a homotopy limit of

a diagram whose values are products of spaces of operations in O and P; this follows by (for example) using the alternative model of operads in terms of dendroidal Segal spaces. Applied to $O = P = E_2$, one sees that $\mathrm{Map}^h(E_2, E_2)$ is a homotopy limit of $K(\pi, 1)$, so it is contractible after looping twice.

4 Embedding calculus and the Johnson filtration

This section serves to introduce the filtration (2) of the mapping class group $\pi_0 \mathrm{Diff}_\partial(\Sigma_{g,1})$, and to prove in Theorem 4.2 that it is contained in the Johnson filtration.

4.1 The cardinality filtration

Returning to the general setting of manifold calculus of Section 2.2.1 with a fixed $(d-1)$–manifold K, possibly with boundary, we consider the filtration

$$(19) \qquad \mathrm{Disc}_{\partial_0, \leq 0} \subset \mathrm{Disc}_{\partial_0, \leq 1} \subset \cdots \subset \mathrm{Disc}_{\partial_0, \leq \infty} := \mathrm{Disc}_{\partial_0}$$

of the topologically enriched category $\mathrm{Disc}_{\partial_0}$ by its full subcategories $\mathrm{Disc}_{\partial_0, \leq k}$ on triads that are diffeomorphic to $K \times [0, 1) \sqcup T \times \mathbb{R}^d$ for finite sets T of bounded cardinality $\leq k$. Localising the categories $\mathrm{PSh}(\mathrm{Disc}_{\partial_0, \leq k})$ at the objectwise weak equivalences as we did for $k = \infty$ in Section 2.2.1, given a presheaf $F \in \mathrm{PSh}(\mathrm{Disc}_{\partial_0})$ we obtain presheaves on $\mathrm{Man}_{\partial_0}$ by

$$T_k F(M) := \mathrm{Map}_{\mathrm{PSh}(\mathrm{Disc}_{\partial_0, \leq k})^{\mathrm{loc}}}(\mathrm{Emb}_{\partial_0}(-, M), F)$$

which are related by maps of presheaves

$$(20) \qquad T_\infty F(M) \to \cdots \to T_2 F(M) \to T_1 F(M)$$

induced by restriction along the inclusions (19). If F is the restriction of a presheaf on $\mathrm{Man}_{\partial_0}$, we can precompose this tower with the canonical map $F(M) \to T_\infty F(M)$ from (5).

4.1.1 Sheaf-theoretic point of view The tower (20) can also be seen from the point of view of $\mathcal{J}_k$–sheaves as described in Section 2.2.3(b): by [1, Theorem 1.2] the functor

$$\mathrm{PSh}(\mathrm{Man}_{\partial_0}) \ni F \mapsto T_k F \in \mathrm{PSh}(\mathrm{Man}_{\partial_0})$$

together with the natural transformation $\mathrm{id}_{\mathrm{PSh}(\mathrm{Man}_{\partial_0})} \Rightarrow T_k$ is a model for the homotopy $\mathcal{J}_k$–sheafification. From this point of view the maps (20) are induced by the universal property of homotopy sheafification, using the fact that any $\mathcal{J}_{k+1}$–sheaf is in particular a $\mathcal{J}_k$–sheaf.

In particular, in the case of embedding calculus, ie for presheaves $F(-) = \mathrm{Emb}_{\partial_0}(-, N)$ for triads M and N and a boundary condition $e_{\partial_0}: \partial_0 M \hookrightarrow \partial_0 N$ (see Example 2.1), this implies that there is a factorisation of the map from the discussion in Section 2.3(d) of the form

$$(21) \qquad T_\infty \mathrm{Emb}_{\partial_0}(M, N) \to \cdots \to T_1 \mathrm{Emb}_{\partial_0}(M, N) \to \mathrm{Bun}_{\partial_0}(TM, TN) \to \mathrm{Map}_{\partial_0}(M, N).$$

4.2 $H\mathbb{Z}$–embedding calculus

Much of the above goes through for presheaves valued in categories other than spaces. We have use for one such generalisation, which we discuss now.

It involves the topologically enriched category Sp of spectra and the topologically enriched category $H\mathbb{Z}\mathsf{mod}$ of module spectra over the Eilenberg–Mac Lane spectrum $H\mathbb{Z}$, both modelled for example using symmetric spectra in spaces as in [20]. We denote by $\mathsf{PSh}^{H\mathbb{Z}}(\mathsf{Disc}_{\partial_0})$ the category of $H\mathbb{Z}$–module spectrum-valued enriched presheaves on $\mathsf{Disc}_{\partial_0}$, and its localisation at the objectwise stable equivalences by $\mathsf{PSh}^{H\mathbb{Z}}(\mathsf{Disc}_{\partial_0})^{\mathrm{loc}}$. The composition of the left-adjoints $\Sigma_+^\infty \colon \mathsf{Top} \to \mathsf{Sp}$ and $- \wedge H\mathbb{Z} \colon \mathsf{Sp} \to H\mathbb{Z}\mathsf{mod}$ induces the vertical arrows in the commutative diagram

$$
(22) \qquad
\begin{array}{ccc}
\mathsf{PSh}(\mathsf{Disc}_{\partial_0}) & \longrightarrow & \mathsf{PSh}(\mathsf{Disc}_{\partial_0})^{\mathrm{loc}} \\
{\scriptstyle (-)_+ \wedge H\mathbb{Z}}\big\downarrow & & \big\downarrow{\scriptstyle (-)_+ \wedge H\mathbb{Z}} \\
\mathsf{PSh}^{H\mathbb{Z}}(\mathsf{Disc}_{\partial_0}) & \longrightarrow & \mathsf{PSh}^{H\mathbb{Z}}(\mathsf{Disc}_{\partial_0})^{\mathrm{loc}}
\end{array}
$$

For a presheaf $F \in \mathsf{PSh}(\mathsf{Disc}_{\partial_0})$ we define presheaves

$$
T_k^{H\mathbb{Z}} F(M) := \mathrm{Map}_{\mathsf{PSh}^{H\mathbb{Z}}(\mathsf{Disc}_{\partial_0, \le k})^{\mathrm{loc}}}(\mathrm{Emb}_{\partial_0}(-, M)_+ \wedge H\mathbb{Z}, F_+ \wedge H\mathbb{Z})
$$

for $1 \le k \le \infty$, giving rise to an extension of the tower (20) to a map of towers

$$
(23) \qquad
\begin{array}{ccccccc}
T_\infty F(M) & \longrightarrow & \cdots & \longrightarrow & T_2 F(M) & \longrightarrow & T_1 F(M) \\
\big\downarrow & & & & \big\downarrow & & \big\downarrow \\
T_\infty^{H\mathbb{Z}} F(M) & \longrightarrow & \cdots & \longrightarrow & T_2^{H\mathbb{Z}} F(M) & \longrightarrow & T_1^{H\mathbb{Z}} F(M)
\end{array}
$$

whose vertical maps are induced by (22) and horizontal maps are induced by restriction along (19). Note that for $F(-) = \mathrm{Emb}_{\partial_0}(-, M)$, composition induces an E_1–structure on $T_k F(M) = T_k \mathrm{Emb}_{\partial_0}(M, M)$ and $T_k^{H\mathbb{Z}} F(M) = T_k^{H\mathbb{Z}} \mathrm{Emb}_{\partial_0}(M, M)$ which upgrades (23) to a diagram of E_1–spaces.

Remark 4.1 In [28], Weiss considers manifold calculus applied to the space-valued presheaf

$$
\Omega^\infty(\mathrm{Emb}_{\partial_0}(-, M)_+ \wedge H\mathbb{Z}).
$$

This agrees with the above $H\mathbb{Z}$–embedding calculus since the adjunctions $\Sigma_+^\infty \dashv \Omega^\infty$ and $- \wedge H\mathbb{Z} \dashv U$, with $U \colon H\mathbb{Z}\mathsf{mod} \to \mathsf{Sp}$ the forgetful functor, induce adjunctions on presheaf categories, which in turn induces for $F \in \mathsf{PSh}(\mathsf{Man}_{\partial_0})$ and $1 \le k \le \infty$ an identification

$$
\mathrm{Map}_{\mathsf{PSh}(\mathsf{Disc}_{\partial_0 M, \le k})^{\mathrm{loc}}}(\mathrm{Emb}_{\partial_0}(-, M), \Omega^\infty(F_+ \wedge H\mathbb{Z}))
$$

$$
\downarrow{\simeq}
$$

$$
T_k^{H\mathbb{Z}} F(M) = \mathrm{Map}_{\mathsf{PSh}^{H\mathbb{Z}}(\mathsf{Disc}_{\partial_0 M, \le k})^{\mathrm{loc}}}(\mathrm{Emb}_{\partial_0}(-, M)_+ \wedge H\mathbb{Z}, F_+ \wedge H\mathbb{Z})
$$

which is compatible with the restrictions maps.

4.3 An $H\mathbb{Z}$–embedding calculus filtration of $\pi_0 \mathrm{Diff}_\partial(\Sigma)$

We fix a compact orientable Σ of genus g with a single boundary component. A naive attempt at a filtration of the mapping class group $\pi_0 \mathrm{Diff}_\partial(\Sigma)$ as promised in the introductory Section 1.2 would be to consider the kernels of the maps $\pi_0 \mathrm{Diff}_\partial(\Sigma) = \pi_0 \mathrm{Emb}_\partial(\Sigma, \Sigma) \to \pi_0 T_k \mathrm{Emb}_\partial(\Sigma, \Sigma)$ for varying k, but these turn out to be trivial for all $k \geq 1$ simply because the composition of the above map with the map to $\pi_0 \mathrm{Map}_\partial(\Sigma, \Sigma)$ from (21) is injective (see Remark 2.6). To obtain a more interesting filtration, we perform two modifications.

Firstly, we change the triad structure of Σ. Instead of $\partial_0 \Sigma = \partial\Sigma$ we choose $\partial_0 \Sigma \subset \partial\Sigma$ to be an embedded interval. We think of $\partial_0 \Sigma$ as "half the boundary" and abbreviate $\partial/2 := \partial_0 \Sigma \cong [0, 1]$. Note that the inclusion $\mathrm{Diff}_\partial(\Sigma) \subset \mathrm{Emb}_{\partial/2}(\Sigma, \Sigma)$ is a homotopy equivalence since its homotopy fibres are equivalent to $\mathrm{Diff}_\partial(D^2) \simeq *$. The maps $\pi_0 \mathrm{Diff}_\partial(\Sigma) = \pi_0 \mathrm{Emb}_{\partial/2}(\Sigma, \Sigma) \to \pi_0 T_k \mathrm{Emb}_{\partial/2}(\Sigma, \Sigma)$ still do not give rise to an interesting filtration, for a similar reason as above since the map $\pi_0 \mathrm{Map}_\partial(\Sigma, \Sigma) \to \pi_0 \mathrm{Map}_{\partial/2}(\Sigma, \Sigma)$ is injective. The filtration becomes more interesting after the second modification: we switch from embedding calculus to embedding calculus in $H\mathbb{Z}$–modules as described above. More precisely, we consider the filtration

$$(24) \qquad \pi_0 \mathrm{Diff}_\partial(\Sigma) = T\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(0) \supset T\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(1) \supset T\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(2) \supset \cdots$$

defined by

$$T\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(k) := \ker[\pi_0 \mathrm{Diff}_\partial(\Sigma) \to \pi_0 T_k^{H\mathbb{Z}} \mathrm{Emb}_{\partial/2}(\Sigma, \Sigma)],$$

where we formally set $\pi_0 T_k^{H\mathbb{Z}} \mathrm{Emb}_{\partial/2}(\Sigma, \Sigma) := *$. Denoting by

$$(25) \qquad \pi_0 \mathrm{Diff}_\partial(\Sigma) = \mathcal{J}(0) \supset \mathcal{J}(1) \supset \mathcal{J}(2) \supset \cdots$$

the usual Johnson filtration

$$\mathcal{J}(k) := \ker\left[\pi_0 \mathrm{Diff}_\partial(\Sigma) \to \mathrm{Aut}\left(\frac{\pi_1(\Sigma, *)}{\Gamma_k(\pi_1(\Sigma, *))}\right)\right],$$

where $\Gamma_i(-)$ is the ith stage in the lower central series of a group (so $\Gamma_0(G) = G$ and $\Gamma_1(G)$ is the derived subgroup of G), the purpose of this section is to relate the filtrations (24) and (25) as follows.

Theorem 4.2 *For a compact orientable surface Σ with a single boundary component, the subgroup*

$$T\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(k) = \ker[\pi_0 \mathrm{Diff}_\partial(\Sigma) \to \pi_0 T_k^{H\mathbb{Z}} \mathrm{Emb}_{\partial/2}(\Sigma, \Sigma)]$$

is contained in the kth stage $\mathcal{J}(k)$ of the Johnson filtration for $k \geq 0$.

Remark 4.3   (i)   The group $\pi_1(\Sigma, *)$ is free, so it is residually nilpotent (ie $\bigcap_k \Gamma_k(\pi_1(\Sigma, *)) = \{1\}$), which implies that the Johnson filtration is exhaustive, ie $\bigcap_k \mathcal{J}(k) = \{\mathrm{id}\}$. By Theorem 4.2, the same holds for $\{T\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(k)\}$ so in particular the map $\pi_0 \mathrm{Diff}_\partial(\Sigma) \to \pi_0 T_\infty^{H\mathbb{Z}} \mathrm{Emb}_{\partial/2}(\Sigma, \Sigma)$ is injective.

(ii)   If the genus of Σ is at least 3, then the inclusion $T\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(1) \subset \mathcal{J}(1)$ is strict. Indeed, an element of the mapping class group lies in $T\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(1)$ if and only if induced the identity on the homology of frame bundle $\mathrm{Fr}(T\Sigma)$. By [25, Theorem 2.2 and Corollary 2.7], this is the case if and only if it lies in the *Chillingworth subgroup* of the Torelli subgroup $\mathcal{J}(1)$ [5; 6].

Theorem 4.2 and the final part of the previous remark suggest:

Question 4.4   What is the precise relationship between the Johnson filtration $\mathcal{J}(k)$ and the filtration $\mathcal{J}_{\partial/2}^{H\mathbb{Z}}(k)$ arising from the $H\mathbb{Z}$–embedding calculus tower?

We will deduce Theorem 4.2 from Moriyama's work [21]. The key step for this deduction is not special to surfaces and applies to a general d–dimensional manifold triad M, so we will formulate it in this generality. To do so, we fix a presheaf $F \in \mathsf{PSh}(\mathsf{Disc}_{\partial_0, \leq k})$, restrict it to $\mathsf{Disc}_{\partial_0, \leq k-1}$ and homotopy left Kan extending it back along the inclusion $\iota_k \colon \mathsf{Disc}_{\partial_0, \leq k-1} \subset \mathsf{Disc}_{\partial_0, \leq k}$ to obtain a presheaf $\mathrm{hLan}_{\iota_k} F$ with a natural map $\mathrm{hLan}_{\iota_k} F \to F$. Evaluating it at

$$\partial_0 \sqcup \mathbb{R}^d_{\underline{k}} := \partial_0 M \times [0,1) \sqcup \underline{k} \times \mathbb{R}^d$$

where $\underline{k} := \{1, \ldots, k\}$ we get a map of $\Sigma_k \wr \mathrm{O}(d)$–spaces $(\mathrm{hLan}_{\iota_k} F)(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}) \to F(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}})$, and then taking homotopy quotients by the subgroup $\mathrm{O}(d)^k \subset \Sigma_k \wr \mathrm{O}(d)$ gives a map

$$(26) \qquad (\mathrm{hLan}_{\iota_k} F)(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}) /\!/ \mathrm{O}(d)^k \to F(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}) /\!/ \mathrm{O}(d)^k.$$

In Proposition 4.5 below, we relate this map for $F(-) = \mathrm{Emb}_{\partial_0}(-, M)$ to a certain "boundary inclusion" of the ordered configuration spaces $\mathrm{Emb}(\underline{k}, M)$. For this, recall the *Fulton–MacPherson compactification* $\mathrm{FM}_k(M)$ of $\mathrm{Emb}(\underline{k}, M)$ (eg from [23]) which comes with a natural inclusion $\mathrm{Emb}(\underline{k}, M) \hookrightarrow \mathrm{FM}_k(M)$ that is homotopy equivalence, and a "macroscopic location" map $\mu \colon \mathrm{FM}_k(M) \to M^k$ that extends the inclusion $\mathrm{Emb}(\underline{k}, M) \hookrightarrow M^k$. We write $\partial_0 \mathrm{FM}_k(M)$ for the preimage $\mu^{-1}(\Delta_k \cup A_k)$ of the union of the subspace $A_k \subset M^k$ where at least one point lies in $\partial_0 M$ and the fat diagonal

$$\Delta_k := \{(m_1, \ldots, m_i) \in M^k \mid m_i = m_j \text{ for some } i \neq j\} \subset M^k.$$

The key step in the proof of Theorem 4.2 is to identify the map (26) for $F(-) = \mathrm{Emb}_{\partial_0}(-, M)$ with the boundary inclusion $\partial_0 \mathrm{FM}_k(M) \subset \mathrm{FM}_k(M)$ in the following sense:

Proposition 4.5   *There is zigzag of compatible weak equivalences*

$$
\begin{array}{ccccc}
(\mathrm{hLan}_{\iota_k} \mathrm{Emb}_{\partial_0}(-, M))(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}) /\!/ \mathrm{O}(d)^k & \xrightarrow{\simeq} & \cdots & \xleftarrow{\simeq} & \partial_0 \mathrm{FM}_k(M) \\
\downarrow & & \downarrow & & \downarrow \\
\mathrm{Emb}(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}, M) /\!/ \mathrm{O}(d)^k & \xrightarrow{\simeq} & \cdots & \xleftarrow{\simeq} & \mathrm{FM}_k(M)
\end{array}
$$

which, when varying M, defines a zigzag of weak equivalences in the arrow category of $\mathsf{Fun}(\mathsf{Man}_{\partial_0}, \mathsf{S})$.

Before turning to the proof of Proposition 4.5, we explain how it implies Theorem 4.2.

Proof of Theorem 4.2 An element $\phi \in \text{Diff}_\partial(\Sigma)$ induces a commutative diagram

(27)
$$
\begin{array}{ccc}
\partial_0 \, \text{FM}_k(\Sigma)_+ \wedge H\mathbb{Z} & \xrightarrow{\phi_*} & \partial_0 \, \text{FM}_k(\Sigma)_+ \wedge H\mathbb{Z} \\
\downarrow & & \downarrow \\
\text{FM}_k(\Sigma)_+ \wedge H\mathbb{Z} & \xrightarrow{\phi_*} & \text{FM}_k(\Sigma)_+ \wedge H\mathbb{Z}
\end{array}
$$

Abbreviating $E_\Sigma := \text{Emb}_{\partial_0}(-, \Sigma)$, this agrees by Proposition 4.5 with the square

$$
\begin{array}{ccc}
((\text{hLan}_{\iota_k} E_\Sigma)(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}) /\!/ \text{O}(d)^k)_+ \wedge H\mathbb{Z} & \xrightarrow{\phi_*} & ((\text{hLan}_{\iota_k} E_\Sigma)(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}) /\!/ \text{O}(d)^k)_+ \wedge H\mathbb{Z} \\
\downarrow & & \downarrow \\
(\text{Emb}(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}, \Sigma) /\!/ \text{O}(d)^k)_+ \wedge H\mathbb{Z} & \xrightarrow{\phi_*} & (\text{Emb}(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}, \Sigma) /\!/ \text{O}(d)^k)_+ \wedge H\mathbb{Z}
\end{array}
$$

up to a zigzag of weak equivalences of maps of squares. As $(-)_+ \wedge H\mathbb{Z}$ commutes with taking homotopy orbits and left Kan extensions, we conclude that the square (27) depends up to natural weak equivalences only on the endomorphism $\phi_* \colon \text{Emb}_{\partial_0}(-, \Sigma)_+ \wedge H\mathbb{Z} \to \text{Emb}_{\partial_0}(-, \Sigma)_+ \wedge H\mathbb{Z}$ in $\mathsf{PSh}^{H\mathbb{Z}}(\mathsf{Disc}_{\leq k})$ and moreover, as homotopy orbits and homotopy left Kan extensions preserve weak equivalences, only on its image in $\mathsf{PSh}^{H\mathbb{Z}}(\mathsf{Disc}_{\leq k})^{\text{loc}}$. Taking vertical cofibres in (27) and homotopy groups, we conclude that the map

(28)
$$
\phi_* \colon H_*(\text{FM}_k(\Sigma), \partial_0 \, \text{FM}_k(\Sigma); \mathbb{Z}) \to H_*(\text{FM}_k(\Sigma), \partial_0 \, \text{FM}_k(\Sigma); \mathbb{Z})
$$

depends only on the image of ϕ under the map $\pi_0 \text{Diff}_\partial(\Sigma) \to \pi_0 T_k^{H\mathbb{Z}} \text{Emb}_{\partial_0}(\Sigma, \Sigma)$. In particular, if ϕ lies in the kernel $T\mathscr{I}_{\partial/2}^{H\mathbb{Z}}(k)$ of this map, then (28) is the identity. Using excision as in [19, Section 5.4.1] one see that the macroscopic location map $\mu \colon (\text{FM}_k(M), \partial_0 \, \text{FM}_k(M)) \to (M^k, \Delta_k \cup A_k)$ is a homology isomorphism, so ϕ induces the identity on $H_*(M^k, \Delta_k \cup A_k; \mathbb{Z})$. But the subgroup of mapping classes with this property is exactly $\mathscr{I}(k)$, by [21, Theorem A, Proposition 3.3]. $\qquad\square$

Remark 4.6 It might be interesting to study the various filtrations of the mapping class group obtained by replacing $H\mathbb{Z}$ in the definition of $T\mathscr{I}_{\partial/2}^{H\mathbb{Z}}(k)$ with HR for any ring R, such as $\mathbb{Q}$ or $\mathbb{F}_p$.

As long as R has characteristic 0, the resulting filtration is contained in the Johnson filtration. This follows from the proof for $\mathbb{Z}$ we gave above, together with the fact from [21, Proposition 3.3] that $H_*(\Sigma^k, \Delta_k \cup A_k; \mathbb{Z})$ is trivial if $* \neq k$ and free abelian for $* = k$.

4.4 The proof of Proposition 4.5

It will be convenient for us to work with an explicit model for the homotopy left Kan extension as a bar construction, which we recall next.

4.4.1 The enriched bar construction Given enriched space-valued functors F and G on a topologically enriched category C where F is contravariant and G is covariant, the *bar construction* $B_\bullet(G, \mathsf{C}, F)$ is the

semisimplicial space given by

$$[p] \mapsto \bigsqcup_{c_0,\dots,c_p} \left(G(c_0) \times \prod_{i=1}^{p} \mathsf{C}(c_{i-1}, c_i) \times F(c_p) \right)$$

where the coproduct is taking over ordered collections $c_0, \dots, c_p$ of objects in C and face maps are induced by the composition in C and the functoriality of F and G. We denote the geometric realisation of this semisimplicial space by omitting the $\bullet$–subscript. Since geometric realisations of levelwise weak equivalences of semisimplicial spaces are weak equivalences, the object $B(G, \mathsf{C}, F)$ is weakly homotopy invariant in triples (G, C, F), in the appropriate sense.

Given an enriched functor $\iota \colon \mathsf{C} \to \mathsf{D}$ and $d \in \mathsf{D}$, the space $B\big(\mathsf{D}(d, \iota(-)), \mathsf{C}, F\big)$ agrees, naturally in d, with the homotopy left Kan extension $\mathrm{hLan}_\iota F(d)$ (see eg [22, Example 9.2.11]; the cofibrancy conditions are not relevant for us as we consider the bar construction as a *semi*simplicial space and geometric realisations of *semi*simplicial spaces preserve weak equivalences). Moreover, if F extends to a functor on D, then there is a natural augmentation map

$$(29) \qquad\qquad B_\bullet\big(\mathsf{D}(d, \iota(-)), \mathsf{C}, F\big) \to F(d)$$

induced by composition and evaluation, which agrees upon geometric realisations with the canonical map $\mathrm{hLan}_\iota F(d) \to F(d)$ (or rather, it provides a model thereof).

In particular, using the notation introduced above, the left vertical map in the statement of Proposition 4.5 is given by the map induced by (29) and taking homotopy orbits

$$(30) \quad B\big(\mathrm{Emb}_{\partial_0}(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}, -), \mathrm{Disc}_{\partial_0, \leq k-1}, \mathrm{Emb}_{\partial_0}(-, M)\big) /\!\!/ \mathrm{O}(d)^k \xrightarrow{\ \epsilon\ } \mathrm{Emb}_{\partial_0}(\partial_0 \sqcup \mathbb{R}^d_{\underline{k}}, M) /\!\!/ \mathrm{O}(d)^k.$$

To compare (30) to the boundary inclusion $\partial_0 \mathrm{FM}_k(M) \subset \mathrm{FM}_k(M)$, we first show the following.

Lemma 4.7 *The map induced by the augmentation*

$$B\big(\partial_0 \mathrm{FM}_k(-), \mathrm{Disc}_{\partial_0, \leq k-1}, \mathrm{Emb}_{\partial_0}(-, M)\big) \to \partial_0 \mathrm{FM}_k(M)$$

is a weak equivalence.

Proof sketch The strategy is to show that this map is a Serre microfibration and has weakly contractible fibres, which implies the statement by a lemma of Weiss [29, Lemma 2.2]. This is a standard argument, so we will explain the idea somewhat informally and avoid spelling out lengthy but routine technical details that are similar to eg [16, Section 4].

To verify that the map is a Serre microfibration the task is to show that in a commutative diagram

$$
\begin{array}{ccc}
D^i \times \{0\} & \longrightarrow & B(\partial_0 \mathrm{FM}_k(-), \mathrm{Disc}_{\partial_0, \leq k-1}, \mathrm{Emb}_{\partial_0}(-, M)) \\
\big\downarrow & \nearrow & \big\downarrow \\
D^i \times [0, \varepsilon] \subset D^i \times [0, 1] & \longrightarrow & \partial_0 \mathrm{FM}_k(M)
\end{array}
$$
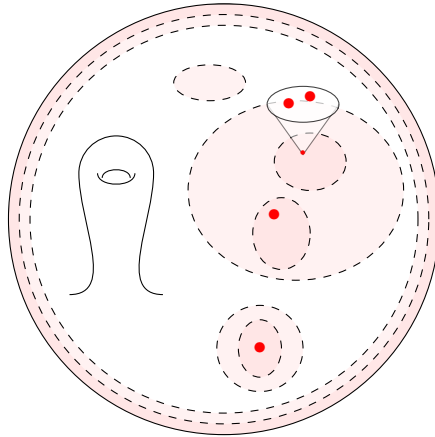
Figure 6: An element of $B(\partial_0 \, \mathrm{FM}_4(-), \mathrm{Disc}_{\partial_0, \leq 3}, \mathrm{Emb}_{\partial_0}(-, M))$ for $\partial_0 M = \partial M$, consisting of a configuration $x \in \partial_0 \, \mathrm{FM}_4(M)$ where two points are infinitesimally close, so that its macroscopic image $\mu(x)$ consists of three points, and two levels of discs and collars indicated by the orange and light-orange coloured regions. We suppressed the weights $(t_0, t_1) \in \Delta^1$.

whose solid arrows are given, there exists an $\varepsilon > 0$ and dashed lift. To see why this holds, it is helpful to think of the space $B(\partial_0 \, \mathrm{FM}_k(-), \mathrm{Disc}_{\partial_0, \leq k-1}, \mathrm{Emb}_{\partial_0}(-, M))$ as the subspace of

$$\partial_0 \, \mathrm{FM}_k(M) \times B(*, \mathrm{Disc}_{\partial_0, \leq k-1}, \mathrm{Emb}_{\partial_0}(-, M))$$

consisting of pairs $(x, [\vec{e}, \vec{t}])$ of element x in $\partial_0 \, \mathrm{FM}_k(M)$ and an equivalence class of a collection $\vec{e}$ of $p + 1$ levels of nested embedded discs in M with weight $\vec{t} \in \Delta^p$. The pair $[\vec{e}, \vec{t}]$ must have the property that the image $\mu(x)$ of x under the macroscopic location map is contained in the interior of the deepest level (see Figure 6 for an example) and the equivalence relation is that if a coordinate of $\vec{t} \in \Delta^p = \{(t_0, \ldots, t_p) \in [0, 1]^{p+1} \mid t_0 + \cdots + t_p = 1\}$ is 0 then we may forget it and the corresponding level of discs.

In these terms, the right vertical map in the diagram sends $(x, [\vec{e}, \vec{t}])$ to x. The map

$$D^i \to B(\partial_0 \, \mathrm{FM}_k(-), \mathrm{Disc}_{\partial_0, \leq k}, \mathrm{Emb}_{\partial_0}(-, M))$$

provides for each $s \in D^i$ a configuration $x(s) \in \partial_0 \, \mathrm{FM}_k(M)$ together with nested embedded discs and weights $[\vec{\epsilon}(s), \vec{t}(s)]$. The map $D^i \times [0, 1] \to \partial_0 \, \mathrm{FM}_k(M)$ defines a homotopy $x_t(s)$ with $t \in [0, 1]$ starting at $x(s)$. If t is small enough then this remains within the deepest level of the discs for $x(s, 0)$, and by compactness of D^i we find a single $\varepsilon > 0$ such that this is the case for all (s, t) with $t \leq \varepsilon$. The dashed lift is then given by sending (s, t) to $(x(s, t), [\vec{e}(s), \vec{t}(s)])$.

To see that the fibre over $x \in \partial_0 \, \mathrm{FM}_k(M)$ is weakly contractible, ie any map from S^i to the fibre extends over D^{i+1}, we observe that given an equivalence class $[\vec{e}, \vec{t}]$ represented by a family of nested embedded discs in M with weights, whose deepest level contains x, we find a smaller collection of $\leq (k-1)$ discs around points in the macroscopic image $\mu(x)$ of x and contained in the deepest level. By compactness

we can find a single such small collection which works for all images of $s \in S^i$. Adding this collection and transferring all weight to this collection provides an extension to D^{i+1}.   □

Proof of Proposition 4.5 For brevity, we abbreviate

$$\mathsf{D}_k := \mathsf{Disk}_{\partial_0, \leq k}, \quad E_M := \mathrm{Emb}_{\partial_0}(-, M), \quad E^{\partial_0 \sqcup \mathbb{R}_{\underline{k}}^d} := \mathrm{Emb}_{\partial_0}(\partial_0 \sqcup \mathbb{R}_{\underline{k}}^d, -),$$

$$\mathrm{FM}_k = \mathrm{FM}_k(-), \quad \partial_0 \mathrm{FM}_k = \partial_0 \mathrm{FM}_k(-).$$

We claim that the commutative diagram

$$B(E^{\partial_0 \sqcup \mathbb{R}_{\underline{k}}^d}, \mathsf{D}_{k-1}, E_M) /\!\!/ \mathrm{O}(d)^k \overset{①}{\to} B(\mathrm{FM}_k, \mathsf{D}_{k-1}, E_M) \overset{③}{\leftarrow} B(\partial_0 \mathrm{FM}_k, \mathsf{D}_{k-1}, E_M) \overset{④}{\to} \partial_0 \mathrm{FM}_k(M)$$

provides a zigzag as claimed. Here all vertical arrows are induced by the augmentation (29) or the inclusion $\partial_0 \mathrm{FM}_k \subset \mathrm{FM}_k$. Maps ① and ② are induced by the composition

$$(31) \qquad\qquad \mathrm{Emb}_{\partial_0}(\partial_0 \sqcup \mathbb{R}_{\underline{k}}^d, -) \to \mathrm{Emb}(\mathbb{R}_{\underline{k}}^d, -) \to \mathrm{Emb}(\underline{k}, -) \to \mathrm{FM}_k(-)$$

induced by restriction and inclusion, ③ is induced by inclusion, and ④ is another instance of (29). As the diagram is natural in M and the leftmost vertical map agrees with the left vertical map in the statement by the discussion around (30), it remains to show that ①–④ are weak equivalences.

The map ① factors as a composition

$$B(E^{\partial_0 \sqcup \mathbb{R}_{\underline{k}}^d}, \mathsf{D}_{k-1}, E_M) /\!\!/ \mathrm{O}(d)^k \to B(E^{\partial_0 \sqcup \mathbb{R}_{\underline{k}}^d} /\!\!/ \mathrm{O}(d)^k, \mathsf{D}_{k-1}, E_M) \to B(\mathrm{FM}_k, \mathsf{D}_{k-1}, E_M)$$

whose first map is a weak equivalence since left Kan extensions commute with homotopy orbits. To show that the second map in this composition (and also the map ②) is a weak equivalence, we argue that the composition (31) consists of weak equivalences upon applying $(-) /\!\!/ \mathrm{O}(d)^k$ to the first two spaces. For the first map this follows by shrinking the collar, for the second map it holds because the derivative $\mathrm{Emb}(\mathbb{R}_{\underline{k}}^d, N) \to \underline{k} \times \mathrm{Fr}(N)$ is a weak equivalence for any manifold N where $\mathrm{Fr}(N)$ is the frame bundle, and for the third map it is clear.

The map ③ is a weak equivalence because $\partial_0 \mathrm{FM}_k(-) \subset \mathrm{FM}_k(-)$ is a weak equivalence when evaluated on objects U of $\mathsf{D}_{\leq k-1}$. Indeed, if U consists of a collar and $l \leq k-1$ discs,

$$\mathrm{FM}_k(U) \cong \bigsqcup_{n_0 + \cdots + n_l = k} \mathrm{FM}_{n_0}(\partial_0 M \times [0,1)) \times \mathrm{FM}_{n_1}(\mathbb{R}^d) \times \cdots \times \mathrm{FM}_{n_l}(\mathbb{R}^d)$$

and $\partial_0 \mathrm{FM}_k(U)$ is the union of such terms where one FM_{n_i} is replaced by $\partial_0 \mathrm{FM}_{n_i}$. By the pigeonhole principle we have $n_0 \geq 1$ or $n_i \geq 2$ for some $1 \leq i \leq l$, so it suffices to observe that in these cases

$\partial_0 \, \mathrm{FM}_{n_0}(\partial M \times [0, 1)) \hookrightarrow \mathrm{FM}_{n_0}(M \times [0, 1))$ or $\partial_0 \, \mathrm{FM}_{n_i}(\mathbb{R}^d) \hookrightarrow \mathrm{FM}_{n_i}(\mathbb{R}^d)$ are inclusions of deformation retracts, either by modifying configurations such that one has a macroscopic location in $\partial_0 M \times \{0\} \subset \partial_0 M \times [0, 1)$ or such that all have macroscopic location at $\{0\} \in \mathbb{R}^d$. Finally, ④ is a weak equivalence by Lemma 4.7. □

References

[1] **P Boavida de Brito**, **M Weiss**, *Manifold calculus and homotopy sheaves*, Homology Homotopy Appl. 15 (2013) 361–383 MR Zbl

[2] **P Boavida de Brito**, **M Weiss**, *Spaces of smooth embeddings and configuration categories*, J. Topol. 11 (2018) 65–143 MR Zbl

[3] **S K Boldsen**, *Different versions of mapping class groups of surfaces*, preprint (2009) arXiv 0908.2221

[4] **J Cerf**, *Théorèmes de fibration des espaces de plongements: applications*, from "Séminaire Henri Cartan, 1962/1963", École Norm. Sup., Paris (1964) Exposé 8 MR Zbl

[5] **D R J Chillingworth**, *Winding numbers on surfaces, I*, Math. Ann. 196 (1972) 218–249 MR Zbl

[6] **D R J Chillingworth**, *Winding numbers on surfaces, II*, Math. Ann. 199 (1972) 131–153 MR Zbl

[7] **W G Dwyer**, **D M Kan**, *Function complexes in homotopical algebra*, Topology 19 (1980) 427–440 MR Zbl

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

[9] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Math. Ser. 49, Princeton Univ. Press (2012) MR Zbl

[10] **C D Feustel**, *Homotopic arcs are isotopic*, Proc. Amer. Math. Soc. 17 (1966) 891–896 MR Zbl

[11] **T G Goodwillie**, **J R Klein**, *Multiple disjunction for spaces of smooth embeddings*, J. Topol. 8 (2015) 651–674 MR Zbl

[12] **T G Goodwillie**, **M Weiss**, *Embeddings from the point of view of immersion theory, II*, Geom. Topol. 3 (1999) 103–118 MR Zbl

[13] **A Gramain**, *Le type d'homotopie du groupe des difféomorphismes d'une surface compacte*, Ann. Sci. École Norm. Sup. 6 (1973) 53–66 MR Zbl

[14] **A Hatcher**, *A short exposition of the Madsen–Weiss theorem*, preprint (2011) arXiv 1103.5223

[15] **G Horel**, *Profinite completion of operads and the Grothendieck–Teichmüller group*, Adv. Math. 321 (2017) 326–390 MR Zbl

[16] **I Klang**, **A Kupers**, **J Miller**, *The May–Milgram filtration and $\mathcal{E}_k$–cells*, Algebr. Geom. Topol. 21 (2021) 105–136 MR Zbl

[17] **B Knudsen**, **A Kupers**, *Embedding calculus and smooth structures*, Geom. Topol. 28 (2024) 353–392 MR Zbl

[18] **M Krannich**, **A Kupers**, *The $\mathcal{D}isc$–structure space*, preprint (2022) arXiv 2205.01755

[19] **A Kupers**, **O Randal-Williams**, *The cohomology of Torelli groups is algebraic*, Forum Math. Sigma 8 (2020) art. id. e64 MR Zbl

[20] **M A Mandell**, **J P May**, **S Schwede**, **B Shipley**, *Model categories of diagram spectra*, Proc. Lond. Math. Soc. 82 (2001) 441–512 MR Zbl

[21] **T Moriyama**, *The mapping class group action on the homology of the configuration spaces of surfaces*, J. Lond. Math. Soc. 76 (2007) 451–466 MR Zbl

[22] **E Riehl**, *Categorical homotopy theory*, New Math. Monogr. 24, Cambridge Univ. Press (2014) MR Zbl

[23] **D P Sinha**, *Manifold-theoretic compactifications of configuration spaces*, Selecta Math. 10 (2004) 391–428 MR Zbl

[24] **S Smale**, *Diffeomorphisms of the 2–sphere*, Proc. Amer. Math. Soc. 10 (1959) 621–626 MR Zbl

[25] **R Trapp**, *A linear representation of the mapping class group $\mathbb{M}$ and the theory of winding numbers*, Topology Appl. 43 (1992) 47–64 MR Zbl

[26] **V Turchin**, *Context-free manifold calculus and the Fulton–MacPherson operad*, Algebr. Geom. Topol. 13 (2013) 1243–1271 MR Zbl

[27] **M Weiss**, *Embeddings from the point of view of immersion theory, I*, Geom. Topol. 3 (1999) 67–101 MR Zbl Correction in 15 (2011) 407–409

[28] **M S Weiss**, *Homology of spaces of smooth embeddings*, Q. J. Math. 55 (2004) 499–504 MR Zbl

[29] **M Weiss**, *What does the classifying space of a category classify?*, Homology Homotopy Appl. 7 (2005) 185–195 MR Zbl

Department of Mathematics, Karlsruhe Institute of Technology
Karlsruhe, Germany

Department of Mathematics, University of Toronto
Toronto, ON, Canada

krannich@kit.edu, a.kupers@utoronto.ca

Vietoris–Rips persistent homology, injective metric spaces, and the filling radius

SUNHYUK LIM
FACUNDO MÉMOLI
OSMAN BERAT OKUTAN

In the applied algebraic topology community, the persistent homology induced by the Vietoris–Rips simplicial filtration is a standard method for capturing topological information from metric spaces. We consider a different, more geometric way of generating persistent homology of metric spaces which arises by first embedding a given metric space into a larger space and then considering thickenings of the original space inside this ambient metric space. In the course of doing this, we construct an appropriate category for studying this notion of persistent homology and show that, in a category-theoretic sense, the standard persistent homology of the Vietoris–Rips filtration is isomorphic to our geometric persistent homology provided that the ambient metric space satisfies a property called injectivity.

As an application of this isomorphism result, we are able to precisely characterize the type of intervals that appear in the persistence barcodes of the Vietoris–Rips filtration of any compact metric space and also to give succinct proofs of the characterization of the persistent homology of products and metric gluings of metric spaces. Our results also permit proving several bounds on the length of intervals in the Vietoris–Rips barcode by other metric invariants, for example the notion of spread introduced by M Katz.

As another application, we connect this geometric persistent homology to the notion of filling radius of manifolds introduced by Gromov and show some consequences related to the homotopy type of the Vietoris–Rips complexes of spheres, which follow from work of Katz, and characterization (rigidity) results for spheres in terms of their Vietoris–Rips persistence barcodes, which follow from work of F Wilhelm.

Finally, we establish a sharp version of Hausmann's theorem for spheres which may be of independent interest.

53C23, 55N31

1 Introduction

The simplicial complex nowadays referred to as the Vietoris–Rips complex was originally introduced by Leopold Vietoris in the early 1900s in order to build a homology theory for metric spaces [80]. Later, Eliyahu Rips and Mikhail Gromov [47] both utilized the Vietoris–Rips complex in their study of hyperbolic groups.

Given a metric space (X, d_X) and $r > 0$, the r–Vietoris–Rips complex $VR_r(X)$ has X as its vertex set, and simplices are all nonempty finite subsets of X whose diameter is strictly less than r. In [50], Hausmann showed that the Vietoris–Rips complex can be used to recover the homotopy type of a Riemannian manifold M. More precisely, he introduced a quantity $r(M)$ (a certain variant of the injectivity radius), and proved that $VR_r(M)$ is homotopy equivalent to M for any $r \in (0, r(M))$.

Since $VR_r(X) \subseteq VR_s(X)$ for all $0 < r \leq s$, this construction then naturally induces the so-called Vietoris–Rips simplicial filtration of X, denoted by $VR_*(X) = (VR_r(X))_{r>0}$. By applying the simplicial homology functor (with coefficients in a given field) one obtains a *persistence module*: a directed system $V_* = (V_r \xrightarrow{v_{rs}} V_s)_{r \leq s}$ of vector spaces and linear maps (induced by the simplicial inclusions). The persistent module obtained from $VR_*(X)$ is referred to as the Vietoris–Rips persistent homology of X.

The notion of *persistent homology* arose from work by Ferri, Frosini, Landi, Verri and Uras, [39; 40; 41; 79], Robins [73], and Delfinado, Edelsbrunner, Letscher and Zomorodian [27; 36]. After that, considering the persistent homology of the simplicial filtration induced from Vietoris–Rips complexes was a natural next step. For example, Carlsson and de Silva [76] applied Vietoris–Rips persistent homology to topological estimation from point cloud data, and Ghrist and de Silva applied it to sensor networks [77]. Its efficient computation has been addressed by Bauer in [11]. A more detailed historical survey and review of general ideas related to persistent homology can be found in Carlsson [16] and Edelsbrunner and Harer [34; 35].

The persistent homology of the Vietoris–Rips filtration of a metric space provides a functorial way[1] of assigning a persistence module to a metric space. Persistence modules are usually represented, up to

---

[1]Where for metric spaces X and Y morphisms are given by 1–Lipschitz maps $\phi: X \to Y$, and for persistence modules V_* and W_* morphisms are systems of linear maps $v_* = (v_r: V_r \to W_r)_{r>0}$ making all squares commute.

isomorphism, as *barcodes*: multisets of intervals each representing the lifetime of a homological feature. In this paper, the barcodes are associated to Vietoris–Rips filtrations, and these barcodes will be denoted by $\mathrm{barc}_*^{\mathrm{VR}}(\cdot)$. In the areas of topological data analysis (TDA) and computational topology, this type of persistent homology is a widely used tool for capturing topological properties of a dataset [11; 76; 77].

Despite its widespread use in applications, little is known in terms of relationships between Vietoris–Rips barcodes and other metric invariants. For instance, whereas it is obvious that the right endpoint of any interval I in $\mathrm{barc}_*^{\mathrm{VR}}(X)$ must be bounded above by the diameter of X, there has been little progress in relating the length of bars to other invariants such as volume (or Hausdorff measure) or curvature (whenever defined).

Contributions One main contribution of this paper is establishing a precise relationship (ie a filtered homotopy equivalence) between the Vietoris–Rips simplicial filtration of a metric space and a more geometric (or extrinsic) way of assigning a persistence module to a metric space, which consists of first isometrically embedding it into a larger space and then considering the persistent homology of the filtration obtained by considering the resulting system of nested neighborhoods of the original space inside this ambient space. These neighborhoods, being also metric (and thus topological) spaces, permit giving a short proof of the Künneth formula for Vietoris–Rips persistent homology.

A particularly nice ambient space inside which one can isometrically embed any given compact metric space (X, d_X) is $L^\infty(X)$; the Banach space consisting of all the bounded real-valued functions on X, together with the ℓ^∞–norm. The embedding is given by $X \ni x \mapsto d_X(x, \cdot)$: it is indeed immediate that this embedding is isometric since $\|d_X(x, \cdot) - d_X(x', \cdot)\|_\infty = d_X(x, x')$ for all $x, x' \in X$. This is usually called the *Kuratowski* isometric embedding of X.

That the Vietoris–Rips filtration of a *finite* metric space produces persistence modules isomorphic to the sublevel set filtration of the distance function

$$\delta_X : L^\infty(X) \to \mathbb{R}_{\geq 0}, \quad L^\infty(X) \ni f \mapsto \inf_{x \in X} \|d_X(x, \cdot) - f\|_\infty,$$

was already used by Chazal, Cohen-Steiner, Guibas, Mémoli and Oudot [19] in order to establish the Gromov–Hausdorff stability of Vietoris–Rips persistence of finite metric spaces.

In this paper we significantly generalize this point of view by proving an isomorphism theorem between the Vietoris–Rips filtration of *any* compact metric space X and its *Kuratowski filtration*,

$$\big(\delta_X^{-1}([0, r))\big)_{r > 0},$$

a fact which immediately implies that their persistent homologies are isomorphic.

We do so by constructing a filtered homotopy equivalence between the Vietoris–Rips filtration and the sublevel set filtration induced by δ_X. Furthermore, we prove that $L^\infty(X)$ above can be replaced with *any injective* (*or equivalently*, *hyperconvex*) *metric space* — see Dress, Huber, Koolen, Moulton and Spillner [31] and Lang [60] — admitting an isometric embedding of X:

Theorem 4.1 (isomorphism theorem) *Let $\eta\colon \mathrm{Met} \to \mathrm{PMet}$ be a metric homotopy pairing (for example, the Kuratowski functor). Then $\mathrm{B}_* \circ \eta\colon \mathrm{Met} \to \mathrm{hTop}_*$ is naturally isomorphic to VR_{2*}.*

Above, Met is the category of compact metric spaces with 1–Lipschitz maps, PMet is the category of metric pairs (X, E) where $X \hookrightarrow E$ isometrically, E is an injective metric space, a metric homotopy pairing is any right adjoint to the forgetful functor (eg the Kuratowski embedding), and B_* is the functor sending a pair (X, E) to the filtration $(B_r(X, E))_{r>0}$; see Sections 3 and 4.

A certain well known construction which involves the isometric embedding $X \hookrightarrow L^\infty(X)$ is that of the *filling radius* of a Riemannian manifold [46] defined by Gromov in the early 1980s. In that construction, given an n–dimensional Riemannian manifold M, one studies for each $r > 0$ the inclusion

$$\iota_r\colon M \hookrightarrow \delta_M^{-1}([0, r))$$

and seeks the infimal $r > 0$ such that the map induced by ι_r at degree n homology level annihilates the fundamental class $[M]$ of M. This infimal value defines $\mathrm{FillRad}(M)$, the filling radius of M. In this paper, we will consider a version of the filling radius associated to the fundamental class with coefficients in a given field $\mathbb{F}$ which will be denoted by $\mathrm{FillRad}(M; \mathbb{F})$.

Via our isomorphism theorem we are able prove that there always exists a bar in the barcode of a manifold whose length is exactly twice its filling radius:

Proposition 9.28 *Let M be a closed connected n–dimensional Riemannian manifold. Then*

$$(0, 2\,\mathrm{FillRad}(M; \mathbb{F})] \in \mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F}),$$

*where $\mathbb{F}$ is an arbitrary field if M is orientable, and $\mathbb{F} = \mathbb{Z}_2$ if M is nonorientable. Moreover, this is the **unique** interval in $\mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F})$ starting at 0, and $\mathrm{FillRad}(M; \mathbb{F}) \leq \mathrm{FillRad}(M)$ whenever M is orientable.*

As a step in his proof of the celebrated systolic inequality, Gromov proved in [46] that the filling radius satisfies $\mathrm{FillRad}(M) \leq c_n (\mathrm{vol}(M))^{1/n}$ for any n–dimensional complete manifold M (where c_n is a universal constant, and Nabutovsky recently proved that c_n can be improved to $\frac{n}{2}$ [69, Theorem 1.2]). This immediately yields a relationship between $\mathrm{barc}_*^{\mathrm{VR}}(M)$ and the volume of M. The fact that the filling radius has already been connected to a number of other metric invariants also permits importing these results to the setting of Vietoris–Rips barcodes (see Section 9.3). This in turn permits relating $\mathrm{barc}_*^{\mathrm{VR}}(M)$ with other metric invariants of M, a research thread which has remained mostly unexplored. See Proposition 9.46 for a certain generalization of Proposition 9.28 to ANR spaces.

In a series of papers [54; 55; 56; 57], M Katz studied both the problem of computing the filling radius of spheres (endowed with the geodesic distance) and complex projective spaces, and the problem of understanding the change in homotopy type of $\delta_X^{-1}([0, r))$ when $X \in \{\mathbb{S}^1, \mathbb{S}^2\}$ as r increases.

Of central interest in topological data analysis has been the question of providing a complete characterization of the Vietoris–Rips persistence barcodes of spheres of different dimensions. Despite the existence

of a complete answer to the question for the case of $\mathbb{S}^1$ due to Adamaszek and Adams [1], relatively little is known for higher-dimensional spheres. In [2], Adamaszek, Adams and Frick consider a variant of the Vietoris–Rips filtration, which they call Vietoris–Rips metric thickening. The authors are able to obtain information about the successive homotopy types of this filtration on spheres of different dimension — see [2, Section 5] — for a certain range of values of the scale parameter.

The authors of [2] conjecture that the open Vietoris–Rips filtration (which is the one considered in the present paper) is filtered homotopy equivalent to their open Vietoris–Rips metric thickening filtration (as a consequence their persistent homologies are isomorphic). This isomorphism was conjectured in [2, Conjecture 6.12] which was recently settled in [7, Corollary 5.10].

Our isomorphism theorem (Theorem 4.1) permits applying Katz's results in order to provide partial answers to the questions mentioned above and also to elucidate other properties of the standard open Vietoris–Rips filtration and its associated persistence barcodes $\mathrm{barc}_*^{\mathrm{VR}}(\,\cdot\,)$. In addition to these results derived from our isomorphism theorem, in Section A.4, we refine certain key lemmas used in the original proof of Hausmann's theorem [50] and establish the homotopy equivalence between $\mathrm{VR}_r(\mathbb{S}^n)$ and $\mathbb{S}^n$ for any $r \in \big(0, \arccos\left(-1/(n+1)\right)\big]$:

Theorem 7.1 *For any $n \in \mathbb{Z}_{>0}$, we have $\mathrm{VR}_r(\mathbb{S}^n) \simeq \mathbb{S}^n$ for any $r \in \big(0, \arccos(-1/(n+1))\big]$.*

Note that this is indeed an improvement since, for spheres, Hausmann's quantity satisfies

$$r(\mathbb{S}^n) = \tfrac{\pi}{2} < \arccos\left(-\frac{1}{n+1}\right).$$

This improvement is obtained with the aid of a refined version of Jung's theorem (see Theorem A.8) which we also establish. Theorem 7.1 also improves upon [54, Remark, page 508]; see the discussion in Section 7.1.

In the direction of characterizing the Vietoris–Rips barcodes of spheres, we are able to provide a complete characterization of the homotopy types of the Vietoris–Rips complexes of round spheres $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ endowed with the (restriction of the) ℓ^∞–metric, which we denote by $\mathbb{S}^{n-1}_\infty$. Two critical observations are that

(1) the r–thickening of $\mathbb{S}^{n-1}_\infty$ inside of $\mathbb{R}^n_\infty$ ($\mathbb{R}^n$ equipped with the ℓ^∞–metric) is homotopy equivalent to the r–thickening of $\mathbb{S}^{n-1}_\infty$ inside of $\mathbb{D}^n_\infty$ (n–dimensional unit ball with ℓ^∞–metric), and

(2) it is easier to find the precise shape of the latter.

Theorem 7.19 *For any $n \in \mathbb{Z}_{>0}$ and $r > 0$,*

$$B_r(\mathbb{S}^{n-1}_\infty, \mathbb{R}^n_\infty) \simeq B_r(\mathbb{S}^{n-1}_\infty, \mathbb{D}^n_\infty) = \mathbb{D}^n_\infty \backslash V_{n,r},$$

where

$$V_{n,r} := \bigcap_{(p_1,\ldots,p_n)\in\{r,-r\}^n} \left\{ (x_1,\ldots,x_n) \in \mathbb{R}^n \,\Big|\, \sum_{i=1}^n (x_i - p_i)^2 \leq 1 \right\}.$$
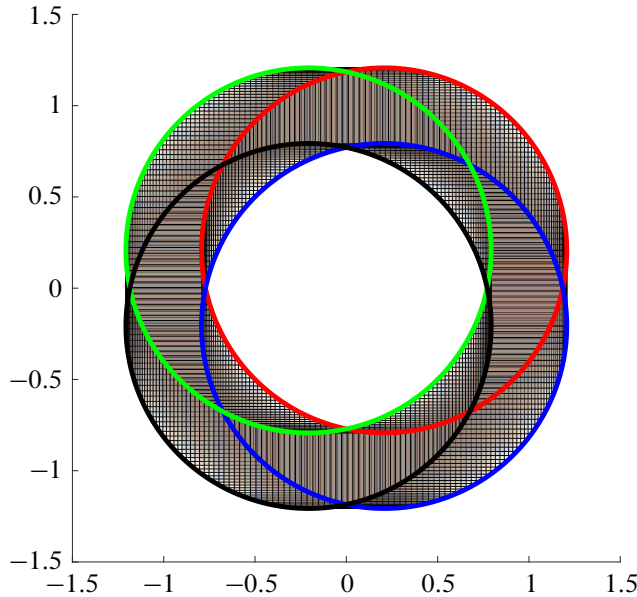
Figure 1: $B_r(\mathbb{S}^1_\infty, \mathbb{D}^2_\infty) = \mathbb{D}^2_\infty \setminus V_{2,r}$ in the plane $\mathbb{R}^2_\infty$. The set $V_{2,r}$ is given by the intersection of the four closed disks shown in the figure. See Theorem 7.19.

In particular, for $r > 1/\sqrt{n}$ we have $V_{n,r} = \varnothing$, so $B_r(\mathbb{S}^{n-1}_\infty, \mathbb{D}^n_\infty) = \mathbb{D}^n_\infty$. As a result, $B_r(\mathbb{S}^{n-1}_\infty, \mathbb{R}^n_\infty)$ is homotopy equivalent to $\mathbb{S}^{n-1}$ for $r \in (0, 1/\sqrt{n}]$ and contractible for $r > 1/\sqrt{n}$ (see Figure 1 for an illustration of the case $n = 2$).

From a different perspective, by appealing to our isomorphism theorem, it is also possible to apply certain results from quantitative topology to the problem of characterization of metric spaces by their Vietoris–Rips persistence barcodes. In applied algebraic topology, a general question of interest is:

Question 1 Assume X and Y are compact metric spaces such that $\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F}) = \mathrm{barc}_k^{\mathrm{VR}}(Y; \mathbb{F})$ for all $k \in \mathbb{Z}_{\geq 0}$. Then how similar are X and Y (in a suitable sense)?

It follows from work by Wilhelm [83] and Yokota [84] on rigidity properties of spheres via the filling radius, and the isomorphism theorem (Theorem 4.1), that any n–dimensional Alexandrov space without boundary and sectional curvature bounded below by 1 such that its Vietoris–Rips persistence barcode agrees with that of $\mathbb{S}^n$ must be *isometric* to $\mathbb{S}^n$. This provides some new information about the inverse problem for persistent homology; see Curry [26] and Gameiro, Hiraoka and Obayashi [43]. More precisely, and for example, we obtain the corollary below, where for an n–dimensional manifold M, $I_{n,\mathbb{F}}^M$ denotes the persistence interval in $\mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F})$ induced by the fundamental class of M (see Proposition 9.28):

Corollary 9.51 ($\mathrm{barc}_*^{\mathrm{VR}}$ rigidity for spheres) *For any closed connected n–dimensional Riemannian manifold M with sectional curvature $\mathrm{K}_M \geq 1$,*

(1) $I_{n,\mathbb{F}}^M \subseteq I_n^{\mathbb{S}^n}$;
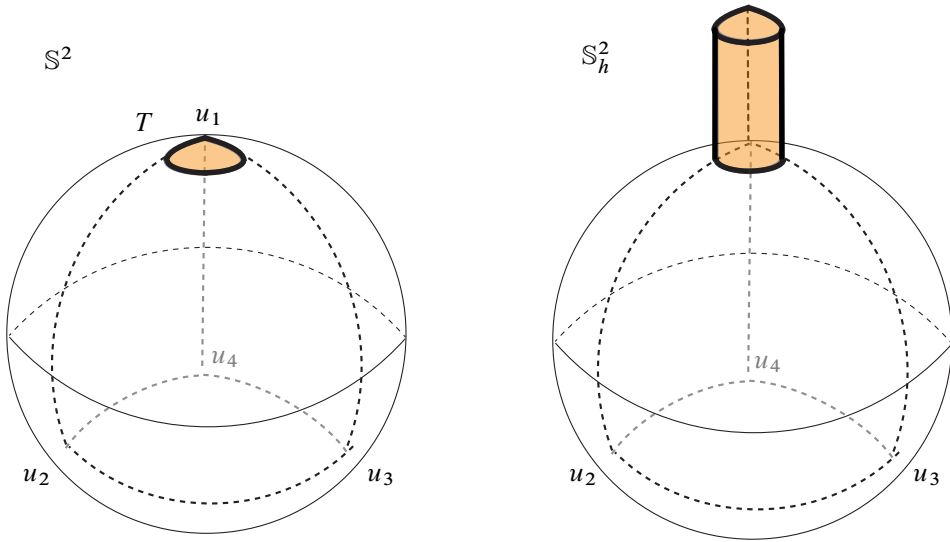
Figure 2: The construction of the one parameter family of surfaces $\mathbb{S}_h^2$ with the same filling radius as $\mathbb{S}^2$. The points u_1, u_2, u_3 and u_4 are vertices of a regular geodesic tetrahedron, and T is a small geodesic triangle, which is used to form a cylinder of height h (left figure). See Example 9.54 for details.

(2) if $I_{n,\mathbb{F}}^M = I_n^{\mathbb{S}^n}$ then M is isometric to $\mathbb{S}^n$;

(3) there exists $\epsilon_n > 0$ such that if $\mathrm{length}(I_n^{\mathbb{S}^n}) - \epsilon_n < \mathrm{length}(I_{n,\mathbb{F}}^M)$, then M is diffeomorphic to $\mathbb{S}^n$;

(4) if $\mathrm{length}(I_{n,\mathbb{F}}^M) > \frac{\pi}{3}$, then M is a twisted n–sphere (and, in particular, homotopy equivalent to the n–sphere).

The lower bound on sectional curvature is crucial — in Example 9.54 we construct a one parameter family of deformations of the sphere $\mathbb{S}^2$ with constant filling radius (see Figure 2).

See Propositions 9.56 and 9.57 for additional related results, and see Question 3 for a relaxation of Question 1.

Lastly, let us address a variant of Question 1 concerning the case when $\mathrm{barc}_k^{\mathrm{VR}}(X;\mathbb{F})$ and $\mathrm{barc}_k^{\mathrm{VR}}(Y;\mathbb{F})$ are possibly different. Recall that there is the bottleneck distance d_{B} measuring the dissimilarity between two barcodes (see Definition 2.12). One of the fundamental results of topological data analysis is the following stability theorem (see Theorems 2.13 and 2.14): for any field $\mathbb{F}$,

$$(1) \qquad \ell^{\mathrm{VR}}(X,Y) := \tfrac{1}{2} \sup_k d_{\mathrm{B}}(\mathrm{barc}_k^{\mathrm{VR}}(X;\mathbb{F}), \mathrm{barc}_k^{\mathrm{VR}}(Y;\mathbb{F})) \leq d_{\mathrm{GH}}(X,Y).$$

Therefore, in order to understand how strong the Vietoris–Rips barcode is as a geometric invariant, it is natural to ask the following question:

Question 2(i) How good is $\ell^{\mathrm{VR}}(X,Y)$ as an estimator of $d_{\mathrm{GH}}(X,Y)$?

For example, one might ask whether the inequality (1) is tight or not. What we know is that this is indeed not tight when X and Y are spheres of different dimension since, in Corollary 9.39, we show that $\ell^{\mathrm{VR}}(\mathbb{S}^m, \mathbb{S}^n) = \frac{1}{4} \arccos(-1/(m+1))$ for any $0 < m < n$. However, in [63, Theorem B] it is proved that $\frac{1}{2} \arccos(-1/(m+1))$ (ie twice $\ell^{\mathrm{VR}}(\mathbb{S}^m, \mathbb{S}^n)$) lower bounds $d_{\mathrm{GH}}(\mathbb{S}^m, \mathbb{S}^n)$ for any $0 < m < n$ and that this bound is tight.

Now, let us ask the following question:

Question 2(ii) For what type of spaces X and Y does inequality (1) become tight?

Or, one might ask the following question too:

Question 2(iii) For what type of spaces X and Y do we have the reverse stability inequality

$$d_{\mathrm{GH}}(X, Y) \leq C \cdot \ell^{\mathrm{VR}}(X, Y)$$

for some $C > 0$?

Note that the reverse stability inequality mentioned in Question 2(iii) cannot hold in general. For example, if we let $X = \mathbb{S}^1$ and Y be $\mathbb{S}^1$ attached with disjoint trees of arbitrary length (regarded as a geodesic metric space), then we can prove $\ell^{\mathrm{VR}}(X, Y) = 0$ whereas $d_{\mathrm{GH}}(X, Y)$ can be arbitrarily large (depending on the length of the attached trees). See Figure 10 and the beginning of Section 9.4 for a more detailed explanation.

The authors hope that this paper can help bridge between the applied algebraic topology and the quantitative topology communities.

Organization In Section 2, we provide some necessary definitions and results about Vietoris–Rips filtration, persistence, and injective metric spaces.

In Section 3, we construct a category of metric pairs. This category will be the natural setting for our extrinsic persistent homology. Although being functorial is trivial in the case of Vietoris–Rips persistence, the type of functoriality which one is supposed to expect in the case of metric embeddings is a priori not obvious. We address this question in Section 3 by introducing a suitable category structure.

In Section 4, we show that the Vietoris–Rips filtration can be (categorically) seen as a special case of persistent homology obtained through metric embeddings via the isomorphism theorem (Theorem 4.1). In this section, we also we also establish the stability of the filtration obtained via metric embeddings.

Sections 5–9 provide applications of our isomorphism theorem to different questions.

In Section 5, we prove that any interval in persistence barcode for open Vietoris–Rips filtration must have open left endpoint and closed right endpoint.

In Section 6, we obtain new proofs of formulas about the Vietoris–Rips persistence of metric products and metric gluings of metric spaces.

In Section 7, we prove a number of results concerning the homotopy types of Vietoris–Rips filtrations of spheres and complex projective spaces. Also, we fully compute the homotopy types of Vietoris–Rips filtration of spheres with ℓ^∞–norm.

In Section 8, we reprove Rips and Gromov's result about the contractibility of the Vietoris–Rips complex of hyperbolic geodesic metric spaces, by using our method consisting of isometric embeddings into injective metric spaces. As a result, we will be able to bound the length of intervals in the Vietoris–Rips persistence barcode by the hyperbolicity of the underlying space.

In Section 9, we give some applications of our ideas to the filling radius of Riemannian manifolds and also study consequences related to the characterization of spheres by their persistence barcodes and some generalizations and novel stability properties of the filling radius.

The appendix contains relegated proofs and some background material.

2 Background

In this section we cover the background needed for proving our main results. We alert readers that, in this paper, the same notation can mean either a simplicial complex itself or its geometric realization, interchangeably. The precise meaning will be made clear in each context.

2.1 Vietoris–Rips filtration and persistence

References for the definitions and results in this subsection are [12; 61].

Definition 2.1 (Vietoris–Rips filtration) Let X be a metric space and $r > 0$. The (*open*) *Vietoris–Rips complex* $\mathrm{VR}_r(X)$ of X is the simplicial complex whose vertices are the points of X and whose simplices are the finite subsets of X with diameter strictly less then r. Note that if $r \leq s$, then $\mathrm{VR}_r(X)$ is contained in $\mathrm{VR}_s(X)$. Hence, the family $\mathrm{VR}_*(X)$ is a filtration, called the *open Vietoris–Rips filtration* of X.

The (geometric realization of) a Vietoris–Rips filtration is a special case of the following more general notion:

Definition 2.2 (persistence family) A *persistence family* is a collection $(U_r, f_{r,s})_{r \leq s \in T}$, where T is a nonempty subset of $\mathbb{R}$ such that, for each $r \leq s \leq t \in T$, U_r is a topological space, $f_{r,s} \colon U_r \to U_s$ is a continuous map, $f_{r,r} = \mathrm{id}_{U_r}$ and $f_{s,t} \circ f_{r,s} = f_{r,t}$.

Given two persistence families $(U_*, f_{*,*})$ and $(V_*, g_{*,*})$ indexed by the same $T \subseteq \mathbb{R}$, a morphism from the first one to the second is a collection $(\phi_r)_{r \in T}$ such that for each $r \leq s$, ϕ_r is a homotopy class of maps $U_r \to V_r$, and $\phi_s \circ f_{r,s}$ is homotopy equivalent to $g_{r,s} \circ \phi_r$.

Definition 2.3 (persistence module) A persistence module $V_* = (V_r, v_{r,s})_{r \leq s \in T}$ over $T \subseteq \mathbb{R}$ is a family of $\mathbb{F}$–vector spaces V_r for some field $\mathbb{F}$ with morphisms $v_{r,s} \colon V_r \to V_s$ for each $r \leq s$ such that

- $v_{r,r} = \mathrm{id}_{V_r}$,
- $v_{s,t} \circ v_{r,s} = v_{r,t}$ for each $r \leq s \leq t$.

In other words, a persistence module is a functor from the poset $(T, \leq)$ to the category of vector spaces. The morphisms $v_{*,*}$ are referred to as the structure maps of V_*.

By 0_* we will denote the zero persistence module.

For any $k \geq 0$, applying the degree k homology functor (with coefficients in a field $\mathbb{F}$) to a persistence family $(U_r, f_{r,s})_{r \leq s \in T}$ produces the persistence module $\mathrm{H}_k(U_*; \mathbb{F})$ where the morphisms are those induced by $(f_{r,s})_{r \leq s}$.

Following the extant literature, we will use the term *persistent homology* of a persistence family (ie a filtration) to refer to the persistence module obtained upon applying the homology functor to this family.

In particular, one can apply the homology functor to the Vietoris–Rips filtration of a metric space X. This induces a persistence module (with $T = \mathbb{R}_{>0}$) where the morphisms are those induced by inclusions. As a persistence module, it is denoted by $\mathrm{PH}_k(\mathrm{VR}_*(X); \mathbb{F})$ and referred to as the *Vietoris–Rips persistent homology* of X.

Definition 2.4 (interval persistence module [17]) Given an interval I in $T \subseteq \mathbb{R}$ (ie if $r \leq s \leq t$ and $r, t \in I$, then $s \in I$) and a field $\mathbb{F}$, the persistence module $\mathbb{F}_*[I]$ over T is defined as follows: the vector space at r is $\mathbb{F}$ if r is in I and zero otherwise; given $r \leq s$, the morphism corresponding to (r, s) is the identity if r and s are in I and zero otherwise.

Definition 2.5 (barcode) For a given persistence module V_*, if there is a multiset of intervals $(I_\lambda)_{\lambda \in \Lambda}$ such that V_* is isomorphic to $\bigoplus_{\lambda \in \Lambda} \mathbb{F}_*[I_\lambda]$, then that multiset is denoted by $\mathrm{barc}(V_*)$ and referred to as a (*persistence*) *barcode* associated to the persistence module V_* (see below). Modules for which there exist such a multiset of intervals are said to be *interval decomposable*.

By Azumaya's theorem [10], persistence barcodes, whenever they exist, are unique: any two persistence barcodes associated to a given V_* must agree (up to reordering). The most important existence result for persistence barcodes is Crawley-Boevey's theorem [25] which guarantees the existence of a persistence barcode associated to $V_* = (V_r, v_{r,s})$ if V_* is pointwise finite-dimensional (ie $\dim(V_r) < \infty$ for all r). However, for many natural persistence modules (eg Vietoris–Rips persistent homology of a nonfinite metric

space X), it is not straightforward to verify the pointwise finite-dimensionality condition. Nevertheless, in Theorem 2.9, we are able to establish that, if X is totally bounded, then its Vietoris–Rips persistent homology has a (unique) persistence barcode. This is achieved without invoking Crawley-Boevey's theorem and instead through combining our main (isomorphism) theorem (see Theorem 4.1) with a recent result by Schmahl [74, Theorem 1.2]. The proof of Theorem 2.9 can be found in the extended (arXiv) version of this paper [62, Section 5]. The totally boundedness condition is required in the theorem in order to guarantee the following notion of regularity:

Definition 2.6 (q–tame persistence module) A persistence module $V_* = (V_r, v_{r,s})_{r \leq s \in T}$ is said to be q–*tame* if $\mathrm{rank}(v_{r,s}) < \infty$ whenever $r < s$.

Remark 2.7 The notions of interval decomposability and q–tameness are not equivalent. Indeed:

(1) [22, Remark 2.9] There exist q–tame modules which are not interval decomposable.

(2) [22, Example 3.30] There exist interval decomposable modules which are not q–tame.

Interval decomposability and q–tameness are however related through a certain notion of weak isomorphism; see [20].

Remark 2.8 In [23, Proposition 5.1], it is proved that if X is a totally bounded metric space, then $\mathrm{PH}_k(\mathrm{VR}_*(X); \mathbb{F})$ is q–tame for any nonnegative integer $k \geq 0$ and any field $\mathbb{F}$.

Theorem 2.9 [62, Section 5] *If X is a totally bounded metric space, then there is a (unique) persistence barcode associated to* $\mathrm{PH}_k(\mathrm{VR}_*(X); \mathbb{F})$.

If X is a totally bounded metric space, then we denote the barcode corresponding to $\mathrm{PH}_k(\mathrm{VR}_*(X); \mathbb{F})$ by $\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$.

From now on, unless specified otherwise, we will always assume that $T = \mathbb{R}$. For a given metric space and integer $k \geq 0$, we will occasionally view $V_* = \mathrm{PH}_k(\mathrm{VR}_*(X); \mathbb{F})$ as a persistence module defined over the whole real line $\mathbb{R}$ by trivially extending it to the left of $0 \in \mathbb{R}$; that is, we set $V_t = 0$ for $t \leq 0$.

We now recall a notion of distance between persistence modules.

Definition 2.10 (interleaving distance) Two persistence modules V_* and W_* are said to be δ–interleaved for some $\delta \geq 0$ if there are natural transformations $f : V_* \to W_{*+\delta}$ and $g : W_* \to V_{*+\delta}$ such that $f \circ g$ and $g \circ f$ are equal to the structure maps $W_* \to W_{*+2\delta}$ and $V_* \to V_{*+2\delta}$, respectively. The interleaving distance between V_* and W_* is defined as

$$d_{\mathrm{I}}(V_*, W_*) := \inf\{\delta \geq 0 \mid V_* \text{ and } W_* \text{ are } \delta\text{–interleaved}\}.$$

It is known [12] that d_{I} is an extended pseudometric on the collection of all persistence modules.

Example 2.11　Consider 0_*, the zero persistence module. Then for any finite-dimensional V_* one has

$$d_{\mathrm{I}}(V_*, 0_*) = \tfrac{1}{2}\sup\{\mathrm{length}(I) \mid I \in \mathrm{barc}(V_*)\}.$$

Definition 2.12　(bottleneck distance)　Let M and M' be two possibly empty multisets of intervals. A subset $P \subseteq M \times M'$ is said to be a partial matching between M and M' if it satisfies that

- every interval $I \in M$ is matched with at most one interval of M', ie there is at most one interval $I' \in M'$ such that $(I, I') \in P$;
- every interval $I' \in M'$ is matched with at most one interval of M, ie there is at most one interval $I \in M$ such that $(I, I') \in P$.

The bottleneck distance between M and M' is defined as

$$d_{\mathrm{B}}(M, M') := \inf_{P \text{ partial matching}} \mathrm{cost}(P),$$

where

$$\mathrm{cost}(P) := \max\Big\{ \sup_{(I,I')\in P} \|I - I'\|_\infty, \sup_{I=\langle a,b\rangle \in M \sqcup M' \text{ unmatched}} \tfrac{1}{2}|a - b| \Big\}$$

and

$$\|I - I'\|_\infty := \max\{|a - a'|, |b - b'|\}$$

for $I = \langle a, b\rangle$, $I' = \langle a', b'\rangle$ (here, $\langle \cdot, \cdot \rangle$ means either open or closed endpoint).

Theorem 2.13　(isometry theorem [22, Theorem 5.14])　*For any two q–tame persistence modules V_* and W_*,*

$$d_{\mathrm{B}}(\mathrm{barc}(V_*), \mathrm{barc}(W_*)) = d_{\mathrm{I}}(V_*, W_*).$$

For the proof of the following theorem, see [23, Lemma 4.3] or [13; 19; 65].

Theorem 2.14　*Let X and Y be compact metric spaces and $\mathbb{F}$ be an arbitrary field. Then, for any $k \in \mathbb{Z}_{\geq 0}$,*

$$d_{\mathrm{I}}\big(\mathrm{PH}_k(\mathrm{VR}_*(X); \mathbb{F}), \mathrm{PH}_k(\mathrm{VR}_*(Y); \mathbb{F})\big) \leq 2d_{\mathrm{GH}}(X, Y).$$

2.2　Injective (hyperconvex) metric spaces

A hyperconvex metric space is one where any collection of balls with nonempty pairwise intersections forces the nonempty intersection of all balls. These were studied by Aronszajn and Panitchpakdi [8] who showed that every hyperconvex space is an absolute 1–Lipschitz retract. Isbell [52] proved that every metric space admits a *smallest* hyperconvex *hull* (see the definition of tight span below). Dress rediscovered this concept in [30] and subsequent work provided much development in the context of phylogenetics [31; 75]. More recently, Joharinad and Jost [53] considered relaxations of hyperconvexity and related it to a certain notion of curvature applicable to general metric spaces.

References for this subsection are [30; 31; 60].

Definition 2.15 (injective metric space) A metric space E is called *injective* if for each 1–Lipschitz map $f: X \to E$ and isometric embedding of X into $\widetilde{X}$, there exists a 1–Lipschitz map $\tilde{f}: \widetilde{X} \to E$ extending f:

$$
\begin{array}{ccc}
X & \lhook\joinrel\longrightarrow & \widetilde{X} \\
& \searrow{\scriptstyle f} & \downarrow{\scriptstyle \tilde{f}} \\
& & E
\end{array}
$$

Definition 2.16 (hyperconvex space) A metric space X is called *hyperconvex* if for every family $(x_i, r_i)_{i \in I}$ of x_i in X and $r_i \geq 0$ such that $d_X(x_i, x_j) \leq r_i + r_j$ for each $i, j \in I$, there exists a point x such that $d_X(x_i, x) \leq r_i$ for each $i \in I$.

The following lemma is easy to deduce from the definition of hyperconvex space:

Lemma 2.17 *Any nonempty intersection of closed balls in hyperconvex space is hyperconvex.*

For a proof of the following proposition, see [8] or [60, Proposition 2.3].

Proposition 2.18 *A metric space is injective if and only if it is hyperconvex.*

Moreover, every injective metric space is a contractible geodesic metric space, as one can see in Lemma 2.20 and Corollary 2.21.

Definition 2.19 (geodesic bicombing) By a *geodesic bicombing* γ on a metric space (X, d_X), we mean a continuous map $\gamma: X \times X \times [0, 1] \to X$ such that, for every pair $(x, y) \in X \times X$, $\gamma(x, y, \cdot)$ is a geodesic from x to y with constant speed. In other words, γ satisfies

(1) $\gamma(x, y, 0) = x$ and $\gamma(x, y, 1) = y$;

(2) $d_X(\gamma(x, y, s), \gamma(x, y, t)) = (t - s) \cdot d_X(x, y)$ for any $0 \leq s \leq t \leq 1$.

Lemma 2.20 [60, Proposition 3.8] *Every injective metric space (E, d_E) admits a geodesic bicombing γ such that, for any $x, y, x', y' \in E$ and $t \in [0, 1]$, it is:*

(1) ***Conical*** $d_E(\gamma(x, y, t), \gamma(x', y', t)) \leq (1 - t)d_E(x, x') + t d_E(y, y')$.

(2) ***Reversible*** $\gamma(x, y, t) = \gamma(y, x, 1 - t)$.

(3) ***Equivariant*** $L \circ \gamma(x, y, \cdot) = \gamma(L(x), L(y), \cdot)$ *for every isometry L of E.*

Corollary 2.21 *Every injective metric space E is contractible.*

Proof By Lemma 2.20, there is a geodesic bicombing γ on E. Fix an arbitrary point $x_0 \in E$. Then restricting γ to $E \times \{x_0\} \times [0, 1]$ gives a deformation retraction of E onto x_0; hence E is contractible. □

Example 2.22 For any set S, the Banach space $L^\infty(S)$ consisting of all the bounded real-valued functions on S with the ℓ^∞–norm is injective.

Definition 2.23 For a compact metric space (X, d_X), the map $\kappa \colon X \to L^\infty(X)$, defined by $x \mapsto d_X(x, \cdot)$, is an isometric embedding and it is called the *Kuratowski embedding*. Hence every compact metric space can be isometrically embedded into an injective metric space.

Let us introduce some notation which will be used throughout this paper. Suppose that X is a subspace of a metric space (E, d_E). For any $r > 0$, let $B_r(X, E) := \{z \in E \mid \exists x \in X \text{ with } d_E(z, x) < r\}$ denote the open r–neighborhood of X in E. In particular, if $X = \{x\}$ for some $x \in E$, it is just denoted by $B_r(x, E)$, the usual open r–ball around x in E.

As one more convention, whenever there is an isometric embedding $\iota \colon X \hookrightarrow E$, we will use the notation $B_r(X, E)$ instead of $B_r(\iota(X), E)$. For instance, in the sequel we will use $B_r(X, L^\infty(X))$ rather than $B_r(\kappa(X), L^\infty(X))$.

Definition 2.24 For any metric space E, a nonempty subspace X, and $r > 0$, the Čech complex $\check{C}_r(X, E)$ is defined as the nerve of the open covering $\mathcal{U}_r := \{B_r(x, E) \mid x \in X\}$. In other words, $\check{C}_r(X, E)$ is the simplicial complex whose vertices are the points of X, and $\{x_0, \ldots, x_n\} \subseteq X$ is a simplex in $\check{C}_r(X, E)$ if and only if $\bigcap_{i=0}^n B_r(x_i, E) \neq \varnothing$.

The following observation is simple, yet it plays an important role in our paper:

Proposition 2.25 If (E, d_E) is an injective metric space and $\varnothing \neq X \subseteq E$ then, for any $r > 0$,
$$\check{C}_r(X, E) = \mathrm{VR}_{2r}(X).$$

Remark 2.26 Proposition 2.25 is optimal in the sense that if $\check{C}_r(X, E) = \mathrm{VR}_{2r}(X)$ holds true *for all* $\varnothing \neq X \subseteq E$, then this condition itself resembles hyperconvexity of E (see Definition 2.16).

Also note that Proposition 2.25 is a generalization of both [45, Lemma 4] and [19, Lemma 2.9] in that those papers only consider the case when X is finite and $E = \ell^\infty(X)$.

Proof of Proposition 2.25 Because of the triangle inequality, it is obvious that $\check{C}_r(X, E)$ is a subcomplex of $\mathrm{VR}_{2r}(X)$. Now, fix an arbitrary simplex $\{x_0, \ldots, x_n\} \in \mathrm{VR}_{2r}(X)$. Then $d_E(x_i, x_j) < 2r$ for any $i, j = 0, \ldots, n$. Since E is hyperconvex, by Proposition 2.18, there exists $\bar{x} \in E$ such that $d_X(x_i, \bar{x}) < r$ for any $i = 0, \ldots, n$ (note that, since $\{x_0, \ldots, x_n\}$ is finite, one can use $<$ instead of $\leq$ when invoking the hyperconvexity property). Therefore, $\{x_0, \ldots, x_n\} \in \check{C}_r(X, E)$. Hence $\mathrm{VR}_{2r}(X)$ is a subcomplex of $\check{C}_r(X, E)$. $\qquad\square$

In particular, Proposition 2.25 implies the following result:

Proposition 2.27 Let X be a subspace of an injective metric space (E, d_E). Then, for any $r > 0$, the Vietoris–Rips complex $\mathrm{VR}_{2r}(X)$ is homotopy equivalent to $B_r(X, E)$.

The proof of Proposition 2.27 will use the following lemma:

Lemma 2.28 *In an injective metric space E, every nonempty intersection of open balls is contractible.*

Proof Let γ be a geodesic bicombing on E, whose existence is guaranteed by Lemma 2.20. Then, for each x, y, x', y' in E and t in $[0, 1]$,

$$d_E(\gamma(x, y, t), \gamma(x', y', t)) \leq (1 - t)d_E(x, x') + t d_E(y, y').$$

In particular, by letting $x' = y' = z$, we obtain

$$d_E(\gamma(x, y, t), z) \leq \max\{d_E(x, z), d_E(y, z)\}$$

for any $t \in [0, 1]$. Hence, if x and y are contained in an open ball with center z, then $\gamma(x, y, t)$ is contained in the same ball for each t in $[0, 1]$. Therefore, if U is a nonempty intersection of open balls in E, then γ restricts to $U \times U \times [0, 1] \to U$, which implies that U is contractible. □

Proof of Proposition 2.27 Let $\mathcal{U}_r := \{B_r(x, E) \mid x \in X\}$. By Lemma 2.28, $\mathcal{U}_r$ is a good cover of $B_r(X, E)$. Hence, by the nerve lemma [49, Corollary 4G.3], $B_r(X, E)$ is homotopy equivalent to the nerve of $\mathcal{U}_r$, which is the same as the Čech complex $\check{C}_r(X, E)$. By Proposition 2.25, $\check{C}_r(X, E) = \mathrm{VR}_{2r}(X)$. □

3 Persistence via metric pairs

One of the insights leading to the notion of persistent homology associated to metric spaces was considering neighborhoods of a metric space in a nice (for example Euclidean) embedding [70]. In this section we formalize this idea in a categorical way.

Definition 3.1 (category of metric pairs) • A *metric pair* is an ordered pair (X, E) of metric spaces such that X is a metric subspace of E.

- Let (X, E) and (Y, F) be metric pairs. A 1–Lipschitz map from (X, E) to (Y, F) is a 1–Lipschitz map from E to F mapping X into Y.

- Let (X, E) and (Y, F) be metric pairs and f and g be 1–Lipschitz maps from (X, E) to (Y, F). We say that f and g are *equivalent* if there exists a continuous family $(h_t)_{t \in [0,1]}$ of 1–Lipschitz maps from E to F and a 1–Lipschitz map $\phi \colon X \to Y$ such that $h_0 = f$, $h_1 = g$ and $h_t|_X = \phi$ for each t.

- We define PMet as the category whose objects are metric pairs and whose morphisms are defined as follows: given metric pairs (X, E) and (Y, F), the morphisms from (X, E) to (Y, F) are equivalence classes of 1–Lipschitz maps from (X, E) to (Y, F).

Recall the definition of persistence families, Definition 2.2. We let hTop_* denote the category of persistence families with morphisms specified as in Definition 2.2.

Remark 3.2 Let (X, E) and (Y, F) be persistent pairs and let f be a 1–Lipschitz morphism between them. Then f maps $B_r(X, E)$ into $B_r(Y, F)$ for each $r > 0$. Furthermore, if g is equivalent to f, then they reduce to homotopy equivalent maps from $B_r(X, E)$ to $B_r(Y, F)$ for each $r > 0$.

By the remark above, we obtain the following functor from PMet to hTop$_*$:

Definition 3.3 (persistence functor)   Define the *persistence functor* B$_*$: PMet $\to$ hTop$_*$ sending (X, E) to the persistence family obtained by the filtration $(B_r(X, E))_{r>0}$ and sending a morphism between metric pairs to the homotopy classes of maps it induces between the filtrations.

Remark 3.4   Suppose a metric pair (X, E) is given. For any $k \geq 0$, one can apply the degree k homology functor (with coefficients in a given field $\mathbb{F}$) to a persistence family B$_*(X, E)$. This induces a persistence module where the morphisms are induced by inclusions. As a persistence module, it is denoted by PH$_k$(B$_*(X, E)$; $\mathbb{F}$).

Let Met be the category of metric spaces where morphisms are given by 1–Lipschitz maps. There is a forgetful functor from PMet to Met mapping (X, E) to X and mapping a morphism defined on (X, E) to its restriction to X. Although forgetful functors often have left adjoints, we are going to see that this one has a right adjoint.

Theorem 3.5   *The forgetful functor from* PMet *to* Met *has a right adjoint.*

First we need to prove a few results. The reader should consult Section 2.2 for background on injective metric spaces.

Lemma 3.6   *Let (X, E) and (Y, F) be metric pairs such that F is an injective metric space. Let f and g be 1–Lipschitz maps from (X, E) to (Y, F). Then f is equivalent to g if and only if $f|_X \equiv g|_X$.*

Proof   The "only if" part is obvious from Definition 3.1. Now assume that $f|_X \equiv g|_X$. By Lemma 2.20, there exists a geodesic bicombing $\gamma \colon F \times F \times [0, 1] \to F$ such that for each $x, y, x', y' \in F$ and $t \in [0, 1]$,

$$d_F(\gamma(x, y, t), \gamma(x', y', t)) \leq (1-t)d_F(x, x') + t\, d_F(y, y').$$

For $t \in [0, 1]$, define $h \colon E \times [0, 1] \to F$ by $h_t(x) = \gamma(f(x), g(x), t)$. Note that $h_0 = f$, $h_1 = g$ and $(h_t)|_X$ is the same map for all t. The inequality above implies that h_t is 1–Lipschitz for all t.   □

Lemma 3.7   *Let (X, E) and (Y, F) be metric pairs such that F is an injective metric space. Then, for each 1–Lipschitz map $\phi \colon X \to Y$, there exists a unique (up to equivalence) 1–Lipschitz map from (X, E) to (Y, F) extending ϕ.*

Proof   The uniqueness up to equivalence part follows from Lemma 3.6. The existence part follows from the injectivity of F.   □

Proof of Theorem 3.5   Let $\kappa \colon$ Met $\to$ PMet be the functor sending X to $(X, L^\infty(X))$ where $L^\infty(X)$ is the Banach space consisting of all the bounded real-valued functions on X with ℓ^∞–norm (see Definition 2.23 in Section 2.2). A 1–Lipschitz map $f \colon X \to Y$ is sent to the unique morphism (see Lemma 3.7) extending f. This functor κ is said to be the *Kuratowski functor*.

There is a natural morphism

$$\mathrm{Hom}\big((X, E), (Y, L^\infty(Y))\big) \to \mathrm{Hom}(X, Y),$$

sending a morphism to its restriction to X. By Lemma 3.7, this is a bijection. Hence κ is a right adjoint to the forgetful functor. □

Recall that any two right adjoints of a same functor must be isomorphic [9, Proposition 9.9].

Definition 3.8 (metric homotopy pairing) A functor $\eta\colon \mathrm{Met} \to \mathrm{PMet}$ is called a *metric homotopy pairing* if it is a right adjoint to the forgetful functor.

Example 3.9 Let (X, d_X) be a metric space. $L^\infty(X)$ is an injective space associated to X; see Section 2.2 for the precise definition. Consider also the following additional spaces associated to X:

$$\Delta(X) := \{ f \in L^\infty(X) \mid f(x) + f(x') \geq d_X(x, x') \text{ for all } x, x' \in X \},$$

$$E(X) := \{ f \in \Delta(X) \mid \text{if } g \in \Delta(X) \text{ and } g \leq f \text{ then } g = f \},$$

$$\Delta_1(X) := \Delta(X) \cap \mathrm{Lip}_1(X, \mathbb{R}),$$

with ℓ^∞–metrics for all of them; see [60, Section 3]. Then

$$(X, L^\infty(X)), \quad (X, E(X)), \quad (X, \Delta(X)), \quad (X, \Delta_1(X))$$

are all metric homotopy pairings, since the second element in each pair is an injective metric space [60, Section 3] into which X isometrically embeds via the map $\kappa\colon x \mapsto d_X(x, \cdot)$. Here, $E(X)$ is said to be the *tight span* of X [30; 52] and it is a especially interesting space. $E(X)$ is the smallest injective metric space into which X can be embedded and it is unique up to isometry. Furthermore, if X is a tree metric space (ie a metric space with 0–hyperbolicity; see Definition 8.1), then $E(X)$ is the smallest metric tree containing X. This special property has recently been used to the application of phylogenetics [31].

4 Isomorphism and stability

Recall that Met is the category of metric spaces with 1–Lipschitz maps as morphisms. We have the functor $\mathrm{VR}_*\colon \mathrm{Met} \to \mathrm{hTop}_*$ induced by the Vietoris–Rips filtration. The main theorem we prove in this section is the following:

Theorem 4.1 (isomorphism theorem) *Let $\eta\colon \mathrm{Met} \to \mathrm{PMet}$ be a metric homotopy pairing* (for example the Kuratowski functor). *Then $\mathrm{B}_* \circ \eta\colon \mathrm{Met} \to \mathrm{hTop}_*$ is naturally isomorphic to VR_{2*}.*

Recall the precise definitions of $\mathcal{U}_r$ and $\check{\mathrm{C}}_r(X, E)$ from Definition 2.24. We denote the filtration of Čech complexes $(\check{\mathrm{C}}_r(X, E))_{r>0}$ by $\check{\mathrm{C}}_*(X, E)$.

The following theorem is the main tool for the proof of Theorem 4.1. Its proof, being fairly long, is relegated to Section A.3.

Theorem 4.2 (generalized functorial nerve lemma) *Let X and Y be two paracompact spaces, $\rho\colon X \to Y$ be a continuous map, $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ and $\mathcal{V} = \{V_\beta\}_{\beta \in B}$ be good open covers (every nonempty finite intersection is contractible) of X and Y, respectively, based on arbitrary index sets A and B, and $\pi\colon A \to B$ be a map such that*

$$\rho(U_\alpha) \subseteq V_{\pi(\alpha)} \quad \text{for any } \alpha \in A.$$

Let $\mathrm{N}\mathcal{U}$ and $\mathrm{N}\mathcal{V}$ be the nerves of $\mathcal{U}$ and $\mathcal{V}$, respectively. Observe that, since $U_{\alpha_0} \cap \cdots \cap U_{\alpha_n} \neq \varnothing$ implies $V_{\pi(\alpha_0)} \cap \cdots \cap V_{\pi(\alpha_n)} \neq \varnothing$, π induces the canonical simplicial map $\bar{\pi}\colon \mathrm{N}\mathcal{U} \to \mathrm{N}\mathcal{V}$.

Then there exist homotopy equivalences $X \to \mathrm{N}\mathcal{U}$ and $Y \to \mathrm{N}\mathcal{V}$ that commute with ρ and $\bar{\pi}$ up to homotopy:

$$
\begin{array}{ccc}
X & \longrightarrow & \mathrm{N}\mathcal{U} \\
\rho \downarrow & & \downarrow \bar{\pi} \\
Y & \longrightarrow & \mathrm{N}\mathcal{V}
\end{array}
$$

The next corollary is an important special case of Theorem 4.2.

Corollary 4.3 (functorial nerve lemma) *Let $X \subseteq X'$ be two paracompact spaces. Let $\mathcal{U} = \{U_\alpha\}_{\alpha \in \Lambda}$ and $\mathcal{U}' = \{U'_\alpha\}_{\alpha \in \Lambda}$ be good open covers (every nonempty finite intersection is contractible) of X and X', respectively, based on the same index set Λ, such that $U_\alpha \subseteq U'_\alpha$ for all $\alpha \in \Lambda$. Let $\mathrm{N}\mathcal{U}$ and $\mathrm{N}\mathcal{U}'$ be the nerves of $\mathcal{U}$ and $\mathcal{U}'$, respectively.*

Then there exist homotopy equivalences $X \to \mathrm{N}\mathcal{U}$ and $X' \to \mathrm{N}\mathcal{U}'$ that commute with the canonical inclusions $X \hookrightarrow X'$ and $\mathrm{N}\mathcal{U} \hookrightarrow \mathrm{N}\mathcal{U}'$, up to homotopy:

$$
\begin{array}{ccc}
X & \longrightarrow & \mathrm{N}\mathcal{U} \\
\downarrow & & \downarrow \\
X' & \longrightarrow & \mathrm{N}\mathcal{U}'
\end{array}
$$

Proof Choose the canonical inclusion map $X \hookrightarrow X'$ as ρ, the identity map on Λ as π, and apply Theorem 4.2.   □

Remark 4.4 A result similar to Corollary 4.3 was already proved in [21, Lemma 3.4] for finite-index sets, whereas in our version index sets can have arbitrary cardinality. In [24, Theorems 25 and 26], the authors prove a simplicial complex version of Corollary 4.3 for finite-index sets and invoke a certain functorial version of Dowker's theorem.

Finally, recently we became aware of [82, Lemma 5.1], which is similar to Theorem 4.2. The author considers spaces with *numerable* covers (ie the spaces admit locally finite partition of unity subordinate to the covers), whereas in our version that condition is automatically satisfied since we only consider paracompact spaces. Our proof technique differs from that of [82] in that whereas [82] relies on a result from [29], our proof follows the traditional proof of the nerve lemma [49].

Proposition 4.5 *For each metric pair $(X, E) \in \mathrm{PMet}$, there exist homotopy equivalences*

$$\phi_*^{(X,E)} \colon B_*(X, E) \to \check{C}_*(X, E)$$

such that, for any $0 < r \leq s$, the diagram

$$
\begin{array}{ccc}
B_r(X, E) & \xrightarrow{\ \phi_r^{(X,E)}\ } & \check{C}_r(X, E) \\
\Big\downarrow & & \Big\downarrow \\
B_s(X, E) & \xrightarrow{\ \phi_s^{(X,E)}\ } & \check{C}_s(X, E)
\end{array}
$$

commutes up to homotopy, where $B_r(X, E) \hookrightarrow B_s(X, E)$ and $\check{C}_r(X, E) \hookrightarrow \check{C}_s(X, E)$ are the canonical inclusions.

Proof Observe that $\check{C}_r(X, E)$ is the nerve of the open cover $\mathcal{U}_r$ for any $r > 0$, and apply Corollary 4.3. □

Proposition 4.6 *Let (X, E) and (Y, F) be metric pairs in PMet, and $f \colon (X, E) \to (Y, F)$ be a 1–Lipschitz map. Let $\phi_*^{(X,E)} \colon B_*(X, E) \to \check{C}_*(X, E)$ and $\phi_*^{(Y,F)} \colon B_*(Y, F) \to \check{C}_*(Y, F)$ be the homotopy equivalences guaranteed by Proposition 4.5. Then, for any $r > 0$, the diagram*

$$
\begin{array}{ccc}
B_r(X, E) & \xrightarrow{\ \phi_r^{(X,E)}\ } & \check{C}_r(X, E) \\
f_r \Big\downarrow & & \Big\downarrow f_r \\
B_r(Y, F) & \xrightarrow{\ \phi_r^{(Y,F)}\ } & \check{C}_r(Y, F)
\end{array}
$$

commutes up to homotopy, where $f_* \colon B_r(X, E) \to B_r(Y, F)$ and $f_* \colon \check{C}_r(X, E) \to \check{C}_r(Y, F)$ are the canonical maps induced from f.

Furthermore, if we substitute f with an equivalent map, then the homotopy types of the vertical maps remain unchanged.

Proof Since f is 1–Lipschitz, $f(B_r(x, E)) \subseteq B_r(f(x), F)$. Hence, if we choose $f|_{B_r(X,E)}$ as ρ, and $f|_X$ as π, the commutativity of the diagram is the direct result of Theorem 4.2.

Furthermore, if f and g are equivalent, then the homotopy (h_t) between f and g induces the homotopy between $f_r \colon B_r(X, E) \to B_r(Y, F)$ and $g_r \colon B_r(X, E) \to B_r(Y, F)$. Moreover, since $f|_X = g|_X$, both of the induced maps $f_r \colon \check{C}_r(X, E) \to \check{C}_r(Y, F)$ and $g_r \colon \check{C}_r(X, E) \to \check{C}_r(Y, F)$ are exactly the same. □

We are now ready to prove the main theorem of this section.

Proof of Theorem 4.1 Since all metric homotopy pairings are naturally isomorphic, without loss of generality we can assume that $\eta = \kappa$, the Kuratowski functor. Note that, by Proposition 2.25, $\check{C}_r(X, E) = \mathrm{VR}_{2r}(X)$ for any $X \in \mathrm{Met}$ and $r > 0$.

Let's construct the natural transformation τ from $B_* \circ \kappa \colon \mathrm{Met} \to \mathrm{hTop}_*$ to VR_{2*} in the following way: Fix an arbitrary metric space $X \in \mathrm{Met}$, and let τ_X be the homotopy equivalences

$$\phi_*^{(X, L^\infty(X))} \colon B_*(X, L^\infty(X)) \to \mathrm{VR}_{2*}(X)$$

guaranteed by Proposition 4.5. Then, when $f \colon X \to Y$ is 1–Lipschitz, the functoriality between τ_X and τ_Y is the direct result of Proposition 4.6. So τ is indeed a natural transformation. Finally, since each $\phi_r^{(X, L^\infty(X))}$ is a homotopy equivalence for any $X \in \mathrm{Met}$ and $r > 0$, τ is natural isomorphism. □

4.1 Stability of metric homotopy pairings

In this subsection, we consider a distance between metric pairs by invoking the homotopy interleaving distance introduced by Blumberg and Lesnick [13] and then show that metric homotopy pairings are 1–Lipschitz with respect to this distance and the Gromov–Hausdorff distance.

Let us give a quick review of homotopy interleaving distance between $\mathbb{R}$–spaces. For more details, please see [13, Section 3.3]. An $\mathbb{R}$–space is a functor from the poset $(\mathbb{R}, \leq)$ to the category of topological spaces. Note that given a metric pair (X, E), the filtration of open neighborhoods $B_*(X, E)$ is an $\mathbb{R}$–space. Two $\mathbb{R}$–spaces A_* and B_* are said to be δ–interleaved for some $\delta > 0$ if there are natural transformations $f \colon A_* \to B_{*+\delta}$ and $g \colon B_* \to A_{*+\delta}$ such that $f \circ g$ and $g \circ f$ are equal to the structure maps $B_* \to B_{*+2\delta}$ and $A_* \to A_{*+2\delta}$, respectively.

A natural transformation $f \colon R_* \to A_*$ is called a weak homotopy equivalence if f induces a isomorphism between homotopy groups at each index. Two $\mathbb{R}$–spaces A_* and A'_* are said to be weakly homotopy equivalent if there exists an $\mathbb{R}$–space R_* and weak homotopy equivalences $f \colon R_* \to A_*$ and $f' \colon R_* \to A'_*$. The homotopy interleaving distance $d_{\mathrm{HI}}(A_*, B_*)$ is then defined as the infimal $\delta > 0$ such that there exists δ–interleaved $\mathbb{R}$–spaces A'_* and B'_* with the property that A'_* and B'_* are weakly homotopy equivalent to A_* and B_*, respectively.

We now adapt this construction to metric pairs. Given metric pairs (X, E) and (Y, F), we define the homotopy interleaving distance between them by

$$d_{\mathrm{HI}}((X, E), (Y, F)) := d_{\mathrm{HI}}(B_*(X, E), B_*(Y, F)).$$

The main theorem that we are going to prove in this section is the following. Below, d_{GH} denotes the Gromov–Hausdorff distance between metric spaces (see [15]) and d_{I} denotes the interleaving distance between persistence modules (see Section 2.1).

Theorem 4.7 *Let $\eta \colon \mathrm{Met} \to \mathrm{PMet}$ be a metric homotopy pairing. Then for any compact metric spaces X and Y,*

$$d_{\mathrm{HI}}(\eta(X), \eta(Y)) \leq d_{\mathrm{GH}}(X, Y).$$

Remark 4.8 By combining Theorem 4.7 and the isomorphism theorem (Theorem 4.1), one obtains another proof of Theorem 2.14: for any compact metric spaces X and Y, a field $\mathbb{F}$, and $k \in \mathbb{Z}_{\geq 0}$,

$$d_{\mathrm{I}}\big(\mathrm{PH}_k(\mathrm{VR}_*(X); \mathbb{F}), \mathrm{PH}_k(\mathrm{VR}_*(Y); \mathbb{F})\big) \leq 2d_{\mathrm{GH}}(X, Y).$$

Lemma 4.9 *If (X, E) and (Y, F) are isomorphic in* PMet, *then* $d_{\mathrm{HI}}((X, E), (Y, F)) = 0$.

Proof Let $f : (X, E) \to (Y, F)$ and $g : (Y, F) \to (X, E)$ be 1–Lipschitz maps such that $f \circ g$ and $g \circ f$ are equivalent to the respective identities. Then the result follows since f and g induce an isomorphism between the $\mathbb{R}$–spaces $B_*(X, E)$ and $B_*(Y, F)$. $\qquad\square$

Lemma 4.10 *Let E and F be injective metric spaces containing X. Then (X, E) is isomorphic to (X, F) in* PMet.

Proof By injectivity of E and F, there are 1–Lipschitz maps $f : E \to F$ and $g : F \to E$ such that $f|_X$ and $g|_X$ are equal to id_X. Hence, by Lemma 3.6, $f \circ g : (X, F) \to (X, F)$ and $g \circ f : (X, E) \to (X, F)$ are equivalent to the identity. $\qquad\square$

Proof of Theorem 4.7 Since all metric homotopy pairings are naturally isomorphic, by Lemma 4.9, without loss of generality we can assume that $\eta = \kappa$, the Kuratowski functor. Let $r > d_{\mathrm{GH}}(X, Y)$. Let us show that

$$d_{\mathrm{HI}}\big((X, L^\infty(X)), (Y, L^\infty(Y))\big) \leq r.$$

By assumption (see [15]), there exists a metric space Z containing X and Y such that the Hausdorff distance between X and Y as subspaces of Z is less than or equal to r. Hence, the $\mathbb{R}$–spaces $B_*(X, L^\infty(Z))$ and $B_*(Y, L^\infty(Z))$ are r–interleaved as

$$B_\epsilon(X, L^\infty(Z)) \subseteq B_{r+\epsilon}(Y, L^\infty(Z)) \quad \text{and} \quad B_\epsilon(Y, L^\infty(Z)) \subseteq B_{r+\epsilon}(X, L^\infty(Z))$$

for each ϵ. Now, by Lemma 4.10,

$$d_{\mathrm{HI}}\big((X, L^\infty(X)), (Y, L^\infty(Y))\big) = d_{\mathrm{HI}}\big((X, L^\infty(Z)), (Y, L^\infty(Z))\big) \leq r. \qquad\square$$

5 Application: endpoints of intervals in $\mathrm{barc}_k^{\mathbf{VR}}(X)$

It is known that, in some cases, the intervals in the Vietoris–Rips barcode of a metric space are of the form $(u, v]$ or (u, ∞) for $0 \leq u < v < \infty$.

Example 5.1 In the following examples, any $I \in \mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$ has a form of $(u, v]$ or (u, ∞) for some $0 \leq u < v < \infty$:

(1) when X is a finite metric space, for any $k \geq 0$;

(2) when $X = \mathbb{S}^1$, for any $k \geq 0$ (see [1, Theorem 7.4]);

(3) when X is a compact geodesic metric space, for $k = 1$ (see [81, Theorem 8.2]).

As far as we know, the general statement given in Theorem 5.2 below is first proved in this paper. Our proof crucially exploits the isomorphism theorem (Theorem 4.1).

Theorem 5.2 *Suppose a compact metric space (X, d_X), a field $\mathbb{F}$, and a nonnegative integer k are given. Then, for any $I \in \mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$, I must be of the form $(u, v]$ or (u, ∞) for some $0 \le u < v < \infty$.*

We first state and prove two lemmas which will be combined in order to furnish the proof of Theorem 5.2.

Lemma 5.3 *Let X be a topological space and G be an abelian group. Then, for any $k \ge 0$ and any k–dimensional singular chain c of X with coefficients in G, there exist a compact subset $K_c \subseteq X$ and k–dimensional singular chain c' of K_c with coefficients in G such that*

$$(\iota)_\sharp(c') = c,$$

where $\iota \colon K_c \hookrightarrow X$ is the canonical inclusion map.

Proof Recall that one can express c as a sum of finitely many k–dimensional singular simplices with coefficients in G. In other words,

$$c = \sum_{i=1}^{l} \alpha_i \sigma_i,$$

where $\alpha_i \in G$ and $\sigma_i \colon \Delta_k \to X$ is a continuous map for each $i = 1, \ldots, l$. Next, let $K_c := \bigcup_{i=1}^{l} \sigma_i(\Delta_k)$. This K_c is the compact subspace that we required. $\square$

For the remainder of this section, given any field $\mathbb{F}$ and a metric pair (X, E), for each $0 < r < \infty$ we will denote by $(\mathrm{SC}_*^{(r)}, \partial_*^{(r)})$ the singular chain complex of $B_r(X, E)$ with coefficients in $\mathbb{F}$. For each $0 < r \le s < \infty$ we will denote by $i_{r,s}$ the canonical inclusion map $B_r(X, E) \subseteq B_s(X, E)$. By $(i_{r,s})_\sharp$ we will denote the (injective) map induced at the level of singular chain complexes.

Lemma 5.4 *Suppose that a compact metric space (X, d_X), a field $\mathbb{F}$, a metric homotopy pairing η, and a nonnegative integer k are given. Then, for every $I \in \mathrm{barc}\big(\mathrm{PH}_k(\mathrm{B}_* \circ \eta(X); \mathbb{F})\big)$:*

  (i) *If $u \in [0, \infty)$ is the left endpoint of I, then $u \notin I$ (ie I is left-open).*

  (ii) *If $v \in [0, \infty)$ is the right endpoint of I, then $v \in I$ (ie I is right-closed).*

Proof (i) Let $\eta(X) = (X, E)$. The fact that $I \in \mathrm{barc}\big(\mathrm{PH}_k(\mathrm{B}_* \circ \eta(X); \mathbb{F})\big)$ implies that, for each $r \in I$, there exists a singular k–cycle c_r on $B_r(X, E)$ with coefficients in $\mathbb{F}$ satisfying

  (1) $[c_r] \in \mathrm{H}_k(B_r(X, E); \mathbb{F})$ is nonzero for any $r \in I$,

  (2) $(i_{r,s})_*([c_r]) = [c_s]$ for any $r \le s$ in I.

Now, suppose that u is a closed left endpoint of I (so $u \in I$). In particular, by the above there exists a singular k–cycle c_u on $B_u(X, E)$ with coefficients in $\mathbb{F}$ with the above two properties.

Then, by Lemma 5.3, we know that there is a compact subset $K_{c_u} \subseteq B_u(X, E)$ and a singular k–cycle c'_u on K_{c_u} with coefficients in $\mathbb{F}$ such that $(\iota)_\sharp(c'_u) = c_u$ where $\iota \colon K_{c_u} \to B_u(X, E)$ is the canonical inclusion. Moreover, since K_{c_u} is compact, there exists a small $\varepsilon > 0$ such that

$$K_{c_u} \subseteq B_{u-\varepsilon}(X, E).$$

Now, define $c_{u-\varepsilon} := (\iota')_\sharp(c'_u)$ where $\iota' \colon K_{c_u} \to B_{u-\varepsilon}(X, E)$ is the canonical inclusion. Then, this singular chain satisfies

$$(i_{u-\varepsilon,u})_\sharp(c_{u-\varepsilon}) = (i_{u-\varepsilon,u})_\sharp \circ (\iota')_\sharp(c'_u) = (\iota)_\sharp(c'_u) = c_u.$$

Moreover, $c_{u-\varepsilon}$ cannot be null-homologous. Otherwise, there would exist a singular $(k+1)$–chain $d_{u-\varepsilon}$ on $B_{u-\varepsilon}(X, E)$ with coefficients in $\mathbb{F}$ such that $\partial^{(u-\varepsilon)}_{k+1} d_{u-\varepsilon} = c_{u-\varepsilon}$. However, this would imply

$$\partial^{(u)}_{k+1} \circ (i_{u-\varepsilon,u})_\sharp(d_{u-\varepsilon}) = (i_{u-\varepsilon,u})_\sharp \circ \partial^{(u-\varepsilon)}_{k+1}(d_{u-\varepsilon}) = (i_{u-\varepsilon,u})_\sharp(c_{u-\varepsilon}) = c_u,$$

by the naturality of the boundary operators $\partial^{(u-\varepsilon)}_{k+1}$ and $\partial^{(u)}_{k+1}$. This would in turn contradict the property $[c_u] \neq 0$.

So, we must have $[c_{u-\varepsilon}] \neq 0$. But, the existence of such $c_{u-\varepsilon}$ contradicts the fact that u is the left endpoint of I. Therefore, one concludes that u cannot be a closed left endpoint, so it must be an open endpoint.

(ii) Now, suppose that v is an open right endpoint of I (so that $v \notin I$ and therefore c_v is not defined by the above two conditions). Choose a small enough $\varepsilon > 0$ that $v - \varepsilon \in I$, and let

$$c_v := (i_{v-\varepsilon,v})_\sharp(c_{v-\varepsilon}).$$

Then c_v must be null-homologous.

This means that there is a singular $(k+1)$–dimensional chain d_v on $B_v(X, E)$ with coefficients in $\mathbb{F}$ such that $\partial^{(v)}_{k+1} d_v = c_v$. By Lemma 5.3, we know that there is a compact subset $K_{d_v} \subseteq B_v(X, E)$ and a singular $(k+1)$–chain d'_v of K_{d_v} with coefficients in $\mathbb{F}$ such that $(\iota)_\sharp(d'_v) = d_v$ where $\iota \colon K_{d_v} \to B_v(X, E)$ is the canonical inclusion. Moreover, since K_{d_v} is compact, there exists $\varepsilon' \in (0, \varepsilon]$ such that $K_{d_v} \subseteq B_{v-\varepsilon'}(X, E)$.

Let $d_{v-\varepsilon'} := (\iota')_\sharp(d'_v)$ where $\iota' \colon K_{d_v} \hookrightarrow B_{v-\varepsilon'}(X, E)$ is the canonical inclusion. Then, again by the naturality of boundary operators,

$$(i_{v-\varepsilon',v})_\sharp \circ \partial^{(v-\varepsilon')}_{k+1}(d_{v-\varepsilon'}) = \partial^{(v)}_{k+1} \circ (i_{v-\varepsilon',v})_\sharp(d_{v-\varepsilon'})$$
$$= \partial^{(v)}_{k+1} \circ (i_{v-\varepsilon',v})_\sharp \circ (\iota')_\sharp(d'_v) = \partial^{(v)}_{k+1} \circ (\iota)_\sharp(d'_v) = \partial^{(v)}_{k+1} d_v = c_v.$$

Since $(i_{v-\varepsilon',v})_\sharp$ is injective and $(i_{v-\varepsilon',v})_\sharp \circ (i_{v-\varepsilon,v-\varepsilon'})_\sharp(c_{v-\varepsilon}) = (i_{v-\varepsilon,v})_\sharp(c_{v-\varepsilon}) = c_v$, one can conclude that $\partial^{(v-\varepsilon')}_{k+1}(d_{v-\varepsilon'}) = (i_{v-\varepsilon,v-\varepsilon'})_\sharp(c_{v-\varepsilon})$. This indicates that

$$0 = [(i_{v-\varepsilon,v-\varepsilon'})_\sharp(c_{v-\varepsilon})] = (i_{v-\varepsilon,v-\varepsilon'})_*([c_{v-\varepsilon}]) = [c_{v-\varepsilon'}],$$

but it contradicts the fact that $[c_{v-\varepsilon'}] \neq 0$. Therefore, v must be a closed endpoint. $\square$

Finally, the proof of Theorem 5.2 follows from the lemmas above.

Proof of Theorem 5.2 Apply Lemma 5.4 and Theorem 4.1. $\square$

A (false) conjecture Actually, we first expected the following conjecture to be true. Observe that, if true, the conjecture would imply Theorem 5.2. Also, it is obvious that this conjecture is true when X is a finite metric space.

Conjecture 5.5 (lower semicontinuity of the homotopy type of Vietoris–Rips complexes) Suppose X is a compact metric space. Then, for any $r \in \mathbb{R}_{>0}$, $\mathrm{VR}_r(X)$ is homotopy equivalent to $\mathrm{VR}_{r-\varepsilon}(X)$ whenever $\varepsilon > 0$ is small enough.

However, the following example shows that this conjecture is false:

Example 5.6 By [1, Theorem 7.4], we know that $\mathrm{VR}_r(\mathbb{S}^1)$ is homotopy equivalent to $\mathbb{S}^{2m+1}$ if $r \in (2\pi m/(2m+1), 2\pi(m+1)/(2m+3)]$ for $m = 0, 1, \ldots$. Observe that $\lim_{m\to\infty} 2\pi m/(2m+1) = \pi$. Therefore, $\mathrm{VR}_\pi(\mathbb{S}^1)$ cannot be homotopy equivalent to $\mathrm{VR}_{\pi-\varepsilon}(\mathbb{S}^1)$ for all small enough ε, since for r in the interval $[\pi - \varepsilon, \pi]$, $\mathrm{VR}_r(\mathbb{S}^1)$ attains infinitely many different homotopy types.

Then, one might now wonder whether the conjecture holds when we restrict the range of r to $(0, \mathrm{diam}(X))$. But, again this new conjecture is false, as the following example shows:

Example 5.7 Let $X := \mathbb{S}^1 \vee \alpha \cdot \mathbb{S}^1$ for some $\alpha \in (0, 1)$. Observe that $\mathrm{diam}(X) = \pi$. Also, by Lemma 6.6, $E \vee F$ will be an injective metric space containing X whenever E is an injective metric space containing $\mathbb{S}^1$ (eg $E(\mathbb{S}^1)$) and F is an injective metric space containing $\alpha \cdot \mathbb{S}^1$ (eg $E(\alpha \cdot \mathbb{S}^1)$). Hence, by Proposition 2.27, $\mathrm{VR}_{2r}(X) \simeq B_r(X, E \vee F) = B_r(\mathbb{S}^1, E) \vee B_r(\alpha \cdot \mathbb{S}^1, F)$ and $\mathrm{VR}_{2r}(\alpha \cdot \mathbb{S}^1) \simeq B_r(\alpha \cdot \mathbb{S}^1, F)$ for any $r > 0$. Therefore, $\mathrm{VR}_{\alpha\pi}(X)$ cannot be homotopy equivalent to $\mathrm{VR}_{\alpha\pi-\varepsilon}(X)$ for small enough ε, since $\mathrm{VR}_r(\alpha \cdot \mathbb{S}^1)$ attains infinitely many homotopy types for $r \in [\alpha\pi - \varepsilon, \alpha\pi]$.

6 Application: products and metric gluings

The following statement regarding products of filtrations are obtained at the simplicial level (and in more generality) in [71, Proposition 2.6; 42; 72]. The statement about metric gluings appeared in [3, Proposition 4; 66, Proposition 4.4]. These proofs operate at the simplicial level.

Here we give alternative proofs through the consideration of neighborhoods in an injective metric space via Theorem 4.1.

We first recall the notion of metric gluing: given two metric spaces X and Y and points $p \in X$ and $q \in Y$, the *metric gluing* $X \vee Y := X \sqcup Y / p \sim q$ is defined with the metric

$$d_{X \vee Y}(z, z') := \begin{cases} d_X(z, z') & \text{if } z, z' \in X, \\ d_Y(z, z') & \text{if } z, z' \in Y, \\ d_X(z, p) + d_Y(z', q) & \text{if } z \in X \text{ and } z' \in Y. \end{cases}$$

Theorem 6.1 (persistent Künneth formula) *Let X and Y be metric spaces, and $\mathbb{F}$ be a field.*

(1) **Persistent Künneth formula** *Let $X \times Y$ denote the ℓ^∞–product of X and Y. Then*

$$\mathrm{PH}_*(\mathrm{VR}_*(X \times Y); \mathbb{F}) \cong \mathrm{PH}_*(\mathrm{VR}_*(X); \mathbb{F}) \otimes \mathrm{PH}_*(\mathrm{VR}_*(Y); \mathbb{F}).$$

(2) *Let p and q be points in X and Y respectively. Let $X \vee Y$ denote the metric gluing of metric spaces X and Y along p and q. Then[2]*

$$\mathrm{PH}_*(\mathrm{VR}_*(X \vee Y); \mathbb{F}) \cong \mathrm{PH}_*(\mathrm{VR}_*(X); \mathbb{F}) \oplus \mathrm{PH}_*(\mathrm{VR}_*(Y); \mathbb{F}).$$

Remark 6.2 Corollaries 5.2 and 5.8 of [5] establish results analogous to Theorem 6.1 for the products and metric gluings of Vietoris–Rips metric thickenings.

Remark 6.3 The tensor product of two simple persistence modules corresponding to intervals I and J is the simple persistence module corresponding to the interval $I \cap J$. Therefore, the first part of Theorem 6.1 implies that

$$\mathrm{barc}_k^{\mathrm{VR}}(X \times Y; \mathbb{F}) := \{I \cap J \mid I \in \mathrm{barc}_i^{\mathrm{VR}}(X; \mathbb{F}), \ J \in \mathrm{barc}_j^{\mathrm{VR}}(Y; \mathbb{F}), \ i + j = k\}$$

for any nonnegative integer k.

Example 6.4 (tori) For a given choice of $\alpha_1, \ldots, \alpha_n > 0$, let X be the ℓ^∞–product $\prod_{i=1}^n (\alpha_i \cdot \mathbb{S}^1)$. Then, by [1, Theorem 7.4] and Remark 6.3,

$$\mathrm{barc}_0^{\mathrm{VR}}(X; \mathbb{F}) = \{(0, \infty)\},$$

and

$\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$

$$= \left\{ \left(\max_{1 \le j \le m} \frac{2\pi \alpha_{i_j} l_{i_j}}{2l_{i_j} + 1}, \ \min_{1 \le j \le m} \frac{2\pi \alpha_{i_j} (l_{i_j} + 1)}{2l_{i_j} + 3} \right] \ \middle| \ \{i_j\}_{j=1}^m \subseteq \{1, \ldots, n\}, \ l_{i_j} \in \mathbb{Z}_{\ge 0}, \ \sum_{j=1}^m (2l_{i_j} + 1) = k \right\}$$

for any $k \in \mathbb{Z}_{>0}$.

Note that above we are defining a multiset; hence if an element appears more than once in the definition, then it will appear more than once in the multiset. In particular, in the case of $X = \mathbb{S}^1 \times \mathbb{S}^1$, for all integers $k \ge 0$,

$$\mathrm{barc}_0^{\mathrm{VR}}(X; \mathbb{F}) = \{(0, \infty)\},$$
$$\mathrm{barc}_{2k+1}^{\mathrm{VR}}(X; \mathbb{F}) = \left\{ \left(\frac{2\pi k}{2k+1}, \frac{2\pi(k+1)}{2k+3} \right], \left(\frac{2\pi k}{2k+1}, \frac{2\pi(k+1)}{2k+3} \right] \right\},$$
$$\mathrm{barc}_{4k+2}^{\mathrm{VR}}(X; \mathbb{F}) = \left\{ \left(\frac{2\pi k}{2k+1}, \frac{2\pi(k+1)}{2k+3} \right] \right\},$$
$$\mathrm{barc}_{4k+4}^{\mathrm{VR}}(X; \mathbb{F}) = \varnothing.$$

See also the remarks on homotopy types of Vietoris–Rips complexes of tori in [1, Proposition 10.2; 18].

---

[2]We use the "reduced" homology functor for this metric gluing case.

To be able to prove Theorem 6.1, we need the following lemmas:

Lemma 6.5 *If E and F are injective metric spaces, then so is their ℓ^∞–product.*

Proof Let X be a metric space. Note that $(f, g)\colon X \to E \times F$ is 1–Lipschitz if and only if f and g are 1–Lipschitz. Given such f and g and a metric embedding X into Y, we have 1–Lipschitz extensions $\tilde{f}$ and $\tilde{g}$ of f and g from Y to E and F, respectively. Hence, $(\tilde{f}, \tilde{g})\colon Y \to E \times F$ is a 1–Lipschitz extension of (f, g). Therefore $E \times F$ is injective.   □

Lemma 6.6 *If E and F are injective metric spaces, then so is their metric gluing along any two points.*

Proof Let p and q be points in E and F, respectively, and $E \vee F$ denote the metric gluing of E and F along p and q. We are going to show that $E \vee F$ is hyperconvex, hence injective (see Proposition 2.18). We denote the metric on $E \vee F$ by d, the metric on E by d_E and the metric on F by d_F.

Let $(x_i, r_i)_i$ and $(y_j, s_j)_j$ be such that each x_i is in E, each y_j is in F, $r_i \geq 0$, $s_j \geq 0$,

$$d_E(x_i, x_{i'}) \leq r_i + r_{i'}, \quad d_F(y_j, y_{j'}) \leq s_j + s_{j'}, \quad d(x_i, y_j) \leq r_i + s_j$$

for each i, i', j and j'. Define ϵ by

$$\epsilon := \max\{\inf_i(r_i - d_E(x_i, p)), \inf_j(s_j - d_F(y_j, q))\}.$$

Let us show that $\epsilon \geq 0$. If the second element inside the maximum is negative, then there exists j_0 such that $d_F(y_{j_0}, q) - s_{j_0} > 0$. Since $d(x_i, y_{j_0}) = d_E(x_i, p) + d_F(q, y_{j_0})$ for all i,

$$r_i - d_E(x_i, p) = d_F(y_{j_0}, q) + (r_i - d(x_i, y_{j_0})) \geq d_F(y_{j_0}, q) - s_{j_0} > 0.$$

Therefore the first element inside the maximum is nonnegative. Hence $\epsilon \geq 0$.

Without loss of generality, let us assume that

$$\epsilon = \inf_i(r_i - d_E(x_i, p)) \geq 0.$$

This implies that the nonempty closed ball $\bar{B}_\epsilon(q, F)$ is contained in $\bar{B}_{r_i}(x_i, E \vee F)$ for all i. Now, for each j,

$$\epsilon + s_j = \inf_i(r_i - d_E(x_i, p) + s_j) \geq \inf_i(d(x_i, y_j) - d_E(x_i, p)) = d_F(y_j, q).$$

Therefore,

$$\left(\bigcap_i \bar{B}_{r_i}(x_i, E \vee F)\right) \cap \left(\bigcap_j \bar{B}_{s_j}(y_j, E \vee F)\right) \supseteq \bar{B}_\epsilon(q, F) \cap \left(\bigcap_j \bar{B}_{s_j}(y_j, F)\right) \neq \varnothing,$$

where the right-hand side is nonempty by hyperconvexity of F.   □

Proof of Theorem 6.1 (1) Let E and F be injective metric spaces containing X and Y respectively. Let $E \times F$ denote the ℓ^∞–product of E and F. Note that for each $r > 0$,

$$B_r(X \times Y, E \times F) = B_r(X, E) \times B_r(Y, F).$$

Hence, by the (standard) Künneth formula [68, Theorem 58.5],

$$H_*(B_r(X \times Y, E \times F); \mathbb{F}) \cong H_*(B_r(X, E); \mathbb{F}) \otimes H_*(B_r(Y, F); \mathbb{F}).$$

Now, the result follows from Lemma 6.5 and Theorem 4.1.

(2) Let E and F be as above and $E \vee F$ denote metric gluing of E and F along p and q. Note that

$$B_r(X \vee Y, E \vee F) = B_r(X, E) \vee B_r(Y, F).$$

Hence, by [49, Corollary 2.25],

$$H_*(B_r(X \vee Y, E \vee F); \mathbb{F}) \cong H_*(B_r(X, E); \mathbb{F}) \oplus H_*(B_r(Y, F); \mathbb{F}).$$

Now, the result follows from Lemma 6.6 and Theorem 4.1. □

7 Application: homotopy types of $VR_r(X)$ for $X \in \{\mathbb{S}^1, \mathbb{S}^2, \mathbb{CP}^n\}$

In a series of papers [54; 55; 56; 57], Katz studied the filling radius of spheres and complex projective spaces. In this sequence of papers, Katz developed a notion of Morse theory for the diameter function $\mathrm{diam} \colon \mathrm{pow}(X) \to \mathbb{R}$ over a given metric space. By characterizing critical points of the diameter function on each of the spaces $\mathbb{S}^1$, $\mathbb{S}^2$, and $\mathbb{CP}^n$, he was able to prove some results about the different homotopy types attained by $B_r(X, L^\infty(X))$ for $X \in \{\mathbb{S}^1, \mathbb{S}^2, \mathbb{CP}^n\}$ as r increases. Here, we obtain some corollaries that follow from combining the work of Katz [55; 56] with Theorem 4.1.

7.1 The case of spheres with geodesic distance

In [50, Theorem 3.5], Hausmann introduced the quantity $r(M)$ for a Riemannian manifold M, which is the supremum of those $r > 0$ satisfying the following three conditions:

(1) For all $x, y \in M$ such that $d_M(x, y) < 2r$, there is a unique shortest geodesic joining x to y. Its length is $d_M(x, y)$.

(2) Let $x, y, z, w \in M$ with $d_M(x, y), d_M(y, z), d_M(z, x) < r$, and w be any point on the shortest geodesic joining x to y. Then $d_M(z, w) \leq \max\{d_M(y, z), d_M(z, x)\}$.

(3) If γ and γ' are arc-length parametrized geodesics such that $\gamma(0) = \gamma'(0)$, and if $0 \leq s, s' < r$ and $0 \leq t \leq 1$, then $d_M(\gamma(ts), \gamma'(ts')) \leq d_M(\gamma(s), \gamma(s'))$.

In particular, it can be checked that $r(\mathbb{S}^n) = \frac{\pi}{2}$ for any $n \geq 1$. Hausmann then proved that if $r(M) > 0$, $VR_r(M)$ is homotopy equivalent to M for any $r \in (0, r(M))$. This theorem is one of the foundational results in topological data analysis, since it provides theoretical basis for the use of the Vietoris–Rips filtration for recovering the homotopy type of the underlying space.

Then, via Proposition 2.27, we obtain that $B_r(M, L^\infty(M)) \simeq M$ for $r \in \left(0, \frac{1}{2}r(M)\right]$, and therefore $B_r(\mathbb{S}^n, L^\infty(\mathbb{S}^n)) \simeq \mathbb{S}^n$ for all $r \in \left(0, \frac{\pi}{4}\right]$. In [54, Remark, page 508], Katz constructs a retraction from $B_r(\mathbb{S}^n, L^\infty(\mathbb{S}^n))$ to $\mathbb{S}^n$ for r in the range $\left(0, \frac{1}{2}\arccos(-1/(n+1))\right]$, which is a larger range than the
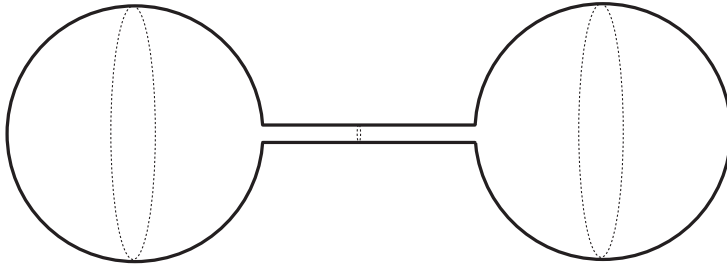
Figure 3: A 2–dimensional sphere with more than one interval in its $\mathrm{barc}_2^{\mathrm{VR}}$.

one guaranteed by Hausmann's result. This suggests that an improvement of Hausmann's results might be possible for the particular case of spheres.

Indeed, in the special case of spheres, by a refinement of Hausmann's method of proof (critically relying upon Jung's theorem) we obtain the following theorem, which also improves the aforementioned claim by Katz:

Theorem 7.1 *For any $n \in \mathbb{Z}_{>0}$, we have $\mathrm{VR}_r(\mathbb{S}^n) \simeq \mathbb{S}^n$ for any $r \in \big(0, \arccos(-1/(n+1))\big]$.*

That this result improves upon Hausmann's follows from the fact that $\arccos(-1/(n+1)) \geq \frac{\pi}{2}$ for all integers $n \geq 1$. The proof follows from the fact that with the aid of Jung's theorem, one can modify the lemmas that Hausmann originally used. See Section A.4 for a detailed proof along these lines which we believe is of independent interest.

Remark 7.2 Proposition 5.3 of [2] establishes a result analogous to Theorem 7.1 for Vietoris–Rips metric thickenings of $\mathbb{S}^n$.

Remark 7.3 The above theorem implies that for every n, $\mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n; \mathbb{F})$ contains an interval I_n of the form $(0, d_n]$ where $d_n \geq \arccos(-1/(n+1))$. This theorem does not, however, guarantee that d_n equals its lower bound, nor that I_n is the unique interval in $\mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n; \mathbb{F})$. See Figure 3 for an example of a 2–dimensional sphere (with nonround metric) having more than one interval in its 2–dimensional persistence barcode, and see Proposition 9.28 for a general result about I_n.

For the particular cases of $\mathbb{S}^1$ and $\mathbb{S}^2$, we have additional information regarding the homotopy types of their Vietoris–Rips r–complexes when r exceeds the range contemplated in the above corollary.

The case of $\mathbb{S}^1$ The complete characterization of the different homotopy types of $\mathrm{VR}_r(\mathbb{S}^1)$ as $r > 0$ grows was obtained by Adamaszek and Adams in [1]. Their proof is combinatorial in nature and takes place at the simplicial level.

Below, by invoking Theorem 4.1, we show how partial results can be obtained from the work of Katz who directly analyzed the filtration $\big(B_r(\mathbb{S}^1, L^\infty(\mathbb{S}^1))\big)_{r>0}$ via a Morse-theoretic argument.

For each integer $k \geq 1$ let $\lambda_k := 2\pi k/(2k+1)$. Katz proved in [56] that $B_r(\mathbb{S}^1, L^\infty(\mathbb{S}^1))$ changes homotopy type only when $r = \frac{1}{2}\lambda_k$ for some k. In particular, his results imply:

Corollary 7.4 *For $r \in \left(\frac{2\pi}{3}, \frac{4\pi}{5}\right)$, $\mathrm{VR}_r(\mathbb{S}^1) \simeq \mathbb{S}^3$.*

Proof $B_r(\mathbb{S}^1, L^\infty(\mathbb{S}^1))$ is homotopy equivalent to $\mathbb{S}^3$ for $r \in \left(\frac{1}{2} \cdot \frac{2\pi}{3}, \frac{1}{2} \cdot \frac{4\pi}{5}\right)$ by [56, Theorem 1.1]. Hence, the result follows from Theorem 4.1. $\square$

The case of $\mathbb{S}^2$ Similar arguments hold for the case of $\mathbb{S}^2$. Whereas the homotopy types of $\mathrm{VR}_r(\mathbb{S}^1)$ for any $r > 0$ are known [1], we are not aware of similar results for $\mathbb{S}^2$. Below, E_6 is the binary tetrahedral group.

Corollary 7.5 *For $r \in \left(\arccos\left(-\frac{1}{3}\right), \arccos(-1/\sqrt{5})\right)$, $\mathrm{VR}_r(\mathbb{S}^2) \simeq \mathbb{S}^2 * \mathbb{S}^3/E_6$.*

Proof $B_r(\mathbb{S}^2, L^\infty(\mathbb{S}^2))$ is homotopy equivalent to the topological join of $\mathbb{S}^2$ and $\mathbb{S}^3/E_6$ for r in the interval $\left(\frac{1}{2} \cdot \arccos\left(-\frac{1}{3}\right), \frac{1}{2} \cdot \arccos(-1/\sqrt{5})\right)$ by [56, Theorem 7.1]. Hence, applying Theorem 4.1 yields the result. $\square$

Remark 7.6 $\mathbb{S}^3/E_6 = \mathrm{SO}(3)/A_4$, where A_4 is the tetrahedral group; see [2, Remark 5.6].

Remark 7.7 As already pointed out in Remark 7.3, by virtue of Theorem 7.1, $(0, d_n] \in \mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n; \mathbb{F})$ for some $d_n \geq \arccos(-1/(n+1))$. Moreover, since for $n = 1$ and $n = 2$ we know (by Corollaries 7.4 and 7.5) that the homotopy type changes after $\arccos(-1/(n+1))$, we conclude that $\mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n; \mathbb{F})$ contains $\left(0, \arccos(-1/(n+1))\right]$ for $n = 1$ and $n = 2$ and that this is the unique interval in $\mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n; \mathbb{F})$ starting at 0. Surprisingly, it is currently unknown how the homotopy type of $\mathrm{VR}_r(\mathbb{S}^n)$ changes after $\arccos(-1/(n+1))$ for $n \geq 3$. But, still, in Section 9 we will be able to show that $\left(0, \arccos(-1/(n+1))\right] \in \mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n; \mathbb{F})$ for general n via arguments involving the filling radius; see Proposition 9.28. In particular, this implies that the homotopy type of $\mathrm{VR}_r(\mathbb{S}^n)$ must change after the critical point $r = \arccos(-1/(n+1))$ since the fundamental class dies after that point, even though we still do not know "how" the homotopy type changes. Moreover, since $\mathrm{VR}_r(\mathbb{S}^n)$ is homotopy equivalent to $\mathbb{S}^n$ for any $r \in \left(0, \arccos(-1/(n+1))\right]$, we know that for any interval $I \in \mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n; \mathbb{F})$ with $I \neq \left(0, \arccos(-1/(n+1))\right]$, the left endpoint of I must be greater than or equal to $\arccos(-1/(n+1))$.

The following subconjecture of [2, Conjecture 5.7] is still open except for the $n = 1$ and $n = 2$ cases; see also [2, Theorem 5.4].

Conjecture 7.8 For any $n \in \mathbb{Z}_{>0}$, there exists an $\varepsilon > 0$ such that

$$\mathrm{VR}_r(\mathbb{S}^n) \simeq \mathbb{S}^n * (\mathrm{SO}(n+1)/A_{n+2})$$

for any $r \in \left(\arccos(-1/(n+1)), \arccos(-1/(n+1)) + \varepsilon\right)$, where A_{n+2} is the alternating group of degree $n+2$.

Remark 7.9 To see that Conjecture 7.8 is a subconjecture of [2, Conjecture 5.7], observe that

$$\mathbb{S}^n * (\mathrm{SO}(n+1)/A_{n+2}) \cong \Sigma^{n+1}(\mathrm{SO}(n+1)/A_{n+2})$$

for any nonnegative integer n. It is a special case of the more general homeomorphism

$$\mathbb{S}^n * X \cong \Sigma^{n+1} X$$

for any Hausdorff and locally compact space X. This fact can be proved by induction on n and the associativity of the topological join (see [38, Lecture 2.4]).

7.2 The case of $\mathbb{CP}^n$

Partial information can be provided for the case of $\mathbb{CP}^n$ as well. First of all, recall that the complex projective line $\mathbb{CP}^1$ with its canonical metric actually coincides with the sphere $\mathbb{S}^2$. Hence, one can apply Theorem 7.1 and Corollary 7.5 to $\mathbb{CP}^1$. The following results can be derived for general $\mathbb{CP}^n$:

Corollary 7.10 *Let $\mathbb{CP}^n$ be the complex projective space with sectional curvature between $\frac{1}{4}$ and 1 with canonical metric. Then*:

(1) *There exist $\alpha_n \in \left(0, \arccos\left(-\frac{1}{3}\right)\right]$ such that $\mathrm{VR}_r(\mathbb{CP}^n)$ is homotopy equivalent to $\mathbb{CP}^n$ for any $r \in (0, \alpha_n]$.*

(2) *Let A be the space of equilateral 4–tuples in projective lines of $\mathbb{CP}^n$. Let X be the partial join of A and $\mathbb{CP}^n$ where $x \in \mathbb{CP}^n$ is joined to a tuple $a \in A$ by a line segment if x is contained in the projective line determined by a. There exists a constant $\beta_n > 0$ such that if*

$$\arccos\left(-\tfrac{1}{3}\right) < r < \arccos\left(-\tfrac{1}{3}\right) + \beta_n$$

then $\mathrm{VR}_r(\mathbb{CP}^n)$ is homotopy equivalent to X.

Proof By Hausmann's theorem [50, Theorem 3.5], there exist $\alpha_n > 0$ such that $\mathrm{VR}_r(\mathbb{CP}^n)$ is homotopy equivalent to $\mathbb{CP}^n$ for any $r \in (0, \alpha_n]$. Also, by [56, Theorem 8.1], α_n cannot be greater than $\arccos\left(-\frac{1}{3}\right)$. The second claim is a direct result of Theorem 4.1 and [56, Theorem 8.1]. □

7.3 The case of spheres with the ℓ^∞–metric

The Vietoris–Rips filtration of $\mathbb{S}^1$ with the usual geodesic metric is quite challenging to understand [1]. However, it turns out that if we change its underlying metric, the situation becomes very simple. Throughout this section, all metric spaces of interest are embedded in $(\mathbb{R}^n, \ell^\infty)$ and are endowed with the restriction of the ambient space metric. In particular, in this section, for any $n \in \mathbb{Z}_{>0}$,

(1) $\mathbb{R}^n_\infty = (\mathbb{R}^n, \ell^\infty)$,

(2) $\mathbb{D}^n_\infty := \left(\{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 \leq 1\}, \ell^\infty\right)$,

(3) $\mathbb{S}^{n-1}_\infty := \left(\{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 = 1\}, \ell^\infty\right)$,

(4) $\blacksquare_\infty^n := (\{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_i \in [-1, 1] \text{ for every } i = 1, \ldots, n\}, \ell^\infty)$,

(5) $\square_\infty^{n-1} := (\{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid (x_1, \ldots, x_n) \in \blacksquare_\infty^n \text{ and } x_i = \pm 1 \text{ for some } i = 1, \ldots, n\}, \ell^\infty)$.

Note that $\blacksquare_\infty^n$ is just the unit closed ℓ^∞–ball around the origin in $\mathbb{R}_\infty^n$ and $\square_\infty^{n-1}$ is its boundary.

The following theorem by Kılıç and Koçak is the motivation of this subsection:

Theorem 7.11 [59, Theorem 2] *Let X and Y be subspaces of $\mathbb{R}_\infty^2$. If Y contains X, is closed, geodesically convex,[3] and minimal (with respect to inclusion) with these properties, then Y is the tight span of X.*

Theorem 7.11 has a number of interesting consequences.

Lemma 7.12 $\blacksquare_\infty^2$ *is the tight span of* $\square_\infty^1$. *Moreover,*

$$B_r(\square_\infty^1, \blacksquare_\infty^2) = \begin{cases} [-1, 1]^2 \setminus [-(1-r), (1-r)]^2 & \text{if } r \in (0, 1], \\ [-1, 1]^2 & \text{if } r > 1. \end{cases}$$

Proof By Theorem 7.11, the first claim is straightforward. The second claim, namely the explicit expression of $B_r(\square_\infty^1, \blacksquare_\infty^2)$ is also obvious since we are using the ℓ^∞–norm. $\square$

Corollary 7.13 $B_r(\square_\infty^1, \blacksquare_\infty^2)$ *is homotopy equivalent to* $\mathbb{S}^1$ *for* $r \in (0, 1]$ *and contractible for* $r > 1$. *Hence, for any field* $\mathbb{F}$,

$$\mathrm{barc}_k^{\mathrm{VR}}(\square_\infty^1, \mathbb{F}) = \begin{cases} \{(0, \infty)\} & \text{if } k = 0, \\ \{(0, 2]\} & \text{if } k = 1, \\ \varnothing & \text{if } k \geq 2. \end{cases}$$

Proof Apply Lemma 7.12 and Theorem 4.1. $\square$

Interestingly, one can also prove the following result:

Lemma 7.14 $\mathbb{D}_\infty^2$ *is the tight span of* $\mathbb{S}_\infty^1$. *Moreover,*

$$B_r(\mathbb{S}_\infty^1, \mathbb{D}_\infty^2) = \mathbb{D}_\infty^2 \setminus V_r$$

for any $r > 0$, *where*

$$V_r := \bigcap_{(p,q) \in \{r, -r\}^2} \{(x, y) \in \mathbb{R}^2 \mid (x - p)^2 + (y - q)^2 \leq 1\}.$$

In particular, for $r > 1/\sqrt{2}$ *we have* $V_r = \varnothing$, *so* $B_r(\mathbb{S}_\infty^1, \mathbb{D}_\infty^2) = \mathbb{D}_\infty^2$ (*see* Figure 1).

Proof By Theorem 7.11, the first claim is straightforward.

Fix an arbitrary $(z + t, w + s) \in B_r(\mathbb{S}_\infty^1, \mathbb{D}_\infty^2)$, where $z^2 + w^2 = 1$ and $t, s \in (-r, r)$. Suppose $z \geq 0$ and $w \geq 0$. Then

$$(z + t + r)^2 + (w + s + r)^2 = z^2 + w^2 + (t + r)^2 + (s + r)^2 + 2z(t + r) + 2w(s + r) > 1$$

---

[3] That is, for any $p, q \in Y$, there exists at least one geodesic in $\mathbb{R}_\infty^2$ between p and q which is fully contained in Y.

because of the assumptions on z, w, t, and s. Therefore, $(z+t, w+s) \notin V_r$, so $(z+t, w+s) \in \mathbb{D}^2_\infty \backslash V_r$. By symmetry, the same result holds for other possible sign combinations of z and w. Hence, we have $B_r(\mathbb{S}^1_\infty, \mathbb{D}^2_\infty) \subseteq \mathbb{D}^2_\infty \backslash V_r$.

Now, fix an arbitrary $(x, y) \in \mathbb{D}^2_\infty \backslash V_r$. Since $(x, y) \notin V_r$, without loss of generality, one can assume that $(x+r)^2 + (y+r)^2 > 1$. Also, $x^2 + y^2 \le 1$ since $(x, y) \in \mathbb{D}^2_\infty$. Then, there must be some $t \in [0, r)$ such that $(x+t)^2 + (y+t)^2 = 1$. It follows that $(x, y) \in B_r(\mathbb{S}^1_\infty, \mathbb{D}^2_\infty)$. Since (x, y) is an arbitrary point in $\mathbb{D}^2_\infty \backslash V_r$ it follows that $\mathbb{D}^2_\infty \backslash V_r \subseteq B_r(\mathbb{S}^1_\infty, \mathbb{D}^2_\infty)$.

With this we conclude that $B_r(\mathbb{S}^1_\infty, \mathbb{D}^2_\infty) = \mathbb{D}^2_\infty \backslash V_r$, as we wanted. $\qquad \square$

Corollary 7.15 $B_r(\mathbb{S}^1_\infty, \mathbb{D}^2_\infty)$ *is homotopy equivalent to* $\mathbb{S}^1$ *for* $r \in (0, 1/\sqrt{2}]$ *and contractible for* $r > 1/\sqrt{2}$. *Hence, for any field* $\mathbb{F}$,

$$\mathrm{barc}^{\mathrm{VR}}_k(\mathbb{S}^1_\infty, \mathbb{F}) = \begin{cases} \{(0, \infty)\} & \text{if } k = 0, \\ \{(0, \sqrt{2}]\} & \text{if } k = 1, \\ \varnothing & \text{if } k \ge 2. \end{cases}$$

Proof Apply Lemma 7.14 and Theorem 4.1. $\qquad \square$

Moreover, it turns out that, despite the fact that Theorem 7.11 is restricted to subsets of $\mathbb{R}^2$, Lemma 7.12 can be generalized to arbitrary dimensions.

Lemma 7.16 *For any* $n \in \mathbb{Z}_{>0}$, $\blacksquare^n_\infty$ *is the tight span of* $\square^{n-1}_\infty$. *Moreover,*

$$B_r(\square^{n-1}_\infty, \blacksquare^n_\infty) = \begin{cases} [-1, 1]^n \backslash [-(1-r), 1-r]^n & \text{if } r \in (0, 1], \\ [-1, 1]^n & \text{if } r > 1. \end{cases}$$

Proof When $n \ge 3$ one cannot invoke Theorem 7.11 since it does not hold for general n; see [59, Example 5]. We will instead directly prove that $\blacksquare^n_\infty$ is the tight span of $\square^{n-1}_\infty$.

First, observe that $\blacksquare^n_\infty = \bar{B}_1(O, \mathbb{R}^n_\infty)$, where $O = (0, \dots, 0)$ is the origin, is hyperconvex by Lemma 2.17.

Therefore, in order to show that $\blacksquare^n_\infty$ is indeed the tight span of $\square^{n-1}_\infty$, it is enough to show that there is no proper hyperconvex subspace of $\blacksquare^n_\infty$ containing $\square^{n-1}_\infty$. Suppose this is not true. Then there exists a proper hyperconvex subspace X such that $\square^{n-1}_\infty \subseteq X \subsetneq \blacksquare^n_\infty$. Choose $p = (x_1, \dots, x_n) \in \blacksquare^n_\infty \backslash X$. Without loss of generality, one can assume $x_1 \ge \cdots \ge x_n$. Now, let

$$p_0 := (x_1 - (x_n + 1), x_2 - (x_n + 1), \dots, -1),$$
$$p_1 := (1, x_2 + (1 - x_1), \dots, x_n + (1 - x_1)).$$

See Figure 4. Then it is clear that $p_0, p_1 \in \square^{n-1}_\infty \subseteq X$ and

$$\|p_0 - p_1\|_\infty = (x_n + 1) + (1 - x_1) = \|p_0 - p\|_\infty + \|p - p_1\|_\infty.$$
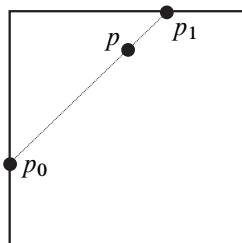
Figure 4: The points p, p_0 and p_1 in the proof of Lemma 7.16.

Therefore, since X is hyperconvex, we know that

$$\overline{B}_{\|p_0-p\|_\infty}(p_0, X) \cap \overline{B}_{\|p-p_1\|_\infty}(p_1, X) \neq \varnothing.$$

However, note that

$$\overline{B}_{\|p_0-p\|_\infty}(p_0, X) \cap \overline{B}_{\|p-p_1\|_\infty}(p_1, X) \subseteq \overline{B}_{\|p_0-p\|_\infty}(p_0, \mathbb{R}^n_\infty) \cap \overline{B}_{\|p-p_1\|_\infty}(p_1, \mathbb{R}^n_\infty) = \{p\}.$$

This means that $p \in X$, which is a contradiction; hence no such X exists. Therefore, $\blacksquare^n_\infty$ is the tight span of $\square^{n-1}_\infty$, as we required.

The second claim, namely the explicit expression of $B_r(\square^{n-1}_\infty, \blacksquare^n_\infty)$ is obvious since we are using the ℓ^∞–norm. $\qquad\square$

Corollary 7.17 *For any $n \in \mathbb{Z}_{>0}$, $B_r(\square^{n-1}_\infty, \blacksquare^n_\infty)$ is homotopy equivalent to $\mathbb{S}^{n-1}$ for $r \in (0, 1]$ and contractible for $r > 1$. Hence, for any field $\mathbb{F}$,*

$$\mathrm{barc}^{\mathrm{VR}}_k(\square^{n-1}_\infty, \mathbb{F}) = \begin{cases} \{(0, \infty)\} & \text{if } k = 0, \\ \{(0, 2]\} & \text{if } k = n-1, \\ \varnothing & \text{otherwise,} \end{cases}$$

for $n \geq 2$, and

$$\mathrm{barc}^{\mathrm{VR}}_k(\square^0_\infty, \mathbb{F}) = \begin{cases} \{(0, \infty), (0, 2]\} & \text{if } k = 0, \\ \varnothing & \text{if } k \geq 1. \end{cases}$$

Proof Apply Lemma 7.16 and Theorem 4.1. $\qquad\square$

Remark 7.18 It seems of interest to study the homotopy types of Vietoris–Rips complexes of ellipsoids with the ℓ^∞–metric; see [4].

Here, observant readers would have already noticed that we do not need to use the tight spans of $\mathbb{S}^1_\infty$ and $\square^{n-1}_\infty$ in order to apply Theorem 4.1 since $\mathbb{R}^n_\infty$ itself is an injective metric space for any $n \in \mathbb{Z}_{>0}$. In particular, the persistent homology of $\square^{n-1}_\infty$ is simpler to compute if we use $\mathbb{R}^n_\infty$ as an ambient space. However, we believe that it is worth clarifying what are the tight spans of $\mathbb{S}^1_\infty$ and $\square^{n-1}_\infty$ since the exact shape of tight spans are largely mysterious in general.

We do not know whether $\mathbb{D}^n_\infty$ is the tight span of $\mathbb{S}^{n-1}_\infty$ for general n. However, if we use $\mathbb{R}^n_\infty$ as an ambient injective metric space, we are still able to compute its persistent homology.

Theorem 7.19 *For any $n \in \mathbb{Z}_{>0}$ and $r > 0$,*

$$B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n) \simeq B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n) = \mathbb{D}_\infty^n \setminus V_{n,r},$$

where

$$V_{n,r} := \bigcap_{(p_1,\dots,p_n) \in \{r,-r\}^n} \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n \;\Big|\; \sum_{i=1}^n (x_i - p_i)^2 \le 1 \right\}.$$

In particular, for $r > 1/\sqrt{n}$ we have $V_{n,r} = \varnothing$, so $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n) = \mathbb{D}_\infty^n$. As a result, $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n)$ is homotopy equivalent to $\mathbb{S}^{n-1}$ for $r \in (0, 1/\sqrt{n}]$ and contractible for $r > 1/\sqrt{n}$ (see Figure 1 for an illustration for the case when $n = 2$).

Proof First, let's prove that $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n)$ is a deformation retract of $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n)$. Consider the map $P_n : \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 \ge 1\} \to \mathbb{S}_\infty^{n-1}$ such that $P_n(x_1, \dots, x_n)$ is the unique point of $\mathbb{S}_\infty^{n-1}$ such that $\|(x_1, \dots, x_n) - P_n(x_1, \dots, x_n)\|_\infty = \inf_{(y_1,\dots,y_n) \in \mathbb{S}_\infty^{n-1}} \|(x_1, \dots, x_n) - (y_1, \dots, y_n)\|_\infty$. Observe that it is easy (but very tedious) to prove that P_n is well-defined, continuous, and that $P_n|_{\mathbb{S}_\infty^{n-1}} = \mathrm{id}_{\mathbb{S}_\infty^{n-1}}$.

Now, for any $r > 0$, consider the homotopy

$$h_{n,r} : B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n) \times [0, 1] \to B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n),$$

$$(x_1, \dots, x_n, t) \mapsto \begin{cases} (x_1, \dots, x_n) & \text{if } (x_1, \dots, x_n) \in \mathbb{D}_\infty^n, \\ (1-t)(x_1, \dots, x_n) + t P_n(x_1, \dots, x_n) & \text{if } (x_1, \dots, x_n) \notin \mathbb{D}_\infty^n. \end{cases}$$

The only subtle point is ascertaining whether the image of this map is contained in $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n)$. For this, note that $\|(x_1, \dots, x_n) - P_n(x_1, \dots, x_n)\|_\infty < r$ by the definition of P_n and the fact that $(x_1, \dots, x_n)$ is in $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n)$. Therefore, both $(x_1, \dots, x_n)$ and $P_n(x_1, \dots, x_n)$ belong to $B_r(P_n(x_1, \dots, x_n), \mathbb{R}_\infty^n)$, so the linear interpolation is also contained in $B_r(P_n(x_1, \dots, x_n), \mathbb{R}_\infty^n) \subset B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n)$. Hence, one can conclude that $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n)$ is a deformation retract of $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n)$.

Next, let's prove that $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n) = \mathbb{D}_\infty^n \setminus V_{n,r}$. Fix an arbitrary $(z_1 + t_1, \dots, z_n + t_n) \in B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n)$ where $\sum_{i=1}^n z_i^2 = 1$ and $t_i \in (-r, r)$ for all $i = 1, \dots, n$. Consider the case of $z_i \ge 0$ for all $i = 1, \dots, n$. Then

$$\sum_{i=1}^n (z_i + t_i + r)^2 = \sum_{i=1}^n z_i^2 + \sum_{i=1}^n (t_i + r)^2 + \sum_{i=1}^n 2z_i(t_i + r) > 1$$

by the assumptions on $\{z_i\}_{i=1}^n$ and $\{t_i\}_{i=1}^n$. Therefore, $(z_1 + t_1, \dots, z_n + t_n) \notin V_{n,r}$, so

$$(z_1 + t_1, \dots, z_n + t_n) \in \mathbb{D}_\infty^n \setminus V_{n,r}.$$

By symmetry, the same result holds for other possible sign combinations of the z_i. Hence, we have $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n) \subseteq \mathbb{D}_\infty^n \setminus V_{n,r}$.

Now, fix arbitrary $(x_1 \dots, x_n) \in \mathbb{D}_\infty^n \setminus V_{n,r}$. Since $(x_1, \dots, x_n) \notin V_{n,r}$, without loss of generality, one can assume that $\sum_{i=1}^n (x_i + r)^2 > 1$. Also, $\sum_{i=1}^n x_i^2 \le 1$ since $(x_1, \dots, x_n) \in \mathbb{D}_\infty^n$. Then, there must

be some $t \in [0, r)$ such that $\sum_{i=1}^{n}(x_i + t)^2 = 1$. It follows that $(x_1, \ldots, x_n) \in B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n)$. Since $(x_1, \ldots, x_n)$ is an arbitrary point in $\mathbb{D}_\infty^n \setminus V_{n,r}$ it follows that $\mathbb{D}_\infty^n \setminus V_{n,r} \subseteq B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n)$.

With this we conclude that $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{D}_\infty^n) = \mathbb{D}_\infty^n \setminus V_{n,r}$. $\qquad\square$

Corollary 7.20 *For any $n \geq 2$, $B_r(\mathbb{S}_\infty^{n-1}, \mathbb{R}_\infty^n)$ is homotopy equivalent to $\mathbb{S}^{n-1}$ for $r \in (0, 1/\sqrt{n}]$ and contractible for $r > 1/\sqrt{n}$. Hence, for any field $\mathbb{F}$,*

$$\mathrm{barc}_k^{\mathrm{VR}}(\mathbb{S}_\infty^{n-1}, \mathbb{F}) = \begin{cases} \{(0, \infty)\} & \text{if } k = 0, \\ \{(0, 2/\sqrt{n}]\} & \text{if } k = n-1, \\ \varnothing & \text{otherwise.} \end{cases}$$

Proof Apply Theorems 7.19 and 4.1. $\qquad\square$

8 Application: hyperbolicity and persistence

One can reap benefits from the fact that one can choose *any* metric homotopy pairing in the statement of Theorem 4.1, not just the Kuratowski functor.

In this section, we will see one such example which arises from the interplay between the *hyperbolicity* of the geodesic metric space X and its tight span $E(X)$ (see Example 3.9 to recall the definition of tight span).

We first recall the notion of hyperbolicity.

Definition 8.1 (δ–hyperbolicity) A metric space (X, d_X) is called δ–hyperbolic, for some constant $\delta \geq 0$, if

$$d_X(w, x) + d_X(y, z) \leq \max\{d_X(w, y) + d_X(x, z), d_X(x, y) + d_X(w, z)\} + \delta$$

for all quadruples of points $w, x, y, z \in X$. If a metric space is δ–hyperbolic for some $\delta \geq 0$, it is said to be hyperbolic.

The *hyperbolicity* $\mathrm{hyp}(X)$ of X is defined as the infimal $\delta \geq 0$ such that X is δ–hyperbolic. A metric space is said to be *hyperbolic* if $\mathrm{hyp}(X)$ is finite.

For a more concrete development on the geometry of hyperbolic metric spaces and its applications (especially to group theory), see [14; 47].

Example 8.2 Here are some examples of hyperbolic spaces:

(1) Metric trees are 0–hyperbolic spaces.

(2) All compact Riemannian manifolds are trivially hyperbolic spaces. More interestingly, among unbounded manifolds, Riemannian manifolds with strictly negative sectional curvature are hyperbolic spaces. Observe that "strictly negative" sectional curvature is a necessary condition (for example, consider the Euclidean plane $\mathbb{R}^2$).

The following proposition guarantees that the tight span $E(X)$ preserves the hyperbolicity of the underlying space X with controlled distortion:

Proposition 8.3 [60, Proposition 1.3] *If X is a δ–hyperbolic geodesic metric space for some $\delta \geq 0$, then its tight span $E(X)$ is also δ–hyperbolic. Moreover,*

$$B_r(X, E(X)) = E(X)$$

for any $r > \delta$.

Remark 8.4 Since X embeds isometrically into $E(X)$, the above implies that

$$\mathrm{hyp}(E(X)) = \mathrm{hyp}(X).$$

The following corollary was already established by Gromov (who attributes it to Rips) in [47, Lemma 1.7.A]. The proof given by Gromov operates at the simplicial level. By invoking Proposition 8.3 we obtain an alternative proof, which instead of operating the simplicial level, exploits the isometric embedding of X into its tight span $E(X)$ (which is a compact contractible space).

Corollary 8.5 *If X is a hyperbolic geodesic metric space, then $\mathrm{VR}_{2r}(X)$ is contractible for any $r > \mathrm{hyp}(X)$.*

Proof Choose an arbitrary $r > \mathrm{hyp}(X)$. Then, there is $\delta \in [\mathrm{hyp}(X), r)$ such that X is δ–hyperbolic.

By Proposition 2.27, $\mathrm{VR}_{2r}(X)$ is homotopy equivalent to $B_r(X, E(X))$. But, by Proposition 8.3, $B_r(X, E(X)) = E(X)$. Since $E(X)$ is contractible by Corollary 2.21, $\mathrm{VR}_{2r}(X)$ is contractible. □

As a consequence one can bound the length of intervals in the persistence barcode of hyperbolic spaces.

Corollary 8.6 *If X is a hyperbolic geodesic metric space, then for any $k \geq 1$ and $I = (u, v] \in \mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$, we have $v \leq 2\,\mathrm{hyp}(X)$. In particular, $\mathrm{length}(I) \leq 2\,\mathrm{hyp}(X)$.*

Proof Apply Corollary 8.5. □

Observe that metric trees are both 0–hyperbolic and hyperconvex. A recent paper by Joharinad and Jost [53] analyzes the persistent homology of metric spaces satisfying the hyperconvexity condition (which is equivalent to injectivity) as well as that of spaces satisfying a relaxed version of hyperconvexity.

9 Application: the filling radius, spread, and persistence

In this section, we recall the notions of spread and filling radius, as well as their relationship. In particular, we prove a number of statements about the filling radius of a closed connected manifold. Moreover, we consider a generalization of the filling radius and also define a strong notion of filling radius which is akin to the so-called *maximal persistence* in the realm of topological data analysis.

9.1 Spread

We recall a metric concept called *spread*. The following definition is a variant of the one given in [54, Lemma 1]:

Definition 9.1 (*N*–spread) For any integer $N \in \mathbb{Z}_{>0}$, the Nth *spread* $\text{spread}_N(X)$ of a metric space (X, d_X) is the infimal $r > 0$ such that there exists a subset A of X with cardinality at most N such that

- $\text{diam}(A) < r$,
- $\sup_{x \in X} \inf_{a \in A} d_X(x, a) < r$.

Finally, the *spread* of X is defined to be $\text{spread}(X) := \inf_N \text{spread}_N(X)$, ie the set A is allowed to have arbitrary (finite) cardinality.

Remark 9.2 Recall that the *radius* of a compact metric space (X, d_X) is

$$\text{rad}(X) := \inf_{p \in X} \max_{x \in X} d_X(p, x).$$

Thus, $\text{rad}(X) = \text{spread}_1(X)$.

Remark 9.3 (the spread of spheres) Katz proves in [54, Theorem 2] that for all integers $n \geq 1$,

$$\text{spread}(\mathbb{S}^n) = \arccos\Big(\frac{-1}{n+1}\Big).$$

For example, $\text{spread}(\mathbb{S}^1) = \frac{2\pi}{3}$. Notice that $\text{spread}(\mathbb{S}^m) \geq \text{spread}(\mathbb{S}^n) \geq \frac{\pi}{2}$ for $m \leq n$. Katz's proof actually yields that

$$\text{spread}_{n+2}(\mathbb{S}^n) = \text{spread}(\mathbb{S}^n)$$

for each n.

9.2 Bounding barcode length via spread

Let (X, d_X) be a compact metric space. Recall that for each integer $k \geq 0$, $\text{barc}_k^{\text{VR}}(X; \mathbb{F})$ denotes the persistence barcode associated to $\text{PH}_k(\text{VR}_*(X); \mathbb{F})$, the kth persistent homology induced by the Vietoris–Rips filtration of X (see Section 2.1).

The following lemma is due to Katz [54, Lemma 1]:

Lemma 9.4 *Let (X, d_X) be a compact metric space. Then, for any $\delta > \frac{1}{2}\text{spread}(X)$, there exists a contractible space U such that $X \subseteq U \subseteq B_\delta(X, L^\infty(X))$.*

Remark 9.5 Via the isomorphism theorem, Katz's lemma implies the fact that whenever $I = (0, v] \in \text{barc}_*^{\text{VR}}(X)$, we have $v \leq \text{spread}(X)$. The lemma does not permit bounding the length of intervals whose left endpoint is strictly greater than zero.

It turns out that we can prove a general version of Lemma 9.4 for closed s–thickenings $\bar{B}_s(X, L^\infty(X))$ for any $s \geq 0$.

Lemma 9.6 Let (X, d_X) be a compact metric space. Then, for any $s \geq 0$ and $\delta > \frac{1}{2}\,\mathrm{spread}(X)$, there exists a contractible space $U_{s,\delta}$ such that $\bar{B}_s(X, L^\infty(X)) \subseteq U_{s,\delta} \subseteq B_{s+\delta}(X, L^\infty(X))$.

Note that Lemma 9.4 can be obtained from the case $s = 0$ in Lemma 9.6. We provide a detailed self-contained proof of this general version in Section 9.2.2.

Armed with Lemma 9.6 and Theorem 4.1, one immediately obtains item (1) in the proposition below:

Proposition 9.7 Let (X, d_X) be a compact metric space, $k \geq 1$, and let I be any interval in $\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$. Then

(1) $\mathrm{length}(I) \leq \mathrm{spread}(X)$, and

(2) if $I = (u, v]$ for some $0 < u < v$, then $v \leq \mathrm{spread}_1(X)$.

Remark 9.8 Item (2) of the proposition above implies that the right endpoint of any interval I (often referred to as the *death time* of I) cannot exceed the radius $\mathrm{rad}(X)$ of X; see Remark 9.2.

Note that by [54, Section 1], when X is a geodesic space (eg a Riemmanian manifold),

$$\mathrm{spread}(X) \leq \tfrac{2}{3}\,\mathrm{diam}(X).$$

This means that we have the following universal bound on the length of intervals in the Vietoris–Rips persistence barcode of a geodesic space X:

Corollary 9.9 (bound on length of bars of geodesic spaces) Let X be a compact geodesic space. Then, for any $k \geq 1$ and any $I \in \mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$,

$$\mathrm{length}(I) \leq \tfrac{2}{3}\,\mathrm{diam}(X).$$

Remark 9.10
- For $k = 1$, $\mathbb{S}^1$ achieves equality in the corollary above. Indeed, this follows from [1] since the longest interval in $\mathrm{barc}_k^{\mathrm{VR}}(\mathbb{S}^1)$ corresponds to $k = 1$ and is exactly $\left(0, \frac{2\pi}{3}\right]$.

- Since $\mathrm{VR}_r(X)$ is contractible for any $r > \mathrm{diam}(X)$, it is clear that $\mathrm{length}(I) \leq \mathrm{diam}(X)$ in general. The corollary above improves this bound by a factor of $\frac{2}{3}$ when X is geodesic.

- In [54], Katz proves that the filling radius of a manifold is bounded above by $\frac{1}{3}$ of its diameter. Our result is somewhat more general than Katz's in two senses: his claim applies to Riemannian manifolds M and only provides information about the interval induced by the fundamental class of the manifold (see Proposition 9.28). In contrast, Corollary 9.9 applies to any compact geodesic space and in this case it provides the same upper bound for the length any interval in $\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$, for any k.

- Besides the proof via Lemma 9.6 and Theorem 4.1 explained above, we provide an alternative direct proof of Proposition 9.7 via simplicial arguments. We believe each proof is interesting in its own right.

Proof of Proposition 9.7 via simplicial arguments Let $\delta > \text{spread}(X)$. It is enough to show that for each $s > 0$, the map

$$H_k(\text{VR}_s(X); \mathbb{F}) \to H_k(\text{VR}_{\delta+s}(X); \mathbb{F})$$

induced by the inclusion is zero. By the definition of spread, we know that there is a nonempty finite subset $A \subseteq X$ such that

- $\text{diam}(A) < \delta$,
- $\sup_{x \in X} \inf_{a \in A} d_X(x, a) < \delta$.

Note that then $H_k(\text{VR}_\delta(A); \mathbb{F}) = 0$ because $\text{VR}_\delta(A)$ is a simplex. Let $\pi \colon X \to A$ be a map sending x to a closest point in A. Then $d_X(x, \pi(x)) < \delta$ for any $x \in X$ because of the second property of A (moreover, $\pi(x) = x$ if $x \in A$). Observe that, since $\text{diam}(\pi(\sigma)) < \delta$ for any simplex $\sigma \in \text{VR}_s(X)$ by the first property of A, this map π induces a simplicial map from $\text{VR}_s(X)$ to $\text{VR}_\delta(A)$. Hence, one can construct a composite map ν from $\text{VR}_s(X)$ to $\text{VR}_{\delta+s}(X)$,

$$\text{VR}_s(X) \xrightarrow{\pi} \text{VR}_\delta(A) \hookrightarrow \text{VR}_\delta(X) \hookrightarrow \text{VR}_{\delta+s}(X),$$

where the second and third maps are induced by the canonical inclusions. Observe that this composition of maps induces a map from $H_k(\text{VR}_s(X))$ to $H_k(\text{VR}_{\delta+s}(X))$, and this induced map is actually the zero map since $H_k(\text{VR}_\delta(A); \mathbb{F}) = 0$. So, it is enough to show that the composite map ν is contiguous to the canonical inclusion $\text{VR}_s(X) \hookrightarrow \text{VR}_{\delta+s}(X)$. Let $\sigma = \{x_0, \ldots, x_n\}$ be a subset of X with diameter strictly less than s. Let $a_i := \pi(x_i)$ for $i = 0, 1, \ldots, n$. Then

$$d_X(x_i, a_j) \le d_X(x_i, x_j) + d_X(x_j, a_j) < \delta + s.$$

Hence the diameter of the subspace $\{x_1, \ldots, x_k, a_1, \ldots, a_k\}$ is strictly less than $\delta + s$. This shows the desired contiguity and completes the proof. The proof of (2) follows similar (but simpler) steps and thus we omit it. $\square$

Remark 9.11 Whereas the proof of Lemma 1 in [54] takes place at the level of $L^\infty(X)$, the proof of Proposition 9.7 given above takes place at the level of simplicial complexes and simplicial maps.

9.2.1 Bounds based on localization of spread One can improve Proposition 9.7 by considering a localized version of spread. Note that, in [6], Adams and Coskunuzer also built some bounds on the length of barcodes based on certain notions of size of homology classes.

For an integer $k \ge 0$, a given field $\mathbb{F}$, and a metric space X, let

$$\text{Spec}_k(X, \mathbb{F}) := \bigcup_{r > 0} \left(H_k(\text{VR}_r(X); \mathbb{F}) \setminus \{0\} \times \{r\} \right)$$

be the kth Vietoris–Rips *homological spectrum* of X (with coefficients in $\mathbb{F}$). Note that we only consider nonzero elements of $H_k(\text{VR}_r(X); \mathbb{F})$ in the definition $\text{Spec}_k(X, \mathbb{F})$ to avoid trivial cases (there can be no positive length bars associated to a zero element).

Example 9.12 Consider $X = \{0, 1\}$ equipped with the metric inherited from $\mathbb{R}$. Then, for any field $\mathbb{F}$,

$$\mathrm{Spec}_0(X, \mathbb{F}) = \left(\bigcup_{r \in (0,1]} \mathrm{Span}_{\mathbb{F}}(\{\mu_r, \nu_r\}) \times \{r\} \right) \cup \left(\bigcup_{r > 1} \mathrm{Span}_{\mathbb{F}}(\{\omega_r\}) \times \{r\} \right),$$

where μ_r and $\nu_r \in \mathrm{H}_0(\mathrm{VR}_r(X); \mathbb{F}) \setminus \{0\}$ are the homology classes homologous to 0 and 1, respectively, for $r \in (0, 1]$, and $\omega_r \in \mathrm{H}_0(\mathrm{VR}_r(X); \mathbb{F}) \setminus \{0\}$ is the homology class homologous to both 0 and 1 for $r > 1$ (ie $\omega_r = (i_{r',r})_*(\mu_{r'}) = (i_{r',r})_*(\nu_{r'})$ for any $r' \in (0, 1]$ and $r > 1$).

Definition 9.13 (prelocalized spread of a homology class) For each $(\omega, s) \in \mathrm{Spec}_k(X, \mathbb{F})$ we define the *prelocalized spread* of (ω, s) as

$$\mathrm{pspread}(X; \omega, s) := \inf_{B \in S(\omega,s)} \mathrm{spread}(B),$$

where $S(\omega, s)$ denotes the collection of all $B \subseteq X$ such that $\omega = \iota_*([c])$, c is a simplicial k–cycle on $\mathrm{VR}_s(B)$, and $\iota \colon B \hookrightarrow X$ is the canonical inclusion.

Any B as in the definition above will be said to *support* the homology class $(\omega, s) \in \mathrm{Spec}_k(X, \mathbb{F})$.

Lemma 9.14 *Suppose* $(\omega, s) \in \mathrm{Spec}_k(X, \mathbb{F})$ *and* $k \geq 1$ *are given. Then for any* $\delta > \mathrm{pspread}(X; \omega, s)$,

$$(i_{s,(s+\delta)})_*(\omega) = 0,$$

where $i_{s,(s+\delta)} \colon \mathrm{VR}_s(X) \hookrightarrow \mathrm{VR}_{s+\delta}(X)$ *is the canonical inclusion.*

Proof By the definition of $\mathrm{pspread}(X; \omega, s)$, there exists $B \subseteq X$ such that $\omega = \iota_*([c])$ where c is a simplicial k–cycle on $\mathrm{VR}_s(B)$ and $\mathrm{spread}(B) < \delta$. Then, as in the proof of Proposition 9.7, one can prove that

$$(j_{s,(s+\delta)})_* \colon \mathrm{H}_k(\mathrm{VR}_s(B); \mathbb{F}) \to \mathrm{H}_k(\mathrm{VR}_{s+\delta}(B); \mathbb{F})$$

is the zero map, where $j_{s,(s+\delta)} \colon \mathrm{VR}_s(B) \hookrightarrow \mathrm{VR}_{s+\delta}(B)$ is the canonical inclusion. Hence,

$$(j_{s,(s+\delta)})_*([c]) = 0.$$

Furthermore, note that the diagram

$$
\begin{array}{ccc}
\mathrm{H}_k(\mathrm{VR}_s(B); \mathbb{F}) & \xrightarrow{(j_{s,(s+\delta)})_*} & \mathrm{H}_k(\mathrm{VR}_{s+\delta}(B); \mathbb{F}) \\
{\scriptstyle \iota_*} \downarrow & & \downarrow {\scriptstyle \iota_*} \\
\mathrm{H}_k(\mathrm{VR}_s(X); \mathbb{F}) & \xrightarrow{(i_{s,(s+\delta)})_*} & \mathrm{H}_k(\mathrm{VR}_{s+\delta}(X); \mathbb{F})
\end{array}
$$

commutes, where all the arrows are maps induced by canonical inclusions. Hence, one can conclude $(i_{s,(s+\delta)})_*(\omega) = 0$ as we required. $\square$

Now, fix an arbitrary $(\omega, s) \in \mathrm{Spec}_k(X, \mathbb{F})$. Then, let

$$u_{(\omega,s)} := \inf\{r > 0 \mid r \leq s \text{ and } \exists \text{ nonzero } \omega_r \in \mathrm{H}_k(\mathrm{VR}_r(X); \mathbb{F}) \text{ such that } (i_{r,s})_*(\omega_r) = \omega\},$$

$$v_{(\omega,s)} := \sup\{t > 0 \mid t \geq s \text{ and } \exists \text{ nonzero } \omega_t \in \mathrm{H}_k(\mathrm{VR}_t(X); \mathbb{F}) \text{ such that } (i_{s,t})_*(\omega) = \omega_t\}.$$
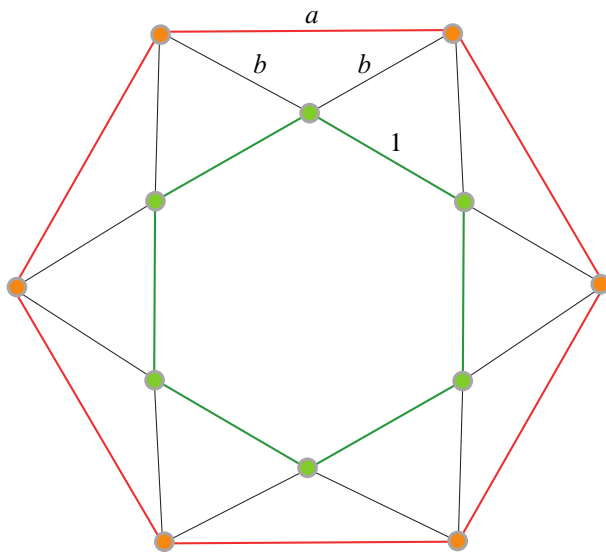
Figure 5: See Example 9.16. In this example, X is the set of vertices of the above metric graph (in green and orange) where $1 < a < b < 2$. Then, $\mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F}) = \{(1, 2], (a, b]\}$, while one can choose $(\omega, s) \in \mathrm{Spec}_1(X, \mathbb{F})$ such that $I_{(\omega,s)} = (a, 2] \notin \mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F})$.

With an argument similar to the one used in Section 5, one can prove that $u_{(\omega,s)} < s \leq v_{(\omega,s)}$. Let

$$I_{(\omega,s)} := \begin{cases} (u_{(\omega,s)}, v_{(\omega,s)}] & \text{if } v_{(\omega,s)} < \infty, \\ (u_{(\omega,s)}, \infty) & \text{otherwise.} \end{cases}$$

Intuitively, the interval $I_{(\omega,s)}$ is the maximal (left open, right closed) interval containing s inside which the class ω can be "propagated".

Remark 9.15 If $(\omega, s), (\omega', s') \in \mathrm{Spec}_k(X, \mathbb{F})$, $s \leq s'$, and $\omega' = (i_{s,s'})_*(\omega)$, then $v_{(\omega,s)} = v_{(\omega',s')}$ and $u_{(\omega,s)} \geq u_{(\omega',s')}$. Furthermore, if $(i_{s,s'})_*$ is injective, then $u_{(\omega,s)} = u_{(\omega',s')}$ so $I_{(\omega,s)} = I_{(\omega',s')}$.

Example 9.16 In general, $I_{(\omega,s)}$ is not necessarily one of the intervals in $\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$. Here is a brief sketch of how to construct such an example. Consider the metric graph consisting of 12 vertices and 24 edges as shown in Figure 5. Assume that the length of the edge between adjacent inner (green) vertices is 1, the length of the edge between adjacent outer (orange) vertices is a, and the length of the edge between adjacent inner and outer vertices is b where $1 < a < b < 2$. Now, let X be the set of vertices of this graph, and let d_X be the shortest path metric between them. Then one can easily check that $\mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F}) = \{(1, 2], (a, b]\}$, where $(1, 2]$ is associated to the homology class induced by the inner cycle and $(a, b]$ is associated to the homology class induced by the outer cycle. Now, if we choose $\{(\omega_s, s)\}_{s \in (a,b]} \subset \mathrm{Spec}_1(X, \mathbb{F})$ corresponding to the interval $(a, b] \in \mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F})$, then $I_{(\omega_s,s)} = (a, 2] \notin \mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F})$ for $s \in (a, b]$.

Despite the above, in the extended (arXiv) version of this paper (see [62, Proposition 9.2]), we prove that, for all $r < s$, the multiplicity of the interval $(r, s]$ in the barcode $\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$ is equal to the maximal

nonnegative integer m for which there exist linearly independent vectors $\omega_1, \ldots, \omega_m \in H_k(VR_s(X); \mathbb{F})$ such that $I_{(\omega_i,s)} = (r, s]$ for all i and any nonzero linear combination of the ω_i does not belong to $Im((i_{r,s})_*)$.

Definition 9.17 (localized spread of a homology class) For each $(\omega, s) \in \mathrm{Spec}_k(X, \mathbb{F})$, we define the *localized spread* of (ω, s) as

$$\mathrm{spread}(X; \omega, s) := \sup\{\mathrm{pspread}(X; \omega', s') \mid s' \leq s \text{ and } \omega = (i_{s',s})_*(\omega')\}.$$

Remark 9.18 It is easy to check that both $\mathrm{pspread}(X; \omega, s)$ and $\mathrm{spread}(X; \omega, s)$ are always bound above by $\mathrm{spread}(X)$.

The following Proposition 9.19 is the "localized" version of Proposition 9.7 we promised in the beginning of this section:

Proposition 9.19 *Let (X, d_X) be a compact metric space and $k \geq 1$. Then for any $(\omega, s) \in \mathrm{Spec}_k(X, \mathbb{F})$,*

$$\mathrm{length}(I_{(\omega,s)}) \leq \mathrm{spread}(X; \omega, s).$$

Proof Fix an arbitrary $\delta > \mathrm{spread}(X; \omega, s)$ and $s' \in (u_{(\omega,s)}, s]$. Then there exists $\omega' \in H_k(VR_{s'}(X); \mathbb{F})$ such that $\omega = (i_{s',s})_*(\omega')$. Hence, by Lemma 9.14, $(i_{s',(s'+\delta)})_*(\omega') = 0$. This indicates $v_{(\omega,s)} < s' + \delta$. Since the choice of δ and s' are arbitrary, one can conclude

$$\mathrm{length}(I_{(\omega,s)}) = v_{(\omega,s)} - u_{(\omega,s)} \leq \mathrm{spread}(X; \omega, s). \qquad \square$$

For an arbitrary $I \in \mathrm{barc}_k^{VR}(X; \mathbb{F})$, a family of nonzero homology classes $\{(\omega_s, s)\}_{s \in I} \subseteq \mathrm{Spec}_k(X, \mathbb{F})$ such that $(i_{s,s'})_*(\omega_s) = \omega_{s'}$ for any $s \leq s'$ in I where $i_{s,s'} : VR_s(X) \hookrightarrow VR_{s'}(X)$ is the canonical inclusion, will be said to *correspond* to I if there is an isomorphism

$$\Phi_* : \mathrm{PH}_k(VR_*(X); \mathbb{F}) \to \bigoplus_{I \in \mathrm{barc}_k^{VR}(X; \mathbb{F})} I_{\mathbb{F}}$$

such that

$$\begin{array}{ccc}
\mathrm{Span}_{\mathbb{F}}(\{\omega_s\}) & \xrightarrow{(i_{s,s'})_*} & \mathrm{Span}_{\mathbb{F}}(\{\omega_{s'}\}) \\
\Phi_s \downarrow & & \downarrow \Phi_{s'} \\
\mathbb{F} & \xrightarrow{\mathrm{id}} & \mathbb{F}
\end{array}$$

Observe that Theorem 2.9 guarantees that at least one such family of nonzero homology classes $\{(\omega_s, s)\}_{s \in I}$ always exists.

Remark 9.20 Now, given an arbitrary $I \in \mathrm{barc}_k^{VR}(X; \mathbb{F})$, there is a family of nonzero homology classes $\{(\omega_s, s)\}_{s \in I} \subseteq \mathrm{Spec}_k(X, \mathbb{F})$ corresponding to I as described above. Then obviously $I \subseteq I_{(\omega_s,s)}$ for each $s \in I$. Hence,

$$\mathrm{length}(I) \leq \inf_{s \in I} \mathrm{length}(I_{(\omega_s,s)}) \leq \inf_{s \in I} \mathrm{spread}(X; \omega_s, s) \leq \mathrm{spread}(X),$$

so one recovers the result in Proposition 9.7. Below we show some examples that highlight cases in which the localized spread is more efficient at estimating the length of bars than its global counterpart.

Example 9.21 Here are some applications of the notion of localized spread.

Let X be a compact metric space. If for a given $I \in \mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$ a corresponding family $\{(\omega_s, s)\}_{s \in I} \subseteq \mathrm{Spec}_k(X, \mathbb{F})$ is supported by a subset $B \subseteq X$, then

(2) $$\mathrm{length}(I) \leq \inf_{s \in I} \mathrm{length}(I_{(\omega_s, s)}) \leq \inf_{s \in I} \mathrm{spread}(X; \omega_s, s) \leq \mathrm{spread}(B),$$

where the first inequality holds as in Remark 9.20, the second inequality holds by Proposition 9.19, and the last inequality follows from Remark 9.18.

Here are three scenarios in which the estimate in inequality (2) is useful:

(1) Suppose a closed Riemannian manifold M and a nonzero homology class $\omega \in \mathrm{H}_1(M; \mathbb{F})$ are given. Also, let $B \subseteq M$ be the shortest loop representing ω. Recall that there is an interval $I \in \mathrm{barc}_1^{\mathrm{VR}}(M; \mathbb{F})$ associated to ω; see Proposition 9.46. Then

$$\mathrm{length}(I) \leq \mathrm{spread}(B) = \tfrac{1}{3}\, \mathrm{length}(B)$$

by inequality (2) and Remark 9.3. Actually, $I = \left(0, \tfrac{1}{3}\, \mathrm{length}(B)\right]$; see [44; 81, Theorem 8.10].

(2) Let X be the metric gluing of a loop of length l_2 and an interval of length l_1 (glued to the circle at one of its endpoints). Then, by Proposition 9.7, $I \leq \mathrm{spread}(X)$ for any $I \in \mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$. However, observe that one can make $\mathrm{spread}(X)$ arbitrarily large by increasing l_1. But, if $J \in \mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F})$ and a family of nonzero homology classes $\{(\omega_s, s)\}_{s \in J} \subseteq \mathrm{Spec}_1(X, \mathbb{F})$ corresponding to J is supported by the loop, then

$$\mathrm{length}(J) \leq \mathrm{spread}(B) = \tfrac{1}{3} l_2$$

by inequality (2) and Remark 9.3. Again, as in the first item, $J = \left(0, \tfrac{1}{3} l_2\right]$. Note that the existence of the interval $\left(0, \tfrac{1}{3} l_2\right]$ in $\mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F})$ can also be proved via the "crushing" technique introduced by Hausmann (see [50, Proposition 2.2]) since X can be crushed onto the loop of length l_2.

(3) An example similar to the one described in the previous item arises from Figure 3. Consider the tube connecting the two blobs to be large: in that case the standard spread of the space will be large yet the lifetime of the individual H_2 classes will be much smaller.

9.2.2 The proof of Lemma 9.6 Let us introduce a technical tool for this subsection. It is easy to check that the usual linear interpolation in $L^\infty(X)$ gives a geodesic bicombing on $L^\infty(X)$ satisfying all three properties mentioned in Lemma 2.20. However, in [54], Katz introduced an alternative way to construct a geodesic bicombing on $L^\infty(X)$:

Definition 9.22 (Katz's geodesic bicombing) Let X be a compact metric space. We define the *Katz geodesic bicombing* γ_K on $L^\infty(X)$ by

$$\gamma_K : L^\infty(X) \times L^\infty(X) \times [0, 1] \to L^\infty(X), \quad (f, g, t) \mapsto \gamma_K(f, g, t),$$

where
$$\gamma_K(f, g, t): X \to \mathbb{R}, \qquad x \mapsto \begin{cases} \max\{f(x) - t\|f - g\|_\infty, g(x)\} & \text{if } f(x) \geq g(x), \\ \min\{f(x) + t\|f - g\|_\infty, g(x)\} & \text{if } f(x) \leq g(x). \end{cases}$$

In other words, $\gamma_K(f, g, \cdot)$ moves from f to g with the same speed at every point.

The following proposition establishes that γ_K is indeed a (continuous) geodesic bicombing, amongst other properties. The proof is relegated to Section A.1.

Proposition 9.23 *Let X be a compact metric space. Then, for any $f, g, h \in L^\infty(X)$ and $0 \leq s \leq t \leq 1$, the Katz geodesic bicombing γ_K on $L^\infty(X)$ satisfies*

(1) $\gamma_K(f, g, 0) = f$ and $\gamma_K(f, g, 1) = g$;

(2) $\|\gamma_K(f, g, s) - \gamma_K(f, g, t)\|_\infty = (t - s) \cdot \|f - g\|_\infty$;

(3) $\|\gamma_K(f, g, t) - \gamma_K(h, g, t)\|_\infty \leq 2\|f - h\|_\infty$;

(4) $\|\gamma_K(f, g, t) - \gamma_K(f, h, t)\|_\infty \leq \|g - h\|_\infty$;

(5) $\gamma_K(\phi, \psi, \lambda) = \gamma_K(f, g, (1 - \lambda)s + \lambda t)$ where $\phi = \gamma_K(f, g, s)$ and $\psi = \gamma_K(f, g, t)$ for any $\lambda \in [0, 1]$ (*this property is called **consistency***);

(6) $\|\gamma_K(f, g, r) - h\|_\infty \leq \max\{\|\gamma_K(f, g, s) - h\|_\infty, \|\gamma_K(f, g, t) - h\|_\infty\}$ for any $r \in [s, t]$.

Properties (2), (3), and (4) of Proposition 9.23 imply the continuity of the Katz geodesic bicombing. In contrast, this bicombing is neither conical nor reversible; see Section A.2 in the appendix.

Proof of Lemma 9.6   By the definition of spread, we know that there is a nonempty finite subset $A \subseteq X$ and $\delta' \in (0, \delta)$ such that $\text{diam}(A) < 2\delta'$ and $\sup_{x \in X} \inf_{a \in A} d_X(x, a) < 2\delta'$.

Next, we define
$$f: X \to \mathbb{R}, \quad x \mapsto d_X(x, A) + \delta'.$$

The main strategy of the proof is depicted in Figure 6.

Claim 1   *For any $a \in A$, $\|d_X(a, \cdot) - f\|_\infty = \delta'$.*

Proof   To prove this, fix arbitrary $x \in X$. Note that
$$d_X(a, x) - f(x) = d_X(a, x) - d_X(x, A) - \delta'.$$

Since $d_X(x, A) \leq d_X(a, x)$, we have $-\delta' \leq d_X(a, x) - d_X(x, A) - \delta'$. Also, because the diameter of A is smaller than $2\delta'$, we have $d_X(a, x) - d_X(x, A) - \delta' < \delta'$. Therefore, $|d_X(a, x) - f(x)| \leq \delta'$. Furthermore, if we put $x = a$, we have that $\|d_X(a, \cdot) - f\|_\infty = \delta'$.   □

Now, let
$$U_{s,\delta} := \{\gamma_K(g, f, t) \mid g \in \bar{B}_s(X, L^\infty(X)), t \in [0, 1]\}.$$

Then $U_{s,\delta}$ obviously contains $\bar{B}_s(X, L^\infty(X))$ and can be contracted to the point f.
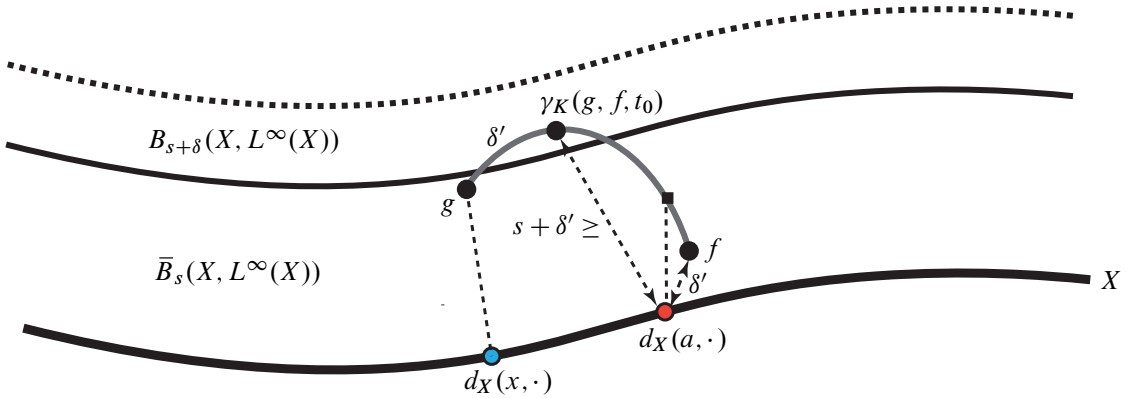
Figure 6: Strategy of the proof of Lemma 9.6. By construction, the distance between f and $d_X(a, \cdot)$ (represented by the red dot) is δ' and the distance between $\gamma_K(g, f, t_0)$ and $d_X(a, \cdot)$ is less than or equal to $s + \delta'$. Hence, by item (6) of Proposition 9.23, we have that the point represented by a square will be at distance at most $s + \delta'$ from $d_X(a, \cdot)$.

The lemma will follow once we establish the following claim:

Claim 2
$$U_{s,\delta} \subseteq B_{s+\delta}(X, L^\infty(X)).$$

Proof To see this, fix an arbitrary $g \in \bar{B}_s(X, L^\infty(X))$ and $t \in [0, 1]$. Note that one can choose $x \in X$ such that $\|g - d_X(x, \cdot)\|_\infty \leq s$.

- If $\|g - f\|_\infty \leq \delta'$, then

$$\|\gamma_K(g, f, t) - d_X(x, \cdot)\|_\infty \leq \|\gamma_K(g, f, t) - g\|_\infty + \|g - d_X(x, \cdot)\|_\infty \leq s + \delta' < s + \delta$$

by the triangle inequality and properties (1) and (2) of Proposition 9.23. So, $\gamma_K(g, f, t) \in B_{s+\delta}(X, L^\infty(X))$.

- Now, assume $\|g - f\|_\infty > \delta'$. Let us denote $t_0 := \delta' / \|g - f\|_\infty$. Now, for $t \in [0, t_0]$, we have $\gamma_K(g, f, t) \in B_{s+\delta}(X, L^\infty(X))$ since

$$\|\gamma_K(g, f, t) - d_X(x, \cdot)\|_\infty \leq \|\gamma_K(g, f, t) - g\|_\infty + \|g - d_X(x, \cdot)\|_\infty \leq t\|g - f\|_\infty + s \leq s + \delta' < s + \delta.$$

Next, we want to show $\gamma_K(g, f, t) \in B_{s+\delta}(X, L^\infty(X))$ for $t \in [t_0, 1]$. To do that, choose $a \in A$ such that $d_X(x, a) < 2\delta'$. We will prove $\|\gamma_K(g, f, t_0) - d_X(a, \cdot)\|_\infty \leq s + \delta'$.

Fix arbitrary $x' \in X$. If $|g(x') - f(x')| \leq \delta'$, then $\gamma_K(g, f, t_0)(x') = f(x')$. Hence,

$$|\gamma_K(g, f, t_0)(x') - d_X(a, x')| = |f(x') - d_X(a, x')| \leq \delta'$$

by Claim 1. If $|g(x') - f(x')| > \delta'$, then $g(x')$ cannot be between $d_X(a, x')$ and $f(x')$ since, by Claim 1, $|d_X(a, x') - f(x')| \leq \delta'$. This implies that either

$$|g(x') - d_X(a, x')| = |d_X(a, x') - f(x')| + |g(x') - f(x')|$$

or

$$|g(x') - f(x')| = |d_X(a, x') - f(x')| + |g(x') - d_X(a, x')|.$$

Either way, it is easy to see that we always have

$$|\gamma_K(g, f, t_0)(x') - d_X(a, x')| = \left| |g(x') - d_X(a, x')| - \delta' \right| \le s + \delta',$$

where the last inequality is true because $|g(x') - d_X(a, x')| \le |g(x') - d_X(x, x')| + d_X(a, x) < s + 2\delta'$. So, one can conclude that

$$\|\gamma_K(g, f, t_0) - d_X(a, \cdot)\|_\infty \le s + \delta'.$$

Therefore, combining this inequality with Claim 1 and property (6) of Proposition 9.23, one finally obtains that

$$\|\gamma_K(d_X(x, \cdot), f, t) - d_X(a, \cdot)\|_\infty \le s + \delta' < s + \delta,$$

so $\gamma_K(g, f, t) \in B_{s+\delta}(X, L^\infty(X))$ for any $t \in [t_0, 1]$. □

This concludes the proof of Lemma 9.6. □

9.3 The filling radius and Vietoris–Rips persistent homology

Now, we recall the notion of *filling radius*, an invariant for closed connected manifolds introduced by Gromov [46, page 8] in the course of proving the systolic inequality (see also [48; 58] for a comprehensive treatment). It turns out to be that this notion can be a bridge between topological data analysis and differential geometry/topology.

Definition 9.24 (filling radius) Let M be a closed connected n–dimensional manifold with compatible metric d_M. One defines the filling radius of M as

$$\mathrm{FillRad}(M; G) := \inf\{r > 0 \mid \mathrm{H}_n(\iota_r; G)([M]) = 0\},$$

where $\iota_r \colon M \hookrightarrow B_r(M, L^\infty(M))$ is the (corestriction of the) Kuratowski isometric embedding, and $[M]$ is the fundamental class of M, with coefficients in G. We will use the shorthand notation $\mathrm{FillRad}(M)$ when either M is orientable and $G = \mathbb{Z}$ or when M is not orientable and $G = \mathbb{Z}_2$.

Remark 9.25 (metric manifolds) The definition of the filling radius does not require the metric d_M on M to be Riemannian — it suffices that d_M generates the manifold topology. We call any (M, d_M) satisfying this condition a *metric manifold*. In particular, one can consider the filling radius of

(1) the ℓ^∞–metric product of (M, d_M) and (N, d_N) when M and N are Riemannian manifolds and d_M and d_N are their corresponding geodesic distances;

(2) $(N, d_M|_{N \times N})$ when N is a submanifold of the Riemannian manifold (M, d_M).

Remark 9.26 (relative filling radius and minimality for injective metric spaces) The relative filling radius can be defined for every metric pair (M, E) by considering r–neighborhoods of M in E — it is denoted by $\mathrm{FillRad}(M, E)$. Gromov [46] showed that we obtain the minimal possible relative filling
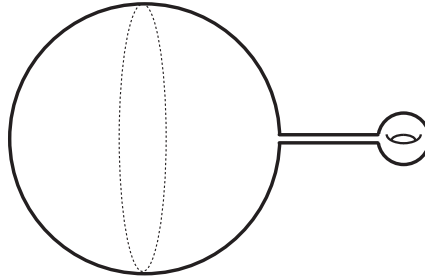
Figure 7: A big sphere X with a small handle. In this case, as $r > 0$ increases, $B_r(X, L^\infty(X))$ changes homotopy type from that of X to that of $\mathbb{S}^2$ as soon as $r > r_0$ for some $r_0 < \text{FillRad}(X)$.

radius through the Kuratowski embedding (that is when $E = L^\infty(M)$). This also follows from our work but in greater generality in the context of embeddings into injective metric spaces. If M can be isometrically embedded into an injective metric space F, then this embedding can be extended to a 1–Lipschitz map $f : E \to F$, which induces a map of filtrations $f_r : B_r(M, E) \to B_r(M, F)$, for each $r > 0$ (see Definition 2.15). Hence, if the fundamental class of M vanishes in $B_r(M, E)$, then it also vanishes in $B_r(M, F)$. Therefore,

$$\text{(3)} \qquad\qquad \text{FillRad}(M, F) \leq \text{FillRad}(M, E).$$

In particular, this implies that $\text{FillRad}(M, E) = \text{FillRad}(M, F)$ whenever E and F are both injective metric spaces admitting isometric embeddings of M.

Remark 9.27 (filling radius and first change in homotopy type) In [54, Theorem 2], Katz proved that $\text{FillRad}(\mathbb{S}^n) = \frac{1}{2} \arccos(-1/(n+1))$. Moreover, in a remark right after the proof of Theorem 2 in that paper he shows that $B_r(\mathbb{S}^n, L^\infty(\mathbb{S}^n))$ is homotopy equivalent to $\mathbb{S}^n$ if $r \in (0, \text{FillRad}(\mathbb{S}^n)]$.

One might then ask whether for any closed connected manifold M it holds that $\text{FillRad}(M)$ is the first value of r where the homotopy type of $B_r(M, L^\infty(M))$ changes. In general, however, this is not true as the following two examples show:

(1) It is known [57, Proposition 0.3] that $\text{FillRad}(\mathbb{CP}^3) > \text{FillRad}(\mathbb{CP}^1) = \frac{1}{2} \arccos\left(-\frac{1}{3}\right)$. Also, by [56, Theorem 8.1], $B_r(\mathbb{CP}^3, L^\infty(\mathbb{CP}^3))$ is not homotopy equivalent to $\mathbb{CP}^3$ for r in the interval $\left(\frac{1}{2} \arccos\left(-\frac{1}{3}\right), \frac{1}{2} \arccos\left(-\frac{1}{3}\right) + \varepsilon_0\right)$, where $\varepsilon_0 > 0$ is a positive constant. In other words, the homotopy type of $B_r(\mathbb{CP}^3, L^\infty(\mathbb{CP}^3))$ already changed before $r = \text{FillRad}(\mathbb{CP}^3)$.

(2) The following example provides geometric intuition for how the homotopy type of Kuratowski neighborhoods may change before r reaches the filling radius. Consider a big sphere with a small handle attached through a long neck (see Figure 7). Since the top-dimensional hole in this space is big, we expect the filling radius to be big. On the other hand, the degree 1 homology class coming from the small handle dies in a small Kuratowski neighborhood, hence the homotopy type changes at that point.

We now relate the filling radius of a closed connected n–dimensional manifold to its n–dimensional Vietoris–Rips persistence barcode.

Proposition 9.28 *Let M be a closed connected n–dimensional Riemannian manifold. Then*

$$(0, 2\,\mathrm{FillRad}(M; \mathbb{F})] \in \mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F}),$$

*where $\mathbb{F}$ is an arbitrary field if M is orientable, and $\mathbb{F} = \mathbb{Z}_2$ if M is nonorientable. Moreover, this is the **unique** interval in $\mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F})$ starting at 0 and $\mathrm{FillRad}(M; \mathbb{F}) \leq \mathrm{FillRad}(M)$ whenever M is orientable.*

The unique interval identified by Proposition 9.28 will be henceforth denoted by

$$I_{n,\mathbb{F}}^M := (0, \mathrm{d}_{n,\mathbb{F}}^M].$$

Proof First, let us consider the case when M is orientable. Observe that the diagram

$$
\begin{array}{ccc}
\mathrm{H}_n(M; \mathbb{Z}) \otimes \mathbb{F} \longrightarrow \mathrm{H}_n\big(B_r(M, L^\infty(M)); \mathbb{Z}\big) \otimes \mathbb{F} \longrightarrow \mathrm{H}_n\big(B_s(M, L^\infty(M)); \mathbb{Z}\big) \otimes \mathbb{F} \\
\Big\downarrow{\scriptstyle j} \qquad\qquad\qquad \Big\downarrow{\scriptstyle j_r} \qquad\qquad\qquad\qquad \Big\downarrow{\scriptstyle j_s} \\
\mathrm{H}_n(M; \mathbb{F}) \longrightarrow \mathrm{H}_n\big(B_r(M, L^\infty(M)); \mathbb{F}\big) \longrightarrow \mathrm{H}_n\big(B_s(M, L^\infty(M)); \mathbb{F}\big)
\end{array}
$$

commutes for any $0 < r \leq s$, where every horizontal arrow is induced by the obvious inclusions, and the vertical arrows (j, j_r, and j_s) must be injective by the universal coefficient theorem for homology (see [68, Theorem 55.1]). Hence, one obtains that

$$\mathrm{FillRad}(M; \mathbb{F}) = \inf\{r > 0 \mid \mathrm{H}_n(\iota_r; \mathbb{F})(j([M])) = 0\}.$$

Therefore, with the aid of Theorems 4.1 and 5.2, one concludes that

$$(0, 2\,\mathrm{FillRad}(M; \mathbb{F})] \in \mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F}).$$

Also, by Hausmann's theorem [50, Theorem 3.5], $\mathrm{VR}_r(M)$ is homotopy equivalent to M for $r > 0$ small enough. Therefore, $(0, 2\,\mathrm{FillRad}(M; \mathbb{F})]$ must be the unique interval in $\mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F})$ with left endpoint 0.

The proof of the nonorientable case is similar, so we omit it. $\qquad\square$

Remark 9.29 $\mathrm{FillRad}(\mathbb{S}^n; \mathbb{F}) = \mathrm{FillRad}(\mathbb{S}^n)$ for any field $\mathbb{F}$. This can be verified via Proposition 9.28 and Remark 7.3. Alternatively, a more direct proof can be obtained via Jung's theorem (Theorem A.8) following an idea similar to the one used in the proofs of [46, Lemmas 1.2.B and 4.5.A; 54, Theorem 2]. Details of this direct proof can be found in the extended (arXiv) version of this paper [62, Remark 9.13]. With this observation, from now on we will drop $\mathbb{F}$ from the notation $I_{n,\mathbb{F}}^{\mathbb{S}^n}$ and $d_{n,\mathbb{F}}^{\mathbb{S}^n}$, and respectively use $I_n^{\mathbb{S}^n}$ and $d_n^{\mathbb{S}^n}$ instead.

Remark 9.30 Actually, one can generalize Proposition 9.28 to metric manifolds. See Proposition 9.46 for the full generalization.

Remark 9.31 Let M and N be closed connected metric manifolds. Let $M \times N$ denote the ℓ^∞–product of M and N (as metric spaces). By Theorem 6.1, Remark 6.3, and Proposition 9.28,

$$\mathrm{FillRad}(M \times N; \mathbb{F}) = \min\{\mathrm{FillRad}(M; \mathbb{F}), \mathrm{FillRad}(N; \mathbb{F})\}.$$

A similar result is true for the ℓ^∞–product of more than two metric manifolds.

9.3.1 Bounding the filling radius and consequences for Vietoris–Rips persistent homology Using Proposition 9.28, we can estimate certain properties of the barcode $\mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F})$ of an n–dimensional manifold M.

Injectivity radius and persistence barcodes If $I_{n,\mathbb{F}}^M = (0, \mathrm{d}_{n,\mathbb{F}}^M]$ is the unique interval in $\mathrm{barc}_n^{\mathrm{VR}}(M; \mathbb{F})$ identified by Proposition 9.28, then

(4)
$$\mathrm{d}_{n,\mathbb{F}}^M \geq \frac{\mathrm{Inj}(M)}{n+2}.$$

This follow from the fact that $\mathrm{FillRad}(M; \mathbb{F}) \geq \mathrm{Inj}(M)/(2(n+2))$ for any field $\mathbb{F}$, where $\mathrm{Inj}(M)$ is the injectivity radius of M [46, Proof of Lemma 4.5.A]. Since the injectivity radius of the sphere is $\frac{\pi}{2}$, equation (4) implies that $\mathrm{d}_n^{\mathbb{S}^n} \geq \pi/(2(n+2))$. Note that Proposition 9.28 indicates that this estimate is not tight in general since

$$\mathrm{d}_n^{\mathbb{S}^n} = 2\,\mathrm{FillRad}(\mathbb{S}^n) = \arccos\left(\frac{-1}{n+1}\right) \geq \frac{\pi}{2}.$$

Systole and persistence barcodes The *systole* $\mathrm{sys}_1(M)$ of a Riemannian manifold M is defined to be the infimal length over noncontractible loops of M. In [46, Lemma 1.2.B], Gromov proved that

$$\mathrm{sys}_1(M) \leq 6\,\mathrm{FillRad}(M)$$

for any closed *essential* Riemannian manifold M.[4] Note that, by slightly modifying the proof of [46, Lemma 1.2.B], one can also verify that $\mathrm{sys}_1(M) \leq 6\,\mathrm{FillRad}(M; \mathbb{F})$ whenever M is orientable and $\mathbb{F}$ is an arbitrary field. Moreover, one can also define the *homological systole* $\mathrm{sysh}_1(M; G)$ to be the infimal length over non null-homologous (with coefficients in a given group G) loops of M. We will use the shorthand notation $\mathrm{sysh}_1(M)$ whenever $G = \mathbb{Z}$. In general, $\mathrm{sys}_1(M) \leq \mathrm{sysh}_1(M) \leq \mathrm{sysh}_1(M; G)$ for any group G since any contractible loop is null-homologous (see [49, 2.A]). See Figure 8 for a space on which the notions differ. In [81, Theorem 8.10], Ž Virk proved that

$$\left(0, \tfrac{1}{3}\,\mathrm{sysh}_1(M; \mathbb{F})\right] \in \mathrm{barc}_1^{\mathrm{VR}}(M; \mathbb{F})$$

for any closed Riemannian manifold M. Observe that the n–dimensional torus $\mathbb{T}^n$ is an aspherical, hence essential, manifold. Also, observe that $\mathrm{sys}_1(\mathbb{T}^n) = \mathrm{sysh}_1(\mathbb{T}^n) = \mathrm{sysh}_1(\mathbb{T}^n; \mathbb{F})$ for any field of coefficients $\mathbb{F}$ since the fundamental group $\pi_1(\mathbb{T}^n)$ is abelian and the homology group of $\mathbb{T}^n$ is free abelian. Therefore, this permits relating the top-dimensional persistence barcode with the 1–dimensional barcode of any n–dimensional Riemannian torus. We summarize this via the following:

---

[4]See [46] for the definition of essential manifolds. For this paper it suffices to keep in mind that aspherical manifolds are essential.
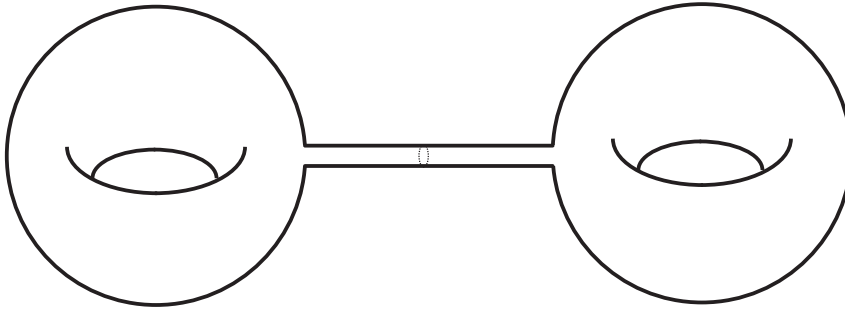
Figure 8: A space X for which $\mathrm{sys}(X) \neq \mathrm{sysh}(X)$.

Corollary 9.32 *For **any** Riemannian metric on the n–dimensional torus $\mathbb{T}^n$,*

- *the interval $I_1^{\mathbb{T}^n} := \left(0, \frac{1}{3}\,\mathbf{sys}_1(\mathbb{T}^n)\right]$ is an element of $\mathrm{barc}_1^{\mathrm{VR}}(\mathbb{T}^n; \mathbb{F})$,*
- *the interval $I_{n,\mathbb{F}}^{\mathbb{T}^n} := \left(0, 2\,\mathrm{FillRad}(\mathbb{T}^n; \mathbb{F})\right]$ is an element of $\mathrm{barc}_n^{\mathrm{VR}}(\mathbb{T}^n; \mathbb{F})$, and*
- $I_1^{\mathbb{T}^n} \subseteq I_{n,\mathbb{F}}^{\mathbb{T}^n}.$

Finally, observe that if the metric on $\mathbb{T}^n$ is the ℓ^∞–product metric, then $\mathrm{FillRad}(\mathbb{T}^n; \mathbb{F}) = \mathrm{FillRad}(\mathbb{T}^n)$ for any field $\mathbb{F}$ by Remark 9.31.

Volume and persistence barcodes An inequality proved by Gromov in [46, Main Theorem 1.2.A] states that for each n natural there exists a constant $c_n > 0$ such that if M is any n–dimensional complete Riemannian manifold, then

$$(5) \qquad\qquad \mathrm{FillRad}(M) \leq c_n (\mathrm{vol}(M))^{1/n}.$$

It then follows that

$$(6) \qquad\qquad \mathrm{d}_{n,\mathbb{F}}^M \leq 2c_n (\mathrm{vol}(M))^{1/n}.$$

In particular, this bound improves upon the one given by Corollary 9.9, $\mathrm{d}_{n,\mathbb{F}}^M \leq \frac{2}{3}\,\mathrm{diam}(M)$, when M is "thin" like in the case of a thickened embedded graph [67].

Spread and persistence barcodes The following proposition is proved in [54, Lemma 1]. Here we provide a different proof, which easily follows from the persistent homology perspective that we have adopted in this paper.

Proposition 9.33 *Let M be a closed connected metric manifold. Then*

$$\mathrm{FillRad}(M; \mathbb{F}) \leq \tfrac{1}{2}\,\mathrm{spread}(M).$$

Proof This follows from Propositions 9.7 and 9.28. $\qquad\square$

Remark 9.34 One can also use Lemma 9.4 to prove Proposition 9.33.

Remark 9.35 The inequality in the statement above becomes an equality for spheres [54].

By Corollary 9.9, Proposition 9.28, and the fact that $\text{FillRad}(\mathbb{S}^1) = \frac{\pi}{3}$, we know that

$$\text{length}(I) \leq \frac{2\pi}{3} = \text{length}(I_1^{\mathbb{S}^1})$$

for any $k \geq 1$, and any $I \in \text{barc}_k^{\text{VR}}(\mathbb{S}^1; \mathbb{F})$. This motivates the following conjecture:

Conjecture 9.36 Let M be a closed connected n–dimensional metric manifold. Then

$$\text{length}(I) \leq \text{length}(I_{n,\mathbb{F}}^M)$$

for any $I \in \text{barc}_k^{\text{VR}}(M; \mathbb{F})$ and any $k \geq 1$.

However, this conjecture is not true in general, as the following example shows:

Remark 9.37 Consider the ℓ^∞–product $X = \mathbb{S}^1 \times \mathbb{S}^2$. Then, by Remark 9.31,

$$\text{FillRad}(X) = \min(\text{FillRad}(\mathbb{S}^1), \text{FillRad}(\mathbb{S}^2)) = \tfrac{1}{2} \arccos\left(-\tfrac{1}{3}\right).$$

This implies that $\text{length}(I_3^X) = 2\,\text{FillRad}(X) = \arccos\left(-\tfrac{1}{3}\right)$. Now, we will prove that there is a longer interval in $\text{barc}_1^{\text{VR}}(X; \mathbb{F})$. First, observe that there is an infinite length interval in $\text{barc}_0^{\text{VR}}(\mathbb{S}^2; \mathbb{F})$. Also, $I_1^{\mathbb{S}^1} = (0, 2\,\text{FillRad}(\mathbb{S}^1)] = \left(0, \frac{2\pi}{3}\right]$. Therefore, by the persistent Künneth formula (Theorem 6.1(1)), and Remark 6.3, the interval $I = \left(0, \frac{2\pi}{3}\right]$ exists in $\text{barc}_1^{\text{VR}}(X; \mathbb{F})$.

Therefore, since $\frac{2\pi}{3} > \arccos\left(-\tfrac{1}{3}\right)$, Conjecture 9.36 is false.

9.3.2 Application to obtaining lower bounds for the Gromov–Hausdorff distance With the aid of the stability of barcodes (Theorem 2.14) and the notion of filling radius, one can obtain the following result:

Proposition 9.38 *Let M be a closed connected m–dimensional orientable (resp. nonorientable) Riemannian manifold, and let X be a compact metric space such that*

(1) *$\text{H}_m(X; \mathbb{F}) = 0$ for some arbitrary field $\mathbb{F}$ (resp. $\text{H}_m(X; \mathbb{F}) = 0$ for $\mathbb{F} = \mathbb{Z}_2$), and*

(2) *$\text{VR}_r(X) \simeq X$ for every $r \in (0, \text{FillRad}(M; \mathbb{F})]$.*

Then

$$d_{\text{B}}(\text{barc}_m^{\text{VR}}(M; \mathbb{F}), \text{barc}_m^{\text{VR}}(X; \mathbb{F})) \geq \text{FillRad}(M; \mathbb{F})$$

and, as a consequence,

$$d_{\text{GH}}(M, X) \geq \tfrac{1}{2}\,\text{FillRad}(M; \mathbb{F}).$$

Proof Observe that by Theorems 2.13 and 2.14,

$$d_{\text{GH}}(M, X) \geq \tfrac{1}{2} d_{\text{B}}(\text{barc}_m^{\text{VR}}(M; \mathbb{F}), \text{barc}_m^{\text{VR}}(X; \mathbb{F})).$$

Hence, it is enough to establish that

$$d_{\text{B}}(\text{barc}_m^{\text{VR}}(M; \mathbb{F}), \text{barc}_m^{\text{VR}}(X; \mathbb{F})) \geq \text{FillRad}(M; \mathbb{F}).$$

Recall that the special interval $I_{m,\mathbb{F}}^M := (0, 2\,\text{FillRad}(M; \mathbb{F})]$ belongs to $\text{barc}_m^{\text{VR}}(M; \mathbb{F})$ by Proposition 9.28. Moreover, if $I := (u, v] \in \text{barc}_m^{\text{VR}}(X; \mathbb{F})$, then $u \geq \text{FillRad}(M; \mathbb{F})$ by the two assumptions on X.

Now, fix an arbitrary partial matching P between $\text{barc}_m^{\text{VR}}(M;\mathbb{F})$ and $\text{barc}_m^{\text{VR}}(X;\mathbb{F})$. If I_m^M is unmatched to any interval in $\text{barc}_m^{\text{VR}}(X;\mathbb{F})$, then

$$\text{cost}(P) \geq \tfrac{1}{2}|0 - 2\,\text{FillRad}(M;\mathbb{F})| = \text{FillRad}(M;\mathbb{F}).$$

If $(I_{m,\mathbb{F}}^M, I := (u,v]) \in P$, then $\text{cost}(P) \geq |0 - u| = u \geq \text{FillRad}(M;\mathbb{F})$. Since P is arbitrary, one can conclude $d_{\text{B}}(\text{barc}_m^{\text{VR}}(M;\mathbb{F}), \text{barc}_m^{\text{VR}}(X;\mathbb{F})) \geq \text{FillRad}(M;\mathbb{F})$, as we required. □

By combining [Proposition 9.38](#) with [Proposition 9.7](#) we now obtain the *exact value* of the lower bound for $d_{\text{GH}}(\mathbb{S}^m, \mathbb{S}^n)$ given by invoking the stability of Vietoris–Rips barcodes:

Corollary 9.39 *For any positive integers* $1 \leq m < n$,

$$\sup_k d_{\text{B}}\big(\text{barc}_k^{\text{VR}}(\mathbb{S}^m;\mathbb{F}), \text{barc}_k^{\text{VR}}(\mathbb{S}^n;\mathbb{F})\big) = \text{FillRad}(\mathbb{S}^m) = \tfrac{1}{2}\arccos\Big(\frac{-1}{m+1}\Big)$$

and, as a consequence,

$$d_{\text{GH}}(\mathbb{S}^m, \mathbb{S}^n) \geq \tfrac{1}{4}\arccos\Big(\frac{-1}{m+1}\Big) \geq \tfrac{\pi}{8}.$$

Proof Notice that $\mathbb{S}^m$ is orientable, $\text{H}_m(\mathbb{S}^n;\mathbb{F}) = 0$ for any field $\mathbb{F}$, $\text{VR}_r(\mathbb{S}^n) \simeq \mathbb{S}^n$ for any r in the interval $\big(0, \arccos(-1/(n+1))\big]$ by [Theorem 7.1](#), and

$$\arccos\Big(\frac{-1}{n+1}\Big) \geq \tfrac{\pi}{2} \geq \tfrac{1}{2}\arccos\Big(\frac{-1}{m+1}\Big) = \text{FillRad}(\mathbb{S}^m).$$

Hence, by [Proposition 9.38](#),

$$\sup_k d_{\text{B}}\big(\text{barc}_k^{\text{VR}}(\mathbb{S}^m;\mathbb{F}), \text{barc}_k^{\text{VR}}(\mathbb{S}^n;\mathbb{F})\big) \geq \text{FillRad}(\mathbb{S}^m) = \tfrac{1}{2}\arccos\Big(\frac{-1}{m+1}\Big).$$

The reverse inequality follows from [Proposition 9.7](#) and Remarks [9.3](#) and [9.27](#) relating the spread to the filling radius of spheres. Indeed, by basic properties of the bottleneck distance,[5] for every integer $k \geq 0$,

$$d_{\text{B}}(\text{barc}_k^{\text{VR}}(\mathbb{S}^m;\mathbb{F}), \text{barc}_k^{\text{VR}}(\mathbb{S}^n;\mathbb{F})) \leq \tfrac{1}{2}\max\Big(\max_{I \in \text{barc}_k^{\text{VR}}(\mathbb{S}^m;\mathbb{F})} \text{length}(I), \max_{J \in \text{barc}_k^{\text{VR}}(\mathbb{S}^n;\mathbb{F})} \text{length}(J)\Big).$$

Now, by [Proposition 9.7](#), the right-hand side is bounded above by $\tfrac{1}{2}\max(\text{spread}(\mathbb{S}^m), \text{spread}(\mathbb{S}^n))$ which, by [Remark 9.3](#), is equal to $\tfrac{1}{2}\arccos(-1/(m+1))$ and in turn equal to $\text{FillRad}(\mathbb{S}^m)$ by [Remark 9.27](#). □

Remark 9.40 The lower bounds provided by [Corollary 9.39](#) are nonoptimal; see [63] for improved lower bounds via considerations based on a certain version of the Borsuk–Ulam theorem. In fact, there the factor $\tfrac{1}{2}$ is removed, leading, for example, to the bound $d_{\text{GH}}(\mathbb{S}^m, \mathbb{S}^n) \geq \text{FillRad}(\mathbb{S}^{\min\{m,n\}})$ for all $0 \leq m < n \leq \infty$. This bound is therein shown to be tight when $(m,n) \in \{(1,2),(1,3),(2,3)\}$ via the construction of suitable correspondences. Via [Example 2.11](#), [Theorem 2.14](#), and [Proposition 9.28](#), one can directly see that for any geodesic, compact, and simply connected space Y,

$$d_{\text{GH}}(\mathbb{S}^1, Y) \geq \tfrac{\pi}{6}.$$

This, taken together with the comments above leads to the following conjecture:

---

[5]The cost of the empty matching upper bounds the bottleneck distance.
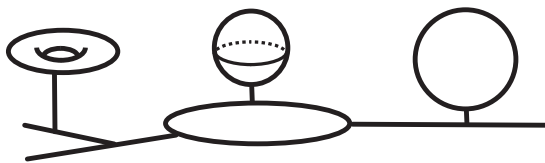
Figure 9: An ANR space contemplated by Proposition 9.46.

Conjecture 9.41 For any geodesic, compact, and simply connected space Y,

$$d_{\mathrm{GH}}(\mathbb{S}^1, Y) \geq \tfrac{\pi}{3}.$$

9.3.3 A generalization of the filling radius The goal of this section is to provide some partial results regarding the structure of $\mathrm{barc}_*^{\mathrm{VR}}(\cdot)$ for nonsmooth spaces; see Figure 9. In order to do so we consider a generalization of the notion of filling radius for arbitrary compact ANR metric spaces and arbitrary homology dimension. See [51] for an introduction to the general theory of ANRs.

Definition 9.42 (absolute neighborhood retract) A metric space (X, d_X) is said to be *ANR* (*absolute neighborhood retract*) if, whenever X is a subspace of another metric space Y, there exists an open set $X \subset U \subseteq Y$ such that X is a retract of U.

It is known that every topological manifold with compatible metric (so, a metric manifold) is an ANR. Not only that, every locally Euclidean metric space is an ANR (see [51, Theorem III.8.1]). Also, every compact, (topologically) finite-dimensional, and locally contractible metric space is ANR (see [33, Section 1]). The following example is one application of this fact:

Example 9.43 Let $\mathcal{G}$ be a compact metric graph and $M_1, \ldots, M_n$ be closed connected metric manifolds. Choose points $v_1, \ldots, v_n \in G$ and $p_i \in M_i$ for each $i = 1, \ldots, n$ and consider the geodesic metric space $X := \mathcal{G} \vee M_1 \vee \cdots \vee M_n$ arising from metric gluings via $v_1 \sim p_1, \ldots, v_n \sim p_n$. Since X is compact, (topologically) finite-dimensional, and locally contractible, it is an ANR. See Figure 9.

Finally, we are ready to define a generalized filling radius.

Definition 9.44 (generalized filling radius) Let (X, E) be a metric pair where X is a compact ANR metric space. For any integer $k \geq 1$, any abelian group G, and any $\omega \in \mathrm{H}_k(X; G)$, we define the generalized filling radius as

$$\mathrm{FillRad}_k((X, E), G, \omega) := \inf\{r > 0 \mid \mathrm{H}_k(\iota_r^E; G)(\omega) = 0\},$$

where $\iota_r^E : X \hookrightarrow B_r(X, E)$ is the (corestriction of the) isometric embedding. In other words, we have the map

$$\mathrm{FillRad}_k((X, E), G, \cdot) : \mathrm{H}_k(X; G) \to \mathbb{R}_{\geq 0}.$$

Remark 9.45 Following the discussion in Remark 9.26 after equation (3), one can also prove that the smallest possible value of the generalized filling radius is attained when E is an injective metric space. Hence, we write $\mathrm{FillRad}_k(X, G, \omega)$ instead of $\mathrm{FillRad}_k((X, E), G, \omega)$ whenever E is injective, for simplicity.

Let M be an n–dimensional metric manifold. Then, note that we have $\mathrm{FillRad}_n(M, G, [M]) = \mathrm{FillRad}(M)$ in the following two cases: when M is orientable and $G = \mathbb{Z}$, and when M is nonorientable and $G = \mathbb{Z}_2$.

A priori, one can define the generalized filling radius for any metric space X. However, we believe that the context of ANR metric spaces is the right level of generalization for our purposes because of the following proposition, analogous to Proposition 9.28:

Proposition 9.46 *Let X be a compact ANR metric space. Then, for any $k \geq 1$ and nonzero $\omega \in \mathrm{H}_k(X; \mathbb{F})$, we have $\mathrm{FillRad}_k(X, \mathbb{F}, \omega) > 0$, and*

$$(0, 2\,\mathrm{FillRad}_k(X, \mathbb{F}, \omega)] \in \mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F}),$$

where $\mathbb{F}$ is an arbitrary field.

Proof First, note that one cannot apply Hausmann's theorem since X is not necessarily a Riemannian manifold. However, since X is ANR and a closed subset of $L^\infty(X)$, there exists an open $U \subseteq L^\infty(X)$ such that $X \subset U$ and U retracts onto X. Let $\rho \colon U \to X$ be the retraction. Now, since U is open there exists an $r > 0$ such that $B_r(X, L^\infty(X)) \subseteq U$. Observe that the restriction

$$\rho_r := \rho|_{B_r(X, L^\infty(X))} \colon B_r(X, L^\infty(X)) \to X$$

is still a retraction. It means that $\rho_r \circ \iota_r = \mathrm{id}_X$. Therefore,

$$\mathrm{H}_k(\iota_r; \mathbb{F}) \colon \mathrm{H}_k(X; \mathbb{F}) \to \mathrm{H}_k\big(B_r(X, L^\infty(X)); \mathbb{F}\big)$$

is injective. This implies that $\mathrm{FillRad}_k(X, \mathbb{F}, \omega) > 0$ and that there exists some interval in $\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})$ corresponding to the nonzero homology class $\omega \in \mathrm{H}_k(X; \mathbb{F})$.

The remaining part of proof is essentially the same as the proof of Proposition 9.28, so we omit it. □

Example 9.47 For any nonzero $\omega \in \mathrm{H}_1(M; \mathbb{F})$ with an arbitrary field $\mathbb{F}$, because of the result in [81, Theorem 8.10], one has that $2\,\mathrm{FillRad}_1(M, \mathbb{F}, \omega) = \frac{1}{3}\,\mathrm{length}(\gamma)$, where γ is a shortest closed curve representing the homology class ω.

A refinement for the case $k = 1$ We now prove that when $k = 1$, the intervals given by Proposition 9.46 are the *only* bars in $\mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F})$.

Lemma 9.48 *Let X be a compact geodesic metric space, which is a subspace of an injective metric space (E, d_E). Then, for any $r > 0$, the canonical inclusion $\iota_r \colon X \hookrightarrow B_r(X, E)$ induces a surjection at the level of fundamental groups. In particular, this also implies ι_r induces a surjection at the level of first degree of homology.*

Proof Let $\gamma \colon [0,1] \to B_r(X,E)$ be an arbitrary continuous path with endpoints x and x' in X. It is enough to show that γ is homotopy equivalent to a path in X relative to its endpoints. By the Lebesgue number lemma, one can choose $0 = t_0 < t_1 < \cdots < t_n = 1$ such that there exists $x_i \in X$ satisfying $\gamma([t_{i-1}, t_i]) \subseteq B_r(x_i, E)$ for each $i \in \{1, \ldots, n\}$. Let $y_i := \gamma(t_i)$ for $i \in \{0, \ldots, n\}$. Since each $B_r(x_i, E)$ is contractible by Lemma 2.28, one can choose continuous paths α_i and β_i contained in $B_r(x_i, E)$ such that α_i is from y_{i-1} to x_i and β_i is from x_i to y_i. As $B_r(x_i, E)$ is contractible, $\gamma|_{[t_{i-1}, t_i]}$ is homotopy equivalent to $\alpha_i \cdot \beta_i$ relative to endpoints. Hence

$$\gamma \simeq (\alpha_1 * \beta_1) * \cdots * (\alpha_n * \beta_n)$$

relative to endpoints. Note that α_1 and β_n can be chosen as geodesics in X as they connect x and x_1 in $B_r(x_1, E)$ and x_n and x' in $B_r(x_n, E)$, respectively. Hence it is enough to show that

$$(\beta_1 * \alpha_2) * \cdots * (\beta_{n-1} * \alpha_n)$$

is homotopy equivalent to a path in X relative to endpoints. Let us show that $\beta_i \cdot \alpha_{i+1}$ is homotopy equivalent to a path in X for each i. Let p be a midpoint of x_i and x_{i+1} in X. Note that p and y_i are contained in $B_r(x_i, E) \cap B_r(x_{i+1}, E)$, which is contractible (again by Lemma 2.28). Let θ be a path in that intersection from y_i to p. Let $\gamma_{x_i, p}$ be a shortest geodesic in X from x_i to p and $\gamma_{p, x_{i+1}}$ be a shortest geodesic in X from p to x_{i+1}. Note that $\gamma_{x_i, p} \cdot \bar{\theta}$ is contained in $B_r(x_i)$ and has endpoints x_i and y_i; hence it is homotopy equivalent to β_i relative to endpoints. Similarly $\theta \cdot \gamma_{p, x_{i+1}}$ is homotopy equivalent to α_{i+1} relative to endpoints. Hence

$$\beta_i \cdot \alpha_{i+1} \simeq \gamma_{x_i, p} \cdot \bar{\theta} \cdot \theta \cdot \gamma_{p, x_{i+1}} \simeq \gamma_{x_i, p} \cdot \gamma_{p, x_{i+1}}$$

relative to endpoints. This completes the proof of the first claim.

For the second claim, exploit [49, Theorem 2A.1]. □

In [81, Theorem 8.10], Virk provided a proof of the corollary below which takes place at the simplicial level. The proof we give below exploits the hyperconvexity properties of $L^\infty(X)$ and also our isomorphism theorem, Theorem 4.1. Given our main results, we can give a more concise proof. See [28, Section 3] for related results.

Corollary 9.49 *Let X be a compact geodesic metric space. Then, for any $I \in \mathrm{barc}_1^{\mathrm{VR}}(X; \mathbb{F})$, there exists $\omega \in \mathrm{H}_1(X; \mathbb{Z})$ such that $I = (0, 2\,\mathrm{FillRad}_1(X, \mathbb{Z}, \omega)]$.*

Proof Apply Lemma 9.48 and Theorem 4.1. □

A conjecture After seeing the proof of Proposition 9.46, some readers might wonder whether one can prove a version of Hausmann's theorem [50, Theorem 3.5] for compact ANR metric spaces. This leads to formulating the conjecture below:

Conjecture 9.50 *Let (X, d_X) be a compact ANR metric space. Then, there exists $r(X) > 0$ such that $\mathrm{VR}_r(X)$ is homotopy equivalent to X for any $r \in (0, r(X)]$.*

9.4 Rigidity of spheres

A problem of interest in the area of persistent homology is that of deciding how much information from a metric space is captured by its associated persistent homology invariants. One basic (admittedly imprecise) question that we posed on page 1024 is:

Question 1 *Assume X and Y are compact metric spaces such that $\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F}) = \mathrm{barc}_k^{\mathrm{VR}}(Y; \mathbb{F})$ for all $k \in \mathbb{Z}_{\geq 0}$. Then how similar are X and Y (in a suitable sense)?*

As proved in [66] via the notion of *core* of a metric graph or as a consequence of [50, Proposition 2.2], the unit circle $\mathbb{S}^1$ and the join X of $\mathbb{S}^1$ with disjoint trees of arbitrary length (regarded as a geodesic metric space) have the same Vietoris–Rips persistence barcodes (for all dimensions); see Figure 10. However, by increasing the length of the trees attached these two spaces are at arbitrarily large Gromov–Hausdorff distance, as shown in Figure 10. This means that, in full generality, Question 1 does not admit a reasonable answer if "similarity" is measure in a strict metric sense via the Gromov–Hausdorff distance.

A related type of questions one might pose are of the type:

Question 3 *Let $\mathcal{C}$ be a given class of compact metric spaces. Does there exist $\epsilon_{\mathcal{C}} > 0$ such that whenever $d_{\mathrm{B}}(\mathrm{barc}_*^{\mathrm{VR}}(X), \mathrm{barc}_*^{\mathrm{VR}}(Y)) < \epsilon_{\mathcal{C}}$ for some $X, Y \in \mathcal{C}$, then X and Y are homotopy equivalent?*

Answers to questions such as Questions 1 and 3 above (together with Questions 2(i), 2(ii), and 2(iii) on page 1025) are not currently known in full generality. One might then consider "localized" versions of the above questions: fix some special compact metric space X_0, and then assume Y satisfies the respective conditions stipulated in the above question statements.

In this regard, from work by Wilhelm [83, Main Theorem 2] and Proposition 9.28 we immediately obtain the following corollary for the case of Riemannian manifolds:

Corollary 9.51 ($\mathrm{barc}_*^{\mathrm{VR}}$ rigidity for spheres) *For any closed connected n–dimensional Riemannian manifold M with sectional curvature $\mathrm{K}_M \geq 1$:*

(1) $I_{n,\mathbb{F}}^M \subseteq I_n^{\mathbb{S}^n}$.

(2) *If $I_{n,\mathbb{F}}^M = I_n^{\mathbb{S}^n}$ then M is isometric to $\mathbb{S}^n$.*

(3) *There exists $\epsilon_n > 0$ such that if $\mathrm{length}(I_n^{\mathbb{S}^n}) - \epsilon_n < \mathrm{length}(I_{n,\mathbb{F}}^M)$, then M is diffeomorphic to $\mathbb{S}^n$.*

(4) *If $\mathrm{length}(I_{n,\mathbb{F}}^M) > \frac{\pi}{3}$, then M is a twisted n–sphere (and, in particular, homotopy equivalent to the n–sphere).*

Remark 9.52 The case of $n = 1$ is simpler. Let M be an arbitrary closed connected 1–dimensional Riemannian manifold. Then, M is isometric to $r \cdot \mathbb{S}^1$ for some $r > 0$ and $I_{1,\mathbb{F}}^M = \left(0, \frac{2\pi}{3} r\right]$. Hence, $I_{1,\mathbb{F}}^M = I_1^{\mathbb{S}^1}$ obviously implies M is isometric to $\mathbb{S}^1$.
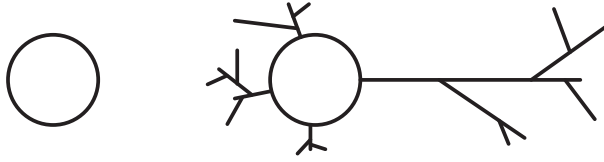
Figure 10: Two geodesic spaces with the same Vietoris–Rips persistence barcodes. Notice that these spaces are at a large Gromov–Hausdorff distance.

Remark 9.53 Wilhelm's method of proof does not yield an explicit value for the parameter ϵ_n given in item (2) above. Wilhelm's rigidity result was extended to Alexandrov spaces by Yokota [84], so Corollary 9.51 can be generalized to that context.

Example 9.54 (a one-parameter family of surfaces with the same filling radius as $\mathbb{S}^2$) If we ignore the sectional curvature condition in Corollary 9.51, then for each $\varepsilon > 0$ small enough one can construct a one-parameter family $\{\mathbb{S}_h^2 \mid h \in [0, \text{FillRad}(\mathbb{S}^2) - \varepsilon]\}$ of surfaces with the same filling radius as $\mathbb{S}^2$ such that $\mathbb{S}_0^2 = \mathbb{S}^2$ but $\mathbb{S}_h^2$ is not isometric to $\mathbb{S}^2$ for any $h > 0$. This phenomenon is analogous to the one depicted in Figure 10.

Here is the construction (see Figure 2):

Let u_1, u_2, u_3, and u_4 be the vertices of a regular tetrahedron inscribed in $\mathbb{S}^2$. Hence, $d_{\mathbb{S}^2}(u_i, u_j) = 2\,\text{FillRad}(\mathbb{S}^2)$ for any $i \neq j$. Now, let T be a very small spherical triangle contained inside the spherical triangle determined by the points u_1, u_2, and u_3 as in Figure 2, left. In other words, we choose $\varepsilon := \text{diam}(T) \ll 2\,\text{FillRad}(\mathbb{S}^2)$.

Now, for any $h \geq 0$, we define $\mathbb{S}_h^2$ by

$$\mathbb{S}_h^2 := (\mathbb{S}^2\backslash\text{Int}(T) \times \{0\}) \cup (\partial T \times [0, h]) \cup (T \times \{h\}) \subsetneq \mathbb{S}^2 \times [0, h]$$

with the metric

$$d_{\mathbb{S}_h^2}((x, s), (y, t)) := d_{\mathbb{S}^2}(x, y) + |s - t|.$$

Then $\mathbb{S}_h^2$ is a 2–dimensional metric manifold. See Figure 2, right, for the description of $\mathbb{S}_h^2$. Also, note that the map

$$P_h \colon \mathbb{S}_h^2 \to \mathbb{S}^2, \quad (x, s) \mapsto x,$$

is 1–Lipschitz.

Claim 1 *First, we claim that* $\text{FillRad}(\mathbb{S}_h^2) \geq \text{FillRad}(\mathbb{S}^2)$ *for any* $h \geq 0$.

Proof Note that, since P_h is 1–Lipschitz, the diagram

$$
\begin{array}{ccc}
\mathbb{S}_h^2 & \longhookrightarrow & B_r(\mathbb{S}_h^2, L^\infty(\mathbb{S}_h^2)) \\
P_h\downarrow & & \downarrow \widetilde{P}_h \\
\mathbb{S}^2 & \longhookrightarrow & B_r(\mathbb{S}^2, L^\infty(\mathbb{S}^2))
\end{array}
$$

commutes for any $r > 0$. Since $(P_h)_*([\mathbb{S}_h^2]) = [\mathbb{S}^2]$, this implies $\text{FillRad}(\mathbb{S}_h^2) \geq \text{FillRad}(\mathbb{S}^2)$. □

Claim 2 *Next, we claim that* $\mathrm{FillRad}(\mathbb{S}_h^2) \leq \mathrm{FillRad}(\mathbb{S}^2)$ *whenever* $h + \varepsilon \leq 2\,\mathrm{FillRad}(\mathbb{S}^2)$.

Proof For this we will prove that the spread of $\mathbb{S}_h^2$ is bounded by twice the filling radius of $\mathbb{S}^2$.

Note that the set $\{(u_1, 0), (u_2, 0), (u_3, 0), (u_4, 0)\} \subset \mathbb{S}_h^2$ satisfies

(1) $\mathrm{diam}\big(\{(u_1, 0), (u_2, 0), (u_3, 0), (u_4, 0)\}\big) = 2\,\mathrm{FillRad}(\mathbb{S}^2)$, and

(2) $\min_{i \in \{1,2,3,4\}} d_{\mathbb{S}_h^2}((x, s), (u_i, 0)) \leq 2\,\mathrm{FillRad}(\mathbb{S}^2)$ for any $(x, s) \in \mathbb{S}_h^2$.

Observe that the second condition holds because if $(x, s) \in \partial T \times [0, h] \cup T \times \{h\}$ (the triangular cylinder with its cap), $d_{\mathbb{S}_h^2}((x, s), (u_1, 0)) = d_{\mathbb{S}^2}(x, u_1) + s \leq h + \varepsilon \leq 2\,\mathrm{FillRad}(\mathbb{S}^2)$.

Hence, by Proposition 9.33, $\mathrm{FillRad}(\mathbb{S}_h^2) \leq \frac{1}{2}\,\mathrm{spread}(\mathbb{S}_h^2) \leq \mathrm{FillRad}(\mathbb{S}^2)$. $\square$

We then conclude that $\mathrm{FillRad}(\mathbb{S}^2) = \mathrm{FillRad}(\mathbb{S}_h^2)$ whenever $h \in [0, 2\,\mathrm{FillRad}(\mathbb{S}^2) - \varepsilon]$.

Remark 9.55 The above construction can be generalized to $\mathbb{S}^n$ for $n \geq 3$. Also, the small subset T need not be a spherical triangle in general, though the argument becomes more involved in that case. For example, one can choose T to be a small geodesic disk on $\mathbb{S}^2$.

Rigidity theorems with respect to the bottleneck distance Propositions 9.56 and 9.57 below provide rigidity results with respect to the bottleneck distance (see Definition 2.12).

For the rest of this subsection we will assume that an arbitrary constant $c \geq 1$ is fixed.

Proposition 9.56 *Suppose M is a closed connected n–dimensional Riemannian manifold with sectional curvature $\mathrm{K}_M \in [1, c]$ and injectivity radius $\mathrm{Inj}(M) \geq \pi/2\sqrt{c}$, then*:

(1) *There exists $\varepsilon_n > 0$ such that, if*

$$d_\mathrm{B}\big(\mathrm{barc}_n^\mathrm{VR}(M; \mathbb{F}), \mathrm{barc}_n^\mathrm{VR}(\mathbb{S}^n; \mathbb{F})\big) < \min\Big\{\varepsilon_n, \frac{1}{\pi\sqrt{c}} \cdot \mathrm{FillRad}(\mathbb{S}^n)\Big\},$$

  then M is diffeomorphic to $\mathbb{S}^n$.

(2) *If $d_\mathrm{B}\big(\mathrm{barc}_n^\mathrm{VR}(M; \mathbb{F}), \mathrm{barc}_n^\mathrm{VR}(\mathbb{S}^n; \mathbb{F})\big) < \min\{2\,\mathrm{FillRad}(\mathbb{S}^n) - \frac{\pi}{3}, (1/\pi\sqrt{c}) \cdot \mathrm{FillRad}(\mathbb{S}^n)\}$, then M is a twisted n–sphere (and, in particular, homotopy equivalent to the n–sphere).*

If M is even-dimensional, then we can drop the assumption on the injectivity radius.

Proposition 9.57 *Suppose M is a closed connected n–dimensional Riemannian manifold with sectional curvature $\mathrm{K}_M \in [1, c]$ for even n, then*:

(1) *There exists $\varepsilon_n > 0$ such that, if*

$$d_\mathrm{B}\big(\mathrm{barc}_n^\mathrm{VR}(M; \mathbb{F}), \mathrm{barc}_n^\mathrm{VR}(\mathbb{S}^n; \mathbb{F})\big) < \min\Big(\varepsilon_n, \frac{1}{4\sqrt{c}} \cdot \mathrm{FillRad}(\mathbb{S}^n)\Big),$$

  then M is diffeomorphic to $\mathbb{S}^n$.

(2) *If $d_\mathrm{B}\big(\mathrm{barc}_n^\mathrm{VR}(M; \mathbb{F}), \mathrm{barc}_n^\mathrm{VR}(\mathbb{S}^n; \mathbb{F})\big) < \min\{2\,\mathrm{FillRad}(\mathbb{S}^n) - \frac{\pi}{3}, (1/4\sqrt{c}) \cdot \mathrm{FillRad}(\mathbb{S}^n)\}$, then M is a twisted n–sphere (and, in particular, homotopy equivalent to the n–sphere).*

Lemma 9.58 *Let M be a closed connected n–dimensional Riemannian manifold. If*

$$d_{\mathrm{B}}\big(\mathrm{barc}_n^{\mathrm{VR}}(M;\mathbb{F}), \mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n;\mathbb{F})\big) < \varepsilon$$

for some $\varepsilon \in (0, 2\,\mathrm{FillRad}(\mathbb{S}^n)]$, then either

(1) $\mathrm{FillRad}(M;\mathbb{F}) < \varepsilon$, *or*

(2) $2|\mathrm{FillRad}(M;\mathbb{F}) - \mathrm{FillRad}(\mathbb{S}^n)| < \varepsilon$.

Proof By Proposition 9.28, we know that $(0, 2\,\mathrm{FillRad}(M;\mathbb{F})] \in \mathrm{barc}_n^{\mathrm{VR}}(M;\mathbb{F})$. Suppose

$$d_{\mathrm{B}}\big(\mathrm{barc}_n^{\mathrm{VR}}(M;\mathbb{F}), \mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n;\mathbb{F})\big) < \varepsilon$$

for some $\varepsilon \in (0, 2\,\mathrm{FillRad}(\mathbb{S}^n)]$. Then there is a partial matching (see Definition 2.12) R_ε between $\mathrm{barc}_n^{\mathrm{VR}}(M;\mathbb{F})$ and $\mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n;\mathbb{F})$ such that $\mathrm{cost}(R_\varepsilon) < \varepsilon$. Consider the following two cases:

(1) Suppose the interval $(0, 2\,\mathrm{FillRad}(M;\mathbb{F})]$ is not matched to any interval in $\mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n;\mathbb{F})$. Then

$$\mathrm{FillRad}(M;\mathbb{F}) \leq \mathrm{cost}(R_\varepsilon) < \varepsilon.$$

(2) Suppose $(0, 2\,\mathrm{FillRad}(M;\mathbb{F})]$ is matched to some interval $(u, v] \in \mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n;\mathbb{F})$ in the partial matching R_ε. Then we claim that $(u, v] = (0, 2\,\mathrm{FillRad}(\mathbb{S}^n)]$. Suppose not. Since we know that $\mathrm{VR}_r(\mathbb{S}^n) \simeq \mathbb{S}^n$ for any $r \in (0, 2\,\mathrm{FillRad}(\mathbb{S}^n)]$ by Theorem 7.1, any interval in $\mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n;\mathbb{F})$ other than $(0, 2\,\mathrm{FillRad}(M;\mathbb{F})]$ must be born after $2\,\mathrm{FillRad}(\mathbb{S}^n)$. In particular, $u \geq 2\,\mathrm{FillRad}(\mathbb{S}^n)$. This implies

$$2\,\mathrm{FillRad}(\mathbb{S}^n) \leq |u - 0| \leq \mathrm{cost}(R_\varepsilon) < \varepsilon \leq 2\,\mathrm{FillRad}(\mathbb{S}^n),$$

which is a contradiction. Hence, $(0, 2\,\mathrm{FillRad}(M;\mathbb{F})]$ is matched to $(0, 2\,\mathrm{FillRad}(\mathbb{S}^n)]$ in the optimal matching. Therefore,

$$2|\mathrm{FillRad}(M;\mathbb{F}) - \mathrm{FillRad}(\mathbb{S}^n)| \leq \mathrm{cost}(R_\varepsilon) < \varepsilon. \qquad \square$$

The proof strategy for Propositions 9.56 and 9.57 is to invoke Wilhelm's result [83, Main Theorem 2] and Lemma 9.58 above. However, if $\mathrm{FillRad}(M)$ were small, one would not be able to apply Wilhelm's theorem. To avoid that, we will invoke a result due to Liu [64].

Proof of Proposition 9.56 Since $c \geq 1$, $(1/\pi\sqrt{c}) \cdot \mathrm{FillRad}(\mathbb{S}^n) \leq 2\,\mathrm{FillRad}(\mathbb{S}^n)$.

(1) By Corollary 9.51(3), there is an $\varepsilon_n > 0$ such that $2|\mathrm{FillRad}(M;\mathbb{F}) - \mathrm{FillRad}(\mathbb{S}^n)| < \varepsilon_n$ implies M is diffeomorphic to $\mathbb{S}^n$.

Suppose $d_{\mathrm{B}}\big(\mathrm{barc}_n^{\mathrm{VR}}(M;\mathbb{F}), \mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n;\mathbb{F})\big) < \min\{\varepsilon_n, (1/\pi\sqrt{c}) \cdot \mathrm{FillRad}(\mathbb{S}^n)\}$. Then

$$\mathrm{FillRad}(M;\mathbb{F}) < \min\Big\{\varepsilon_n, \frac{1}{\pi\sqrt{c}} \cdot \mathrm{FillRad}(\mathbb{S}^n)\Big\}$$

or

$$2|\mathrm{FillRad}(M;\mathbb{F}) - \mathrm{FillRad}(\mathbb{S}^n)| < \min\Big\{\varepsilon_n, \frac{1}{\pi\sqrt{c}} \cdot \mathrm{FillRad}(\mathbb{S}^n)\Big\}$$

by Lemma 9.58. However, the first case is impossible since $\mathrm{FillRad}(M;\mathbb{F}) \geq (1/\pi\sqrt{c})\cdot\mathrm{FillRad}(\mathbb{S}^n)$ by [64, Proofs of Theorem 1.1 and Proposition 1.6]. Therefore,

$$2|\mathrm{FillRad}(M;\mathbb{F}) - \mathrm{FillRad}(\mathbb{S}^n)| < \min\left\{\varepsilon_n, \frac{1}{\pi\sqrt{c}}\cdot\mathrm{FillRad}(\mathbb{S}^n)\right\} \leq \varepsilon_n,$$

so M and $\mathbb{S}^n$ are diffeomorphic.

(2) By basically the same argument,

$$d_{\mathrm{B}}\big(\mathrm{barc}_n^{\mathrm{VR}}(M;\mathbb{F}), \mathrm{barc}_n^{\mathrm{VR}}(\mathbb{S}^n;\mathbb{F})\big) < \min\left\{2\,\mathrm{FillRad}(\mathbb{S}^n) - \frac{\pi}{3}, \frac{1}{\pi\sqrt{c}}\cdot\mathrm{FillRad}(\mathbb{S}^n)\right\}$$

implies $2|\mathrm{FillRad}(M;\mathbb{F}) - \mathrm{FillRad}(\mathbb{S}^n)| < 2\,\mathrm{FillRad}(\mathbb{S}^n) - \frac{\pi}{3}$. Therefore, $\mathrm{length}(I_{n,\mathbb{F}}^M) > \frac{\pi}{3}$, so M is a twisted n–sphere.   □

Proof of Proposition 9.57 The proof is basically the same as the proof of Proposition 9.56. The only difference is we have to use [64, Remark 1.8(3)] instead of [64, Proposition 1.6].   □

9.5 Stability of the filling radius

In [64], Liu studies the mapping properties of the filling radius. His results can be interpreted as providing certain guarantees for how the filling radius changes under *multiplicative* distortion of metrics. Here we study the effect of additive distortion.

Question 4 Under suitable restrictions, does there exists a constant $L > 0$ such that for all closed connected metric manifolds M and N,

$$(7) \qquad\qquad |\mathrm{FillRad}(M) - \mathrm{FillRad}(N)| \leq L \cdot d_{\mathrm{GH}}(M, N)?$$

This question is whether the filling radius could be stable as a map from the collection of all metric manifolds to the real line. The answer is negative, as the following example proves:

Example 9.59 (counterexample for manifolds with different dimension) Fix $\epsilon > 0$ and let $M = \mathbb{S}^1$ and $N_\epsilon = \mathbb{S}^1 \times (\epsilon \cdot \mathbb{S}^1)$, a thin torus. Then, it is clear that $d_{\mathrm{GH}}(M, N_\epsilon) \leq \epsilon$ whereas $\mathrm{FillRad}(\mathbb{S}^1) = \frac{\pi}{3}$ and, by Remark 9.31, $\mathrm{FillRad}(N_\epsilon) = \frac{\pi}{3}\epsilon$. This means that (7) cannot hold in general.

A subsequent possibility is considering only manifolds with the same dimension. The answer in this case is also negative:

Example 9.60 (counterexample for manifolds with the same dimension) Let $n \geq 2$ be any integer and $\epsilon, \delta > 0$; we assume that $\delta \ll \epsilon$ so that a certain tubular neighborhood construction described below works. Consider $M = \mathbb{S}^n \subset \mathbb{R}^{n+1}$. Endow $\mathbb{S}^n$ with the usual round Riemannian metric. Let G_ϵ be a (finite) metric graph embedded in $\mathbb{S}^n$ such that $d_{\mathrm{GH}}(\mathbb{S}^n, G_\epsilon) < \epsilon$; such graphs always exist for compact geodesic

spaces [15, Proposition 7.5.5]. Now, let $N_{\epsilon,\delta}$ be (a suitably smoothed out version of) the boundary of the δ–tubular neighborhood of G_ϵ in $\mathbb{R}^{n+1}$. Then $d_{\mathrm{GH}}(M, N_{\epsilon,\delta}) \leq C \cdot (\epsilon + \delta)$, for some constant $C > 0$ whose exact value is not relevant. However, FillRad$(M) = \frac{1}{2}\arccos(-1/(n+1)) \geq \frac{\pi}{4}$, whereas FillRad$(N_{\epsilon,\delta}) \leq C_n \cdot \delta$ by inequality (5). This means that (7) cannot hold in general, even when the manifolds M and N have the same dimension.

We are however able to establish the following:

Proposition 9.61 (stability of the filling radius) *Let M be a closed connected n–dimensional manifold. Let d_1 and d_2 be two metrics on M compatible with the manifold topology. Then*

$$|\mathrm{FillRad}(M, d_1) - \mathrm{FillRad}(M, d_2)| \leq \|d_1 - d_2\|_\infty.$$

Actually, one can prove a more general result.

Proposition 9.62 (stability of generalized filling radii) *Let M be a closed connected manifold. Let d_1 and d_2 be two metrics on M compatible with the manifold topology. For any integer $k \geq 0$, any abelian group G, and any nonzero $\omega \in \mathrm{H}_k(M;G)$,*

$$|\mathrm{FillRad}_k((M, d_1), G, \omega) - \mathrm{FillRad}_k((M, d_2), G, \omega)| \leq \|d_1 - d_2\|_\infty.$$

Remark 9.63 Proposition 9.61 is just a special case of Proposition 9.62 when $k = n$, $\omega = [M]$, and $G = \mathbb{Z}$ or $\mathbb{Z}_2$.

Proof of Proposition 9.62 Let $i_1 \colon (M, d_1) \to L^\infty(M)$ and $i_2 \colon (M, d_2) \to L^\infty(M))$ be the Kuratowski embeddings of M into $L^\infty(M)$ with respect to d_1 and d_2, respectively. For arbitrary $r > 0$, let $i_1^r \colon (M, d_1) \to B_r(i_1(M), L^\infty(M))$ and $i_2^r \colon (M, d_2) \to B_r(i_2(M), L^\infty(M))$ denote the corresponding isometric embeddings induced from i_1 and i_2. For arbitrary $r > 0$, observe that

$$B_r(i_1(M), L^\infty(M)) \subseteq B_{r+\|d_1-d_2\|_\infty}(i_2(M), L^\infty(M))$$

because, for arbitrary $f \in B_r(i_1(M), L^\infty(M))$, there exist $x \in M$ such that $\|f - d_1(x, \cdot)\|_\infty < r$; hence,

$$\|f - d_2(x, \cdot)\|_\infty \leq \|f - d_1(x, \cdot)\|_\infty + \|d_1(x, \cdot) - d_2(x, \cdot)\|_\infty < r + \|d_1 - d_2\|_\infty.$$

In a similar way, one can prove that $B_r(i_2(M), L^\infty(M)) \subseteq B_{r+\|d_1-d_2\|_\infty}(i_1(M), L^\infty(M))$.

Now, fix arbitrary $r > \mathrm{FillRad}_k((M, d_1), G, \omega)$ and let

$$j^r \colon B_r(i_1(M), L^\infty(M)) \hookrightarrow B_{r+\|d_1-d_2\|_\infty}(i_2(M), L^\infty(M))$$

be the canonical inclusion map. The maps defined above fit into the following (in general noncommutative) diagram:

$$
\begin{array}{ccc}
M & \xrightarrow{\;\;i_1^r\;\;} & B_r(i_1(M), L^\infty(M)) \\
& \searrow{\scriptstyle i_2^{r+\|d_1-d_2\|_\infty}} & \downarrow{\scriptstyle j^r} \\
& & B_{r+\|d_1-d_2\|_\infty}(i_2(M), L^\infty(M))
\end{array}
$$

Next, we prove that $j^r \circ i_1^r$ is homotopic to $i_2^{r+\|d_1-d_2\|_\infty}$ via the linear interpolation

$$H : M \times [0,1] \to B_{r+\|d_1-d_2\|_\infty}(i_2(M), L^\infty(M)), \quad (x,t) \mapsto (1-t)d_1(x,\cdot) + td_2(x,\cdot).$$

The only subtle point is whether this linear interpolation is always contained in the thickening

$$B_{r+\|d_1-d_2\|_\infty}(i_2(M), L^\infty(M))$$

or not. To ascertain this, for arbitrary $x \in M$ and $t \in [0,1]$, compute the distance between $H(x,t)$ and $d_2(x,\cdot)$ as

$$\|(1-t)d_1(x,\cdot)+td_2(x,\cdot)-d_2(x,\cdot)\|_\infty = |1-t|\cdot\|d_1(x,\cdot)-d_2(x,\cdot)\|_\infty \le \|d_1-d_2\|_\infty < r+\|d_1-d_2\|_\infty.$$

Hence, H is a well-defined homotopy between $j^r \circ i_1^r$ and $i_2^{r+\|d_1-d_2\|_\infty}$. Therefore,

$$(j^r)_* \circ (i_1^r)_* = (i_2^{r+\|d_1-d_2\|_\infty})_*.$$

From the assumption on r, we know that $(i_1^r)_*(\omega) = 0$. By the above, this implies that

$$(i_2^{r+\|d_1-d_2\|_\infty})_*(\omega) = 0.$$

Hence,

$$\mathrm{FillRad}_k((M,d_2),G,\omega) \le \mathrm{FillRad}_k((M,d_1),G,\omega) + \|d_1-d_2\|_\infty$$

since $r > \mathrm{FillRad}_k((M,d_1),G,\omega)$ is arbitrary. In a similar way, one can also show

$$\mathrm{FillRad}_k((M,d_1),G,\omega) \le \mathrm{FillRad}_k((M,d_2),G,\omega) + \|d_1-d_2\|_\infty. \qquad \square$$

9.5.1 The strong filling radius Examples 9.59 and 9.60 suggest that the setting of Proposition 9.61 might be a suitable one for studying stability of the filling radius.

In this section we consider a certain strong variant of the filling radius satisfying (7) which arises from the notion of persistent homology.

Definition 9.64 (strong filling radius) Given a closed connected n–dimensional metric manifold M and a field $\mathbb{F}$, we define the *strong filling radius* $\mathrm{sFillRad}(M;\mathbb{F})$ as half the length of the largest interval in the nth Vietoris–Rips persistence barcode of M,

$$\mathrm{sFillRad}(M;\mathbb{F}) := \tfrac{1}{2}\max\{\mathrm{length}(I) \mid I \in \mathrm{barc}_n^{\mathrm{VR}}(M;\mathbb{F})\}.$$

The reader familiar with concepts from applied algebraic topology will have noticed that the definition of strong filling radius of an n–dimensional metric manifold coincides with (one half of) the *maximal persistence* of its associated Vietoris–Rips persistence module. In fact, for each nonnegative integer k one can define the k–dimensional version of strong filling radius of any compact metric space X.

Definition 9.65 (generalized strong filling radius) Given a compact metric space X, a field $\mathbb{F}$, and a nonnegative integer $k \geq 0$, we define the *generalized strong filling radius* $\mathrm{sFillRad}_k(X; \mathbb{F})$ as half the length of the largest interval in the kth Vietoris–Rips persistence barcode of X,

$$\mathrm{sFillRad}_k(X; \mathbb{F}) := \tfrac{1}{2} \max\{\mathrm{length}(I) \mid I \in \mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F})\}.$$

Remark 9.66   • When X is isometric to a metric manifold M with dimension n, we of course have $\mathrm{sFillRad}_n(X) = \mathrm{sFillRad}(M)$.

• In general, $\mathrm{sFillRad}_k$ and $\mathrm{FillRad}_k$ are obviously related in the sense that

$$\mathrm{sFillRad}_k(X; \mathbb{F}) \geq \sup\{\mathrm{FillRad}_k(X, \mathbb{F}, \omega) \mid \omega \in \mathrm{H}_k(X; \mathbb{F})\}$$

for any nonnegative integer k.

The following remark follows directly from Propositions 9.7 and 9.28:

Remark 9.67 $\mathrm{FillRad}(M; \mathbb{F}) \leq \mathrm{sFillRad}(M; \mathbb{F}) \leq \tfrac{1}{2} \mathrm{spread}(M)$ for any field $\mathbb{F}$ when M is orientable, and $\mathbb{F} = \mathbb{Z}_2$ when M is nonorientable.

Definition 9.68 ($\mathbb{F}$–regularly filled manifold) Let (M, d_M) be a closed connected metric manifold and $\mathbb{F}$ be a field. We say that M is $\mathbb{F}$–*regularly filled* if $\mathrm{FillRad}(M; \mathbb{F}) = \mathrm{sFillRad}(M; \mathbb{F})$.

Remark 9.69 For each $n \geq 1$, the n–dimensional unit sphere with the intrinsic metric is $\mathbb{F}$–regularly filled for any field $\mathbb{F}$. Indeed, by [54, Proof of Theorem 2], $\mathrm{FillRad}(\mathbb{S}^n) = \tfrac{1}{2} \mathrm{spread}(\mathbb{S}^n)$. Hence, the result follows from Remark 9.67.

As a consequence of the remark above and Remark 9.3 we have:

Corollary 9.70 *For all integers $n \geq 1$, $\mathrm{FillRad}(\mathbb{S}^n) = \mathrm{sFillRad}(\mathbb{S}^n; \mathbb{F}) = \tfrac{1}{2} \arccos(-1/(n+1))$.*

There exist, however, nonregularly filled metric manifolds. We present two examples: the first one arises from our study of the Künneth formula in Section 6, whereas the second one is a direct construction. Both examples make use of results from [1] about homotopy types of Vietoris–Rips complexes of $\mathbb{S}^1$.

Example 9.71 (a nonregularly filled metric manifold) Fix $r > 1$ and let X be the ℓ^∞–product $\mathbb{S}^1 \times \mathbb{S}^1 \times (r \cdot \mathbb{S}^1)$. By Remark 9.31, $\mathrm{FillRad}(X) = \mathrm{FillRad}(\mathbb{S}^1) = \tfrac{2\pi}{3}$. By Example 6.4, $\mathrm{barc}_3^{\mathrm{VR}}(X; \mathbb{F})$ contains the interval $\left(\tfrac{2\pi}{3} r, \tfrac{4\pi}{5} r\right]$, which has length $\tfrac{2\pi}{15} r$. Hence, if $r > 5$, X is not $\mathbb{F}$–regularly filled.

Example 9.72 (a nonregularly filled Riemannian manifold) Take any embedding of $\mathbb{S}^1$ into $\mathbb{R}^4$ and let $\epsilon > 0$ be small. Consider the boundary C_ϵ of the ϵ–tubular neighborhood around $\mathbb{S}^1$. This will be a 3–dimensional submanifold of $\mathbb{R}^4$. As a submanifold it inherits the ambient inner product and

C_ϵ can be regarded as a Riemannian manifold in itself. Then, as a metric space, with the geodesic distance, C_ϵ will be ϵ–close to $\mathbb{S}^1$ (with geodesic distance) in the Gromov–Hausdorff sense. Because we know that for $r \in \left(\frac{2\pi}{3}, \frac{4\pi}{5}\right]$, $\mathrm{VR}_r(\mathbb{S}^1) \simeq \mathbb{S}^3$, and because of the Gromov–Hausdorff stability of barcodes (Theorem 2.14), it must be that $\mathrm{barc}_3^{\mathrm{VR}}(C_\epsilon; \mathbb{F})$ contains an interval I which itself contains $\left(\frac{2\pi}{3} + \epsilon, \frac{4\pi}{5} - \epsilon\right)$. This latter interval is nonempty whenever $\epsilon > 0$ is small enough, so $\mathrm{sFillRad}(C_\epsilon; \mathbb{F}) \approx \frac{2\pi}{15} - 2\epsilon$. However, $\mathrm{FillRad}(C_\epsilon) \approx \epsilon$.

By invoking the relationship between the Vietoris–Rips persistent homology and the strong filling radius, one can verify that the strong filling radii of two n–dimensional metric manifolds M and N are close if these two manifolds are similar in the Gromov–Hausdorff distance sense.

Proposition 9.73 *Let X and Y be compact metric spaces. Then, for any integer $k \geq 0$,*

$$|\mathrm{sFillRad}_k(X; \mathbb{F}) - \mathrm{sFillRad}_k(Y; \mathbb{F})| \leq 2d_{\mathrm{GH}}(X, Y).$$

Proof By Remark 4.8 one has

$$2d_{\mathrm{GH}}(X, Y) \geq d_{\mathrm{I}}(\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F}), \mathrm{barc}_k^{\mathrm{VR}}(Y; \mathbb{F})) \geq |d_{\mathrm{I}}(\mathrm{barc}_k^{\mathrm{VR}}(X; \mathbb{F}), 0_*) - d_{\mathrm{I}}(\mathrm{barc}_k^{\mathrm{VR}}(Y; \mathbb{F}), 0_*)|,$$

where the last inequality follows from the triangle inequality for the interleaving distance. The conclusion now follows from Example 2.11. □

Remark 9.74 Albeit for the notation $\mathrm{sFillRad}_k$, the above stability result should be well known to readers familiar with applied algebraic topology concepts — we state and prove it here however to provide some background for those readers who are not.

Appendix

A.1 Proof of Proposition 9.23

Proposition 9.23 *Let X be a compact metric space. Then, for any $f, g, h \in L^\infty(X)$ and $0 \leq s \leq t \leq 1$, the Katz geodesic bicombing γ_K on $L^\infty(X)$ satisfies*

(1) $\gamma_K(f, g, 0) = f$ *and* $\gamma_K(f, g, 1) = g$;

(2) $\|\gamma_K(f, g, s) - \gamma_K(f, g, t)\|_\infty = (t - s) \cdot \|f - g\|_\infty$;

(3) $\|\gamma_K(f, g, t) - \gamma_K(h, g, t)\|_\infty \leq 2\|f - h\|_\infty$;

(4) $\|\gamma_K(f, g, t) - \gamma_K(f, h, t)\|_\infty \leq \|g - h\|_\infty$;

(5) $\gamma_K(\phi, \psi, \lambda) = \gamma_K(f, g, (1-\lambda)s + \lambda t)$ *where* $\phi = \gamma_K(f, g, s)$ *and* $\psi = \gamma_K(f, g, t)$ *for any* $\lambda \in [0, 1]$ (*this property is called* **consistency**);

(6) $\|\gamma_K(f, g, r) - h\|_\infty \leq \max\{\|\gamma_K(f, g, s) - h\|_\infty, \|\gamma_K(f, g, t) - h\|_\infty\}$ *for any* $r \in [s, t]$.

Proof (1) The first claim trivially follows from the definition of γ_K and that of the ℓ^∞–norm.

(2) For the second claim, observe that it is enough to show

$$\|\gamma_K(f, g, s) - \gamma_K(f, g, t)\|_\infty \le (t - s) \cdot \|f - g\|_\infty$$

for any $f, g \in L^\infty(X)$ and $0 \le s \le t \le 1$.

Fix an arbitrary $x \in X$. Without loss of generality, one can assume that $f(x) \ge g(x)$. Then

$$|\gamma_K(f, g, s)(x) - \gamma_K(f, g, t)(x)| = \max\{f(x) - s\|f - g\|_\infty, g(x)\} - \max\{f(x) - t\|f - g\|_\infty, g(x)\}.$$

Observe that, if $s \in [0, (f(x) - g(x))/\|f - g\|_\infty]$,

$$\max\{f(x) - s\|f - g\|_\infty, g(x)\} = f(x) - s\|f - g\|_\infty.$$

Hence,

$$\begin{aligned}
|\gamma_K(f, g, s)(x) - \gamma_K(f, g, t)(x)| &= (f(x) - s\|f - g\|_\infty) - \max\{f(x) - t\|f - g\|_\infty, g(x)\} \\
&\le (f(x) - s\|f - g\|_\infty) - (f(x) - t\|f - g\|_\infty) \\
&= (t - s)\|f - g\|_\infty.
\end{aligned}$$

Also, if $s \in [(f(x) - g(x))/\|f - g\|_\infty, 1]$,

$$\max\{f(x) - s\|f - g\|_\infty, g(x)\} = \max\{f(x) - t\|f - g\|_\infty, g(x)\} = g(x)$$

so $|\gamma_K(f, g, s)(x) - \gamma_K(f, g, t)(x)| = 0 \le (t - s)\|f - g\|_\infty$.

Since x is arbitrary, we obtain $\|\gamma_K(f, g, s) - \gamma_K(f, g, t)\|_\infty \le (t - s) \cdot \|f - g\|_\infty$.

(3) Fix an arbitrary $x \in X$. We will prove that

$$|\gamma_K(f, g, t)(x) - \gamma_K(h, g, t)(x)| \le 2\|f - h\|_\infty.$$

Unfortunately, we have to do tedious case-by-case analysis.

(a) If $f(x) \ge g(x)$ and $h(x) \ge g(x)$, then, for $t \in [0, (f(x) - g(x))/\|f - g\|_\infty]$,

$$\gamma_K(f, g, t)(x) = f(x) - t\|f - g\|_\infty.$$

Hence,

$$\begin{aligned}
\gamma_K(f, g, t)(x) - \gamma_K(h, g, t)(x) &= (f(x) - t\|f - g\|_\infty) - \max\{h(x) - t\|h - g\|_\infty, g(x)\} \\
&\le (f(x) - t\|f - g\|_\infty) - (h(x) - t\|h - g\|_\infty) \\
&= f(x) - h(x) - t(\|f - g\|_\infty - \|h - g\|_\infty) \\
&\le |f(x) - h(x)| + t|\|f - g\|_\infty - \|h - g\|_\infty| \\
&\le 2\|f - h\|_\infty.
\end{aligned}$$

Now, for $t \in [(f(x) - g(x))/\|f - g\|_\infty, 1]$, $\gamma_K(f, g, t)(x) = g(x)$. Hence,

$$\begin{aligned}
\gamma_K(f, g, t)(x) - \gamma_K(h, g, t)(x) &= g(x) - \max\{h(x) - t\|h - g\|_\infty, g(x)\} \\
&\le g(x) - g(x) = 0 \le 2\|f - h\|_\infty.
\end{aligned}$$

In a similar way, one can also obtain

$$\gamma_K(h, g, t)(x) - \gamma_K(f, g, t)(x) \leq 2\|f - h\|_\infty$$

for any $t \in [0, 1]$. Hence,

$$|\gamma_K(f, g, t)(x) - \gamma_K(h, g, t)(x)| \leq 2\|f - h\|_\infty.$$

(b) If $f(x) \leq g(x)$ and $h(x) \leq g(x)$, this case is similar to the previous one so we omit it.

(c) If $f(x) \geq g(x)$ and $h(x) \leq g(x)$, then $\gamma_K(f, g, t)(x), \gamma_K(h, g, t)(x) \in [h(x), f(x)]$. Therefore,

$$|\gamma_K(f, g, t)(x) - \gamma_K(h, g, t)(x)| \leq f(x) - h(x) \leq \|f - h\|_\infty \leq 2\|f - h\|_\infty.$$

(d) For $f(x) \leq g(x)$ and $h(x) \geq g(x)$, this is similar to the previous case.

Since x is arbitrary, we finally have

$$\|\gamma_K(f, g, t) - \gamma_K(h, g, t)\|_\infty \leq 2\|f - h\|_\infty.$$

(4) Fix an arbitrary $x \in X$. We will prove that

$$|\gamma_K(f, g, t)(x) - \gamma_K(f, h, t)(x)| \leq \|g - h\|_\infty.$$

Let's do case-by-case analysis.

(a) If $f(x) \geq g(x)$ and $f(x) \geq h(x)$, then, for $t \in [0, (f(x) - g(x))/\|f - g\|_\infty]$,

$$\gamma_K(f, g, t)(x) = f(x) - t\|f - g\|_\infty.$$

Hence,

$$\begin{aligned}
\gamma_K(f, g, t)(x) - \gamma_K(f, h, t)(x) &= (f(x) - t\|f - g\|_\infty) - \max\{f(x) - t\|f - h\|_\infty, h(x)\} \\
&\leq (f(x) - t\|f - g\|_\infty) - (f(x) - t\|f - h\|_\infty) \\
&= t(\|f - h\|_\infty - \|f - g\|_\infty) \\
&\leq \|g - h\|_\infty.
\end{aligned}$$

Now, for $t \in [(f(x) - g(x))/\|f - g\|_\infty, 1]$, $\gamma_K(f, g, t)(x) = g(x)$. Hence,

$$\gamma_K(f, g, t)(x) - \gamma_K(f, h, t)(x) = g(x) - \max\{f(x) - t\|f - h\|_\infty, h(x)\} \leq g(x) - h(x) \leq \|g - h\|_\infty.$$

In a similar way, one can also obtain

$$\gamma_K(f, h, t)(x) - \gamma_K(f, g, t)(x) \leq \|g - h\|_\infty$$

for any $t \in [0, 1]$. Hence,

$$|\gamma_K(f, g, t)(x) - \gamma_K(f, h, t)(x)| \leq \|g - h\|_\infty.$$

(b) For $f(x) \leq g(x)$ and $f(x) \leq h(x)$, this is similar to the previous case.

(c)   If $f(x) \geq g(x)$ and $f(x) \leq h(x)$, then

$$\gamma_K(f, g, t)(x), \gamma_K(f, h, t)(x) \in [g(x), h(x)].$$

Therefore,

$$|\gamma_K(f, g, t)(x) - \gamma_K(f, h, t)(x)| \leq h(x) - g(x) \leq \|g - h\|_\infty.$$

(d)   For $f(x) \leq g(x)$ and $f(x) \geq h(x)$, this is similar to the previous case.

Since x is arbitrary, we finally have

$$\|\gamma_K(f, g, t) - \gamma_K(f, h, t)\|_\infty \leq \|g - h\|_\infty.$$

(5)   Fix arbitrary $x \in X$. Suppose $f(x) \geq g(x)$. Then

$$\phi(x) = \max\{f(x) - s\|f - g\|_\infty, g(x)\}, \quad \psi(x) = \max\{f(x) - t\|f - g\|_\infty, g(x)\}.$$

By property (1) of this proposition, we know $\|\phi - \psi\|_\infty = (t - s)\|f - g\|_\infty$. Moreover, since $\phi(x) \geq \psi(x)$,

$$\gamma_K(\phi, \psi, \lambda)(x) = \max\{\phi(x) - \lambda\|\phi - \psi\|_\infty, \psi(x)\}.$$

Observe that

$$\phi(x) - \lambda\|\phi - \psi\|_\infty = \max\{f(x) - s\|f - g\|_\infty, g(x)\} - \lambda(t - s)\|f - g\|_\infty$$
$$= \max\{f(x) - ((1 - \lambda)s + \lambda t)\|f - g\|_\infty, g(x) - \lambda(t - s)\|f - g\|_\infty\}.$$

Since $f(x) - ((1 - \lambda)s + \lambda t)\|f - g\|_\infty \geq f(x) - t\|f - g\|_\infty$ and $g(x) \geq g(x) - \lambda(t - s)\|f - g\|_\infty$, we finally have

$$\gamma_K(\phi, \psi, \lambda)(x) = \max\{f(x) - ((1 - \lambda)s + \lambda t)\|f - g\|_\infty, g(x)\} = \gamma_K(f, g, (1 - \lambda)s + \lambda t)(x).$$

One can do a similar proof for the case when $f(x) \leq g(x)$. Hence,

$$\gamma_K(\phi, \psi, \lambda) = \gamma_K(f, g, (1 - \lambda)s + \lambda t).$$

(6)   Consider the special case $s = 0$ and $t = 1$. Fix an arbitrary $x \in X$. Observe that $\gamma_K(f, g, r)(x)$ is between $f(x)$ and $g(x)$. Therefore,

$$|\gamma_K(f, g, r)(x) - h(x)| \leq \max\{|f(x) - h(x)|, |g(x) - h(x)|\} \leq \max\{\|f - h\|_\infty, \|g - h\|_\infty\}.$$

Since x is arbitrary,

$$\|\gamma_K(f, g, r) - h\|_\infty \leq \max\{\|f - h\|_\infty, \|g - h\|_\infty\} = \max\{\|\gamma_K(f, g, 0) - h\|_\infty, \|\gamma_K(f, g, 1) - h\|_\infty\}.$$

For general s and t, combine this result with property (5). □

A.2 Some properties of γ_K

Example A.1   In this example, we will see that some of the nice properties considered in Lemma 2.20 do not hold for the Katz geodesic bicombing (see Definition 9.22). Consider X to be a two-point space. Then $L^\infty(X)$ can be regarded as $\mathbb{R}^2$ with the ℓ^∞–norm.

(1) **Katz's geodesic bicombing is not conical in general** We will find $f, f', g, g' \in L^\infty(X)$ and $t \in [0, 1]$ such that

$$\|\gamma_K(f, g, t) - \gamma_K(f', g', t)\|_\infty > (1-t)\|f - f'\|_\infty + t\|g - g'\|_\infty.$$

Let $f = f' = (0, 0)$, $g = (c, d)$ for some $0 < c < d$, and $g = (c', d')$ for some $0 < c' < d'$. Then

$$\gamma_K(0, g, t) = \begin{cases} (td, td) & \text{if } t \in [0, c/d], \\ (c, td) & \text{if } t \in [c/d, 1], \end{cases}$$

and we have a similar expression for $\gamma_K(0, g', t)$. Hence, for any $t \in [\max\{c/d, c'/d'\}, 1)$,

$$\gamma_K(0, g, t) = (c, td) \quad \text{and} \quad \gamma_K(0, g', t) = (c', td').$$

Therefore, if we choose $|c - c'| > |d - d'|$ (for example, $(c, d) = (4, 5)$ and $(c', d') = (1, 5)$), then

$$\|\gamma_K(0, g, t) - \gamma_K(0, g', t)\|_\infty = |c - c'|$$

and

$$t\|g - g'\|_\infty = t|c - c'|.$$

Hence,

$$\|\gamma_K(0, g, t) - \gamma_K(0, g', t)\|_\infty > t\|g - g'\|_\infty.$$

So the Katz geodesic bicombing is not conical. In particular, this implies it is not convex.

(2) **Katz geodesic bicombing is not reversible in general** We will find $f, g \in L^\infty(X)$ and $t \in [0, 1]$ such that

$$\gamma_K(f, g, t) \neq \gamma_K(g, f, 1-t).$$

Let $f = (0, 0)$ and $g = (c, d)$ for some $0 < c < d$ as before. Then

$$\gamma_K(0, g, t) = \begin{cases} (td, td) & \text{if } t \in [0, c/d], \\ (c, td) & \text{if } t \in [c/d, 1], \end{cases}$$

and

$$\gamma_K(g, 0, t) = \begin{cases} (c - td, (1-t)d) & \text{if } t \in [0, c/d], \\ (0, (1-t)d) & \text{if } t \in [c/d, 1]. \end{cases}$$

Now, if we choose $t \in (0, \min\{c/d, 1-c/d\}]$, we have $\gamma_K(0, g, t) = (td, td)$ and $\gamma_K(g, 0, 1-t) = (0, td)$. Hence,

$$\gamma_K(0, g, t) \neq \gamma_K(g, 0, 1-t).$$

A.3 Proof of the generalized functorial nerve lemma

Theorem 4.2 (generalized functorial nerve lemma) *Let X and Y be two paracompact spaces, $\rho \colon X \to Y$ be a continuous map, $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ and $\mathcal{V} = \{V_\beta\}_{\beta \in B}$ be good open covers (every nonempty finite intersection is contractible) of X and Y, respectively, based on arbitrary index sets A and B, and $\pi \colon A \to B$ be a map such that*

$$\rho(U_\alpha) \subseteq V_{\pi(\alpha)} \quad \text{for any } \alpha \in A.$$

Let N$\mathcal{U}$ and N$\mathcal{V}$ be the nerves of $\mathcal{U}$ and $\mathcal{V}$, respectively. Observe that, since $U_{\alpha_0} \cap \cdots \cap U_{\alpha_n} \neq \varnothing$ implies $V_{\pi(\alpha_0)} \cap \cdots \cap V_{\pi(\alpha_n)} \neq \varnothing$, π induces the canonical simplicial map $\bar{\pi} : \mathrm{N}\mathcal{U} \to \mathrm{N}\mathcal{V}$.

Then there exist homotopy equivalences $X \to \mathrm{N}\mathcal{U}$ and $Y \to \mathrm{N}\mathcal{V}$ that commute with ρ and $\bar{\pi}$ up to homotopy:

$$
\begin{array}{ccc}
X & \longrightarrow & \mathrm{N}\mathcal{U} \\
\rho \downarrow & & \downarrow \bar{\pi} \\
Y & \longrightarrow & \mathrm{N}\mathcal{V}
\end{array}
$$

Our proof of Theorem 4.2 invokes many elements of [49, Section 4.G], which provides a proof of the classical nerve lemma.

Definition A.2 Let X be a topological space and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in \Lambda}$ be an open covering of X (Λ is an arbitrary index set). For any $\sigma = \{\alpha_0, \ldots, \alpha_n\} \in \mathrm{N}\mathcal{U}$, the nonempty intersection $U_{\alpha_0} \cap \cdots \cap U_{\alpha_n}$ is denoted by U_σ. Note that, when $\sigma = \{\alpha_0, \ldots, \alpha_n\} \in \mathrm{N}\mathcal{U}$ and σ' is an n'–face of σ, there are the canonical inclusions

$$
i_{\sigma\sigma'} : U_\sigma \hookrightarrow U_{\sigma'} \quad \text{and} \quad j_{\sigma\sigma'} : \Delta_{n'} \hookrightarrow \Delta_n.
$$

Then, the *complex of spaces* corresponding to $\mathcal{U}$ consists of the set of all U_σ and the set of all canonical inclusions $i_{\sigma\sigma'}$ over all possible $\sigma' \subseteq \sigma \in \mathrm{N}\mathcal{U}$.

The *realization* of this complex of spaces, denoted by $\Delta X_\mathcal{U}$, is defined as

$$
\Delta X_\mathcal{U} := \bigsqcup_{\sigma = \{\alpha_0, \ldots, \alpha_n\} \in \mathrm{N}\mathcal{U}} U_\sigma \times \Delta_n / \sim,
$$

where $(x, p) \sim (x', p')$ whenever $i_{\sigma\sigma'}(x) = x'$ and $j_{\sigma\sigma'}(p') = p$.

We need the following slight improvements of Propositions 4G.1 and 4G.2 of [49]. These improved claims are actually implicit in their respective proofs; see [49, pages 458–459].

Proposition A.3 [49, Proposition 4G.1] *Let X be a topological space and $\mathcal{U} = \{U_\alpha\}_{\alpha \in \Lambda}$ be a good open cover of X (every nonempty finite intersection is contractible). Then*

$$
f : \Delta X_\mathcal{U} \to \mathrm{N}\mathcal{U}, \qquad (x, p) \mapsto p \quad \text{if } (x, p) \in U_\sigma \times \Delta_n,
$$

is a homotopy equivalence between $\Delta X_\mathcal{U}$ and $\mathrm{N}\mathcal{U}$.

Proof First of all, since U_σ is contractible whenever $\sigma \in \mathrm{N}\mathcal{U}$, note that there is a homotopy equivalence $\phi_\sigma : U_\sigma \to \{*\}$ for any $\sigma \in \mathrm{N}\mathcal{U}$.

The homotopy equivalence between $\Delta X_\mathcal{U}$ and $\mathrm{N}\mathcal{U}$ is just a special case of [49, Proposition 4G.1]. The choice of f is implicit in the fact that both of $\Delta X_\mathcal{U}$ and $\mathrm{N}\mathcal{U}$ are deformation retracts of $\Delta M X_\mathcal{U}$ where $\Delta M X_\mathcal{U}$ is the realization of the complex of spaces consisting of the mapping cylinders $M\phi_\sigma$ for any $\sigma \in \mathrm{N}\mathcal{U}$ and the canonical inclusions between them. $\qquad \square$

Proposition A.4 [49, Proposition 4G.2] *Let X be a paracompact space, $\mathcal{U} = \{U_\alpha\}_{\alpha \in \Lambda}$ be an open cover of X, and $\{\psi_\alpha\}_{\alpha \in \Lambda}$ be a partition of unity subordinate to the cover $\mathcal{U}$ (it must exist since X is paracompact). Then*

$$g \colon X \to \Delta X_{\mathcal{U}}, \quad x \mapsto \big(x, (\psi_\alpha(x))_{\alpha \in \Lambda}\big),$$

is a homotopy equivalence between X and $\Delta X_{\mathcal{U}}$.

Proof The proof is the same as [49, Proposition 4G.2].   □

Lemma A.5 *Let X and Y be two topological spaces, $\rho \colon X \to Y$ be a continuous map, $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ and $\mathcal{V} = \{V_\beta\}_{\beta \in B}$ be good open covers (every nonempty finite intersection is contractible) of X and Y respectively, based on arbitrary index sets A and B, and $\pi \colon A \to B$ be a map such that*

$$\rho(U_\alpha) \subseteq V_{\pi(\alpha)}$$

for any $\alpha \in A$.

Let $N\mathcal{U}$ and $N\mathcal{V}$ be the nerves of $\mathcal{U}$ and $\mathcal{V}$, respectively. Observe that π induces the canonical simplicial map $\bar{\pi} \colon N\mathcal{U} \to N\mathcal{V}$ since $U_{\alpha_0} \cap \cdots \cap U_{\alpha_n} \neq \varnothing$ implies $V_{\pi(\alpha_0)} \cap \cdots \cap V_{\pi(\alpha_n)} \neq \varnothing$, and ρ induces the canonical map $\bar{\rho} \colon \Delta X_{\mathcal{U}} \to \Delta Y_{\mathcal{V}}$ mapping (x, p) to $(\rho(x), \bar{\pi}(p))$.

Then there exist homotopy equivalences $f \colon \Delta X_{\mathcal{U}} \to N\mathcal{U}$ and $f' \colon \Delta Y_{\mathcal{V}} \to N\mathcal{V}$ which commute with $\bar{\rho}$ and $\bar{\pi}$:

$$
\begin{array}{ccc}
\Delta X_{\mathcal{U}} & \xrightarrow{\ f\ } & N\mathcal{U} \\
{\scriptstyle \bar{\rho}}\downarrow & & \downarrow{\scriptstyle \bar{\pi}} \\
\Delta Y_{\mathcal{V}} & \xrightarrow{\ f'\ } & N\mathcal{V}
\end{array}
$$

Proof By Proposition A.3,

$$f \colon \Delta X_{\mathcal{U}} \to N\mathcal{U}, \qquad (x, p) \mapsto p \quad \text{if } (x, p) \in U_\sigma \times \Delta_n,$$

is a homotopy equivalence between $\Delta X_{\mathcal{U}}$ and $N\mathcal{U}$. Also,

$$f' \colon \Delta Y_{\mathcal{V}} \to N\mathcal{V}, \qquad (y, q) \mapsto q \quad \text{if } (y, q) \in V_\sigma \times \Delta_n,$$

is a homotopy equivalence between $\Delta Y_{\mathcal{V}}$ and $N\mathcal{V}$.

To check the commutativity of the diagram, fix an arbitrary $(x, p) \in U_\sigma \times \Delta_n \subseteq \Delta X_{\mathcal{U}}$. Then,

$$\bar{\pi} \circ f(x, p) = \bar{\pi}(p) = f'(\rho(x), \bar{\pi}(p)) = f' \circ \bar{\rho}(x, p).$$

Hence, $\bar{\pi} \circ f = f' \circ \bar{\rho}$, as we wanted.   □

Lemma A.6 *Let X and Y be two paracompact spaces, $\rho \colon X \to Y$ be a continuous map, $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ and $\mathcal{V} = \{V_\beta\}_{\beta \in B}$ be open covers of X and Y respectively, based on arbitrary index sets A and B, and $\pi \colon A \to B$ be a map such that*

$$\rho(U_\alpha) \subseteq V_{\pi(\alpha)}$$

for any $\alpha \in A$.

Let $\mathrm{N}\mathcal{U}$ and $\mathrm{N}\mathcal{V}$ be the nerves of $\mathcal{U}$ and $\mathcal{V}$, respectively. Observe that π induces the canonical simplicial map $\bar{\pi}\colon \mathrm{N}\mathcal{U} \to \mathrm{N}\mathcal{V}$ since $U_{\alpha_0} \cap \cdots \cap U_{\alpha_n} \neq \varnothing$ implies $V_{\pi(\alpha_0)} \cap \cdots \cap V_{\pi(\alpha_n)} \neq \varnothing$, and ρ induces the canonical map $\bar{\rho}\colon \Delta X_{\mathcal{U}} \to \Delta Y_{\mathcal{V}}$ mapping (x, p) to $(\rho(x), \bar{\pi}(p))$.

Then there exist homotopy equivalences $g\colon X \to \Delta X_{\mathcal{U}}$ and $g'\colon Y \to \Delta Y_{\mathcal{V}}$ which commute with ρ and $\bar{\rho}$ up to homotopy:

$$\begin{array}{ccc} X & \xrightarrow{\ g\ } & \Delta X_{\mathcal{U}} \\ {\scriptstyle \rho}\downarrow & & \downarrow{\scriptstyle \bar{\rho}} \\ Y & \xrightarrow{\ g'\ } & \Delta Y_{\mathcal{V}} \end{array}$$

Proof By Proposition A.4,

$$g\colon X \to \Delta X_{\mathcal{U}}, \quad x \mapsto \big(x, (\psi_\alpha(x))_{\alpha\in\Lambda}\big),$$

is a homotopy equivalence between X and $\Delta X_{\mathcal{U}}$, where $\{\psi_\alpha\}_{\alpha\in A}$ is a partition of unity subordinate to the cover $\mathcal{U}$. And,

$$g'\colon Y \to \Delta Y_{\mathcal{V}}, \quad y \mapsto \big(y, (\psi'_\beta(y))_{\beta\in B}\big),$$

is a homotopy equivalence between Y and $\Delta Y_{\mathcal{V}}$ where $\{\psi'_\beta\}_{\beta\in B}$ is a partition of unity subordinate to the cover $\mathcal{V}$.

Finally, we will show that $\bar{\rho}\circ g \simeq g'\circ\rho$. Observe that, for arbitrary $x \in X$,

$$\bar{\rho}\circ g(x) = \bar{\rho}\big(x, (\psi_\alpha(x))_{\alpha\in\Lambda}\big) = \big(\rho(x), \bar{\pi}\big((\psi_\alpha(x))_{\alpha\in A}\big)\big)$$

and

$$g'\circ\rho(x) = g'(\rho(x)) = \big(\rho(x), \big(\psi'_\beta(\rho(x))\big)_{\beta\in B}\big).$$

Hence, one can just construct a homotopy between $\bar{\rho}\circ g$ and $g'\circ\rho$ by

$$h\colon X \times [0,1] \to \Delta Y_{\mathcal{V}}, \quad (x,t) \mapsto \big(\rho(x), (1-t)\bar{\pi}\big((\psi_\alpha(x))_{\alpha\in A}\big) + t\big(\psi'_\beta(\rho(x))\big)_{\beta\in B}\big).$$

Here, note that the linear interpolation between $\bar{\pi}\big((\psi_\alpha(x))_{\alpha\in A}\big)$ and $\big(\psi'_\beta(\rho(x))\big)_{\beta\in B}$ is well defined since, because of the properties of partition of unity and the assumption that $\rho(U_\alpha) \subseteq V_{\pi(\alpha)}$,

$$\rho(x) \in \bigcap_{\alpha:\psi_\alpha(x)>0} V_{\pi(\alpha)} \cap \bigcap_{\beta:\psi'_\beta(\rho(x))>0} V_\beta,$$

so

$$\{\pi(\alpha) \in B \mid \psi_\alpha(x) > 0\} \cup \{\beta \in B \mid \psi'_\beta(\rho(x)) > 0\}$$

forms a simplex in $\mathrm{N}\mathcal{V}$. □

Finally, one can prove the functorial nerve lemma.

Proof of Theorem 4.2 Combine Lemmas A.5 and A.6. □

A.4 Proof of $\mathrm{VR}_r(\mathbb{S}^n) \simeq \mathbb{S}^n$ for $r \in \big(0, \arccos(-1/(n+1))\big]$

Theorem 7.1 *For any $n \in \mathbb{Z}_{>0}$, we have $\mathrm{VR}_r(\mathbb{S}^n) \simeq \mathbb{S}^n$ for any $r \in \big(0, \arccos(-1/(n+1))\big]$.*

Since the case of $\mathbb{S}^1$ is already proved in [1], it is enough to prove the above theorem for $\mathbb{S}^n$ with $n \geq 2$. Moreover, unlike the other parts of the paper, in this section we discriminate between the simplicial complex $\mathrm{VR}_r(\mathbb{S}^n)$ and its realization $|\mathrm{VR}_r(\mathbb{S}^n)|$.

To prove Theorem 7.1, we will basically emulate the proof strategy of Hausmann in [50]. However, a crucial modification will be necessary, which requires the following version of Jung's theorem:

Definition A.7 Given a nonempty subset $A \subset \mathbb{S}^n$, its *geodesic convex hull* $\mathrm{conv}_{\mathbb{S}^n}(A)$ is defined to be the set consisting of the union of all minimizing geodesics between pairs of points in A. It is clear that when A is contained in an open hemisphere, $\mathrm{conv}_{\mathbb{S}^n}(A) = \{\Pi_{\mathbb{S}^n}(c) \mid c \in \mathrm{conv}(A)\}$ where $\Pi_{\mathbb{S}^n}(p) := p/\|p\|$ for $p \neq 0$ and $\Pi_{\mathbb{S}^n}(p) := 0$ otherwise.

Theorem A.8 (a version of Jung's theorem for spheres) *For any $n \geq 1$, if $A \subset \mathbb{S}^n$ satisfies $D := \mathrm{diam}(A) < \arccos(-1/(n+1))$, then there must be $u \in \mathrm{conv}_{\mathbb{S}^n}(A)$ such that $A \subseteq \bar{B}_{\psi(D)}(u, \mathbb{S}^n)$, where*

$$\psi : \left[0, \arccos\left(-\frac{1}{n+1}\right)\right] \to \mathbb{R}_{\geq 0}, \quad D \mapsto \arccos\left(\sqrt{\frac{1+(n+1)\cos D}{n+2}}\right).$$

The version of Jung's theorem stated above is different from the one considered by Katz [54, Lemma 2] in the following two senses:

(1) We provide a precise formula for the radius $\psi(D)$ of the closed ball covering A, depending on $D = \mathrm{diam}(A)$. In particular, our version is stronger when D is small.

(2) On the contrary, if D is large (close to $\arccos(-1/(n+1))$), then the radius $\psi(D)$ can be as large as $\frac{\pi}{2}$. But $\frac{\pi}{2}$ is strictly greater number than $\pi - \arccos(-1/(n+1))$ which is the radius guaranteed by Katz's version. So, for the case when D is large, Katz's version is stronger.

The proof of our version is somewhat similar to the classical proof in [37].

Remark A.9 The map ψ satisfies:

(1) $\psi(D) \leq \frac{\pi}{2}$ for any $D \in [0, \arccos(-1/(n+1))]$.

(2) ψ is an increasing function.

(3) $\lim_{D \to 0+} \psi(D) = 0$.

Proof of Theorem A.8 Without loss of generality, one can assume A is compact. Recall that one can view $\mathbb{S}^n$ as a subset of $\mathbb{R}^{n+1}$,

$$\mathbb{S}^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\}.$$

Also, for any $x, y \in \mathbb{S}^n$, the Euclidean norm $\|x - y\|$ and the geodesic distance $d_{\mathbb{S}^n}(x, y)$ satisfy

$$\|x - y\| = \sqrt{2 - 2\cos(d_{\mathbb{S}^n}(x, y))}.$$

Now, if we apply [37, Lemma 2.10.40] with $P := A \times \{1\}$, there are $p \in \mathbb{R}^{n+1}$ and $c \geq 0$ such that

(1) for all $a \in A$, $\|a - p\| \leq c$;

(2) p belongs to the convex hull of $\{a \in A \mid \|a - p\| = c\}$.

Therefore, there are nonnegative numbers $\lambda_1, \ldots, \lambda_{n+2}$ and $a_1, \ldots, a_{n+2} \in \{a \in A \mid \|a - x\| = c\}$ such that

(1) $p = \sum_{i=1}^{n+2} \lambda_i a_i$;

(2) $1 = \sum_{i=1}^{n+2} \lambda_i$.

Hence, one can easily check $\|p\| \leq 1$. Also, since

$$\|a_i - a_j\| = \sqrt{2 - 2\cos(d_{\mathbb{S}^n}(x, y))} \leq \sqrt{2 - 2\cos D} < \sqrt{2 + \frac{2}{n+1}},$$

$p \neq 0$ by [32, Lemma 1]. Furthermore, for each $j \in \{1, \ldots, n+2\}$,

$$\begin{aligned}
2c^2 &= \sum_{i=1}^{n+2} \lambda_i \left(2c^2 - 2\langle (a_i - p), (a_j - p) \rangle\right) \\
&= \sum_{i=1}^{n+2} \lambda_i \|(a_i - p) - (a_j - p)\|^2 \\
&= \sum_{i=1}^{n+2} \lambda_i \|a_i - a_j\|^2 \\
&\leq \sum_{i \neq j} \lambda_i (2 - 2\cos D) \\
&= (1 - \lambda_j)(2 - 2\cos D).
\end{aligned}$$

So, by summation with respect to j, we have $2(n+2)c^2 \leq (n+1)(2 - 2\cos D)$. Therefore,

$$c \leq \sqrt{\frac{(n+1)(1 - \cos D)}{n+2}} < 1.$$

Finally, let $u := p/\|p\|$. Then, $u \in \text{conv}_{\mathbb{S}^n}(A)$ since $p \in \text{conv}(A)$. Also, one can check that

$$\|a - u\| \leq \sqrt{2 - 2\sqrt{1 - c^2}} \leq \sqrt{2 - 2\sqrt{\frac{1 + (n+1)\cos D}{n+2}}}$$

for all $a \in A$. This implies

$$d_{\mathbb{S}^n}(a, u) \leq \arccos\left(\sqrt{\frac{1 + (n+1)\cos D}{n+2}}\right) = \psi(D)$$

for any $a \in A$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

A.4.1 The proof of Theorem 7.1 Choose a total ordering on the points of $\mathbb{S}^n$. From now on, whenever we describe a finite subset of $\mathbb{S}^n$ by $\{x_0, \ldots, x_q\}$, we suppose that $x_0 < x_1 < \cdots < x_q$. Let r be in the interval $\big(0, \arccos(-1/(n+1))\big]$. We shall associate to each q–simplex $\sigma := \{x_0, \ldots, x_q\} \in \mathrm{VR}_r(\mathbb{S}^n)$ a singular q–simplex $T_\sigma \colon \Delta_q \to \mathbb{S}^n$. Recall that the standard Euclidean q–simplex Δ_q is defined as

$$\Delta_q := \left\{ \sum_{i=0}^{q} t_i e_i \ \Big| \ t_i \in [0, 1] \text{ and } \sum_{i=0}^{q} t_i = 1 \right\}.$$

This map T_σ is defined inductively as follows: Set $T(e_0) = x_0$. Suppose that $T_\sigma(z)$ is defined for $y = \sum_{i=0}^{p-1} s_i e_i$. Let $z := \sum_{i=0}^{p} t_i e_i$. If $t_p = 1$, we pose $T_\sigma(z) = x_p$. Otherwise, let

$$x := T_\sigma\left(\frac{1}{1-t_p} \sum_{i=0}^{p-1} t_i e_i \right).$$

We define $T_\sigma(z)$ as the point on the unique shortest geodesic joining x to x_p with

$$d_{\mathbb{S}^n}(x, T_\sigma(z)) = t_p \cdot d_{\mathbb{S}^n}(x, x_p)$$

(the unique shortest geodesic exists since $\mathrm{conv}_{\mathbb{S}^n}(\{x_0, \ldots, x_q\})$ must be contained in some open ball of radius smaller than $\frac{\pi}{2}$ by Theorem A.8). To sum up, T_σ is defined inductively on Δ_p for $p \leq q$ as the *geodesic join* of $T_\sigma(\Delta_{p-1})$ with x_p.

If σ' is a face of σ of dimension p, we form the euclidean sub–p–simplex Δ' of Δ_q formed by the points $\sum_{i=0}^{q} t_i e_i \in \Delta_q$ with $t_i = 0$ if $x_i \notin \sigma'$. One can check by induction on $\dim \sigma'$ that

$$(8) \qquad\qquad\qquad\qquad\qquad T_{\sigma'} = T_\sigma|_{\Delta'}.$$

By (8), the correspondence $\sigma \mapsto T_\sigma$ gives rise to a continuous map

$$T \colon |\mathrm{VR}_r(\mathbb{S}^n)| \to \mathbb{S}^n.$$

Here is a quick overview of how we will prove Theorem 7.1. Through Lemmas A.10 (which enables the application of Hausmann's "crushings" on sufficiently small subsets of spheres), A.11 and A.12, we will prove that T induces an isomorphism at homology level. Also, by Lemma A.13, we will prove that T also induces an isomorphism at the level of fundamental groups. Finally, the proof of Theorem 7.1 will follow by invoking the homology Whitehead theorem.

Lemma A.10 *Let $x \in \mathbb{S}^n$ and $y, z \in B_{\pi/2}(x, \mathbb{S}^n)$. Let $\gamma_y \colon [0, 1] \to \mathbb{S}^n$ and $\gamma_z \colon [0, 1] \to \mathbb{S}^n$ be the unique shortest geodesics from x to y and x to z. Then*

$$d_{\mathbb{S}^n}(\gamma_y(s), \gamma_z(s)) \leq d_{\mathbb{S}^n}(\gamma_y(t), \gamma_z(t))$$

for any $0 \leq s \leq t \leq 1$.

Proof Let $d_{\mathbb{S}^n}(x, y) = a$ and $d_{\mathbb{S}^n}(x, z) = b$. Without loss of generality, one can assume $a \geq b$. By the spherical law of cosine, one can compute

$$\cos(d_{\mathbb{S}^n}(\gamma_y(t), \gamma_z(t))) = \cos(ta)\cos(tb) + \sin(ta)\sin(tb)\cos\theta$$

for any $t \in [0, 1]$, where θ is the angle between γ_y and γ_z at x.

Now, consider the map

$$f : [0, 1] \to \mathbb{R}_{\geq 0}, \quad t \mapsto \cos(ta)\cos(tb) + \sin(ta)\sin(tb)\cos\theta.$$

To complete the proof, it is enough to show this f is nonincreasing. Observe that

$$
\begin{aligned}
f'(t) &= -a\sin(ta)\cos(tb) - b\cos(ta)\sin(tb) + a\cos(ta)\sin(tb)\cos\theta + b\sin(ta)\cos(tb)\cos\theta \\
&\leq -a\sin(ta)\cos(tb) - b\cos(ta)\sin(tb) + a\cos(ta)\sin(tb) + b\sin(ta)\cos(tb) \\
&\qquad\qquad\qquad\qquad\qquad -(a-b)\sin(ta)\cos(tb) + (a-b)\cos(ta)\sin(tb) \\
&= -(a-b)\sin(t(a-b)) \\
&\leq 0.
\end{aligned}
$$

Hence, f is nonincreasing. □

The following lemma is an analogue of [50, Proposition 3.3]:

Lemma A.11 *Let $0 < r' \leq r \leq \arccos(-1/(n+1))$. Then the canonical inclusion $\mathrm{VR}_{r'}(\mathbb{S}^n) \subset \mathrm{VR}_r(\mathbb{S}^n)$ induces an isomorphism on homology.*

Proof Let $\sigma = \{x_0, \ldots, x_q\}$ be a simplex of $\mathrm{VR}_r(\mathbb{S}^n)$ and let I_σ be the image of T_σ. If σ' is a face of σ then $I_{\sigma'} \subseteq I_\sigma$, and thus $\mathrm{VR}_\delta(I_{\sigma'})$ is a subcomplex of $\mathrm{VR}_\delta(I_\sigma)$ for all $\delta > 0$. On the other hand, $\mathrm{VR}_\delta(I_\sigma)$ is acyclic for all $\delta > 0$. Indeed, by Theorem A.8, there exists $u \in I_\sigma$ such that $I_\sigma \subset B_{\pi/2}(u, \mathbb{S}^n)$. So, one can consider the obvious crushing from I_σ to $\{x\}$ via the shortest geodesics. So, $\mathrm{VR}_\delta(I_\sigma)$ must be contractible by Lemma A.10 and [50, Corollary 2.3]. These considerations show that for $0 < \delta' \leq \delta \leq \arccos(-1/(n+1))$, the correspondence

$$\sigma \mapsto \mathrm{VR}_{\delta'}(I_\sigma)$$

is an acyclic carrier $\Phi_{\delta,\delta'}$ from $\mathrm{VR}_\delta(\mathbb{S}^n)$ to $\mathrm{VR}_{\delta'}(\mathbb{S}^n)$ (see [68, Section 13]).

We now use the acyclic carrier theorem [68, Theorem 13.3]. This implies that there exists an augmentation preserving chain map $\nu : C_*(\mathrm{VR}_r(\mathbb{S}^n)) \to C_*(\mathrm{VR}_{r'}(\mathbb{S}^n))$ which is carried by $\Phi_{r,r'}$. Let μ denote the canonical inclusion from $\mathrm{VR}_{r'}(\mathbb{S}^n)$ into $\mathrm{VR}_r(\mathbb{S}^n)$. Then $\phi_{r',r'}$ is an acyclic carrier for both $\nu \circ \mu_*$ and the identity of $C_*(\mathrm{VR}_{r'}(\mathbb{S}^n))$. By the acyclic carrier theorem again, these two maps are chain homotopic and thus $\nu \circ \mu_*$ induces the identity on $\mathrm{H}_*(\mathrm{VR}_{r'}(\mathbb{S}^n))$. The same argument shows that $\mu_* \circ \nu$ induces the identity on $\mathrm{H}_*(\mathrm{VR}_r(\mathbb{S}^n))$ (using the acyclic carrier $\Phi_{r,r}$). □

We will now compare the simplicial homology of $\mathrm{VR}_r(\mathbb{S}^n)$ with the singular homology of M. Formula (8) shows that the correspondence $\sigma \mapsto T_\sigma$ gives rise to a chain map

$$T_\sharp^r : C_*(\mathrm{VR}_r(\mathbb{S}^n)) \to \mathrm{SC}_*(\mathbb{S}^n),$$

where $\mathrm{SC}_*(\mathbb{S}^n)$ denotes the singular chain complex of $\mathbb{S}^n$.

The following lemma is an analogue of [50, Proposition 3.4]:

Lemma A.12 *If $0 < r \leq \arccos(-1/(n+1))$ then the chain map $T_\sharp$ induces an isomorphism on homology.*

Proof   We shall need a few accessory chain complexes. For $\delta > 0$, denote by $\mathrm{SC}_*(\mathbb{S}^n; \delta)$ the subchain complexes of $\mathrm{SC}_*(\mathbb{S}^n)$ based on singular simplexes τ such that there exists $u \in \mathbb{S}^n$ with the image of τ contained in the open ball $B_\delta(u, \mathbb{S}^n)$. Recall that the inclusion $\mathrm{SC}_*(\mathbb{S}^n; \delta) \hookrightarrow \mathrm{SC}_*(\mathbb{S}^n)$ induces an isomorphism on homology [50, Theorem 31.5].

We shall also use the ordered chain complex $C'_*(\mathrm{VR}_r(\mathbb{S}^n))$. the group $C'_q(\mathrm{VR}_r(\mathbb{S}^n))$ is free abelian group on $(q+1)$–tuples $(x_0, \ldots, x_q)$ such that $\{x_0\} \cup \cdots \cup \{x_q\}$ is a simplex of $\mathrm{VR}_r(\mathbb{S}^n)$. One can view that $C_*(\mathrm{VR}_r(\mathbb{S}^n))$ as a subchain complex of $C'_*(\mathrm{VR}_r(\mathbb{S}^n))$ by associating a q–simplex $\{x_0, \ldots, x_q\}$ of $\mathrm{VR}_r(\mathbb{S}^n)$ (with our convention that $x_0 < x_1 < \cdots < x_q$ for the well-ordering on $\mathbb{S}^n$) the $(q+1)$–tuple $(x_0, \ldots, x_q)$. It is also classical that this inclusion is homology isomorphism [50, Theorem 3.6]. Observe that the construction $\sigma \mapsto T_\sigma$ does not require that the vertices of σ are all distinct. One can then define T_σ for a basis element of $C'_*(\mathrm{VR}_r(\mathbb{S}^n))$ and thus extend to a chain map $T_\sharp^r : C'_*(\mathrm{VR}_r(\mathbb{S}^n)) \to \mathrm{SC}_*(\mathbb{S}^n; \psi(r))$. Now, choose $r' < r$ such that $\psi(r') \leq \frac{1}{2}r$. One then has the commutative diagram

$$
\begin{CD}
C'_*(\mathrm{VR}_{r'}(\mathbb{S}^n)) @>{T_\sharp^{r'}}>> \mathrm{SC}_*(\mathbb{S}^n; \psi(r')) \\
@VVV @VVV \\
C'_*(\mathrm{VR}_r(\mathbb{S}^n)) @>{T_\sharp^r}>> \mathrm{SC}_*(\mathbb{S}^n; \psi(r))
\end{CD}
$$

Let $\tau : \Delta_q \to \mathbb{S}^n$ be a singular simplex whose image is contained in some open ball of radius $\psi(r')$. The $(q+1)$–tuple $(\tau(e_0), \ldots, \tau(e_q))$ is element of $C'_q(\mathrm{VR}_r(\mathbb{S}^n))$. This correspondence gives rise to a chain map

$$R : \mathrm{SC}_*(\mathbb{S}^n; \psi(r')) \to C'_*(\mathrm{VR}_r(\mathbb{S}^n)).$$

The composition $R \circ T_\sharp^{r'}$ is equal to the canonical inclusion $C'_*(\mathrm{VR}_{r'}(\mathbb{S}^n)) \subset C'_*(\mathrm{VR}_r(\mathbb{S}^n))$ which induces a homotopy isomorphism by Lemma A.11. Let us now understand the composition

$$T_\sharp^r \circ R : \mathrm{SC}_*(\mathbb{S}^n; \psi(r')) \to \mathrm{SC}_*(\mathbb{S}^n; \psi(r)).$$

Let $\tau : \Delta_q \to \mathbb{S}^n$ be a singular simplex such that $\tau(\Delta_q) \subset B_{\psi(r')}(y, \mathbb{S}^n)$ for some $y \in \mathbb{S}^n$. Therefore, $\tau' := T_\sharp^r \circ R(\tau)$ also satisfies $\tau'(\Delta_q) \subset B_{\psi(r')}(y, \mathbb{S}^n)$ since $\psi(r') < \frac{\pi}{2}$. Hence, τ and τ' are canonically homotopic (following, for each $s \in \Delta_q$, the shortest geodesic joining $\tau(s)$ to $\tau'(s)$). As in the proof of

the homotopy axiom for singular homology [50, Section 30], these provide a chain homotopy between $T_\sharp^r \circ R$ and the inclusion $\mathrm{SC}_*(\mathbb{S}^n; \psi(r')) \subset \mathrm{SC}_*(\mathbb{S}^n; \psi(r))$. As said before, this inclusion is known to induce a homology isomorphism. Therefore, $T_\sharp^r \circ R$ induces an isomorphism on homology.

We have shown that both $R \circ T_\sharp^{r'}$ and $T_\sharp^r \circ R$ induce homology isomorphisms. Therefore, $T_\sharp^r$ induces a morphism both injective and surjective, hence a homology isomorphism. $\qquad \square$

Lemma A.13 *If $0 < r \le \arccos(-1/(n+1))$, the map*

$$T : |\mathrm{VR}_r(\mathbb{S}^n)| \to \mathbb{S}^n$$

induces an isomorphism on the fundamental groups.

Proof Let $\gamma : [0,1] \to \mathbb{S}^n$ represent an element of $\pi_1(\mathbb{S}^n)$. Choose large enough positive integer N such that $1/N$ is smaller than the Lebesgue number for the covering $\{\gamma^{-1}(B_{r/2}(x, \mathbb{S}^n))\}_{x \in \mathbb{S}^n}$. Then $d_{\mathbb{S}^n}(\gamma(k/N), \gamma((k+1)/N)) < r$ for any $k = 0, \dots, N-1$. Hence the path $\gamma|_{[(k/N),((k+1)/N)]}$ is then canonically homotopic to a parametrization of the shortest geodesic joining $\gamma(k/N)$ to $\gamma((k+1)/N)$. This shows that γ is homotopic to a composition γ' of geodesics in open balls of radius $\frac{1}{2}r$. Such a path γ' represents the image of T of an element of $\pi_1(|\mathrm{VR}_r(\mathbb{S}^n)|)$, the latter being identified with the edge-path group of the simplicial complex $\mathrm{VR}_r(\mathbb{S}^n)$ [78, pages 134–139]. Thus, $\pi_1 T : \pi_1(|\mathrm{VR}_r(\mathbb{S}^n)|) \to \pi_1(\mathbb{S}^n)$ is surjective.

Now, to prove injectivity, suppose $\pi_1 T([\alpha]) = 0$ where $\alpha : [0,1] \to |\mathrm{VR}_r(\mathbb{S}^n)|$ is a continuous map satisfying $\alpha(0) = \alpha(1)$. Moreover, again by [78, pages 134–139], one can assume α is induced by an edge-path of $\mathrm{VR}_r(\mathbb{S}^n)$. In other words, there is a positive integer N, and $x_0, \dots, x_{N-1}, x_N = x_0 \in \mathbb{S}^n$ such that $d_{\mathbb{S}^n}(x_i, x_{i+1}) < r$ and $\alpha(i/N) = x_i$ for $i = 0, \dots, N-1$ (here, we view x_i as a 0–simplex). Next, by the assumption, $[T \circ \alpha] = \pi_1 T([\alpha]) = 0$. This implies that there is a homotopy map

$$H : [0,1] \times [0,1] \to \mathbb{S}^n$$

such that $H(t, 1) = T \circ \alpha(t)$ and $H(t, 0) = H(0, s) = H(1, s) = x_0$ for any $t, s \in [0,1]$. Next, choose a large enough positive integer N' such that if we triangulate $[0,1] \times [0,1]$ with vertices $(k/N', l/N')$ for $k, l = 0, \dots, N'$, each triangle is contained in one of $\{H^{-1}(B_{r/2}(x, \mathbb{S}^n))\}_{x \in \mathbb{S}^n}$. Then

$$d_{\mathbb{S}^n}(H(k/N', l/N'), H(k'/N', l'/N')) < r$$

whenever $((k/N', l/N'), (k'/N', l'/N'))$ is an edge of the triangulation. Because of this observation, one can prove that the edge path $H(0, 1), H(1/N', 1), \dots, H((N'-1)/N', 1), H(1, 1)$ is equivalent to x_0. Also, it is easy to check that two edge paths $H(0, 1), H(1/N', 1), \dots, H((N'-1)/N', 1), H(1, 1)$ and $x_0, x_1, \dots, x_{N-1}, x_N$ are equivalent. This means that $[\alpha] = 0$. So $\pi_1 T$ is injective. $\qquad \square$

We are now in position to prove Theorem 7.1.

Proof of Theorem 7.1 As mentioned in the beginning of this section, one can assume $n \geq 2$. Hence, $\mathbb{S}^n$ is simply connected. Therefore, by Lemma A.13, $|\mathrm{VR}_r(\mathbb{S}^n)|$ is also simply connected. Also, by Lemma A.12 and the isomorphism between simplicial and singular homology [68, Section 34], T induces an isomorphism on homology. Therefore, T is a homotopy equivalence by [49, Corollary 4.33]. $\qquad\square$

References

[1] **M Adamaszek**, **H Adams**, *The Vietoris–Rips complexes of a circle*, Pacific J. Math. 290 (2017) 1–40 MR Zbl

[2] **M Adamaszek**, **H Adams**, **F Frick**, *Metric reconstruction via optimal transport*, SIAM J. Appl. Algebra Geom. 2 (2018) 597–619 MR Zbl

[3] **M Adamaszek**, **H Adams**, **E Gasparovic**, **M Gommel**, **E Purvine**, **R Sazdanovic**, **B Wang**, **Y Wang**, **L Ziegelmeier**, *Vietoris–Rips and Čech complexes of metric gluings*, from "34th international symposium on computational geometry" (B Speckmann, C D Tóth, editors), Leibniz Int. Proc. Inform. 99, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern (2018) art. id. 3 MR Zbl

[4] **M Adamaszek**, **H Adams**, **S Reddy**, *On Vietoris–Rips complexes of ellipses*, J. Topol. Anal. 11 (2019) 661–690 MR Zbl

[5] **H Adams**, **J Bush**, **J Mirth**, *Operations on metric thickenings*, from "Proceedings of the 3rd annual international applied category theory conference 2020" (D I Spivak, J Vicary, editors), Electron. Proc. Theor. Comput. Sci. 333, EPTCS (2021) 261–275 MR Zbl

[6] **H Adams**, **B Coskunuzer**, *Geometric approaches to persistent homology*, SIAM J. Appl. Algebra Geom. 6 (2022) 685–710 MR Zbl

[7] **H Adams**, **F Mémoli**, **M Moy**, **Q Wang**, *The persistent topology of optimal transport based metric thickenings*, preprint (2021) arXiv 2109.15061

[8] **N Aronszajn**, **P Panitchpakdi**, *Extension of uniformly continuous transformations and hyperconvex metric spaces*, Pacific J. Math. 6 (1956) 405–439 MR Zbl

[9] **S Awodey**, *Category theory*, 2nd edition, Oxford Logic Guides 52, Oxford Univ. Press (2010) MR Zbl

[10] **G Azumaya**, *Corrections and supplementaries to my paper concerning Krull–Remak–Schmidt's theorem*, Nagoya Math. J. 1 (1950) 117–124 MR Zbl Correction to "On generalized semi-primary rings and Krull–Remak–S chmidt's theorem", Jpn. J. Math. 19 (1948) 525–547

[11] **U Bauer**, *Ripser*, C++ code (2015) Available at https://github.com/Ripser/ripser

[12] **U Bauer**, **M Lesnick**, *Induced matchings of barcodes and the algebraic stability of persistence*, from "Computational geometry" (S-W Cheng, O Devillers, editors), ACM, New York (2014) 355–364 MR Zbl

[13] **A Blumberg**, **M Lesnick**, *Universality of the homotopy interleaving distance*, Trans. Amer. Math. Soc. 376 (2023) 8269–8307 MR Zbl

[14] **M Bonk**, **O Schramm**, *Embeddings of Gromov hyperbolic spaces*, Geom. Funct. Anal. 10 (2000) 266–306 MR Zbl

[15] **D Burago**, **Y Burago**, **S Ivanov**, *A course in metric geometry*, Graduate Studies in Math. 33, Amer. Math. Soc., Providence, RI (2001) MR Zbl

[16] **G Carlsson**, *Topology and data*, Bull. Amer. Math. Soc. 46 (2009) 255–308 MR Zbl

[17] **G Carlsson**, **V de Silva**, *Zigzag persistence*, Found. Comput. Math. 10 (2010) 367–405 MR Zbl

[18] **G Chaparro Sumalave**, *Vietoris–Rips complexes of the circle and the torus*, master's thesis, Universidad de los Andes (2016)

[19] **F Chazal**, **D Cohen-Steiner**, **L J Guibas**, **F Mémoli**, **S Y Oudot**, *Gromov–Hausdorff stable signatures for shapes using persistence*, Computer Graphics Forum 28 (2009) 1393–1403

[20] **F Chazal**, **W Crawley-Boevey**, **V de Silva**, *The observable structure of persistence modules*, Homology Homotopy Appl. 18 (2016) 247–265 MR Zbl

[21] **F Chazal**, **S Y Oudot**, *Towards persistence-based reconstruction in Euclidean spaces*, from "Computational geometry" (M Teillaud, editor), ACM, New York (2008) 232–241 MR Zbl

[22] **F Chazal**, **V de Silva**, **M Glisse**, **S Oudot**, *The structure and stability of persistence modules*, Springer (2016) MR Zbl

[23] **F Chazal**, **V de Silva**, **S Oudot**, *Persistence stability for geometric complexes*, Geom. Dedicata 173 (2014) 193–214 MR Zbl

[24] **S Chowdhury**, **F Mémoli**, *A functorial Dowker theorem and persistent homology of asymmetric networks*, J. Appl. Comput. Topol. 2 (2018) 115–175 MR Zbl

[25] **W Crawley-Boevey**, *Decomposition of pointwise finite-dimensional persistence modules*, J. Algebra Appl. 14 (2015) art. id. 1550066 MR Zbl

[26] **J Curry**, *The fiber of the persistence map for functions on the interval*, J. Appl. Comput. Topol. 2 (2018) 301–321 MR Zbl

[27] **C J A Delfinado**, **H Edelsbrunner**, *An incremental algorithm for Betti numbers of simplicial complexes on the 3–sphere*, Comput. Aided Geom. Design 12 (1995) 771–784 MR Zbl

[28] **T K Dey**, **F Mémoli**, **Y Wang**, *Topological analysis of nerves, Reeb spaces, mappers, and multiscale mappers*, from "33rd international symposium on computational geometry" (B Aronov, M J Katz, editors), Leibniz Int. Proc. Inform. 77, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern (2017) art. no. 36 MR Zbl

[29] **T tom Dieck**, *Partitions of unity in homotopy theory*, Compositio Math. 23 (1971) 159–167 MR Zbl

[30] **A W M Dress**, *Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: a note on combinatorial properties of metric spaces*, Adv. in Math. 53 (1984) 321–402 MR Zbl

[31] **A Dress**, **K T Huber**, **J Koolen**, **V Moulton**, **A Spillner**, *Basic phylogenetic combinatorics*, Cambridge Univ. Press (2012) MR Zbl

[32] **L E Dubins**, **G Schwarz**, *Equidiscontinuity of Borsuk–Ulam functions*, Pacific J. Math. 95 (1981) 51–59 MR Zbl

[33] **J Dugundji**, *Absolute neighborhood retracts and local connectedness in arbitrary metric spaces*, Compositio Math. 13 (1958) 229–246 MR Zbl

[34] **H Edelsbrunner**, **J Harer**, *Persistent homology: a survey*, from "Surveys on discrete and computational geometry" (J E Goodman, J Pach, R Pollack, editors), Contemp. Math. 453, Amer. Math. Soc., Providence, RI (2008) 257–282 MR Zbl

[35] **H Edelsbrunner**, **J L Harer**, *Computational topology: an introduction*, Amer. Math. Soc., Providence, RI (2010) MR Zbl

[36] **H Edelsbrunner**, **D Letscher**, **A Zomorodian**, *Topological persistence and simplification*, Discrete Comput. Geom. 28 (2002) 511–533 MR Zbl

[37] **H Federer**, *Geometric measure theory*, Grundl. Math. Wissen. 153, Springer (1969) MR Zbl

[38] **A Fomenko**, **D Fuchs**, *Homotopical topology*, 2nd edition, Graduate Texts in Math. 273, Springer (2016) MR Zbl

[39] **P Frosini**, *A distance for similarity classes of submanifolds of a Euclidean space*, Bull. Austral. Math. Soc. 42 (1990) 407–416 MR Zbl

[40] **P Frosini**, *Measuring shapes by size functions*, from "Intelligent robots and computer vision, X: Algorithms and techniques" (D P Casasent, editor), SPIE Proceedings 1607, SPIE (1992) 122–133

[41] **P Frosini**, **C Landi**, *Size functions and morphological transformations*, Acta Appl. Math. 49 (1997) 85–104 MR Zbl

[42] **H Gakhar**, **J A Perea**, *Künneth formulae in persistent homology*, preprint (2019) arXiv 1910.05656

[43] **M Gameiro**, **Y Hiraoka**, **I Obayashi**, *Continuation of point clouds via persistence diagrams*, Phys. D 334 (2016) 118–132 MR Zbl

[44] **E Gasparovic**, **M Gommel**, **E Purvine**, **R Sazdanovic**, **B Wang**, **Y Wang**, **L Ziegelmeier**, *A complete characterization of the one-dimensional intrinsic Čech persistence diagrams for metric graphs*, from "Research in computational topology" (E W Chambers, B T Fasy, L Ziegelmeier, editors), Assoc. Women Math. Ser. 13, Springer (2018) 33–56 MR Zbl

[45] **R Ghrist**, **A Muhammad**, *Coverage and hole-detection in sensor networks via homology*, from "Fourth international symposium on information processing in sensor networks" (A Savvides, editor), IEEE (2005) 254–260

[46] **M Gromov**, *Filling Riemannian manifolds*, J. Differential Geom. 18 (1983) 1–147 MR Zbl

[47] **M Gromov**, *Hyperbolic groups*, from "Essays in group theory" (S M Gersten, editor), Math. Sci. Res. Inst. Publ. 8, Springer (1987) 75–263 MR Zbl

[48] **M Gromov**, *Metric structures for Riemannian and non-Riemannian spaces*, Birkhäuser, Boston, MA (2007) MR Zbl

[49] **A Hatcher**, *Algebraic topology*, Cambridge Univ. Press (2002) MR Zbl

[50] **J-C Hausmann**, *On the Vietoris–Rips complexes and a cohomology theory for metric spaces*, from "Prospects in topology" (F Quinn, editor), Ann. of Math. Stud. 138, Princeton Univ. Press (1995) 175–188 MR Zbl

[51] **S-T Hu**, *Theory of retracts*, Wayne State Univ. Press, Detroit, MI (1965) MR Zbl

[52] **J R Isbell**, *Six theorems about injective metric spaces*, Comment. Math. Helv. 39 (1964) 65–76 MR Zbl

[53] **P Joharinad**, **J Jost**, *Topological representation of the geometry of metric spaces*, preprint (2020) arXiv 2001.10262

[54] **M Katz**, *The filling radius of two-point homogeneous spaces*, J. Differential Geom. 18 (1983) 505–511 MR Zbl

[55] **M Katz**, *Diameter-extremal subsets of spheres*, Discrete Comput. Geom. 4 (1989) 117–137 MR Zbl

[56] **M Katz**, *On neighborhoods of the Kuratowski imbedding beyond the first extremum of the diameter functional*, Fund. Math. 137 (1991) 161–175 MR Zbl

[57] **M Katz**, *The rational filling radius of complex projective space*, Topology Appl. 42 (1991) 201–215 MR Zbl

[58] **M G Katz**, *Systolic geometry and topology*, Mathematical Surveys and Monographs 137, Amer. Math. Soc., Providence, RI (2007) MR Zbl

[59] **M Kılıç**, **Ş Koçak**, *Tight span of subsets of the plane with the maximum metric*, Adv. Math. 301 (2016) 693–710 MR Zbl

[60] **U Lang**, *Injective hulls of certain discrete metric spaces and groups*, J. Topol. Anal. 5 (2013) 297–331 MR Zbl

[61] **M Lesnick**, *The theory of the interleaving distance on multidimensional persistence modules*, Found. Comput. Math. 15 (2015) 613–650 MR Zbl

[62] **S Lim**, **F Mémoli**, **O B Okutan**, *Vietoris–Rips persistent homology*, *injective metric spaces*, *and the filling radius*, preprint (2020) arXiv 2001.07588

[63] **S Lim**, **F Mémoli**, **Z Smith**, *The Gromov–Hausdorff distance between spheres*, Geom. Topol. 27 (2023) 3733–3800 MR Zbl

[64] **L Liu**, *The mapping properties of filling radius and packing radius and their applications*, Differential Geom. Appl. 22 (2005) 69–79 MR Zbl

[65] **F Mémoli**, *A distance between filtered spaces via tripods*, preprint (2017) arXiv 1704.03965

[66] **F Mémoli**, **O B Okutan**, *Quantitative simplification of filtered simplicial complexes*, Discrete Comput. Geom. 65 (2021) 554–583 MR Zbl

[67] **F Mémoli**, **O B Okutan**, **Q Wang**, *Metric graph approximations of geodesic spaces*, preprint (2018) arXiv 1809.05566

[68] **J R Munkres**, *Elements of algebraic topology*, Addison-Wesley, Menlo Park, CA (1984) MR Zbl

[69] **A Nabutovsky**, *Linear bounds for constants in Gromov's systolic inequality and related results*, Geom. Topol. 26 (2022) 3123–3142 MR Zbl

[70] **P Niyogi**, **S Smale**, **S Weinberger**, *Finding the homology of submanifolds with high confidence from random samples*, Discrete Comput. Geom. 39 (2008) 419–441 MR Zbl

[71] **J A Perea**, *Persistent homology of toroidal sliding window embeddings*, from "2016 IEEE international conference on acoustics, speech and signal processing" (M Dong, T F Zheng, editors), IEEE (2016) 6435–6439

[72] **J A Perea**, *Künneth formulae in persistent homology* (2018) Available at `http://www.birs.ca/events/2018/5-day-workshops/18w5140/videos/watch/201808081206-Perea.html`

[73] **V Robins**, *Towards computing homology from finite approximations*, Topology Proc. 24 (1999) 503–532 MR Zbl

[74] **M Schmahl**, *Structure of semi-continuous q–tame persistence modules*, Homology Homotopy Appl. 24 (2022) 117–128 MR Zbl

[75] **C Semple**, **M Steel**, *Phylogenetics*, Oxford Lecture Series in Mathematics and its Applications 24, Oxford Univ. Press (2003) MR Zbl

[76] **V de Silva**, **G Carlsson**, *Topological estimation using witness complexes*, from "Symposium on point-based graphics 2004" (M Gross, H Pfister, M Alexa, S Rusinkiewicz, editors), The Eurographics Association (2004) 157–166

[77] **V de Silva**, **R Ghrist**, *Coverage in sensor networks via persistent homology*, Algebr. Geom. Topol. 7 (2007) 339–358 MR Zbl

[78] **E H Spanier**, *Algebraic topology*, Springer (1981) MR Zbl

[79] **A Verri**, **C Uras**, **P Frosini**, **M Ferri**, *On the use of size functions for shape analysis*, Biol. Cybern. 70 (1993) 99–107 Zbl

[80] **L Vietoris**, *Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen*, Math. Ann. 97 (1927) 454–472 MR JFM

[81] **Ž Virk**, 1–*dimensional intrinsic persistence of geodesic spaces*, J. Topol. Anal. 12 (2020) 169–207 MR Zbl

[82] **Ž Virk**, *Rips complexes as nerves and a functorial Dowker-nerve diagram*, Mediterr. J. Math. 18 (2021) art. id. 58 MR Zbl

[83] **F H Wilhelm, Jr**, *On the filling radius of positively curved manifolds*, Invent. Math. 107 (1992) 653–668 MR Zbl

[84] **T Yokota**, *On the filling radius of positively curved Alexandrov spaces*, Math. Z. 273 (2013) 161–171 MR Zbl

Department of Mathematics, Sungkyunkwan University
Suwon-si, South Korea

Department of Mathematics, The Ohio State University
Columbus, OH, United States

Department of Mathematics, Florida State University
Tallahassee, FL, United States

lsh3109@skku.edu,   facundo.memoli@gmail.com,   okutan@math.fsu.edu

https://sites.google.com/view/sunhyuklim,   http://facundo-memoli.org,
https://sites.google.com/view/ookutan/home

Slopes and concordance of links

ALEX DEGTYAREV

VINCENT FLORENS

ANA G LECUONA

The slope is an isotopy invariant of colored links with a distinguished component, initially introduced by the authors to describe an extra correction term in the computation of the signature of the splice. It appeared to be closely related to several classical invariants, such as the Conway potential function or the Kojima η–function (defined for two-components links). We prove that the slope is invariant under colored concordance of links. Besides, we present a formula to compute the slope in terms of C–complexes and generalized Seifert forms.

1 Introduction

The slope is an isotopy invariant defined for so-called $(1, \mu)$–colored links $K \cup L$ (with a distinguished component K given color 0) in rational homology spheres. It is closely related to several classical invariants (see Degtyarev, Florens and Lecuona [11; 12; 13]), such as the Conway potential and Kojima–Yamasaki η–function (defined for two-components links; see Cochran [5], Jin [14] and Kojima and Yamasaki [16]). To certain $\mathbb{C}^\times$–valued characters of the group $\pi_1(S \subset L)$, viz those trivial on $[K]$, see (2.2), the slope associates a (possibly infinite) complex number. The torus of characters preserving the coloring is naturally identified with the complex torus $(\mathbb{C}^\times)^\mu$, and the slope is a function on (a Zariski open subset of) the variety $\mathcal{A}(K/L) \subset (\mathbb{C}^\times)^\mu$ of admissible characters. This function is rational away from a certain singular locus determined by the Alexander module of $K \cup L$; however, in general, the *values* of the slope are not determined by the Alexander module.

Our aim here is to show that the slope is invariant under colored topological concordance of links (see Theorem 3.2), and to present a method to compute the slope in terms of the Seifert forms of the colored link L with an extra piece of data; see Theorem 4.3. In the case of algebraically split links of two components, the invariance of the slope under colored concordance was known for certain values, viz those where it coincides with the η–function [13, Corollary 3.24]. We show that, outside a certain subset of $(\mathbb{C}^\times)^\mu$, the *Knotennullstellen* — see Conway, Nagel and Toffoli [7] and Nagel and Powell [19] — (topologically) concordant links have the same slope. More generally, for algebraically split links with an arbitrary number of components, our result implies that a certain quotient of the Conway functions of

$K \cup L$ and L is invariant under colored concordance of $K \cup L$ (see Corollary 3.4), whereas the Conway functions themselves are *not* concordance invariants; see Kawauchi [15].

One can compute the slope directly from the definition using the Fox calculus [13, Section 3.2]. While allowing for easy computer-assisted computations, this approach is not particularly useful when dealing with families of examples. In certain cases, the slope can also be computed as a ratio of the Conway polynomials [13, Theorem 3.1], but this formula is inconclusive at the common roots of the numerator and denominator (l'Hôpital's rule does not work); in particular, it leaves wide open the most interesting case, where both polynomials vanish identically. We suggest yet another method of computing the slope, using C–complexes. These were introduced by Cooper [8] and extended, in very recent years, by different groups to compute many link invariants (Cimasoni [3], Cimasoni and Florens [4], Conway, Friedl and Toffoli [6] and Merz [18] among others) and to study their properties (Amundsen, Anderson, Davis and Guyer [1], Davis, Martin and Otto [9] and Davis and Roth [10] among others).

The computation of the slope using C–complexes is particularly powerful when dealing with families of examples as in [12, Example 5.5; 13, Example 3.28]. For the moment, our formula only works in the special case of K algebraically unlinked from each monochrome sublink L_i. For an algebraically split two-component link, the C–complex used in the computation is merely a Seifert surface.

The paper is organized as follows. In Section 2 we recall the construction and the basic properties of the slope. Section 3 is devoted to the proof of the concordance invariance. In Section 4 the computation of the slope in terms of (generalized) Seifert forms is given, and the main formula is proved in Section 5.

Acknowledgements

2 Slopes

A μ–*colored link* is an oriented link L in S^3 equipped with a surjective map $\pi_0(L) \twoheadrightarrow \{1, \ldots, \mu\}$, called a *coloring*. The union of the components of L given the color i is a monochrome sublink denoted by L_i for all $i = 1, \ldots, \mu$. Each link has a canonical *maximal coloring*, where each component is given a separate color. In this special case, each L_i is a knot.

We denote by $X := S^3 \smallsetminus T_L$ the complement of a small open tubular neighborhood of L. The group $H_1(X)$ is free abelian, generated by the classes m_C of the meridians of the components $C \subset L$. By

convention, m_C is oriented so that $m_C \circ \ell_C = 1$ in ∂T_C, where ℓ_C is a longitude and the orientation on ∂T_C is that induced from X. The coloring induces an epimorphism

$$\varphi \colon \pi_1(X) \twoheadrightarrow H := \bigoplus_{i=1}^{\mu} \mathbb{Z} t_i$$

sending m_C to t_i whenever $C \subset L_i$. A multiplicative character $\omega \colon \pi_1(X) \to \mathbb{C}^\times$ is determined by its values on the meridians, and the torus of characters preserving the coloring (those that factor through φ) is naturally identified with the complex torus $(\mathbb{C}^\times)^\mu$. Through this identification, we set $\omega_i := \omega(\varphi(t_i))$ and, with a certain abuse of the language, speak about a character $\omega = (\omega_1, \ldots, \omega_\mu)$. We define

$$\omega^{-1} := (\omega_1^{-1}, \ldots, \omega_\mu^{-1}), \quad \bar{\omega} := (\bar{\omega}_1, \ldots, \bar{\omega}_\mu) \quad \text{and} \quad \omega^* := \bar{\omega}^{-1}.$$

A character ω is called *unitary* if $\omega^* = \omega$, ie $|\omega_i| = 1$ for all $i = 1, \ldots, \mu$. Unitary characters constitute a torus $(S^1)^\mu \subset (\mathbb{C}^\times)^\mu$.

Given a topological space X and a multiplicative character $\omega \colon \pi_1(X) \to \mathbb{C}^\times$, we denote by $H_*(X; \mathbb{C}(\omega))$ the homology of X with coefficients in the local system $\mathbb{C}(\omega)$ twisted by ω; see [13, Section 2] for more details.

We consider mainly colored links with a distinguished component. They are $(1, \mu)$–*colored links*, defined as $(1+\mu)$–colored links of the form

$$K \cup L = K \cup L_1 \cup \cdots \cup L_\mu,$$

where the *knot* K is the only component given the distinguished color 0. The *linking vector* of a $(1, \mu)$–colored link is $\overline{\ell k}(K, L) := (\lambda_1, \ldots, \lambda_\mu) \in \mathbb{Z}^\mu$, where $\lambda_i := \ell k(K, L_i)$.

Definition 2.1 A character $\omega \colon \pi_1(X) \to \mathbb{C}^\times$ on a $(1, \mu)$–colored link $K \cup L$ is called *admissible* if $\omega([K]) = 1$; it is called *nonvanishing* if $\omega_i \neq 1$ for all $i = 1, \ldots, \mu$.

The variety of admissible characters is denoted by $\mathcal{A}(K/L)$, and $\mathcal{A}^\circ(K/L) \subset \mathcal{A}(K/L)$ is the (Zariski) open subset of admissible nonvanishing characters. Letting $\lambda := \overline{\ell k}(K, L)$ we have

(2.2) $\qquad \mathcal{A}(K/L) = \{\omega \in (\mathbb{C}^\times)^\mu \mid \omega^\lambda = 1\} \quad \text{and} \quad \mathcal{A}^\circ(K/L) = \mathcal{A}(K/L) \cap (\mathbb{C}^\times \smallsetminus 1)^\mu,$

where $\omega^\lambda := \prod \omega_i^{\lambda_i}$. In particular, if $\lambda = 0$, then $\mathcal{A}^\circ(K/L) = (\mathbb{C}^\times \smallsetminus 1)^\mu$.

Let $X_K = S^3 \smallsetminus T_{K \cup L}$ be the complement of an open tubular neighborhood of $K \cup L$. We abbreviate $m := m_K$ and $\ell := \ell_K$, where ℓ_K is the *preferred* longitude, also called *Seifert longitude*, that is, the unique longitude with zero linking number with K.

Remark 2.3 Any character $\omega \in (\mathbb{C}^\times)^\mu$ extends to a natural character $\pi_1(X_K) \to \mathbb{C}^\times$ sending m to 1; for short, this extension is also denoted by ω. In this language, the original character ω is admissible if and only if $\omega(\ell) = 1$.

We denote by $\partial_K X_K = \partial T_K$ the intersection of ∂X_K with the closure of T_K and consider the inclusion

$$i : \partial_K X_K \hookrightarrow \partial X_K \hookrightarrow X_K.$$

If $\omega \in \mathcal{A}^\circ(K/L)$, the homomorphism

(2.4) $i_* : H_1(\partial_K X_K; \mathbb{C}(\omega)) \xrightarrow{\cong} H_1(\partial X_K; \mathbb{C}(\omega)) \to H_1(X_K; \mathbb{C}(\omega))$

can be regarded as that induced by the inclusion $\partial X_K \hookrightarrow X_K$ of the boundary, and $H_1(\partial_K X_K; \mathbb{C}(\omega)) \simeq \mathbb{C}^2$ is generated by the meridian m and Seifert longitude ℓ.

Definition 2.5 [13] If $\mathrm{Ker}\, i_*$ in (2.4) has dimension one, it is generated by a single vector $am + b\ell$ for some $[a : b] \in \mathbb{P}^1(\mathbb{C})$, and the *slope* of $K \cup L$ at $\omega \in \mathcal{A}^\circ(K/L)$ is defined as the quotient

$$(K/L)(\omega) := -\frac{a}{b} \in \mathbb{C} \cup \infty.$$

This notion is extended to all characters $\omega \in \mathcal{A}(K/L)$ by "patching" the components L_i on which $\omega_i = 1$. (This operation results in patching with solid tori the corresponding boundary components of the manifold $X := S^3 \smallsetminus T_L$.)

Proposition 2.6 [13] *The slope at a character $\omega \in \mathcal{A}^\circ(K/L)$ is well defined if and only if the two inclusion homomorphisms $H_1(K; \mathbb{C}(\zeta)) \to H_1(S^3 \smallsetminus L; \mathbb{C}(\zeta))$, for $\zeta = \omega$ or ω^*, are either both trivial or both nontrivial. The slope is finite, $(K/L)(\omega) \in \mathbb{C}$, if and only if both homomorphisms are trivial.*

Note also (see [13, Section 2.4] for details) that the slope is always defined on a *unitary* character $\omega \in (S^1)^\mu$: in this case, by twisted Poincaré duality, $\mathrm{Ker}\, i_*$ is a Lagrangian subspace of

$$H_1(\partial_K X_K; \mathbb{C}(\omega)) = H_1(\partial X_K; \mathbb{C}(\omega)),$$

see (2.4), with respect to the twisted intersection form and, hence, $\dim \mathrm{Ker}\, i_* = 1$.

Recall (see eg [17]) that the *characteristic varieties* associated with a μ–colored link L are the jump loci

$$\mathcal{V}_r(L) := \{\omega \in (\mathbb{C}^\times)^\mu \mid \dim H_1(X; \mathbb{C}(\omega)) \geqslant r\} \quad \text{for } r \geqslant 0.$$

They are indeed nested algebraic subvarieties:

(2.7) $(\mathbb{C}^\times)^\mu = \mathcal{V}_0 \supset \mathcal{V}_1 \supset \mathcal{V}_2 \supset \cdots \quad$ with $\mathcal{V}_1(L) = \{\omega \mid \Delta_L(\omega) = 0\}.$

The first *proper* characteristic variety, ie the first member $\mathcal{V}_r$ of the sequence (2.7) such that $\mathcal{V}_r \neq (\mathbb{C}^\times)^\mu$, is denoted by $\mathcal{V}_{\max} := \mathcal{V}_{\max}(L)$. This variety depends on L only, and, if $\lambda := \overline{\ell k}(K, L) = 0$, it is a proper algebraic subvariety of the torus $\mathcal{A}(K/L) = (\mathbb{C}^\times)^\mu$ of admissible characters.

Remark 2.8 If $\lambda := \overline{\ell k}(K, L) \neq 0$, the situation is slightly more involved. Let $\lambda = n\lambda'$, where $\lambda' \in \mathbb{Z}^\mu$ is a primitive vector. In view of (2.2), the variety $\mathcal{A}(K/L)$ of admissible characters (depending on λ only) splits over $\mathbb{Q}$ into irreducible components

$$\mathcal{A}_d := \{\Phi_d(\omega^{\lambda'}) = 0\} \quad \text{for } d \mid n,$$

where Φ_d stands for the cyclotomic polynomial, and we should speak about a separate *first proper characteristic variety* $\mathcal{V}_{\max}^{\lambda,d}(L) \subsetneq \mathcal{A}_d$ for each component $\mathcal{A}_d$. In general, $\mathcal{V}_{\max}^{\lambda,d}(L) \neq \mathcal{V}_{\max}(L) \cap \mathcal{A}_d$ as $\mathcal{V}_{\max}(L)$ may contain $\mathcal{A}_d$. To keep the notation uniform, we occasionally extend it to the case $\lambda = 0$ via $\mathcal{A}_0 := \mathcal{A}(K/L)$ and $\mathcal{V}_{\max}^{0,0}(L) := \mathcal{V}_{\max}(L)$.

Theorem 2.9 [13, Theorems 3.19 and 3.21] *Let* $\lambda := \overline{\ell k}(K, L)$. *For each rational component* $\mathcal{A}_d \subset \mathcal{A}(K/L)$, *the slope restricts to a rational function, possibly identical* ∞, *on the complement* $\mathcal{A}_d^\circ \smallsetminus \mathcal{V}_{\max}^{\lambda,d}(L)$. *In other words, the slope gives rise to an element of the extended function field* $\mathbb{Q}(\mathcal{A}_d) \cup \infty$.

If $\mathcal{V}_{\max}^{\lambda,d}(L) = \mathcal{V}_1(L) \cap \mathcal{A}_d$, *ie* Δ_L *does not vanish identically on* $\mathcal{A}_d$, *one has*

$$(K/L)(\omega) = -\frac{\nabla'(1, \sqrt{\omega})}{2\nabla_L(\sqrt{\omega})} \in \mathbb{C} \cup \infty,$$

where ∇' *is the derivative of* $\nabla_{K\cup L}(t, \cdot)$ *with respect to* t.

3 Concordance of links

Two oriented μ–colored links L^0 and L^1 are *concordant* if there exists a collection of properly embedded disjoint locally flat cylinders $A := A_1 \sqcup \cdots \sqcup A_\mu$ in $S^3 \times [0, 1]$ such that

$$\partial A_i \cap (S^3 \times 0) = -L_i^0 \quad \text{and} \quad \partial A_i \cap (S^3 \times 1) = L_i^1$$

for all $i = 1, \ldots, \mu$. (In general, each A_i is a union of cylinders.)

3.1 The concordance invariance

In the study of knot and link concordance, there is a subset of the complex numbers of particular relevance, the so-called *Knotennullstellen*. This was first introduced in [19] for knots and extended to the multicomponent link case in [7]. For our purposes, we only need the following definition. Consider the subset of Laurent polynomials

$$U := \{p \in \mathbb{Z}[t_1^{\pm 1}, \ldots, t_\mu^{\pm 1}] \mid p(1, \ldots, 1) = \pm 1\}.$$

An element $\omega \in \mathcal{A}(K/L)$ is called a *concordance root* if there is a polynomial $p \in U$ such that $p(\omega) = 0$. We denote by $\mathcal{A}_c(K/L) \subset \mathcal{A}(K/L)$ the subset of admissible characters that are *not* concordance roots, and abbreviate $\mathcal{A}_c^\circ(K/L) := \mathcal{A}_c(K/L) \cap \mathcal{A}^\circ(K/L)$. Note that these sets are larger than the set $\mathbb{T}_!$ used in [7], since we allow for nonunitary characters.

Remark 3.1 If $\overline{\ell k}(K, L) = 0$, the set $\mathcal{A}_c(K/L)$ is dense in $\mathcal{A}(K/L) = (\mathbb{C}^\times)^\mu$, as it is a countable intersection of Zariski open sets. In general, $\mathcal{A}_c(K/L)$ is only dense in the components $\mathcal{A}_d$ (see Remark 2.8) for which d is a prime power (or $d = 1$ as a special case). Indeed, if d is *not* a prime power, then $\Phi_d(\cdot) \in U$ and, hence each point of $\mathcal{A}_d$ is a concordance root.

Theorem 3.2 *Let $K^0 \cup L^0$ and $K^1 \cup L^1$ be two concordant $(1, \mu)$–colored links. Then $\mathcal{A}_c(K^0/L^0)$ and $\mathcal{A}_c(K^1/L^1)$ coincide as subsets of $(\mathbb{C}^\times)^\mu$ and*

$$(K^0/L^0)(\omega) = (K^1/L^1)(\omega)$$

for any character $\omega \in \mathcal{A}_c(K^0/L^0)$.

The proof of Theorem 3.2 is postponed till Section 3.2. The next few corollaries are direct consequences of Theorems 3.2 and 4.3.

Corollary 3.3 *Let $K^0 \cup L^0$ and $K^1 \cup L^1$ be concordant $(1, \mu)$–colored links such that $\overline{\ell k}(K^s, L^s) = 0$ for $s = 0, 1$. Then the slopes K^0/L^0 and K^1/L^1 are equal as elements of the extended function field $\mathbb{Q}((\mathbb{C}^\times)^\mu) \cup \infty$. In particular, $(K^0/L^0)(\omega) = (K^1/L^1)(\omega)$ for each character ω in the complement of the (common) first proper characteristic variety $\mathcal{V}_{\max}(L^0) = \mathcal{V}_{\max}(L^1)$.*

Proof If L^0 and L^1 are concordant, their nullities coincide (see [4, Theorem 7.1]); hence, so do their first proper characteristic varieties. Therefore, the statement is an immediate consequence of Theorem 3.2, the rationality of the slope given by Theorem 2.9, and the density of $\mathcal{A}_c(K/L)$ discussed in Remark 3.1. $\square$

Corollary 3.4 (of Corollary 3.3 and Theorem 2.9) *Let $K^0 \cup L^0$ and $K^1 \cup L^1$ be two concordant $(1, \mu)$–colored links such that $\overline{\ell k}(K^s, L^s) = 0$ and $\Delta_{L^s} \not\equiv 0$ for $s = 0, 1$. Then*

$$\frac{\nabla'_{K^0 \cup L^0}(1, \bar{t})}{\nabla_{L^0}(\bar{t})} = \frac{\nabla'_{K^1 \cup L^1}(1, \bar{t})}{\nabla_{L^1}(\bar{t})} \quad \text{for } \bar{t} := (t_1, \ldots, t_\mu).$$

Remark 3.5 A priori, the conclusions of Corollaries 3.3 or 3.4 do not need to hold if $\lambda := \overline{\ell k}(K^s, L^s) \neq 0$: it is not even obvious that the first proper varieties $\mathcal{V}_{\max}^{\lambda, d}(L^s)$ or even their indices in (2.7) should coincide if d is not a prime power. (Note though that we do not know any counterexample, as that would require going far beyond the known link tables.) The precise statements, based on Remarks 2.8 and 3.1 and Theorems 3.2 and 2.9, are left to the reader.

Recall that a link is *slice* if it is concordant to an unlink. It is a *boundary link* if the components bound a collection of mutually disjoint Seifert surfaces in S^3. For any coloring of the link L, the slope obstruct L being slice, or concordant to any boundary link. Indeed, the two following corollaries are available for any coloring:

Corollary 3.6 *If $K \cup L$ is a slice link, then $(K/L)(\omega) = 0$ for all ω in $\mathcal{A}_c(K/L)$.*

Corollary 3.7 *If $K \cup L$ is concordant to a boundary link, then $(K/L)(\omega) = 0$ for all ω in $\mathcal{A}_c(K/L)$.*

Corollary 3.7 is in fact a particular case of the following statement (see [4] or Section 4.1 for the definition of a C–complex):

Corollary 3.8 *If $K \cup L$ is concordant to a $(1, \mu)$–colored link $K' \cup L'$ admitting a C–complex F for L and a Seifert surface S for K disjoint from F, then $(K/L)(\omega) = 0$ for all $\omega \in \mathcal{A}_c(K/L)$.*

[Corollary 3.8](#) is actually a consequence of both Theorems [3.2](#) and [4.3](#); see [Example 4.5](#).

The following example illustrates that the values of the slope at concordance roots, that is outside the set $\mathcal{A}_c(K/L)$, might not be invariant under concordance. We observe a similar pattern with knot signatures: Knotennullstelle unitary characters are precisely where they fail to be concordance invariants [2; 19]. See [7] for the case of colored links.

Example 3.9 Let $K \cup L$ be the $(1, 1)$–colored two-component slice link [L10n36](#), where K is the unknotted component. Then $\nabla_{K \cup L}(t, t_1) = 0$ and $\nabla_L(t_1) = (t_1 - 1 + t_1^{-1})^2$, so by [13, Theorem 3.21], $(K/L)(\omega) = 0$ unless ω is one of the two roots $\alpha_{\pm}$ of ∇_L, which agrees with [Theorem 3.2](#) and [Corollary 3.4](#). (By definition, $\alpha_{\pm} \notin \mathcal{A}_c(K/L)$.) A computation using Fox calculus (see [13, Section 3.2]) gives us $(K/L)(\alpha_{\pm}) = \infty$.

In the proof of [Theorem 3.2](#) we will need the following lemma. We state it in our more general setting of arbitrary, not necessarily unitary, characters, but the proof found in [7] extends literally as it relies on simple homological algebra.

Lemma 3.10 [7, Lemma 2.16] *Let $k \geqslant 0$ be an integer. If (X, Y) is a CW–pair over $B\mathbb{Z}^{\mu}$ such that $H_i(X, Y; \mathbb{Z}) = 0$ for all $0 \leqslant i \leqslant k$, then also $H_i(X, Y; \mathbb{C}(\omega)) = 0$ for all $0 \leqslant i \leqslant k$ and any character $\omega \in (\mathbb{C}^{\times})^{\mu}$ that is not a concordance root.*

3.2 Proof of [Theorem 3.2](#)

To save space, we abbreviate $H_*^{\omega}(-) := H_*(-; \mathbb{C}(\omega))$.

Let $D \cup A \subset S^3 \times [0, 1]$ be the concordance, $\partial D = -K^0 \sqcup K^1$, and consider an open tubular neighborhood $T_{D \cup A}$ of $D \cup A$ with a fixed trivialization extending Seifert framings (in the tubular neighborhoods $T_{K^s \cup L^s} := T_{D \cup A} \cap (S^3 \times s)$ for $s = 0, 1$) of the links. Define

$$U := S^3 \times [0, 1] \smallsetminus T_A \quad \text{and} \quad U_K := S^3 \times [0, 1] \smallsetminus T_{D \cup A},$$

and let

$$X^s := U \cap (S^3 \times s) \quad \text{and} \quad X_K^s := U_K \cap (S^3 \times s)$$

for $s = 0, 1$. The inclusions $X_K^s \hookrightarrow U_K$ send the meridians of $K^s \cup L^s$ to those of $D \cup A$. The relative Mayer–Vietoris exact sequences applied to

$$(S^3 \times I, S^3 \times s) = (U_K, X_K^s) \cup (\overline{T}_{D \cup A}, \overline{T}_{K^s \cup L^s}) = (U, X^s) \cup (\overline{T}_A, \overline{T}_{L^s})$$

(where $\overline{T}_*$ stands for the closure of a tubular neighborhood T_*) give us

(3.11) $$H_*(U_K, X_K^s) = H_*(U, X^s) = 0$$

for $s = 0, 1$. In particular, the inclusions $X_K^s \hookrightarrow U_K$ induce isomorphisms

(3.12) $$H_1(X_K^0) \xrightarrow{\cong} H_1(U_K) \xleftarrow{\cong} H_1(X_K^1)$$

preserving the meridians, and thus identify the three character tori. Since the trivialization of T_D homotopes ℓ^0 to ℓ^1, we have $\mathcal{A}_c(K^0/L^0) = \mathcal{A}_c(K^1/L^1)$; see [Remark 2.3](#).

From now on, patching, if necessary, a few components of both links (and the concordance), we can assume the character ω nonvanishing, ie $\omega \in \mathcal{A}_c^\circ(K^0/L^0)$. Referring to Remark 2.3 and using the above identification of the character tori, we can regard ω as a homomorphism $\pi_1(U_K) \to \mathbb{C}^\times$. The twisted Mayer–Vietoris sequence applied to the pairs

$$(U, X^s) = (U_K, X_K^s) \cup (\overline{T}_D, \overline{T}_{K^s})$$

gives us, for all i,

$$\to H_i^\omega(D \times S^1, K^s \times S^1) \to H_i^\omega(U_K, X_K^s) \oplus H_i^\omega(\overline{T}_D, \overline{T}_{K^s}) \to H_i^\omega(U, X^s) \to,$$

where $\{\cdot\} \times S^1$ are the meridians of K^s and D, on which ω is trivial. Since

$$H_*^\omega(D \times S^1, K^s \times S^1) = 0 \quad \text{and} \quad H_*^\omega(U_K, X_K^s) = H_*^\omega(U, X^s) = 0,$$

the latter by Lemma 3.10 and (3.11), we obtain $H_*^\omega(U_K, X_K^s) = 0$ and the inclusions $X_K^s \hookrightarrow U_K$ induce isomorphisms

$$H_1^\omega(X_K^0) \xrightarrow{\cong} H_1^\omega(U_K) \xleftarrow{\cong} H_1^\omega(X_K^1)$$

preserving the meridians and, similar to (3.12), taking the class of ℓ^0 to that of ℓ^1. It follows that $am^0 + b\ell^0 = 0 \in H_1^\omega(X_K^0)$ if and only if $am^1 + b\ell^1 = 0 \in H_1^\omega(X_K^1)$. $\qquad\square$

4 Computation with Seifert forms

For the remainder of the paper, unless specified otherwise, we abbreviate

$$H_*(-) := H_*(-; \mathbb{C}), \quad H^*(-) := H^*(-; \mathbb{C}) \quad \text{and} \quad H_*^\omega(-) = H_*(-; \mathbb{C}(\omega)).$$

For a character $\omega \in (\mathbb{C}^\times \smallsetminus 1)^\mu$, we also abbreviate $\widetilde{\omega}_i := (1 - \omega_i^{-1})$ for $1 \leqslant i \leqslant \mu$.

4.1 Seifert forms

Let $L = L_1 \cup \cdots \cup L_\mu \subset$ be an oriented μ–colored link in S^3. A *C–complex* F for L [3] is a collection of Seifert surfaces $F_1, \ldots, F_\mu$ for the sublinks $L_1, \ldots, L_\mu$ that intersect only along (a finite number of) *clasps*. Each class in $H_1(F; \mathbb{Z})$ can be represented by a union of *proper loops*, ie loops $\alpha \colon S^1 \to F$ such that the pullback of each clasp is a single (possibly empty) segment. We routinely identify classes, loops and their images.

Given a vector $\varepsilon \in \{\pm 1\}^\mu$, the *push-off* α^ε of a proper loop α is the loop in $S^3 \smallsetminus F$ obtained by a slight shift of α off each surface F_i in the direction of ε_i. (If α runs along a clasp $\mathfrak{c} \subset F_i \cap F_j$, the shift respects both directions ε_i and ε_j.) Due to [4], this operation gives rise to a well-defined homomorphism

$$\Theta^\varepsilon \colon H_1(F; \mathbb{Z}) \to H_1(S^3 \smallsetminus F; \mathbb{Z}) = H^1(F; \mathbb{Z})$$

(we use Alexander duality), which can be computed by means of the *Seifert forms*

$$\theta^\varepsilon \colon H_1(F; \mathbb{Z}) \otimes H_1(F; \mathbb{Z}) \to \mathbb{Z} \quad \text{given by} \quad \alpha \otimes \beta \mapsto \ell k(\alpha, \beta^\varepsilon).$$

Now, given a character $\omega \in (\mathbb{C}^{\times} \smallsetminus 1)^{\mu}$, we define

$$\Pi(\omega) := \prod_{i=1}^{\mu}(1 - \omega_i) \in \mathbb{C}^{\times} \quad \text{and} \quad A(\omega) := \sum_{\varepsilon \in \{\pm 1\}^{\mu}} \prod_{i=1}^{\mu} \varepsilon_i \omega_i^{(1-\varepsilon_i)/2} \Theta^{\varepsilon} : H_1(F) \to H^1(F)$$

and let

(4.1) $$E(\omega) := \Pi(\omega^{-1})^{-1} A(\omega^{-1}) : H_1(F) \to H^1(F).$$

Throughout the text we will use the shorthand $\mathrm{Ker}\, E(\omega)^{\perp}$ to denote the subset of $H^1(F)$ defined as $\mathrm{Ann}\, \mathrm{Ker}\, E(\omega)$. It is straightforward that

$$E^*(\omega) = E(\omega^{-1}) \quad \text{and} \quad \bar{E}(\omega) = E(\bar{\omega}),$$

where E^* is the adjoint in the sense of linear algebra over an arbitrary field, and for a linear map $L : U \otimes \mathbb{C} \to V \otimes \mathbb{C}$ between two complexified real vector spaces, we let $\bar{L} : u \mapsto \overline{L(\bar{u})}$. In particular, if $\omega \in (S^1 \smallsetminus 1)^{\mu}$ is unitary, the operator $E(\omega)$ is Hermitian, ie $\bar{E}^*(\omega) = E(\omega)$; thus it has a well-defined signature. Furthermore, if ω is unitary, the operator $E(\omega^{-1})$ differs from $H(\omega)$ considered in [4] by the positive real constant $\Pi(\omega)^{-1}\Pi(\bar{\omega})^{-1}$; hence, the two have the same signature and nullity and E can be used instead of H in the following theorem:

Theorem 4.2 [4] *If* $\omega \in (S^1 \smallsetminus 1)^{\mu}$ *is a nonvanishing unitary character, then* $\sigma_L(\omega) = \mathrm{sign}\, E(\omega)$ *and* $\eta_L(\omega) = \dim \mathrm{Ker}\, E(\omega) + b_0(F) - 1$.

In the case of a 1–colored link L, the C–complex reduces to a single Seifert surface F, so that $\theta := \theta^+$ and $\Theta := \Theta^+$ are the classical Seifert form and operator, respectively. Since, in this case, we obviously have $\theta^- = \theta^*$ and hence $\Theta^- = \Theta^*$, the operator E takes the classical form

$$E(\omega^{-1}) = (1 - \omega)^{-1}(\Theta - \omega\Theta^*).$$

4.2 The statement

Let $K \cup L$ be a $(1, \mu)$–colored link. Assume that λ, the linking vector between K and L, vanishes and fix a C–complex F for L disjoint from K. By Alexander duality $H_1(S^3 \smallsetminus F; \mathbb{Z}) = H^1(F; \mathbb{Z})$, there is a well-defined cohomology class

$$\kappa := [K] \in H^1(F; \mathbb{Z}) \subset H^1(F), \quad \kappa : \alpha \mapsto \ell k(\alpha, K).$$

Theorem 4.3 *Under the above assumptions, for any character* $\omega \in \mathcal{A}^{\circ}(K/L)$*, consider the operator* $E(\omega) : H_1(F) \to H^1(F)$*; see* (4.1)*. Then*

$$(K/L)(\omega) = \begin{cases} -\langle \alpha, \kappa \rangle & \text{if } \kappa \in \mathrm{Im}\, E(\omega) \cap \mathrm{Ker}\, E(\omega)^{\perp}, \\ \infty & \text{if } \kappa \notin \mathrm{Im}\, E(\omega) \cup \mathrm{Ker}\, E(\omega)^{\perp}; \end{cases}$$

otherwise, $(K/L)(\omega)$ *is undefined. In the first case,* $\alpha \in H_1(F)$ *is any class such that* $E(\omega)(\alpha) = \kappa$.

Example 4.4 Consider the Whitehead link $K \cup L$ with the C–complex F depicted in Figure 1, which is simply a genus-one Seifert surface for the knot L. We want to compute the slope $(K/L)(\omega)$ using
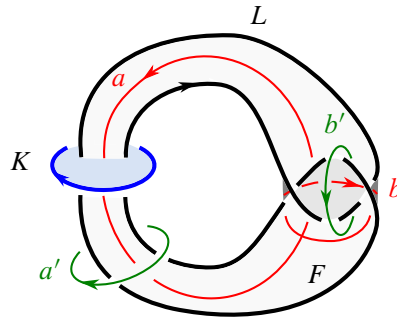
Figure 1: The Whitehead link $K \cup L$ with a C–complex F for L (a Seifert surface in this case) and chosen bases $\{a, b\}$ and $\{a', b'\}$ of $H_1(F)$ and $H_1(S^3 \smallsetminus F) = H^1(F)$, respectively.

Theorem 4.3, and to this end we fix the basis $\{a, b\}$ of $H_1(F)$ and $\{a', b'\}$ of $H_1(S^3 \smallsetminus F) = H^1(F)$ which are illustrated in Figure 1. With respect to these bases,

$$\theta^+ = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \quad A(\omega) = \begin{bmatrix} 0 & -\omega \\ 1 & 1-\omega \end{bmatrix} \quad \text{and} \quad E(\omega) = \begin{bmatrix} 0 & (1-\omega)^{-1} \\ (1-\omega^{-1})^{-1} & 1 \end{bmatrix}.$$

It is evident from the figure that κ is the same class as a'. One can easily compute a class $\alpha \in H_1(F)$ such that $E(\omega)(\alpha) = \kappa$:

$$E(\omega) \begin{bmatrix} (1-\omega^{-1})(\omega-1) \\ 1-\omega \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \kappa.$$

Finally, we calculate the slope as $-\langle \alpha, \kappa \rangle$, that is,

$$(K/L)(\omega) = (1-\omega)(1-\omega^{-1}),$$

which coincides with previous computations using Fox calculus; see [13].

Example 4.5 (see Corollary 3.8) Let $K \cup L$ be a $(1, \mu)$–colored link admitting a C–complex F for L and a Seifert surface S for K disjoint from F. Obviously $\kappa = 0$ and then $(K/L)(\omega) = 0$ for all $\omega \in \mathcal{A}^{\circ}(K/L)$. This implies that, by Theorem 3.2, for any $(1, \mu)$–colored link concordant to a $(1, \mu)$–colored link bounding a disjoint C–complex and Seifert surface, the slope vanishes at any $\omega \in \mathcal{A}_c(K/L)$.

5 Proof of Theorem 4.3

5.1 Geometry of C–complexes

The notation and maps introduced in this section are illustrated in Figure 2. Let L be a μ–colored link and F a C–complex for L. Given a pair $i \neq j$ of indices, let $C_{ij} := F_i \cap F_j$ and $\mathfrak{C}_{ij} := \pi_0(C_{ij})$ be the set of clasps in the intersection of the surfaces F_i and F_j. Also define $C := \bigcup C_{ij}$ and $\mathfrak{C} := \bigcup \mathfrak{C}_{ij}$.

By convention, each clasp $\mathfrak{c} \in \mathfrak{C}_{ij}$ is oriented from $\mathfrak{c} \cap L_i$ to $\mathfrak{c} \cap L_j$, if $i < j$. The *sign* of $\mathfrak{c}$, denoted by $\mathrm{sg}\, \mathfrak{c} \in \{\pm 1\}$, is the local intersection index $L_i \circ F_j = L_j \circ F_i$ at the corresponding endpoint of $\mathfrak{c}$.
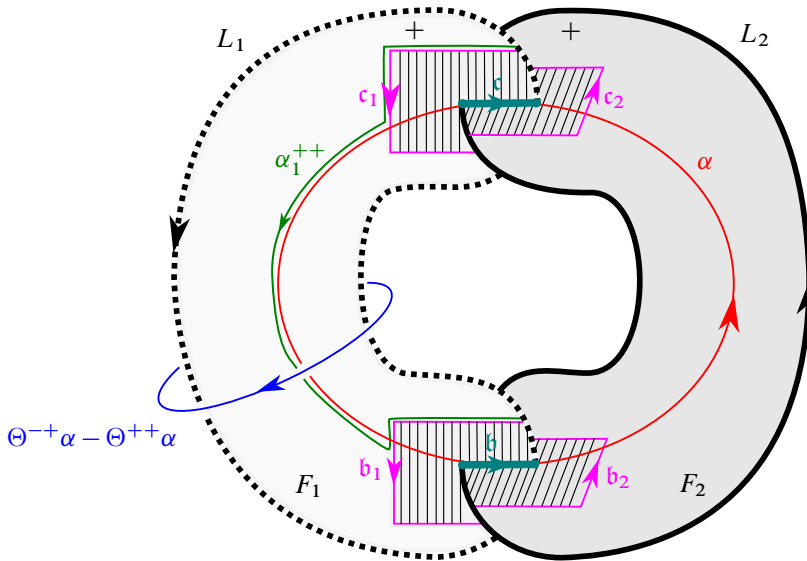
Figure 2: This minimal example shows a two–colored link $L = L_1 \cup L_2$ bounding a C–complex with two positive clasps. In this example $\mathfrak{C} = \mathfrak{C}_{12} = \{\mathfrak{c}, \mathfrak{b}\}$. The lined subset is the open set V with two connected components $V_{\mathfrak{c}}$ and $V_{\mathfrak{b}}$. The relative class $\alpha_1^{++} \in H_1(F_1^{\circ}, \partial_L F_1^{\circ})$ and the element $\Theta^{-+}\alpha - \Theta^{++}\alpha = \mathrm{rel}_1^{++} \alpha \in H^1(F)$ are identified through the isomorphism in Lemma 5.1.

Fix a regular open neighborhood $V \subset F$ of the union of all clasps, denote by $\overline{V}$ its closure, and let $F_i^{\circ} := F_i \smallsetminus V$ for all i. Then $\partial F_i^{\circ} = \partial_L F_i^{\circ} \cup \partial_{\mathfrak{C}} F_i^{\circ}$, where

$$\partial_L F_i^{\circ} := \partial F_i^{\circ} \cap L \quad \text{and} \quad \partial_{\mathfrak{C}} F_i^{\circ} := \partial F_i^{\circ} \cap \overline{V}.$$

Given a clasp $\mathfrak{c} \in \mathfrak{C}_{ij}$, let $\overline{V}_{\mathfrak{c}}$ be the connected component of $\overline{V}$ containing $\mathfrak{c}$, and let $\mathfrak{c}_i \in H_1(F_i^{\circ}, \partial_L F_i^{\circ})$ be the arc $F_i^{\circ} \cap \overline{V}_{\mathfrak{c}}$, with its boundary orientation induced from V, as well as the class realized by this arc.

The following statement is a formalization of the intuitive fact that any class in $H^1(F)$ can be represented as the intersection index with a certain surface $S \subset S^3$ such that $\partial S \cap F = \varnothing$; on the other hand, any such surface can be made disjoint from C and, when doing so, each clasp can be "circumvented" in two ways.

In the lengthy computation that follows, we follow the common practice and treat canonically isomorphic objects as equal, thus simplifying the notation.

Lemma 5.1 *The intersection pairing establishes an isomorphism*

$$H^1(F) = \bigoplus_{i=1}^{\mu} H_1(F_i^{\circ}, \partial_L F_i^{\circ})/\{\mathfrak{c}_i + \mathfrak{c}_j = 0 \mid \mathfrak{c} \in \mathfrak{C}_{ij} \text{ for } 1 \leqslant i < j \leqslant \mu\}.$$

Proof Since all groups involved are torsion free, the statement follows from the exact sequence of the pair $(F, \overline{V})$

$$0 \to H_1(F) \to H_1(F, \overline{V}) \to H_0(\overline{V}) \to H_0(F),$$
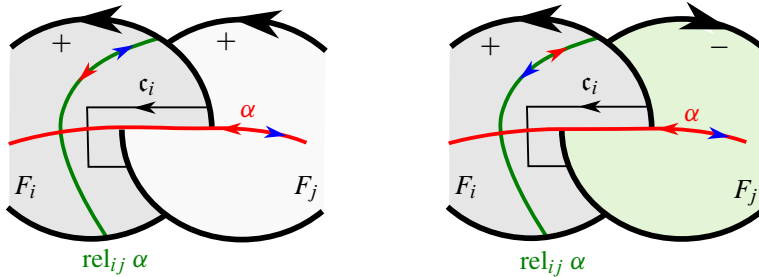
Figure 3: The element $\alpha \in H_1(F)$ is depicted with both possible orientations. The orientation of the element $\mathrm{rel}_{ij}\,\alpha$ depends on the sign of the clasp, as illustrated. Note that the element $\mathrm{rel}_{ij}\,\alpha$ is by definition in $H^1(F)$: the green curve depicted is a representative of that element via Lemma 5.1.

where $H_1(F, \overline{V}) = \bigoplus_i H_1(F_i^\circ, \partial_{\mathfrak{c}} F_i^\circ)$, and applying Poincaré–Lefschetz duality $H^1(F_i^\circ, \partial_{\mathfrak{c}} F_i^\circ) = H_1(F_i^\circ, \partial_L F_i^\circ)$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Let $\varepsilon \in \{\pm 1\}^\mu$. Pick a class $\alpha \in H_1(F)$, represent it by a proper loop, and denote by $\alpha_i^\varepsilon \in H_1(F_i^\circ, \partial_L F_i^\circ)$ the class realized by the arc $\alpha \cap F_i$ pushed off each clasp $\mathfrak{c} \in \mathfrak{C}_{ij}$ in the direction prescribed by ε_j. Passing further to the image in $H^1(F)$, see Lemma 5.1, we obtain a well-defined homomorphism $\mathrm{rel}_i^\varepsilon\colon H_1(F) \to H^1(F)$. It is easily seen that $\mathrm{rel}_i^\varepsilon$ is independent of ε_i. In fact,

$$\mathrm{rel}_i^\varepsilon\,\alpha = \Theta^{\varepsilon[-i]}\alpha - \Theta^{\varepsilon[+i]}\alpha,$$

where $\varepsilon[\pm i]$ is obtained from ε by replacing the ith component by ± 1. Furthermore, for an index $j \neq i$,

(5.2) $$\mathrm{rel}_i^{\varepsilon[+j]}\,\alpha - \mathrm{rel}_i^{\varepsilon[-j]}\,\alpha = \mathrm{rel}_{ij}\,\alpha := \sum_{\mathfrak{c} \in \mathfrak{C}_{ij}} \mathrm{sg}\,\mathfrak{c} \cdot \langle \alpha, \mathfrak{c}_i \rangle \mathfrak{c}_i.$$

For the reader's convenience a local illustration is presented in Figure 3. (Note that $\langle \alpha, \mathfrak{c}_i \rangle \mathfrak{c}_i = \langle \alpha, \mathfrak{c}_j \rangle \mathfrak{c}_j$ for each clasp $\mathfrak{c} \in \mathfrak{C}_{ij}$, and hence $\mathrm{rel}_{ij}\,\alpha = \mathrm{rel}_{ji}\,\alpha$ as elements of $H^1(F)$.) Let $- := [-1, \ldots, -1] \in \{\pm 1\}^\mu$. Then, applying the last two equations inductively, for each $\varepsilon \in \{\pm 1\}^\mu$ we get

(5.3) $$\Theta^\varepsilon \alpha - \Theta^- \alpha = -\sum_{\substack{i \\ \varepsilon_i > 0}} \mathrm{rel}_i^-\,\alpha - \sum_{\substack{i < j \\ \varepsilon_i = \varepsilon_j > 0}} \mathrm{rel}_{ij}\,\alpha.$$

Remark 5.4 It follows from (5.3) that, as in the classical case of a single Seifert surface, all operators Θ^ε are almost determined by any one of them, as the relativization homomorphisms $\mathrm{rel}_i^\varepsilon$ and rel_{ij} are intrinsic to the abstract C–complex F *with prescribed signs* $\mathrm{sg}\,\mathfrak{c}$ of the clasps. In the classical case, (5.3) takes the well-known form

$$\Theta^* - \Theta = \mathrm{rel}\colon H_1(F) \to H_1(F, \partial F) = H^1(F),$$

which explains the notation rel.

Now, given a character $\omega \in (\mathbb{C}^\times \smallsetminus 1)^\mu$, observe that

$$A(\omega) = \Pi(\omega)\Theta^- + \sum_{\varepsilon \in \{\pm 1\}^\mu} \prod_{i=1}^\mu \varepsilon_i \omega_i^{(1-\varepsilon_i)/2} (\Theta^\varepsilon - \Theta^-).$$

Hence, using (5.3), rearranging the terms, and using the definition $\widetilde{\omega}_i = 1 - \omega_i^{-1}$, we arrive at

$$\textbf{(5.5)} \qquad E(\omega) = \Theta^- - R(\omega) \quad \text{for } R(\omega) := \sum_{i=1}^{\mu} \widetilde{\omega}_i^{-1} \operatorname{rel}_i^- + \sum_{1 \leqslant i < j \leqslant \mu} \widetilde{\omega}_i^{-1} \widetilde{\omega}_j^{-1} \operatorname{rel}_{ij}.$$

5.2 Reference sheets

We briefly recall how twisted homology can be computed via coverings. Consider a connected CW–complex X, an abelian group G, and an epimorphism $\varphi \colon \pi_1(X) \twoheadrightarrow H_1(X; \mathbb{Z}) \twoheadrightarrow G$. The kernel of φ, which is a normal subgroup of $\pi_1(X)$, gives rise to a Galois G–covering $\widetilde{X} \to X$, where the deck transformation $g \in G$ sends a point $\tilde{x} \in \widetilde{X}$ to the other endpoint of the arc that begins at $\tilde{x}$ and covers a loop representing an element of $\varphi^{-1}(g)$. This model induces the structure of a $\mathbb{Z}[G]$–module on $C_*(\widetilde{X})$ and, for each multiplicative character $\omega \colon G \to \mathbb{C}^\times$, there is a canonical chain isomorphism of complexes of $\mathbb{C}(\omega)$–modules

$$C_*(X; \mathbb{C}(\omega)) \simeq C_*(\widetilde{X}) \otimes_{\mathbb{Z}G} \mathbb{C}(\omega).$$

Occasionally, the homomorphism $\varphi \colon H_1(X; \mathbb{Z}) \to G$ might not necessarily be surjective. (Typically, this situation occurs when we restrict the construction to a subcomplex $Y \subset X$.) Then, letting $G' := \operatorname{Im} \varphi$, the G–covering $\widetilde{X}$ consists of $[G : G']$ connected components, each isomorphic to the G'–covering $\widetilde{X}'$, and

$$C_*(\widetilde{X}) \simeq C_*(\widetilde{X}') \otimes_{\mathbb{Z}G'} \mathbb{Z}G.$$

However, this isomorphism is no longer canonical; to make it so, we need to fix a *reference component* $\widetilde{X}' \subset \widetilde{X}$. An important special case is that where the restriction of ω to X is trivial. Then we have an isomorphism

$$H_*(C_*(\widetilde{X}) \otimes_{\mathbb{Z}G} \mathbb{C}(\omega)) \simeq H_*^\omega(X) = H_*(X),$$

which is canonical *provided that a **reference sheet** X in the trivial covering $\widetilde{X} \to X$ is fixed*.

Returning to the original setup, when dealing with the twisted homology we need to avoid the ramification locus L. Hence, we fix pairwise disjoint tubular neighborhoods $T_i \supset L_i$ and, denoting by $\overline{T}_i$ the closure of T_i and letting $T := \bigcup_i T_i$ and $\overline{T} := \bigcup_i \overline{T}_i$, introduce

$$S_L := S^3 \smallsetminus T, \quad F_L := (F \cup \overline{T}) \smallsetminus T \subset S_L, \quad C_L := C \smallsetminus T, \quad \overline{V}_L := \overline{V} \smallsetminus T \quad \text{and} \quad \partial_L \overline{V}_L := \overline{V}_L \cap \overline{T};$$

see Figure 4. Here $V \supset C$ is the neighborhood introduced in Section 5.1, and we assume the radius of T is small enough that $F_i \cap \overline{T}_j \subset V$ for each $i \neq j$.

Formally, we also need to shrink the surfaces F_i° to $F_i^\circ \smallsetminus T$, changing the boundary $\partial_L F_i^\circ$ to $(F_i^\circ \smallsetminus T) \cap \overline{T}$; however, using the obvious isomorphisms in (co)homology, we keep the notation $(F_i^\circ, \partial_L F_i^\circ)$ for these new pairs.
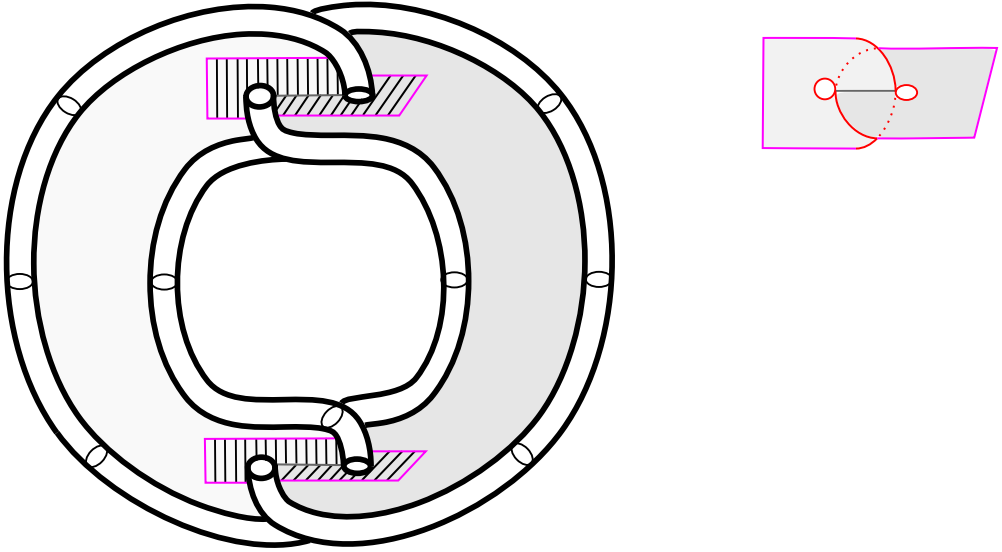
Figure 4: A minimal example of the set $F_L = (F \cup \overline{T}) \smallsetminus T$ consisting of the gray shaded surface together with the two depicted tori. The lined subset is $\overline{V}_L$. To the right we have a copy of a connected component of $\overline{V}_L$ with the subset $\partial_L \overline{V}_L$ highlighted in red.

We make use of the isomorphisms

$$(5.6) \qquad\qquad H_*^\omega(S_L, F_L) \simeq H_*(S_L, F_L) = H_*(S, F),$$

$$(5.7) \qquad\qquad H_*^\omega(F_i^\circ, \partial_L F_i^\circ) \simeq H_*(F_i^\circ, \partial_L F_i^\circ),$$

$$(5.8) \qquad H_*^\omega(\overline{V}_L, \partial_L \overline{V}_L) = H_*^\omega(C_L, \partial C_L) \simeq H_*(C_L, \partial C_L) = H_*(C, \partial C),$$

etc, and, in order to fix the (not quite canonical in the context of a common G–covering) isomorphisms denoted by $\simeq$, we need a coherent choice of reference sheets, upon which we change the notation to $=$. (The other isomorphisms are standard combinations of excision and homotopy equivalences, and thus are canonical.) To this end, we consider a "negative" collar (trace of the push-off in the negative direction) $N := (-\delta, 0) \times (F \smallsetminus T)$ for $\delta \ll \text{radius}(\overline{T})$, and, letting $S'_L := S_L \smallsetminus N$, use excision to identify

$$H_*(S_L, F_L) = H_*(S'_L, \partial S'_L) \quad \text{and} \quad H_*^\omega(S_L, F_L) = H_*^\omega(S'_L, \partial S'_L).$$

Since the covering is obviously trivial over S'_L, we can choose and fix a reference sheet $S'_L \subset \widetilde{S}_L$ and use it for (5.6). Then it remains to observe that this sheet contains a single copy of each of F_i° and C_L, which are used for (5.7) and (5.8), respectively.

Convention 5.9 We have then that $H_2^\omega(S_L, F_L) = H_2(S_L, F_L)$ and $H_1(F_L) = H_1^\omega(F_L)$. For the twisted boundary operators like

$$H_2(S_L, F_L) \to H_1(F_L),$$

we assume that $\partial^\omega = \sum_i (\partial^- + \omega_i^{-1} \partial^+)$, where ∂^+ is the *lower* boundary (the $+$ superscript is related to the orientation conventions).

Convention 5.10 The "reference lift" of a loop is the loop in the covering whose *endpoint* is in the reference sheet.

5.3 The homology of F

Throughout this section, *we assume that F is connected and that $\kappa \neq 0$.* (The general case will be treated later, see Figure 7.) Recall from Lemma 5.1 that $H^1(F)$ is a quotient of $\bigoplus H_1(F_i^\circ, \partial_L F_i^\circ)$ by relations of the form $\mathfrak{c}_i + \mathfrak{c}_j = 0$. We deduce the following description of the twisted homology of F:

Lemma 5.11 *The assignment $\tau \colon H^1(F) \to H_1^\omega(F_L, \partial\overline{T}) = H_1^\omega(F_L)$ given by*

$$\sum_{i=1}^{\mu} \alpha_i \mapsto \text{inclusion}_* \bigoplus_{i=1}^{\mu} \widetilde{\omega}_i \alpha_i \quad \text{for } \alpha_i \in H_1(F_i^\circ, \partial_L F_i^\circ)$$

is a well-defined isomorphism.

Proof The isomorphisms $H_*^\omega(F_L, \partial\overline{T}) = H_*^\omega(F_L)$ follow from the assumption $\omega_i \neq 1$ for each i, and hence $H_*^\omega(\partial\overline{T}) = 0$. We compute $H_1^\omega(F_L, \partial\overline{T})$ using the relative Mayer–Vietoris sequence associated to the decomposition $F \smallsetminus T = \overline{V}_L \cup \left(\bigcup_{i=1}^\mu F_i^\circ\right)$:

(5.12) $\quad H_1^\omega(\partial\overline{V}_L, \partial_L \overline{V}_L) \to H_1^\omega(\overline{V}_L, \partial_L \overline{V}_L) \oplus \bigoplus_{i=1}^{\mu} H_1^\omega(F_i^\circ, \partial_L F_i^\circ) \xrightarrow{p} H_1^\omega(F_L, \partial\overline{T}) \to 0.$

The last term is $H_0^\omega(\partial\overline{V}_L, \partial_L \overline{V}_L) = 0$; see (5.8) and Figure 4. By (5.8), $H_1^\omega(\partial\overline{V}_L, \partial_L \overline{V}_L) = \bigoplus \mathbb{C}\mathfrak{c}_i$, where the sum runs over all $\mathfrak{c} \in \mathfrak{C}_{ij}$ and all pairs $1 \leqslant i \neq j \leqslant \mu$. The inclusions induce the homomorphisms

(5.13)
$$\mathfrak{c}_i \mapsto \mathfrak{c}_i \in H_1^\omega(F_i^\circ, \partial_L F_i^\circ) = H_1(F_i^\circ, \partial_L F_i^\circ) \qquad \text{(see (5.7))},$$
$$\mathfrak{c}_i \mapsto \text{sg}(j-i) \cdot \text{sg}\,\mathfrak{c} \cdot \widetilde{\omega}_j \mathfrak{c} \in H_1^\omega(\overline{V}_L, \partial_L \overline{V}_L) = \bigoplus_{\mathfrak{c} \in \mathfrak{C}} \mathbb{C}\mathfrak{c}.$$

(To follow the above formulas, the reader might find helpful the schematics of the behavior of the twisted homology in Figure 5.) Identifying the two images of each generator $\mathfrak{c}_i$, we conclude that the inclusions $F_i^\circ \hookrightarrow F_L$ induce an isomorphism

$$\bigoplus_{i=1}^{\mu} H_1(F_i^\circ, \partial_L F_i^\circ)/\{\widetilde{\omega}_i \mathfrak{c}_i + \widetilde{\omega}_j \mathfrak{c}_j = 0 \mid \mathfrak{c} \in \mathfrak{C}_{ij}\} = H_1^\omega(F_L, \partial\overline{T}),$$

and the isomorphism in the statement follows from Lemma 5.1. $\qquad\square$

Corollary 5.14 *Given a proper loop $\alpha \subset F$, consider its push-off α^- and its "trace" $S^- \subset S^3$, ie a cylinder contained in a regular neighborhood of α and such that $S^- \cap F = \alpha$ and $\partial S^- = \alpha - \alpha^-$. Then the twisted boundary $\partial^\omega S^- + \alpha^-$ is equal to $\tau(R(\omega)(\alpha)) \in H_1^\omega(F_L)$; see (5.5) and Lemma 5.11.*
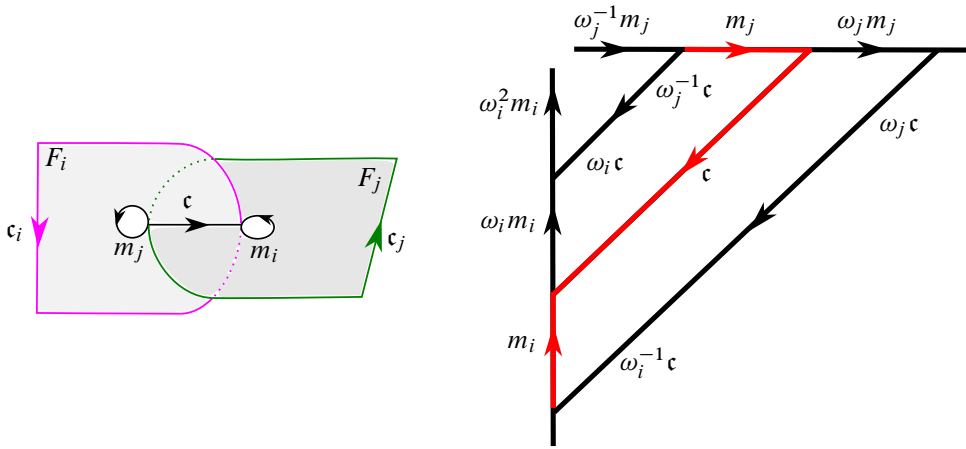
Figure 5: To the left is a local picture of a positive clasp with $i < j$. To the right, the schematics of the behavior of the lifted curves on a covering space. Shown in red are the chosen reference lifts.

Proof Clearly, using Lemma 5.11, $\partial^\omega S^- + \alpha^-$ is homologous to the image under p in (5.12) of the cycle

$$\sum_{i=1}^{\mu} \text{rel}_i^- \alpha + \sum_{1 \leqslant i < j \leqslant \mu} \sum_{\mathfrak{c} \in \mathfrak{C}_{ij}} \langle \alpha, \mathfrak{c}_i \rangle \mathfrak{c};$$

see Figure 6 for a simple example. Then, by (5.13), for all $i < j$ and $\mathfrak{c} \in \mathfrak{C}_{ij}$, we have $\mathfrak{c} = \text{sg}\,\mathfrak{c} \cdot \tilde{\omega}_j^{-1} \mathfrak{c}_i$ in $H_1^\omega(F_L)$ and, using (5.2), we obtain

$$\sum_{i=1}^{\mu} \text{rel}_i^- \alpha + \sum_{1 \leqslant i < j \leqslant \mu} \tilde{\omega}_j^{-1} \sum_{\mathfrak{c} \in \mathfrak{C}_{ij}} \text{sg}\,\mathfrak{c}\langle \alpha, \mathfrak{c}_i \rangle \mathfrak{c}_i \overset{(5.2)}{=} \sum_{i=1}^{\mu} \text{rel}_i^- \alpha + \sum_{1 \leqslant i < j \leqslant \mu} \tilde{\omega}_j^{-1} \text{rel}_{ij} \alpha$$

$$= \sum_{i=1}^{\mu} \tilde{\omega}_i \bigg(\underbrace{\tilde{\omega}_i^{-1} \text{rel}_i^- \alpha + \sum_{j=i+1}^{\mu} \tilde{\omega}_i^{-1} \tilde{\omega}_j^{-1} \text{rel}_{ij} \alpha}_{R_i} \bigg).$$

Now, by (5.5), each R_i is the ith component of (a representative of) $R(\omega)(\alpha)$, and the statement follows from the definition of τ in Lemma 5.11.   □

We proceed with the computation of the twisted homology of S_L and $S_L \smallsetminus K$. We have fixed isomorphisms

$$H_*^\omega(S_L, F_L) = H_*(S, F) \quad \text{and} \quad H_*^\omega(S_L \smallsetminus K, F_L) = H_*(S \smallsetminus K, F);$$

see (5.6). In particular,

$$H_1^\omega(S_L, F_L) = H_1^\omega(S_L \smallsetminus K, F_L) = 0$$

(recall that we assume F is connected and $\kappa \neq 0$) and, by the respective exact sequences of pairs (S, F) and $(S \smallsetminus K, F)$,

$$H_2^\omega(S_L, F_L) = H_1(F) \quad \text{and} \quad H_2^\omega(S_L \smallsetminus K, F_L) = \text{Ker}\,\kappa \subset H_1(F).$$
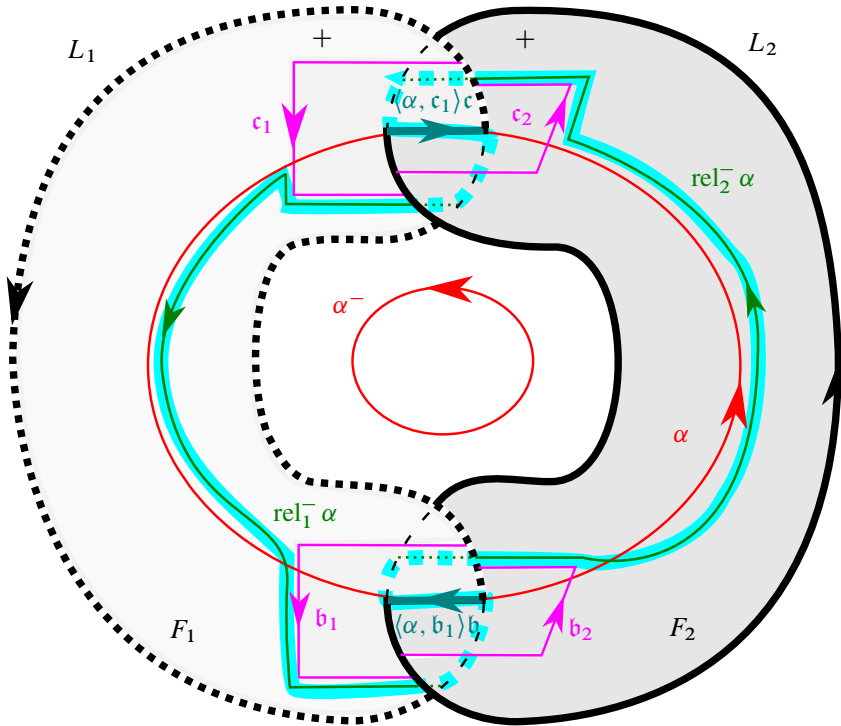
Figure 6: The push-off α^- is to be thought of as located "behind" the surface $F_1 \cup F_2$. With the orientations depicted, together α and $-\alpha^-$ are the obvious boundary of the cylinder S^- (not in the picture). The different elements of the cycle described at the beginning of the proof of Corollary 5.14, $\text{rel}_i^- \alpha$ and $\langle \alpha, c_i \rangle c$, are highlighted.

Now, from the corresponding twisted exact sequences, and with the isomorphism τ given by Lemma 5.11 taken into account, we arrive at

$$(5.15) \qquad H_1^\omega(S_L) = H^1(F)/\operatorname{Im} d \quad \text{and} \quad H_1^\omega(S_L \smallsetminus K) = H^1(F)/d(\operatorname{Ker} \kappa),$$

where d is the composed map

$$(5.16) \qquad d \colon H_1(F) \xrightarrow{\partial^{-1}} H_2(S, F) = H_2^\omega(S_L, F_L) \xrightarrow{\partial^\omega} H_1^\omega(F_L) \xrightarrow{\tau^{-1}} H^1(F).$$

5.4 The twisted homomorphisms

We still assume that F is connected and $\kappa \neq 0$. By (5.15), for $X := S_L$ or $X := S_L \smallsetminus K$, we have natural epimorphisms

$$(5.17) \qquad \pi_X \colon H^1(F) \twoheadrightarrow H_1^\omega(X).$$

Composing the inclusion with Alexander duality, we obtain a homomorphism

$$\mathrm{D} \colon H_1^\omega(X \smallsetminus F_L) = H_1(X \smallsetminus F_L) \to H_1(S^3 \smallsetminus F) \xrightarrow{\cong} H^1(F).$$

Consider also the "orthogonal projection"

$$\mathrm{pr}_X \colon H_1^\omega(X \smallsetminus F_L) \to H_1^\omega(X \smallsetminus F_L) \quad \text{given by} \quad \begin{cases} \alpha \mapsto \alpha & \text{if } X = S_L, \\ \alpha \mapsto \alpha - \ell k(\alpha, K)m & \text{if } X = S_L \smallsetminus K. \end{cases}$$

Lemma 5.18 *For $X = S_L$ or $S_L \smallsetminus K$ and any class $\alpha \in H_1^\omega(X \smallsetminus F_L)$, the image of $\mathrm{pr}_X(\alpha)$ under the inclusion homomorphism $H_1^\omega(X \smallsetminus F_L) \to H_1^\omega(X)$ is $\pi_X(\mathrm{D}(\alpha))$.*

Proof The statement is a geometric version of Lemma 5.11. The class $\alpha' := \mathrm{pr}_X(\alpha)$ is represented by a cycle in $X \smallsetminus F_L$, which bounds a Seifert surface $G \subset S^3 \smallsetminus K$. (This is why we subtract $\ell k(\alpha, K)m$ in the case $X = S_L \smallsetminus K$; we want a Seifert surface disjoint from K.) Set $G_L := G \cap S_L$. We can choose the surface G_L so that it cuts on F a collection of arcs $\alpha_i \subset F_i^\circ$ with $\partial \alpha_i \subset \partial_L F_i^\circ$. Then $\mathrm{D}(\alpha')$ is represented by

$$\sum_{i=1}^\mu \alpha_i \in \bigoplus_{i=1}^\mu H_1(F_i^\circ, \partial_L F_i^\circ) \to H^1(F)$$

(see Lemma 5.1), whereas the twisted boundary is

(5.19) $$\partial^\omega G_L - \alpha' = -\sum_{i=1}^\mu \widetilde{\omega}_i \alpha_i = -\tau(\mathrm{D}(\alpha)),$$

(see Lemma 5.11), implying that $\alpha' = \tau(\mathrm{D}(\alpha))$ in $H_1^\omega(F_L) = H^1(F)$. We complete the proof by passing to the quotient using π_X. $\square$

Corollary 5.20 *For $X = S_L$ or $S_L \smallsetminus K$, let $\alpha \in H_1^\omega(X \smallsetminus F_L)$ be the class of $[K]$ or ℓ, respectively. Then the image of α in $H_1^\omega(X)$ is $\pi_X(\kappa)$.*

Lemma 5.21 *The homomorphism d in (5.16) equals $-E(\omega)$.*

Lemma 5.22 *For each $\alpha \in H_1(F)$, one has*

$$\pi_{S_L \smallsetminus K}(E(\omega)(\alpha)) = -\langle \alpha, \kappa \rangle m$$

in $H_1^\omega(S_L \smallsetminus K)$; see (5.17).

Proof of Lemmas 5.21 and 5.22 Let $\alpha \subset F$ be a proper loop and consider its push-off $\alpha^- \subset S^3 \smallsetminus (K \cup F)$. Let S^- be the trace cylinder as in Corollary 5.14, and let G be a Seifert surface bounded by α^-. (For Lemma 5.22, we replace α^- with its projection $\mathrm{pr}(\alpha^-) = \alpha^- - \langle \alpha, \kappa \rangle m$ in order to keep S in $S^3 \smallsetminus K$; details are left to the reader.)

Defining $G_L := G \cap S_L$ and letting $\overline{S} := G_L \cup S^-$, we have $\partial \overline{S} = \alpha$. On the other hand, the twisted boundary

$$\partial^\omega \overline{S} = (\partial^\omega S^- + \alpha^-) + (\partial^\omega G_L - \alpha^-) = \tau(R(\omega)(\alpha)) - \tau(\Theta^-(\alpha))$$

is given by Corollary 5.14 and (5.19), and the statements follow from (5.5). $\square$
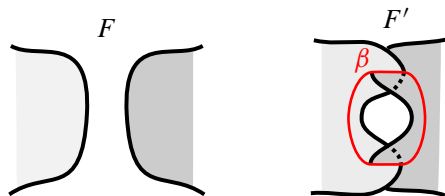
Figure 7: To the left is a local picture of a disconnected C–complex F. To the right, the complex F', obtained by adding a pair of close clasps to F. We have $H_1(F'; \mathbb{Z}) = H_1(F; \mathbb{Z}) \oplus \mathbb{Z}\beta$.

Corollary 5.23 (of Lemma 5.21 and (5.15)) *There are canonical, up to multiplication by integral powers of $\omega_i s$, isomorphisms*

$$H_1^\omega(S_L) = H^1(F)/\operatorname{Im} E(\omega) \quad and \quad H_1^\omega(S_L \smallsetminus K) = H^1(F)/E(\omega)(\operatorname{Ker}\kappa).$$

Proof of Theorem 4.3 If $\kappa = 0$, then K bounds a Seifert surface disjoint from F, and hence $K/L \equiv 0$, which agrees with the statement of the theorem.

Therefore, till the rest of the proof we assume that $\kappa \neq 0$. Assume also that F is connected, so that we can use the results of Sections 5.3 and 5.4. Abbreviate $E := E(\omega)$, so that $E^* = E(\omega^{-1})$ and $\operatorname{Ker} E^\perp = \operatorname{Im} E^*$. Then, in view of Corollary 5.23, the last two cases in the statement, as well as the finiteness of the slope in the first case, are given by Proposition 2.6. To compute this finite slope in the first case, we compare Corollary 5.20 and Lemma 5.22: if $\kappa = E(\alpha)$, then $\ell = -\langle\alpha, \kappa\rangle m$ in $H_1^\omega(S_L \smallsetminus K)$.

Finally, if F is not connected, we can inductively reduce the number of components by introducing pairs of close clasps as in Figure 7. If F' is obtained from F by introducing one such pair connecting two distinct components, then $H_1(F'; \mathbb{Z}) = H_1(F; \mathbb{Z}) \oplus \mathbb{Z}\beta$, where β is a small proper loop running through the two clasps, and, extending the existing pair of dual bases by $\beta \in H_1(F)$ and $\beta^* \in H^1(F)$, the other data are

$$\Theta'^\varepsilon = \Theta^\varepsilon \oplus [0] \quad and \quad \kappa' = \kappa \oplus [0].$$

Obviously, this modification does not affect the result of the computation. $\qquad\square$

References

[1] **J Amundsen**, **E Anderson**, **C W Davis**, **D Guyer**, *The C–complex clasp number of links*, Rocky Mountain J. Math. 50 (2020) 839–850 MR Zbl

[2] **J C Cha**, **C Livingston**, *Knot signature functions are independent*, Proc. Amer. Math. Soc. 132 (2004) 2809–2816 MR Zbl

[3] **D Cimasoni**, *A geometric construction of the Conway potential function*, Comment. Math. Helv. 79 (2004) 124–146 MR Zbl

[4] **D Cimasoni**, **V Florens**, *Generalized Seifert surfaces and signatures of colored links*, Trans. Amer. Math. Soc. 360 (2008) 1223–1264 MR Zbl

[5] **T D Cochran**, *Geometric invariants of link cobordism*, Comment. Math. Helv. 60 (1985) 291–311 MR Zbl

[6] **A Conway**, **S Friedl**, **E Toffoli**, *The Blanchfield pairing of colored links*, Indiana Univ. Math. J. 67 (2018) 2151–2180 MR Zbl

[7] **A Conway**, **M Nagel**, **E Toffoli**, *Multivariable signatures, genus bounds, and 0.5–solvable cobordisms*, Michigan Math. J. 69 (2020) 381–427 MR Zbl

[8] **D Cooper**, *The universal abelian cover of a link*, from "Low-dimensional topology" (R Brown, T L Thickstun, editors), Lond. Math. Soc. Lect. Note Ser. 48, Cambridge Univ. Press (1982) 51–66 MR Zbl

[9] **C W Davis**, **T Martin**, **C Otto**, *Moves relating C–complexes*, Topology Appl. 302 (2021) art. id. 107799 MR Zbl Correction to [3]

[10] **C W Davis**, **G Roth**, *When do links admit homeomorphic C–complexes?*, J. Knot Theory Ramifications 26 (2017) art. id. 1750010 MR Zbl

[11] **A Degtyarev**, **V Florens**, **A G Lecuona**, *The signature of a splice*, Int. Math. Res. Not. 2017 (2017) 2249–2283 MR Zbl

[12] **A Degtyarev**, **V Florens**, **A G Lecuona**, *Slopes of links and signature formulas*, from "Topology, geometry, and dynamics" (A M Vershik, V M Buchstaber, A V Malyutin, editors), Contemp. Math. 772, Amer. Math. Soc., Providence, RI (2021) 93–105 MR Zbl

[13] **A Degtyarev**, **V Florens**, **A G Lecuona**, *Slopes and signatures of links*, Fund. Math. 258 (2022) 65–114 MR Zbl

[14] **G T Jin**, *On Kojima's η–function of links*, from "Differential topology" (U Koschorke, editor), Lecture Notes in Math. 1350, Springer (1988) 14–30 MR Zbl

[15] **A Kawauchi**, *A survey of knot theory*, Birkhäuser, Basel (1996) MR Zbl

[16] **S Kojima**, **M Yamasaki**, *Some new invariants of links*, Invent. Math. 54 (1979) 213–228 MR Zbl

[17] **A Libgober**, *Characteristic varieties of algebraic curves*, from "Applications of algebraic geometry to coding theory, physics and computation" (C Ciliberto, F Hirzebruch, R Miranda, M Teicher, editors), NATO Sci. Ser. II Math. Phys. Chem. 36, Kluwer, Dordrecht (2001) 215–254 MR Zbl

[18] **A Merz**, *An extension of a theorem by Cimasoni and Conway*, preprint (2021) arXiv 2104.02993

[19] **M Nagel**, **M Powell**, *Concordance invariance of Levine–Tristram signatures of links*, Doc. Math. 22 (2017) 25–43 MR Zbl

Department of Mathematics, Bilkent University
Ankara, Turkey

Labaratoire de Mathématiques et leurs Applications, Université de Pau et des Pays de l'Adour
Pau, France

Aix Marseille Université, CNRS, Centrale Marseille, Institut de Mathématiques de Marseille
Marseille, France

Current address: *School of Mathematics and Statistics, University of Glasgow*
Glasgow, United Kingdom

degt@fen.bilkent.edu.tr, vincent.florens@univ-pau.fr, ana.lecuona@glasgow.ac.uk

■■
■msp

Cohomological and geometric invariants of simple complexes of groups

NANSEN PETROSYAN

TOMASZ PRYTUŁA

We investigate cohomological properties of fundamental groups of strictly developable simple complexes of groups X. We obtain a polyhedral complex equivariantly homotopy equivalent to X of the lowest possible dimension. As applications, we obtain a simple formula for proper cohomological dimension of CAT(0) groups whose actions admit a strict fundamental domain; for any building of type (W, S) that admits a chamber transitive action by a discrete group, we give a realisation of the building of the lowest possible dimension equal to the virtual cohomological dimension of W; under general assumptions, we confirm a folklore conjecture on the equality of Bredon geometric and cohomological dimensions in dimension one; finally, we give a new family of counterexamples to the strong form of Brown's conjecture on the equality of virtual cohomological dimension and Bredon cohomological dimension for proper actions.

1 Introduction

Overview

For a finitely generated Coxeter system (W, S), the Davis complex Σ_W is a CAT(0) polyhedral complex on which the Coxeter group W acts properly, cocompactly and by reflections. The complex Σ_W is very useful in understanding properties of W, or more generally of buildings of type (W, S) where it appears as an apartment. However, the Davis complex Σ_W does not in general produce the realisation of these buildings of the lowest possible dimension. There is an associated contractible polyhedral complex $B(W, S)$ of dimension equal to the virtual cohomological dimension vcd W of the Coxeter group W (except possibly when vcd $W = 2$) introduced by Bestvina in [4]. The group W acts by reflections properly and cocompactly on $B(W, S)$. The Bestvina complex $B(W, S)$ is equivariantly homotopy equivalent to the Davis complex Σ_W; see the authors' [26]. Therefore by replacing the apartments with $B(W, S)$ one obtains a realisation of the building of type (W, S) of the lowest possible dimension. In [26], we derived analogous results in the more general setting of strictly developable thin simple complexes of finite groups. In doing so we relied on compactly supported cohomology as a convenient tool for computations. This is certainly the norm, as compactly supported cohomology can be very useful in computations of the cohomology of a G–CW–complex with group ring coefficients; see Bestvina [4], Brown [6], Davis [9],

Degrijse and Martínez-Pérez [12], and Harlander and Meinert [20]. A major drawback of this approach however is that it restricts one to only complexes that are locally finite.

To resolve this difficulty, in this paper we introduce a new approach that bypasses compactly supported cohomology and thus allows us to study nonproper actions admitting a strict fundamental domain, or equivalently, simple complexes of groups whose local groups need not be finite. Given a simple complex of groups $G(\mathcal{Q})$, we first extend the definition of the Bestvina complex to $G(\mathcal{Q})$. Our methods then directly link Bredon cohomology of the Bestvina complex associated to $G(\mathcal{Q})$ with certain coefficients, and the relative integral cohomology of the panel complexes over the poset $\mathcal{Q}$. This enables us to compute the cohomology of the fundamental group of $G(\mathcal{Q})$, determine its cohomological dimension, and identify it with the dimension of the generalisation of Bestvina complex in this context.

Our approach also leads to cohomological computations on more naturally occurring simple complexes of groups without the *thinness* assumption. This is a standing assumption in both [12] and [26]. It states that the cellular structure of a complex is in a sense minimal with respect to the group action, and it is fairly restrictive. In particular, removing this assumption allows us to investigate group actions on CAT(0) polyhedral complexes that admit a strict fundamental domain.

Besides aforementioned applications, another important motivation to study the generalised Bestvina complex comes from the Baum–Connes and Farrell–Jones conjectures (see eg Baum, Connes and Higson [3], and Lück [23]), where it is always desirable to have models for the classifying space of G for the family of stabilisers $\mathcal{F}$ of minimal dimension and cell structure (see eg Fuentes [17] for a direct application of the Bestvina complex).

Statement of results

A *simple complex of groups* $G(\mathcal{Q})$ over a finite poset $\mathcal{Q}$ consists of a collection of groups $\{P_J\}_{J \in \mathcal{Q}}$ and a collection of monomorphisms $\{P_J \to P_T\}_{J \leqslant T \in \mathcal{Q}}$ satisfying the obvious compatibility conditions. We say that $G(\mathcal{Q})$ is *thin* if the monomorphism $P_J \to P_T$ is an isomorphism if and only if $J = T$. The *fundamental group* G of $G(\mathcal{Q})$ is defined as the direct limit of the system $\{P_J\}_{J \in \mathcal{Q}}$. We say that $G(\mathcal{Q})$ is *strictly developable* if for every group P_J the canonical map to the limit G is a monomorphism; in this case, we identify P_J with its image in G and call it a *local subgroup* of G.

A *family* of subgroups of a discrete group G is a collection of subgroups that is closed under conjugation and taking subgroups. Given a collection of subgroups $\{P_J\}_{J \in \mathcal{Q}}$ of G, the family *generated by the collection* $\{P_J\}_{J \in \mathcal{Q}}$ is the smallest family $\mathcal{F}$ of subgroups of G that contains all elements of $\{P_J\}_{J \in \mathcal{Q}}$.

Suppose $G(\mathcal{Q})$ is strictly developable with fundamental group G. We say that $G(\mathcal{Q})$ is *rigid*, if for any $J \in \mathcal{Q}$ no G–conjugate of P_J is properly contained in P_J. Define a *block* $C \subseteq \mathcal{Q}$ as an equivalence class of elements of $\mathcal{Q}$ under the relation $\sim$ generated by $J' \sim J$ if $J' \leqslant J$ and $P_{J'} \to P_J$ is an isomorphism. For a fixed $J \in \mathcal{Q}$ and $g \in G$, let Ω_J^g be the subset of $\mathcal{Q}$ that consists of all $U \in \mathcal{Q}$ for

which $P_U = g^{-1} P_J g$ (seen as subgroups of G). We denote by $C_J^g \subseteq \Omega_J^g$ a block in Ω_J^g. Let $\mathcal{I}_J$ be a complete set of representatives of

$$\{g \in G \mid g^{-1} P_J g = P_U \text{ for some } U \in \mathcal{Q}\}/P_J,$$

where P_J acts by left multiplication.

Let $K = |\mathcal{Q}|$ denote the geometric realisation of the poset $\mathcal{Q}$. For a subset $\Omega \subseteq \mathcal{Q}$ such that $P_U = P_{U'}$ for all $U, U' \in \Omega$, define subcomplexes K_Ω and $K_{>\Omega}$ of K as

$$K_\Omega = \big|\{V \in \mathcal{Q} \mid V \geqslant J \text{ for some } J \in \Omega\}\big|,$$
$$K_{>\Omega} = \big|\{V \in \mathcal{Q} \mid V \geqslant J \text{ for some } J \in \Omega \text{ and } P_V \gneqq P_J\}\big|.$$

The complex $K = |Q|$ is an example of a *panel complex* over the poset $\mathcal{Q}$. For a panel complex Y over $\mathcal{Q}$, the *Basic Construction* is a G–space $D(Y, G(\mathcal{Q}))$ obtained by gluing copies of Y indexed by elements of G, according to the combinatorial information in Y and $G(\mathcal{Q})$.

We denote by $H_{\mathcal{F}}^*(X; M)$ the Bredon cohomology groups of a G–CW–complex X with respect to the family of subgroups $\mathcal{F}$ of G with coefficients a contravariant functor M from the orbit category $\mathcal{O}_{\mathcal{F}}G$ to $\mathbb{Z}$–Mod. In what follows, we will restrict to coefficients $\mathcal{A}_H = \mathbb{Z}[\hom_G(-, G/H)]$ for a subgroup $H \in \mathcal{F}$, and a certain refinement of $\mathcal{A}_H$ which we denote by $\mathcal{B}_H$. Let $\mathrm{cd}_{\mathcal{F}} G$ (resp. $\mathrm{gd}_{\mathcal{F}} G$) denote the Bredon cohomological (resp. geometric) dimension of G with respect to the family $\mathcal{F}$ and let $E_{\mathcal{F}} G$ denote the universal G–CW–complex with stabilisers in $\mathcal{F}$. If $\mathcal{F}$ is the family of all finite subgroups of G, then the respective notions are denoted by $\underline{\mathrm{cd}}\, G$, $\underline{\mathrm{gd}}\, G$, and $\underline{E}G$.

Theorem 1.1 (Theorem 6.1) *Let $G(\mathcal{Q})$ be a strictly developable simple complex of groups with fundamental group G, and let $\mathcal{F}$ be the family generated by local groups. Let $X = D(K, G(\mathcal{Q}))$ be the associated Basic Construction. For $J \in \mathcal{Q}$,*

$$(1\text{-}1) \qquad H_{\mathcal{F}}^*(X; \mathcal{B}_{P_J}) \cong \bigoplus_{g \in \mathcal{I}_J} \bigoplus_{C_J^g \subseteq \Omega_J^g} H^*(K_{C_J^g}, K_{>C_J^g}).$$

If $G(\mathcal{Q})$ is rigid and X is a model for $E_{\mathcal{F}}G$, then

$$(1\text{-}2) \qquad \mathrm{cd}_{\mathcal{F}} G = \max\{n \in \mathbb{N} \mid H^n(K_C, K_{>C}) \neq 0 \text{ for some block } C \subseteq \mathcal{Q}\}.$$

The rigidity assumption holds for example when the local groups are co-Hopfian, and hence in particular when they are finite. We should also remark that the rigidity assumption on local groups in Theorem 1.1 is not superfluous.

Recall that an action of a group on cellular complex is admissible, if the setwise stabiliser of each cell is also its pointwise stabiliser. If a group G acts admissibly on a simply connected cellular complex with a strict fundamental domain Y then it is isomorphic to the fundamental group of a simple complex of groups formed by cells of Y and their stabilisers (see Theorem 3.8). The following corollary of Theorem 1.1 is straightforward.

Corollary 1.2 *Suppose a group G acts properly and admissibly on a CAT(0) polyhedral complex X with a strict fundamental domain Y. Let $\mathcal{Q}$ denote the poset of cells of Y ordered by reverse inclusion (note that we have $|\mathcal{Q}| = K = Y'$). Then*

$$\underline{\mathrm{cd}}\, G = \max\{n \in \mathbb{N} \mid H^n(K_C, K_{>C}) \neq 0 \text{ for some block } C \subseteq \mathcal{Q}\}.$$

This corollary is a generalisation of [12, Theorem 1.2] to nonthin complexes of groups. We remark that nonthinness of a complex of groups resulting from the G–action on X is generic, eg in many cases in order to obtain an admissible action one takes the barycentric subdivision which results in a nonthin complex.

To obtain formula (1-2) of Theorem 1.1, we prove the following general result.

Theorem 1.3 (Theorem 2.5) *Let G be a group and $\mathcal{F}$ be a family of subgroups. Suppose that X is a cocompact model for $E_{\mathcal{F}}G$. Then*

$$\mathrm{cd}_{\mathcal{F}}\, G = \max\{k \in \mathbb{N} \mid H^k_{\mathcal{F}}(X, \mathcal{A}_H) \neq 0 \text{ for some cell stabiliser } H\}.$$

Moreover, if $H^n_{\mathcal{F}}(G; \mathcal{A}_L) \neq 0$ for $n = \mathrm{cd}_{\mathcal{F}}\, G$ and $L \in \mathcal{F}$, then there exists a cell stabiliser $H \leqslant L$ such that $H^n_{\mathcal{F}}(G; \mathcal{A}_H) \neq 0$.

Note that, under the assumptions of the theorem, there are only finitely many conjugacy classes of stabilisers. Thus the theorem reduces the computation of the Bredon cohomological dimension of a given group into a computation of finitely many cohomology groups. Theorem 1.3 together with [12, Theorem 2.4] give us the following strengthening of [12, Theorem 1.1].

Corollary 1.4 (Corollary 2.7) *Let X be a G–CW–complex that is a cocompact model for $\underline{E}G$. Then*

$$\underline{\mathrm{cd}}\, G = \max\{k \in \mathbb{N} \mid H^k_c(X^H, X^H_{\mathrm{sing}}) \neq 0 \text{ for some cell stabiliser } H\},$$

where $X^H_{\mathrm{sing}} \subseteq X^H$ consists of all points whose stabiliser strictly contains H.

Another application of Theorem 1.1 is the construction of new counterexamples to the strong form of Brown's conjecture via the notion of *reflection-like actions*. Here the removal of the thinness assumption is the key to obtaining a systematic approach to constructing such examples. Reflection-like actions are generalisations of groups acting by reflections on Euclidean spaces. The precise definition and examples can be found in Section 9.

Theorem 1.5 (Theorem 9.8) *Let F be a finite group admitting a reflection-like action on a compact, connected, flag simplicial complex L of dimension $n \geqslant 1$. Let W_L be the right-angled Coxeter group associated to L and $G = W_L \rtimes F$ be the associated semidirect product. Suppose that $H^n(L) = 0$. Then*

$$\mathrm{vcd}\, G \leqslant n \quad \text{and} \quad \underline{\mathrm{cd}}\, G = n + 1.$$

We refer the reader to Examples 9.14 and 9.16 for a specific construction of complexes L satisfying the hypothesis of Theorem 1.5 via products of dihedral group actions on 2–dimensional Moore spaces.

Observe that as long as the complex of groups $G(\mathcal{Q})$ is thin, Theorem 1.1 implies that the Bredon cohomological dimension of G depends only on the poset structure of $\mathcal{Q}$. We show that for a strictly developable thin simple complex of groups, there is a model for $E_{\mathcal{F}}G$ of the smallest possible dimension and a simple cell structure. The model is given as the Basic Construction where one replaces panel complex K with the so-called *Bestvina complex B*.

Theorem 1.6 *Let $G(\mathcal{Q})$ be a strictly developable thin complex of groups over a poset $\mathcal{Q}$ with fundamental group G and let $\mathcal{F}$ be the family generated by the local groups. Then*

(i) *the standard development $D(K, G(\mathcal{Q}))$ and the Bestvina complex $D(B, G(\mathcal{Q}))$ are G–homotopy equivalent, and*

$$H^*_{\mathcal{F}}\big(D(K, G(\mathcal{Q})); \mathcal{B}_{P_J}\big) \cong \bigoplus_{g \in \mathcal{I}_J} \bigoplus_{U \in \Omega^g_J} \widetilde{H}^{*-1}(K_{>U});$$

(ii) *if $D(K, G(\mathcal{Q}))$ is a model for $E_{\mathcal{F}}G$, then $D(B, G(\mathcal{Q}))$ is a cocompact model for $E_{\mathcal{F}}G$ satisfying*

$$\dim\big(D\big(B, G(\mathcal{Q})\big)\big) = \begin{cases} \mathrm{cd}_{\mathcal{F}}\, G & \text{if } \mathrm{cd}_{\mathcal{F}}\, G \neq 2, \\ 2 \text{ or } 3 & \text{if } \mathrm{cd}_{\mathcal{F}}\, G = 2, \end{cases}$$

and

(1-3) $$\mathrm{cd}_{\mathcal{F}}\, G = \max\{n \in \mathbb{N} \mid \widetilde{H}^{n-1}(K_{>J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}.$$

Since buildings are CAT(0) and chamber transitive actions on them are thin (see Lemma 8.2), they are ideally suited for applying Theorem 1.6.

Corollary 1.7 *Let G be a group acting chamber transitively on a building of type (W, S). Let $G(\mathcal{Q})$ be the associated simple complex of groups and let $\mathcal{F}$ be the family generated by the stabilisers. Then $D(B, G(\mathcal{Q}))$ is a realisation of the building (and thus a cocompact model for $E_{\mathcal{F}}G$) of dimension*

$$\dim\big(D(B, G(\mathcal{Q}))\big) = \begin{cases} \mathrm{vcd}\, W & \text{if } \mathrm{vcd}\, W \neq 2, \\ 2 \text{ or } 3 & \text{if } \mathrm{vcd}\, W = 2. \end{cases}$$

Moreover,

$$H^*_{\mathcal{F}}(G; \mathcal{B}_{P_J}) \cong \bigoplus_{g \in \mathcal{I}_J} \bigoplus_{U \in \Omega^g_J} \widetilde{H}^{*-1}(K_{>U})$$

and

$$\mathrm{cd}_{\mathcal{F}}\, G = \mathrm{vcd}\, W = \max\{n \in \mathbb{N} \mid \widetilde{H}^{n-1}(K_{>J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}.$$

The formula for Bredon cohomological dimension in Corollary 1.7 extends [12, Corollary 1.4] from finite to arbitrary stabilisers. As a consequence of Corollary 1.7, we obtain one of the main results of Harlander [19].

Corollary 1.8 (Corollary 8.4) *Let G be a virtually torsion-free group acting chamber transitively on a building of type (W, S). Then*

$$\mathrm{vcd}\, G \leqslant \mathrm{vcd}\, W + \max\{\mathrm{vcd}\, P \mid P \text{ is a special parabolic subgroup of } G\}.$$

We point out that in [19, Theorem 2.8] it is proven that, under the assumptions of Theorem 1.6, the dimension of Bestvina complex is minimal among G–complexes which admit a strict fundamental domain with all acyclic panels. Theorem 1.6(ii) is stronger, as it states that the dimension of the Bestvina complex is minimal among all possible models for $E_{\mathcal{F}} G$ (except the case where $\mathrm{cd}_{\mathcal{F}} G = 2$).

The next corollary lists equivalent conditions for fundamental groups of strictly developable thin simple complexes of groups to act on trees with the prescribed family of stabilisers.

Corollary 1.9 (Theorem 7.1) *Let $G(\mathcal{Q})$ be a strictly developable thin simple complex of groups over a poset $\mathcal{Q}$ with the fundamental group G and let $\mathcal{F}$ be the family generated by local groups. Suppose that $D(K, G(\mathcal{Q}))$ is a model for $E_{\mathcal{F}} G$. Then the following are equivalent*:

  (i) $D(B, G(\mathcal{Q}))$ *is a tree and an equivariant deformation retract of $D(K, G(\mathcal{Q}))$.*

  (ii) $\mathrm{cd}_{\mathcal{F}} G \leqslant 1$.

  (iii) $H^n(K_{>J}) = 0$ *for all $J \in \mathcal{Q}$ and $n \geqslant 1$.*

The following corollary is immediate.

Corollary 1.10 *Let G be a group acting chamber transitively on a building of type (W, S). The geometric realisation of the building equivariantly deformation retracts onto a tree if and only if $\mathrm{vcd}\, W \leqslant 1$.*

Corollary 1.9 is a generalisation of [9, Proposition 8.8.5] which deals with the case when $G = W$ is a Coxeter group acting on the Davis complex. It is a special case of the following folklore conjecture.

Conjecture 1.11 *Let G be a group and $\mathcal{F}$ be a family of subgroups. Then $\mathrm{cd}_{\mathcal{F}} G \leqslant 1$ if and only if G acts on a tree with stabilisers generating $\mathcal{F}$.*

This conjecture is wide open in general. When $\mathcal{F}$ is the trivial family, it reduces to the classical theorem of Stallings and Swan. For the family of finite subgroups $\mathcal{F}$, it follows from Dunwoody's accessibility result [16]. Recently, in [11], Degrijse verified the conjecture when $\mathcal{F}$ is the family of virtually cyclic subgroups. Note that Corollary 1.9 confirms the conjecture when G admits a model for $E_{\mathcal{F}} G$ with a strict fundamental domain such that the associated complex of groups is thin.

Organisation

Sections 2 and 3 have a preparatory character. In Section 2, we give a background on classifying spaces for families of subgroups and Bredon cohomology, and we prove Theorem 1.3. In Section 3, we define simple complexes of groups, the Basic Construction and Bestvina complex. We describe the procedure of thinning, and we use it to compute upper bounds for the geometric dimension of the fundamental group of a simple complex of groups.

The next three sections form the technical core of the paper. In Section 4, we prove Proposition 4.1 which allows us to compute the Bredon cohomological dimension of a fundamental group of a thin complex of groups. In Section 5, we prove an analogous Proposition 5.1 for an arbitrary complex of groups. Section 6 contains generalised statements and proofs of Theorems 1.1 and 1.6.

In the remaining sections we discuss applications and consequences of the main theorems. In Section 7, we briefly discuss the case when $\mathrm{cd}_{\mathcal{F}} G = 1$ and we give a proof of Corollary 1.9. In Section 8, we discuss applications of our theory to chamber transitive automorphism groups of buildings and we prove Corollary 1.7 as well as other applications and examples. In Section 9, we define reflection-like actions, establish their basic properties and prove Theorem 1.5. We then give some examples of reflection-like actions. Finally, in Section 10 we pose and discuss some open questions.

Acknowledgements

2 Classifying spaces and Bredon cohomology

2.1 Classifying spaces for families of subgroups

Let G be a countable discrete group. A *family* $\mathcal{F}$ of subgroups of G is a collection of subgroups that is closed under conjugation and taking subgroups. Given a collection of subgroups $\mathcal{P}$ of G, the *family of subgroups generated by* $\mathcal{P}$ is the smallest family of subgroups $\mathcal{F}$ of G containing all subgroups of $\mathcal{P}$.

Definition 2.1 A collection of subgroups $\mathcal{P}$ of G is *rigid* if for every $H \in \mathcal{P}$ no G–conjugate of H is properly contained in H.

Recall that a *polyhedron* (or a *polyhedral complex*) is a CW–complex whose attaching maps are piecewise linear. We say that the action of a group G on a polyhedral (CW, simplicial) complex X is *admissible* if for any cell $e \subset X$ its pointwise stabiliser is equal to its setwise stabiliser. In such case we call X a G–polyhedral (G–CW, G–simplicial) complex. A G–CW–complex X is *cocompact* (or the G–action on X is cocompact) if X/G is compact, ie it has finitely many cells.

Definition 2.2 (classifying space $E_{\mathcal{F}} G$) Given a group G and a family of its subgroups $\mathcal{F}$, a model for the *classifying space of G for the family $\mathcal{F}$* denoted by $E_{\mathcal{F}} G$ is a G–CW–complex X such that

- for any cell $e \subset X$ the stabiliser G_e belongs to the family $\mathcal{F}$,
- for any $H \in \mathcal{F}$ the fixed point set X^H is contractible.

The classifying space $E_{\mathcal{F}}G$ is a terminal object in the homotopy category of G–CW–complexes with stabilisers in $\mathcal{F}$, ie if X is a G–CW–complex with stabilisers in $\mathcal{F}$ then there exists a G–map $X \to E_{\mathcal{F}}G$ which is unique up to G–homotopy. In particular, any two models for $E_{\mathcal{F}}G$ are G–homotopy equivalent. The minimal dimension of a model for $E_{\mathcal{F}}G$ is called the *Bredon geometric dimension of G for the family $\mathcal{F}$* and it is denoted by $\mathrm{gd}_{\mathcal{F}} G$.

Remark 2.3 If $\mathcal{F}$ contains only the trivial subgroup, the classifying space $E_{\mathcal{F}}G$ is the universal space for free actions, commonly denoted by EG. If $\mathcal{F}$ consists of all finite subgroups of G, the classifying space $E_{\mathcal{F}}G$ is called the classifying space for proper actions and it is denoted by $\underline{E}G$.

2.2 Bredon cohomology

The *orbit category* $\mathcal{O}_{\mathcal{F}}G$ is a category defined as follows: the objects are the left coset spaces G/H for all $H \in \mathcal{F}$ and the morphisms are all G–equivariant maps between the objects. Note that every morphism $\varphi\colon G/H \to G/P$ is completely determined by $\varphi(H)$, since $\varphi(xH) = x\varphi(H)$ for all $x \in G$. Moreover, there exists a morphism

$$G/H \to G/P : H \mapsto xP \text{ if and only if } x^{-1}Hx \leqslant P.$$

We denote the morphism $\varphi\colon G/H \to G/P : H \mapsto xP$ by $G/H \xrightarrow{x} G/P$ and note that it is determined by the inclusion $x^{-1}Hx \leqslant P$. Given $H, P \in \mathcal{F}$, we denote by $\hom_G(G/H, G/P)$ the set of morphisms from G/H to G/P.

An $\mathcal{O}_{\mathcal{F}}G$–*module* is a contravariant functor $M\colon \mathcal{O}_{\mathcal{F}}G \to \mathbb{Z}$–Mod. The *category of $\mathcal{O}_{\mathcal{F}}G$–modules*, denoted by Mod-$\mathcal{O}_{\mathcal{F}}G$, is the category whose objects are $\mathcal{O}_{\mathcal{F}}G$–modules and whose morphisms are natural transformations between these objects. The set of morphisms between $M, N \in$ Mod-$\mathcal{O}_{\mathcal{F}}G$ is denoted by $\mathrm{Hom}_{\mathcal{F}}(M, N)$.

A sequence

$$0 \to M' \to M \to M'' \to 0$$

in Mod-$\mathcal{O}_{\mathcal{F}}G$ is called *exact* if it is exact after evaluating in G/H for each $H \in \mathcal{F}$. For any $P \in \mathcal{F}$, the $\mathcal{O}_{\mathcal{F}}G$–module $\mathcal{A}_P = \mathbb{Z}[\hom_G(-, G/P)]$ is a free object in Mod-$\mathcal{O}_{\mathcal{F}}G$. A module $F \in$ Mod-$\mathcal{O}_{\mathcal{F}}G$ is free if and only if $F \cong \bigoplus_{\alpha \in I} \mathcal{A}_{P_\alpha}$ for some collection I of not necessarily distinct subgroups $P_\alpha \in \mathcal{F}$. We will say that F is *based* at the elements $P_\alpha \in \mathcal{F}$ for $\alpha \in I$.

The nth *Bredon cohomology group of G* with coefficients $M \in$ Mod-$\mathcal{O}_{\mathcal{F}}G$ is by definition

$$\mathrm{H}^n_{\mathcal{F}}(G, M) = \mathrm{Ext}^n_{\mathcal{O}_{\mathcal{F}}G}(\underline{\mathbb{Z}}, M),$$

where $\underline{\mathbb{Z}}$ is the functor that maps all objects to $\mathbb{Z}$ and all morphisms to the identity map. The *Bredon cohomological dimension of G* is defined to be

$$\mathrm{cd}_{\mathcal{F}} G = \sup\{n \in \mathbb{N} \mid \mathrm{H}^n_{\mathcal{F}}(G, M) \neq 0 \text{ for some } M \in \text{Mod-}\mathcal{O}_{\mathcal{F}}G\}.$$

Given a G–CW–complex X, an $\mathcal{O}_\mathcal{F}G$–module

$$C_n^\mathcal{F}(X)(-) \colon \mathcal{O}_\mathcal{F}G \to \mathbb{Z}\text{–Mod}$$

is defined as

$$C_n^\mathcal{F}(X)(G/H) = C_n(X^H),$$

where $C_*(-)$ denotes the cellular chains. Note that, in this way, the augmented cellular chain complex of any model for $E_\mathcal{F}G$ yields a free resolution of $\mathbb{Z}$ which can then be used to compute $H_\mathcal{F}^*(G, -)$. It follows that $\mathrm{cd}_\mathcal{F} G \leqslant \mathrm{gd}_\mathcal{F} G$.

We now consider the situation when G admits a cocompact model for $E_\mathcal{F}G$. In this case, Bredon cohomology commutes with arbitrary direct sums of coefficient modules (see eg [25, Proposition 5.2]) and one obtains the following proposition, which is standard (see eg [12, equation 2]).

Proposition 2.4 *Suppose that X is a cocompact model for $E_\mathcal{F}G$. Then*

$$\mathrm{cd}_\mathcal{F} G = \sup\{k \in \mathbb{N} \mid H_\mathcal{F}^k(X, \mathcal{A}_H) \neq 0 \text{ for some } H \in \mathcal{F}\}.$$

Below we derive a strengthening of Proposition 2.4, which is a key ingredient in the proof of Theorem 1.1.

Theorem 2.5 *Suppose that X is a cocompact model for $E_\mathcal{F}G$. Then*

$$\mathrm{cd}_\mathcal{F} G = \max\{k \in \mathbb{N} \mid H_\mathcal{F}^k(X, \mathcal{A}_H) \neq 0 \text{ for some cell stabiliser } H\}.$$

Moreover, if $H_\mathcal{F}^n(G; \mathcal{A}_L) \neq 0$ for $n = \mathrm{cd}_\mathcal{F} G$ and $L \in \mathcal{F}$, then there exists a cell stabiliser $H \leqslant L$ such that $H_\mathcal{F}^n(G; \mathcal{A}_H) \neq 0$.

Proof The chain complex $C_i^\mathcal{F}(X)$ forms a resolution of $\mathbb{Z}$ of finite length by finitely generated free $\mathcal{O}_\mathcal{F}G$–modules. Let $P = \ker\{C_{n-1}^\mathcal{F}(X) \to C_{n-2}^\mathcal{F}(X)\}$. Then P is projective and

$$0 \to P \to C_{n-1}^\mathcal{F}(X) \to \cdots \to C_0^\mathcal{F}(X) \to \mathbb{Z} \to 0$$

is exact. By applying the Bredon analogue of Schanuel's lemma [7, VIII.4.4] to the above two resolutions, it follows that there is a finitely generated free $\mathcal{O}_\mathcal{F}G$–module F based at stabilisers of the action of G on X such that $P \oplus F$ is a finitely generated free $\mathcal{O}_\mathcal{F}G$–module again based at stabilisers of the action of G on X. We can define the resolution $(D_*^\mathcal{F}, \partial_\mathcal{F})$ of $\mathbb{Z}$ by finitely generated free $\mathcal{O}_\mathcal{F}G$–modules

$$D_i^\mathcal{F} = \begin{cases} C_i^\mathcal{F}(X) & \text{if } i \leqslant n-2, \\ C_{n-1}^\mathcal{F}(X) \oplus F & \text{if } i = n-1, \\ P \oplus F & \text{if } i = n, \\ 0 & \text{if } i > n. \end{cases}$$

Since X is cocompact, by Proposition 2.4 there exists $L \in \mathcal{F}$ such that $H_\mathcal{F}^n(X, \mathcal{A}_L) \neq 0$. Then $H_\mathcal{F}^n(D_*^\mathcal{F}, \mathcal{A}_L) \neq 0$, which means that the coboundary map

$$\delta_\mathcal{F}^L \colon \mathrm{Hom}_\mathcal{F}(D_{n-1}^\mathcal{F}, \mathcal{A}_L) \to \mathrm{Hom}_\mathcal{F}(D_n^\mathcal{F}, \mathcal{A}_L)$$

is not onto. Rewriting this more explicitly using the Yoneda lemma,

$$\delta_{\mathcal{F}}^{L} \colon \sum_{i=1}^{k} \mathbb{Z}[\hom_{G}(G/G_{\tau_i}, G/L)] \to \sum_{j=1}^{l} \mathbb{Z}[\hom_{G}(G/G_{\sigma_j}, G/L)]$$

is not onto. This implies that there exists a stabiliser G_σ of some cell σ such that the generator $(G/G_\sigma \xrightarrow{x} G/L)$ of the nth cochain group is not in the image of $\delta_{\mathcal{F}}^{L}$. Let $H = x^{-1}G_\sigma x \leqslant L$. This inclusion induces an $\mathcal{O}_{\mathcal{F}}G$–module map $\mathcal{A}_H \to \mathcal{A}_L$ which in turn induces a map of cochain complexes

$$\Delta_* \colon \operatorname{Hom}_{\mathcal{F}}(D_*^{\mathcal{F}}, \mathcal{A}_H) \to \operatorname{Hom}_{\mathcal{F}}(D_*^{\mathcal{F}}, \mathcal{A}_L)$$

such that

$$\Delta_n(G/G_\sigma \xrightarrow{x} G/H) = (G/G_\sigma \xrightarrow{x} G/L).$$

By the commutativity $\delta_{\mathcal{F}}^{L} \circ \Delta_{n-1} = \Delta_n \circ \delta_{\mathcal{F}}^{H}$, we obtain that $(G/G_\sigma \xrightarrow{x} G/H)$ is not in the image of $\delta_{\mathcal{F}}^{H}$. Therefore,

$$\delta_{\mathcal{F}}^{H} \colon \operatorname{Hom}_{\mathcal{F}}(D_{n-1}^{\mathcal{F}}, \mathcal{A}_H) \to \operatorname{Hom}_{\mathcal{F}}(D_n^{\mathcal{F}}, \mathcal{A}_H)$$

is not onto, which shows that $H_{\mathcal{F}}^{n}(X, \mathcal{A}_H) = H_{\mathcal{F}}^{n}(D_*^{\mathcal{F}}, \mathcal{A}_H) \neq 0$. $\qquad\square$

Define a subset $\operatorname{isom}_G(G/L, G/S) \subseteq \hom_G(G/L, G/S)$ by

$$\operatorname{isom}_G(G/L, G/S) = \{\varphi \colon G/L \to G/S : L \mapsto xS \mid x^{-1}Lx = S\}.$$

Define an $\mathcal{O}_{\mathcal{F}}G$–module $\mathcal{B}_S$ by

$$\mathcal{B}_S(G/L) = \begin{cases} \mathbb{Z}[\operatorname{isom}_G(G/L, G/S)] & \text{if } L =_G S, \\ 0 & \text{if } L \neq_G S, \end{cases}$$

where $L =_G S$ means that L and S are conjugate in G. For each

$$(\varphi \colon G/L \xrightarrow{x} G/S) \in \operatorname{isom}_G(G/L, G/S),$$

we set

$$\mathcal{B}_S(\theta \colon G/H \xrightarrow{y} G/L)(\varphi) = \begin{cases} (\varphi \circ \theta \colon G/H \xrightarrow{yx} G/S) & \text{if } y^{-1}Hy = L, \\ 0 & \text{if } y^{-1}Hy \neq L, \end{cases}$$

which is an element in $\mathcal{B}_S(G/H)$. It is not difficult to check that $\mathcal{B}_S$ is well defined.

Corollary 2.6 *Suppose that X is a cocompact model for $E_{\mathcal{F}}G$ and that the collection of cell stabilisers is rigid. Then*

$$\operatorname{cd}_{\mathcal{F}} G = \max\{k \in \mathbb{N} \mid H_{\mathcal{F}}^{k}(X, \mathcal{B}_P) \neq 0 \text{ for some cell stabiliser } P\}.$$

Proof First, note that the cocompactness of X implies that the set of conjugacy classes of cell stabilisers is finite. By Theorem 2.5, there exists $P \in \mathcal{F}$ that is a stabiliser of a cell in X such that $H_{\mathcal{F}}^{n}(X, \mathcal{A}_P) \neq 0$ where $\operatorname{cd}_{\mathcal{F}} G = n$. By the rigidity of stabilisers and iteration of Theorem 2.5, we can assume that P does not contain a proper subgroup S such that $H_{\mathcal{F}}^{n}(X, \mathcal{A}_S) \neq 0$. Observe that also by the rigidity for $H =_G P$,

$$\hom_G(G/H, G/P) = \operatorname{isom}_G(G/H, G/P).$$

Again using rigidity, we can define an $\mathcal{O}_{\mathcal{F}}G$–submodule $\mathcal{C}_P$ of $\mathcal{A}_P$ by

$$\mathcal{C}_P(G/H) = \begin{cases} 0 & \text{if } H =_G P, \\ \mathbb{Z}[\hom_G(G/H, G/P)] & \text{if } H \neq_G P. \end{cases}$$

Considering the long exact sequence of the resulting short exact sequence

$$0 \to \mathcal{C}_P \to \mathcal{A}_P \to \mathcal{B}_P \to 0,$$

we either have $H^n_{\mathcal{F}}(X, \mathcal{C}_P) \neq 0$ or $H^n_{\mathcal{F}}(X, \mathcal{B}_P) \neq 0$. Considering a module that is a free cover of $\mathcal{C}_P$ consisting of free modules based at proper subgroups of P shows that if $H^n_{\mathcal{F}}(X, \mathcal{C}_P) \neq 0$, then $H^n_{\mathcal{F}}(X, \mathcal{A}_S) \neq 0$ for some $S \lneq P$, which violates the minimality assumption on P; hence $H^n_{\mathcal{F}}(X, \mathcal{B}_P) \neq 0$. $\qquad\square$

The Bredon cohomological and geometric dimensions for proper actions are denoted respectively by $\underline{\mathrm{cd}}\, G$ and $\underline{\mathrm{gd}}\, G$.

Corollary 2.7 *Let X be a G–CW–complex that is a cocompact model for $\underline{E}G$. Then*

$$\underline{\mathrm{cd}}\, G = \max\{k \in \mathbb{N} \mid H^k_c(X^H, X^H_{\mathrm{sing}}) \neq 0 \text{ for some cell stabiliser } H\},$$

where $X^H_{\mathrm{sing}} \subseteq X^H$ consists of all points whose stabiliser strictly contains H.

Proof The claim follows immediately from combining Corollary 2.6 and [12, Theorem 2.4]. $\qquad\square$

3 Simple complexes of groups

3.1 Simple complexes of groups and the Basic Construction

Throughout, let $\mathcal{Q}$ be a finite poset. We denote by $|\mathcal{Q}|$ the *geometric realisation* of $\mathcal{Q}$, ie a simplicial complex whose simplices are chains of elements of $\mathcal{Q}$.

Definition 3.1 (simple complex of groups) A *simple complex of groups $G(\mathcal{Q})$ over $\mathcal{Q}$* consists of the following data:

- for any $J \in Q$ there is a group P_J called a *local group at J*,
- for any two elements $J \leqslant T$ in $\mathcal{Q}$ there is a monomorphism

$$\phi_{TJ} \colon P_J \to P_T$$

  such that if $J \leqslant T \leqslant U$ then

$$\phi_{UT} \circ \phi_{TJ} = \phi_{UJ}.$$

Definition 3.2 (simple morphism) Let $G(\mathcal{Q})$ be a simple complex of groups and let G be a group. A *simple morphism* $\psi \colon G(\mathcal{Q}) \to G$ is a collection of maps $\psi_J \colon P_J \to G$ satisfying

$$\psi_T \circ \phi_{TJ} = \psi_J$$

for all pairs $J \leqslant T$ in $\mathcal{Q}$. We say that $\psi \colon G(\mathcal{Q}) \to G$ is *injective on local groups* if for every $J \in \mathcal{Q}$ the map $\psi_J \colon P_J \to G$ is injective.

Given a simple complex of groups $G(\mathcal{Q})$, the *fundamental group* $\widehat{G(\mathcal{Q})}$ of $G(\mathcal{Q})$ is the direct limit of the resulting direct system of groups

$$\widehat{G(\mathcal{Q})} = \varinjlim_{J \in \mathcal{Q}} P_J.$$

Note that by the universal property of $\widehat{G(\mathcal{Q})}$ there exists a canonical simple morphism $\iota \colon G(\mathcal{Q}) \to \widehat{G(\mathcal{Q})}$ such that for every $J \in \mathcal{Q}$ the map $\iota_J \colon P_J \to G(\mathcal{Q})$ is the canonical map to the limit.

Definition 3.3 (strict developability) We say that a simple complex of groups $G(\mathcal{Q})$ is *strictly developable* if the canonical simple morphism $\iota \colon G(\mathcal{Q}) \to \widehat{G(\mathcal{Q})}$ is injective on local groups.

Note that the strict developability is equivalent to the existence of a simple morphism $\psi \colon G(\mathcal{Q}) \to G$ that is injective on local groups, where G is some group.

Convention 3.4 If $\psi \colon G(\mathcal{Q}) \to G$ is a simple morphism that is injective on local groups then for any $J \in \mathcal{Q}$ we identify the group P_J with its image $\psi(P_J) \leqslant G$.

Definition 3.5 (panel complex) A *panel complex* $(X, \{X_J\}_{J \in \mathcal{Q}})$ over $\mathcal{Q}$ is a compact polyhedron X together with family of subpolyhedra $\{X_J\}_{J \in \mathcal{Q}}$ called *panels* such that

- X is the union of all the panels,
- $X_T \subseteq X_J$ if and only if $J \leqslant T$,
- for any two panels their intersection is either a union of panels or empty.

Definition 3.6 (standard panel complex) Define the panel complex K over $\mathcal{Q}$ as follows. Let $K = |\mathcal{Q}|$ and for $J \in \mathcal{Q}$ let $K_J = |\mathcal{Q}_{\geqslant J}|$ where $\mathcal{Q}_{\geqslant J}$ denotes the subposet of $\mathcal{Q}$ consisting of all the elements greater than or equal to J.

Definition 3.7 (Basic Construction) Suppose that

- $G(\mathcal{Q})$ is a strictly developable complex of groups,
- X is a panel complex over $\mathcal{Q}$,
- $\psi \colon G(\mathcal{Q}) \to G$ is a simple morphism to a group G that is injective on local groups (thus for any $J \in \mathcal{Q}$ we identify P_J with $\psi(P_J)$).

For a point $x \in X$ let $J(x) \in Q$ be such that the panel $X_{J(x)}$ is the intersection of all the panels containing x. Define the Basic Construction $D(X, G(\mathcal{Q}), \psi)$ as

$$D(X, G(\mathcal{Q}), \psi) = G \times X/\sim,$$

where $(g_1, x_1) \sim (g_2, x_2)$ if and only if $x_1 = x_2$ and $g_1^{-1} g_2 \in P_{J(x_1)}$. Let $[g, x]$ denote the equivalence class of (g, x).

The group G acts on $D(X, G(\mathcal{Q}), \psi)$ by $g \cdot [g', x] = [gg', x]$. It is easy to see that $D(X, G(\mathcal{Q}), \psi)$ has the structure of a polyhedral complex and that the G–action preserves that structure. The stabilisers of this action are the conjugates of local groups P_J and the quotient is homeomorphic to

$$X \cong [e, X] \subset D(X, G(\mathcal{Q}), \psi).$$

Moreover, $X \cong [e, X]$ is a so-called *strict fundamental domain* for the G–action in the sense that it is a closed subset of $D(X, G(\mathcal{Q}))$ intersecting every orbit in precisely one point.

In fact, any admissible action with a strict fundamental domain arises in the way described above.

Theorem 3.8 [5, Proposition II.12.20] *Suppose a group G acts admissibly on a connected polyhedral complex X with a strict fundamental domain $Y \subset X$.*

Then there is a strictly developable simple complex of groups $G(\mathcal{Q})$, where $\mathcal{Q}$ is the poset of cells of Y (ordered by the reverse inclusion) and where the local group at cell $e \subset Y$ is its G–stabiliser. The inclusion of cell stabilisers into G defines a simple morphism $\psi \colon G(\mathcal{Q}) \to G$ such that X is G–equivariantly homeomorphic to the Basic Construction $D(K, G(\mathcal{Q}), \psi)$, where K is the standard panel complex associated to $\mathcal{Q}$. Moreover, if X is simply connected then G is isomorphic to the fundamental group of $G(\mathcal{Q})$.

Convention 3.9 In the case when G is isomorphic to the fundamental group of $G(\mathcal{Q})$ and the simple morphism $G(\mathcal{Q}) \to G$ is the canonical simple morphism ι, we will omit the morphism from the notation and simply write $D(X, G(\mathcal{Q}))$ for the associated Basic Construction (where X is a panel complex over $\mathcal{Q}$).

3.2 Thinning procedure

Definition 3.10 We say that a simple complex of groups $G(Q)$ is *thin* if for any pair $J \leqslant T$ in $\mathcal{Q}$, the monomorphism $\phi_{TJ} \colon P_J \to P_T$ is an isomorphism if and only if $J = T$.

Remark 3.11 In [12; 26], the assumption that a simple complex of groups is thin is a part of its definition.

Below we describe a procedure of thinning, which, given a strictly developable simple complex of groups $G(Q)$, results in a thin complex $G(\mathcal{R})$ together with a morphism of simple complexes of groups $G(\mathcal{Q}) \to G(\mathcal{R})$ inducing an isomorphism of fundamental groups.

Definition 3.12 (block poset) Given a simple complex of groups $G(\mathcal{Q})$ with the collection of local groups $\{P_J\}_{J \in \mathcal{Q}}$, let $\sim$ be an equivalence relation on Q generated by

$$J \sim J' \text{ if } J \leqslant J' \text{ and } \phi_{J'J} \colon P_J \to P_{J'} \text{ is an isomorphism.}$$

An equivalence class C of elements of $\mathcal{Q}$ under relation $\sim$ is called a *block*. There is a partial order on the set of blocks given by

$$C \leqslant C' \text{ if and only if there exist } J \in C \text{ and } J' \in C' \text{ with } J \leqslant J'.$$

Denote the associated poset by $\mathcal{R}$ and call it the *block poset*.

Note that there is a surjection of posets $\pi \colon \mathcal{Q} \to \mathcal{R}$ given by $J \in C \mapsto C$.

Definition 3.13 (thinning of a simple complex of groups) Let $G(\mathcal{Q})$ be a strictly developable simple complex of groups with the collection of local groups $\{P_J\}_{J \in \mathcal{Q}}$ and the fundamental group G. Let $\mathcal{R}$ be the block poset associated to $G(\mathcal{Q})$.

Define a simple complex of groups $G(\mathcal{R}) = (\{S_C\}_{C \in \mathcal{R}}, \{\psi_{C'C}\}_{C' \leqslant C \in \mathcal{R}})$ as follows. For a block $C \in R$, let $J \in \mathcal{Q}$ be any element in the preimage $\pi^{-1}(C)$ and set $S_C = P_J$. Observe that S_C is well defined, since for all $J' \in \pi^{-1}(C)$ groups $P_{J'}$ are identified as a single subgroup of G. Now given two blocks $C \leqslant C'$ define the map

$$\psi_{C'C} \colon S_C \to S_{C'}$$

as the inclusion of the corresponding groups $P_J \leqslant P_{J'}$ seen as subgroups of G. Note that $G(\mathcal{R})$ is thin by construction.

One easily verifies that $G(\mathcal{R})$ is strictly developable with fundamental group isomorphic to G. Moreover, the surjection $\pi \colon \mathcal{Q} \to \mathcal{R}$ induces a morphism of simple complexes of groups $G(\mathcal{Q}) \to G(\mathcal{R})$ which in turn induces an isomorphism on the fundamental groups (see [5, Chapter II.12] for a background on morphisms of simple complexes of groups). Finally, if $G(\mathcal{Q})$ is thin, then by definition $\mathcal{R}$ is isomorphic to $\mathcal{Q}$, and the morphism $G(\mathcal{Q}) \to G(\mathcal{R})$ is an isomorphism.

3.3 Bestvina complex

Definition 3.14 Let $(X, \{X_J\}_{J \in \mathcal{Q}})$ be a panel complex over a poset $\mathcal{Q}$. For an element $J \in \mathcal{Q}$ define the subcomplex $X_{>J}$ of X by

$$X_{>J} = \bigcup_{J < J'} X_{J'}.$$

Remark 3.15 In the case where $X = K$ is the standard panel complex over $\mathcal{Q}$,

$$K_{>J} = \big| \{J' \in \mathcal{Q} \mid J' > J\} \big|.$$

Observe that Theorem 3.8 may be seen as evidence that the standard panel complex and the associated Basic Construction occur naturally. However, for computational purposes, a better suited panel complex is the following.

Definition 3.16 (Bestvina complex) The *Bestvina panel complex* $(B, \{B_J\}_{J \in \mathcal{Q}})$ is defined as follows. For every maximal element $J \in \mathcal{Q}$, define B_J to be a point. Now given an element $J \in \mathcal{Q}$ assume that for all J' with $J < J'$ the panel $B_{J'}$ has already been defined. Define B_J to be the compact contractible polyhedron containing $B_{>J} = \bigcup_{J < J'} B_{J'}$ of the smallest possible dimension.

We define $B^{\mathbb{Z}}$ in the same way as B except that we replace "contractible" by "acyclic" polyhedra.

$\mathcal{Q}$ and $G(\mathcal{Q})$

$\mathcal{R}$ and $G(\mathcal{R})$

Bestvina complex for $\mathcal{R}$
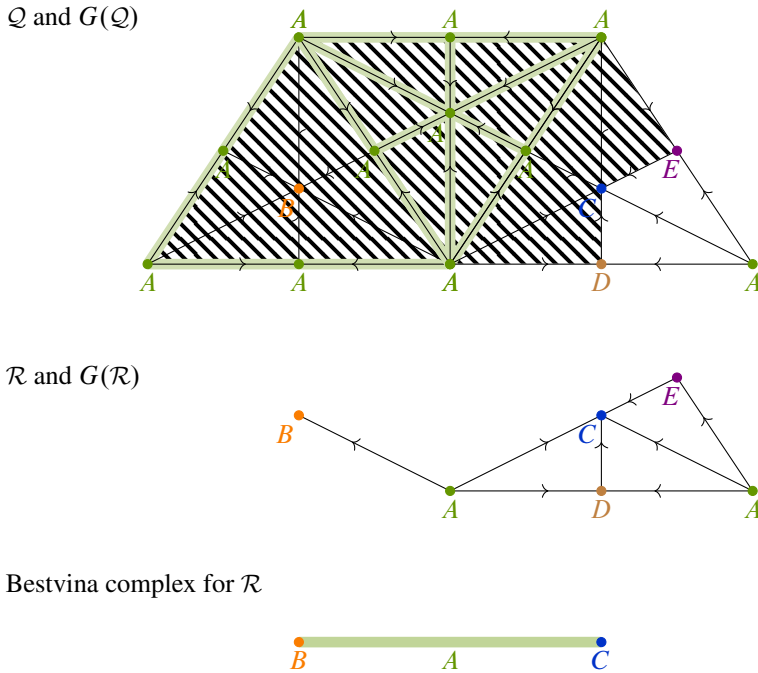
Figure 1: Complex of groups $G(\mathcal{Q})$ together with its thinning $G(\mathcal{R})$ and the Bestvina complex associated to $\mathcal{R}$. Elements of a block $[A]_1 \subset \mathcal{Q}$ with the local group A are connected by green lines. The geometric realisation $|\mathcal{Q}_{\geqslant [A]_1}|$ is in yellow.

Remark 3.17 The panel complex B was introduced by Bestvina in [4] for the poset of special subgroups of a finitely generated Coxeter group. It was extended to graph products of finite groups by Harlander and Meinert in [20] and more generally to buildings that admit a chamber transitive action of a discrete group by Harlander in [19].

Example 3.18 Consider finite groups A, B, and C with two inclusions $A \leqslant B$ and $A \leqslant C$. Consider two subgroups E and D of C, both containing the image of $A \leqslant C$. All inclusions are assumed to be proper. Figure 1 depicts a complex of groups $G(\mathcal{Q})$ (where all the structure maps are the respective inclusions), its thinning $G(\mathcal{R})$ and the Bestvina complex associated to $\mathcal{R}$. The fundamental group of $G(\mathcal{Q})$ (and hence of $G(\mathcal{R})$) is isomorphic to the amalgamated product $B *_A C$. Observe that poset $\mathcal{R}$ has significantly fewer elements than $\mathcal{Q}$. A further simplification is given by the Bestvina complex, whose dimension is lower than the dimension of $|\mathcal{Q}|$ and $|\mathcal{R}|$. The Basic Construction $D(B, G(\mathcal{R}))$ is isomorphic to the Bass–Serre tree of $B *_A C$.

The proof of the following proposition follows directly from [26, Lemmas 2.4 and 2.5].

Proposition 3.19 *Let $G(\mathcal{Q})$ be a strictly developable simple complex of groups and let $\psi \colon G(\mathcal{Q}) \to G$ be a simple morphism that is injective on local groups. Assume that $G(\mathcal{Q})$ is thin. Then:*

(1) *The standard development $D(K, G(\mathcal{Q}), \psi)$ and the Bestvina complex $D(B, G(\mathcal{Q}), \psi)$ are G–homotopy equivalent.*

(2) *The Bredon chain complexes $C_*^{\mathcal{F}}\big(D(K, G(\mathcal{Q}), \psi)\big)$ and $C_*^{\mathcal{F}}\big(D(B^{\mathbb{Z}}, G(\mathcal{Q}), \psi)\big)$ are chain homotopy equivalent.*

Definition 3.20 (local cohomological dimension) For a poset $\mathcal{Q}$, define its *local cohomological dimension* lcd $\mathcal{Q}$ as

$$\operatorname{lcd} \mathcal{Q} = \max\{n \in \mathbb{N} \mid \tilde{H}^{n-1}(K_{>J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}.$$

Proposition 3.21 *We have the equalities*

$$\operatorname{lcd} \mathcal{Q} = \max\{n \in \mathbb{N} \mid H^n(K_J, K_{>J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}$$
$$= \max\{n \in \mathbb{N} \mid \tilde{H}^{n-1}(K_{>J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}$$
$$= \max\{n \in \mathbb{N} \mid \tilde{H}^{n-1}(B_{>J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}$$
$$= \dim(B^{\mathbb{Z}})$$
$$= \max\{n \in \mathbb{N} \mid H^n(B_J^{\mathbb{Z}}, B_{>J}^{\mathbb{Z}}) \neq 0 \text{ for some } J \in \mathcal{Q}\}.$$

Moreover,

$$\dim(B) = \begin{cases} \operatorname{lcd} \mathcal{Q} & \text{if } d \neq 2, \\ 2 \text{ or } 3 & \text{if } d = 2. \end{cases}$$

Proof The proof is essentially the same as the proof of [26, Proposition 3.4]. □

Lemma 3.22 *Let $G(\mathcal{Q})$ be a strictly developable simple complex of groups with fundamental group G and let $\mathcal{F}$ be the family generated by local groups. Suppose $D(K, G(\mathcal{Q}))$ is a model for $E_{\mathcal{F}}G$. Let $\mathcal{R}$ be the corresponding block poset. Then $\operatorname{cd}_{\mathcal{F}} G \leq \operatorname{lcd} \mathcal{R}$. In particular, if $G(\mathcal{Q})$ is thin then $\operatorname{cd}_{\mathcal{F}} G \leq \operatorname{lcd} \mathcal{Q}$.*

Proof Consider the composition of chain maps

$$C_*^{\mathcal{F}}\big(D(K, G(\mathcal{Q}))\big) \to C_*^{\mathcal{F}}\big(D(T, G(\mathcal{R}))\big) \to C_*^{\mathcal{F}}\big(D(B^{\mathbb{Z}}, G(\mathcal{R}))\big),$$

where $K = |\mathcal{Q}|$, $T = |\mathcal{R}|$, and the complex $B^{\mathbb{Z}}$ is taken over the poset $\mathcal{R}$.

The first map is induced by the map of Basic Constructions $D(K, G(\mathcal{Q})) \to D(T, G(\mathcal{R}))$, which is in turn induced by a morphism of simple complexes of groups $G(\mathcal{Q}) \to G(\mathcal{R})$. The second map is constructed in [26, Theorem A.1] (it is straightforward to check that both the statement and the proof of [26, Theorem A.1] carry through for infinite local groups).

Since $D(K, G(\mathcal{Q}))$ is a model for $E_{\mathcal{F}}G$, there is also a classifying G–map that gives a chain map

$$C_*^{\mathcal{F}}\big(D(B^{\mathbb{Z}}, G(\mathcal{R}))\big) \to C_*^{\mathcal{F}}\big(D(K, G(\mathcal{Q}))\big)$$

and the composition of both is chain homotopic to the identity on $C_*^{\mathcal{F}}\big(D(K, G(\mathcal{Q}))\big)$. This shows that $\operatorname{cd}_{\mathcal{F}} G \leq \dim\big(D(B^{\mathbb{Z}}, G(\mathcal{R}))\big) = \dim(B^{\mathbb{Z}})$ and Proposition 3.21 finishes the proof. □

4 Thin complexes of groups

In this section, we will assume that the simple complex of groups $G(\mathcal{Q})$ is thin. We will show that the Bredon cohomological dimension of the fundamental group of $G(\mathcal{Q})$ is equal to the local cohomological dimension of the poset $\mathcal{Q}$.

Proposition 4.1 *Let $G(\mathcal{Q})$ be a thin simple complex of groups, let G be a group, and let $\psi\colon G(\mathcal{Q}) \to G$ be a simple morphism which is injective on local groups. Suppose $(Y, \{Y_J\}_{J\in\mathcal{Q}})$ is a simplicial panel complex over $\mathcal{Q}$, and let $X = D(Y, G(\mathcal{Q}), \psi)$ be the associated Basic Construction. Then for any $J \in \mathcal{Q}$, there is an epimorphism of cochain complexes*

$$\Psi\colon (C_{\mathcal{F}}^*(X; \mathcal{A}_{P_J}), \delta_{\mathcal{F}}) \twoheadrightarrow (C^*(Y_J, Y_{>J}), \delta_J),$$

where P_J is a local group at $J \in \mathcal{Q}$, seen as a subgroup of G.

Proof Fix a dimension i and identify Y with a subcomplex of X. Let $\sigma_j \subset Y$, $j = 1, \ldots, k$ be the i–simplices of Y and denote by $G_{\sigma_j} \in \mathcal{F}$ the stabiliser of σ_j. Then by the Yoneda lemma, we obtain a natural equivalence

$$C_{\mathcal{F}}^i(X; \mathcal{A}_{P_J}) = \operatorname{Hom}_{\mathcal{F}}\left(\bigoplus_{j=1}^k \mathcal{A}_{G_{\sigma_j}}, \mathcal{A}_{P_J}\right) \cong \bigoplus_{j=1}^k \mathbb{Z}[\operatorname{hom}_G(G/G_{\sigma_j}, G/P_J)].$$

Given an i–simplex $\sigma \subset Y$ and a morphism $\varphi\colon G/G_\sigma \xrightarrow{x} G/P_J : G_\sigma \mapsto xP_J$ in the summand indexed by σ, we define

$$\Psi(\varphi) = \begin{cases} c^\sigma & \text{if } \sigma \subset Y_J, G_\sigma = P_J \text{ and } x \in P_J, \quad \text{(type I)} \\ 0 & \text{otherwise}, \quad \text{(type II)} \end{cases}$$

where $c^\sigma \in C^i(Y_J, Y_{>J})$ equals to 1 on σ and vanishes everywhere else.

We claim that Ψ is surjective. To see this, it is enough to note that if $\sigma \subset Y_J$ is an i–simplex with stabiliser $G_\sigma \gneqq P_J$, then by the definition of Basic Construction $\sigma \subset Y_{>J}$.

It is left to check that Ψ commutes with the coboundary map. First, suppose φ is of type II. Then $\delta_J(\Psi(\varphi)) = 0$. On the other hand, $\delta_{\mathcal{F}}(\varphi)$ is a chain based at morphisms which are precomposed with φ and hence of type II. To see this, suppose

$$\phi\colon G/G_\tau \xrightarrow{y^{-1}} G/G_\sigma \xrightarrow{x} G/P_J$$

is such a composition and it is of type I where $\tau \subset Y$ is an $(i{+}1)$–simplex such that $y\tau$ contains σ as a face. Since Y is a strict fundamental domain, observe that $y \in G_\sigma$.

Since ϕ is of type I, we must have $G_\tau = P_J$ and $y^{-1}x \in P_J$, which implies that $G_\sigma = P_J$. Since now $x \in P_J$, this shows that φ is of type I, which is a contradiction. Therefore, $\Psi(\delta_{\mathcal{F}}(\varphi)) = 0$.

Now, suppose φ is of type I, ie $\varphi = \varphi_\sigma \colon G/P_J \xrightarrow{1} G/P_J$ with P_J the stabiliser of σ. Then

$$\tag{4-1} \delta_J(\Psi(\varphi_\sigma)) = \delta_J(c^\sigma) = \sum_{t=1}^{l} (-1)^{\mathrm{sgn}(\tau_t)} c^{\tau_t},$$

where $\tau_t \subset Y_J$ contains σ as a face. On the other hand,

$$\tag{4-2} \Psi(\delta_{\mathcal{F}}(\varphi_\sigma)) = \sum_{s=1}^{r} (-1)^{\mathrm{sgn}(y_s \tau_s)} \Psi(\varphi_{y_s \tau_s})$$

where $y_s \in G$, $\tau_s \subset Y$, $y_s \tau_s$ is an $(i+1)$–simplex containing σ as a face, and $\varphi_{y_s \tau_s} \colon G/G_{\tau_s} \xrightarrow{y_s^{-1}} G/P_J$. Since Y is a strict fundamental domain, $y_s^{-1}\sigma = \sigma$ and hence $y_s \in P_J$. Note that if $\Psi(\varphi_{y_s \tau_s}) \neq 0$, then by definition of Ψ, we have $y_s \tau_s \subset Y_J$ and $G_{\tau_s} = P_J$. Therefore, $\tau_s = y_s \tau_s \subset Y_J$ and $\varphi_{y_s \tau_s} = \varphi_{\tau_s}$. In this case, $\Psi(\varphi_{\tau_s}) = c^{\tau_s}$. The claim now follows from equating (4-1) and (4-2). $\square$

Proposition 4.2 If $D(K, G(\mathcal{Q}), \psi)$ is a model for $E_{\mathcal{F}}G$, then $\mathrm{cd}_{\mathcal{F}} G = \mathrm{lcd}\,\mathcal{Q}$.

Proof Note that by the assumption $D(K, G(\mathcal{Q}), \psi)$ is simply connected, and thus by Theorem 3.8, G is necessarily isomorphic to the fundamental group of $G(\mathcal{Q})$. Consider the panel complex $B^{\mathbb{Z}}$ given in Definition 3.16. By passing to a barycentric subdivision we can assume that $B^{\mathbb{Z}}$ is a simplicial panel complex. Let $X = D(B^{\mathbb{Z}}, G(\mathcal{Q}), \psi)$. By Proposition 3.19(2), $C_*^{\mathcal{F}}\big(D(K, G(\mathcal{Q}), \psi)\big)$ and $C_*^{\mathcal{F}}(X)$ are chain homotopy equivalent and thus the latter can be used to compute $\mathrm{H}_{\mathcal{F}}^n(G, -)$.

Now Proposition 3.21 implies that there exists $J \in \mathcal{Q}$ such that

$$H^{\mathrm{lcd}\,\mathcal{Q}}(B_J^{\mathbb{Z}}, B_{>J}^{\mathbb{Z}}) \neq 0.$$

Since $C_{\mathcal{F}}^i(X; \mathcal{A}_{P_J}) = 0$ for $i > \mathrm{lcd}\,\mathcal{Q}$, by Proposition 4.1, Ψ induces an epimorphism

$$\Psi^* \colon H_{\mathcal{F}}^{\mathrm{lcd}\,\mathcal{Q}}(X; \mathcal{A}_{P_J}) \to H^{\mathrm{lcd}\,\mathcal{Q}}(B_J^{\mathbb{Z}}, B_{>J}^{\mathbb{Z}}).$$

This shows that $H_{\mathcal{F}}^{\mathrm{lcd}\,\mathcal{Q}}(X; \mathcal{A}_{P_J}) \neq 0$ and hence, by Lemma 3.22, we obtain $\mathrm{cd}_{\mathcal{F}} G = \mathrm{lcd}\,\mathcal{Q}$. $\square$

5 Cohomology of simple complexes of groups

Let $G(\mathcal{Q})$ be a simple complex of groups and let $\psi \colon G(\mathcal{Q}) \to G$ be a simple morphism which is injective on local groups. Recall that by Convention 3.4, for any $J \in \mathcal{Q}$ we identify P_J with $\psi(P_J) \leqslant G$.

For $J \in \mathcal{Q}$ let $\mathcal{I}_J$ be a complete set of representatives of the set

$$\{g \in G \mid g^{-1} P_J g = P_U \text{ for some } U \in \mathcal{Q}\}/P_J,$$

where P_J acts by left multiplication.

Suppose $\Omega \subseteq \mathcal{Q}$ is a subset such that $P_U = P_{U'}$ for all $U, U' \in \Omega$. Define subcomplexes K_Ω and $K_{>\Omega}$ of K to be

$$K_\Omega = \big|\{V \in \mathcal{Q} \mid V \geqslant U \text{ for some } U \in \Omega\}\big|, \quad K_{>\Omega} = \big|\{V \in \mathcal{Q} \mid V \geqslant U \text{ for some } U \in \Omega \text{ and } P_V \gneqq P_U\}\big|.$$

For $J \in \mathcal{Q}$ and $g \in G$ define

$$\Omega_J^g = \{U \in \mathcal{Q} \mid P_U = g^{-1} P_J g\}.$$

Proposition 5.1 *Suppose that $G(\mathcal{Q})$ is a strictly developable simple complex of groups. Let $\psi : G(\mathcal{Q}) \to G$ be a simple morphism which is injective on local groups and let $X = D(K, G(\mathcal{Q}), \psi)$ be the associated Basic Construction. Then for any $J \in \mathcal{Q}$, there is an isomorphism of cochain complexes*

$$\Phi : (C_{\mathcal{F}}^*(X; \mathcal{B}_{P_J}), \delta_{\mathcal{F}}) \to \bigoplus_{g \in \mathcal{I}_J} (C^*(K_{\Omega_J^g}, K_{>\Omega_J^g}), \delta).$$

Proof We define $\Phi = \bigoplus_{g \in \mathcal{I}_J} \Psi_g$ with each

$$\Psi_g : C_{\mathcal{F}}^*(X; \mathcal{B}_{P_J}) \to C^*(K_{\Omega_J^g}, K_{>\Omega_J^g})$$

constructed analogously to the map Ψ of Proposition 4.1 where one replaces an arbitrary simplicial panel complex Y with K. Namely, we identify

$$C_{\mathcal{F}}^i(X; \mathcal{B}_{P_J}) = \mathrm{Hom}_{\mathcal{F}}\left(\bigoplus_{\sigma \subset K^{(i)}} \mathcal{A}_{G_\sigma}, \mathcal{B}_{P_J}\right) \cong \bigoplus_{\sigma \subset K^{(i)}} \mathcal{B}_{P_J}(G_\sigma) \cong \bigoplus_{j=1}^{k} \mathbb{Z}[\mathrm{isom}_G(G/G_{\sigma_j}, G/P_J)],$$

where the σ_j are all the i–simplices of K such that $G_{\sigma_j} =_G P_J$.

Now, fix $g \in \mathcal{I}_J$ and suppose σ is an i–simplex with stabiliser $g^{-1} P_J g = P_J^g$. Given an (iso)morphism

$$(\varphi : G/G_\sigma \xrightarrow{x} G/P_J : G_\sigma \mapsto xP_J) \in C_{\mathcal{F}}^i(X; \mathcal{B}_{P_J}),$$

we define

$$\Psi_g(\varphi) = \begin{cases} c^\sigma & \text{if } G_\sigma = P_J^g \text{ and } x \in g^{-1} P_J, \\ 0 & \text{otherwise,} \end{cases}$$

where $c^\sigma \in C^i(K_{\Omega_J^g}, K_{>\Omega_J^g})$ equals to 1 on $\sigma \subset K_{\Omega_J^g}$ and vanishes everywhere else. The proof that Ψ_g commutes with the coboundary maps is analogous to the corresponding argument in the proof of Proposition 4.1 and hence it is omitted. (Alternatively, it also follows from the commutativity of the coboundary maps with sections Δ_g defined below.)

To show that Φ is an isomorphism, we first define a section

$$\Delta_g : C^*(K_{\Omega_J^g}, K_{>\Omega_J^g}) \to C_{\mathcal{F}}^*(X; \mathcal{B}_{P_J}) : c^\sigma \mapsto (\varphi_\sigma : G/P_J^g \xrightarrow{g^{-1}} G/P_J)$$

to each Ψ_g. We need to show that it commutes with the coboundary maps. We have

(5-1) $$\delta_{\mathcal{F}}(\Delta_g(c^\sigma)) = \delta_{\mathcal{F}}(\varphi_\sigma) = \sum_{s=1}^{r} (-1)^{\mathrm{sgn}(y_s \tau_s)} \varphi_{y_s \tau_s}$$

where $\tau_s \subset K$ contains σ as a face and $\varphi_{y_s \tau_s} : G/G_{\tau_s} \xrightarrow{y_s^{-1}} G/P_J^g \xrightarrow{g^{-1}} G/P_J$ with $y_s G_{\tau_s} y_s^{-1} \leqslant P_J^g$ and $y_s \in P_J^g$. Note that if $0 \neq \varphi_{y_s \tau_s} \in C_{\mathcal{F}}^*(X; \mathcal{B}_{P_J})$, then by definition of $\mathcal{B}_{P_J}$, the subgroup G_{τ_s} must be conjugate to P_J and $G_{\tau_s} = P_J^g$. Therefore, $\tau_s = y_s \tau_s \subset K$ and $\varphi_{y_s \tau_s} = \varphi_{\tau_s} \in \mathrm{Im}\, \Delta_g$.

On the other hand,

$$(5\text{-}2) \qquad \Delta_g(\delta_J(c^\sigma)) = \sum_{t=1}^{l} (-1)^{\mathrm{sgn}(\tau_t)} \Delta_g(c^{\tau_t}) = \sum_{t=1}^{l} (-1)^{\mathrm{sgn}(\tau_t)} \varphi_{\tau_t}$$

where $\tau_t \subset K_{\Omega_J^g}$ contains σ as a face. The claim now follows from equating (5-1) and (5-2). It is straightforward to check that Φ and $\Delta = \bigoplus_{g \in \mathcal{I}_J} \Delta_g$ are inverses of each other. $\qquad \square$

Remark 5.2 Proposition 5.1 can be generalised to hold for an arbitrary simplicial panel complex $(X, \{X_J\}_{J \in \mathcal{Q}})$, where one defines

$$X_\Omega = \bigcup_{J \in \Omega} X_J, \qquad X_{>\Omega} = \bigcup_{\{U \in \mathcal{Q} \mid U \geq J \text{ for some } J \in \Omega \text{ and } P_U \gneqq P_J\}} X_U,$$

though this is not necessary for our purposes.

6 Main theorems

In this section we state and prove slightly more general versions of Theorems 1.1 and 1.6 from the introduction. The generalisation concerns the computation of Bredon cohomology of the Basic Construction $D(K, G(\mathcal{Q}), \psi)$. In the statements below, we allow $\psi \colon G(\mathcal{Q}) \to G$ to be a simple morphism to an arbitrary group G, not necessarily the fundamental group of $G(\mathcal{Q})$.

Theorem 6.1 *Let $G(\mathcal{Q})$ be a strictly developable simple complex of groups and let $\psi \colon G(\mathcal{Q}) \to G$ be a simple morphism that is injective on local groups. Let $\mathcal{F}$ be the family of subgroups of G generated by local groups. Let $X = D(K, G(\mathcal{Q}), \psi)$ be the associated Basic Construction. For $J \in \mathcal{Q}$ we then have*

$$(6\text{-}1) \qquad H_{\mathcal{F}}^*(X; \mathcal{B}_{P_J}) \cong \bigoplus_{g \in \mathcal{I}_J} \bigoplus_{C_J^g \subseteq \Omega_J^g} H^*(K_{C_J^g}, K_{>C_J^g}),$$

where $C_J^g \subseteq \Omega_J^g$ denotes a block in Ω_J^g.

If $G(\mathcal{Q})$ is rigid and X is a model for $E_{\mathcal{F}} G$, then

$$(6\text{-}2) \qquad \mathrm{cd}_{\mathcal{F}} G = \max\{n \in \mathbb{N} \mid H^n(K_C, K_{>C}) \neq 0 \text{ for some block } C \subseteq \mathcal{Q}\}.$$

Proof First we prove (6-1). To do this we show that for every $J \in \mathcal{Q}$, $g \in \mathcal{I}_J$ and for any integer $n \geq 0$,

$$(6\text{-}3) \qquad H^n(K_{\Omega_J^g}, K_{>\Omega_J^g}) \cong \bigoplus_{C_J^g \subseteq \Omega_J^g} H^n(K_{C_J^g}, K_{>C_J^g}).$$

To show (6-3), we proceed by induction on the number of blocks $C \subseteq \Omega_J^g$. If Ω_J^g contains only one block then (6-3) is clearly satisfied. Assume now that Ω_J^g contains more than one block. Let $C \subseteq \Omega_J^g$, let $R = \Omega_J^g \smallsetminus C$ and write the pair $(K_{\Omega_J^g}, K_{>\Omega_J^g})$ as

$$(K_{\Omega_J^g}, K_{>\Omega_J^g}) = (K_R \cup K_C, K_{>R} \cup K_{>C}).$$

Consider the relative Mayer–Vietoris sequence for the above pair,

$$H^{n-1}(K_C \cap K_R, K_{>C} \cap K_{>R}) \to H^n(K_{\Omega_J^g}, K_{>\Omega_J^g}) \to H^n(K_C, K_{>C}) \oplus H^n(K_R, K_{>R})$$
$$\to H^n(K_C \cap K_R, K_{>C} \cap K_{>R}).$$

Claim $$K_C \cap K_R = K_{>C} \cap K_{>R}.$$

To prove the claim consider an element $V \in K_C \cap K_R$ (ie we view $V \in \mathcal{Q}$ as a vertex of K). Thus $U \leqslant V$ and $U' \leqslant V$ for some $U \in C$ and $U' \in R$. If $V \notin K_{>C} \cap K_{>R}$, then $P_V = P_U$ or $P_V = P_{U'}$. In either case we get $P_V = g^{-1}P_J g$, which implies that $V \in C$ and $V \in R$. This is a contradiction and the claim follows.

The claim implies that $H^n(K_C \cap K_R, K_{>C} \cap K_{>R}) = 0$ for every $n \geqslant 0$ and therefore the map

$$H^n(K_{\Omega_J^g}, K_{>\Omega_J^g}) \to H^n(K_C, K_{>C}) \oplus H^n(K_R, K_{>R})$$

is an isomorphism. Since by the inductive assumption we have

$$H^n(K_R, K_{>R}) \cong \bigoplus_{C' \subseteq R} H^n(K_{C'}, K_{>C'}),$$

the formula (6-3) is established.

Formula (6-1) follows now easily from Proposition 5.1 and formula (6-3).

We now prove (6-2). Note that here by the assumption X is a cocompact model for $E_{\mathcal{F}}G$ and thus G is isomorphic to the fundamental group of $G(\mathcal{Q})$ (see Theorem 3.8). By Corollary 2.6,

$$\mathrm{cd}_{\mathcal{F}} G = \max\{n \in \mathbb{N} \mid H_{\mathcal{F}}^n(X, \mathcal{B}_{P_J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}.$$

By Proposition 5.1,

$$\max\{n \in \mathbb{N} \mid H_{\mathcal{F}}^n(X, \mathcal{B}_{P_J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}$$
$$= \max\{n \in \mathbb{N} \mid H^n(K_{\Omega_J^g}, K_{>\Omega_J^g}) \neq 0 \text{ for some } J \in \mathcal{Q}, g \in \mathcal{I}_J\}$$
$$= \max\{n \in \mathbb{N} \mid H^n(K_{\Omega_U^1}, K_{>\Omega_U^1}) \neq 0 \text{ for some } U \in \mathcal{Q}\}$$
$$= \max\{n \in \mathbb{N} \mid H^n(K_C, K_{>C}) \neq 0 \text{ for some block } C \subseteq \mathcal{Q}\}. \qquad \square$$

Proof of Theorem 1.6 We first prove part (i). By Proposition 3.19(1), complexes $D(K, G(\mathcal{Q}))$ and $D(B, G(\mathcal{Q}))$ are G–homotopy equivalent.

The formula for cohomology of $D(K, G(\mathcal{Q}))$ follows from formula (6-1) of Theorem 6.1 in the following way (note that in (6-1) one does not assume rigidity). Since by assumption the complex $G(\mathcal{Q})$ is thin, we have that blocks are equal to elements of $\mathcal{Q}$. Moreover, for a single element $U \in \mathcal{Q}$ we have that K_U is contractible, and thus we obtain

$$H^*(K_U, K_{>U}) \cong \widetilde{H}^{*-1}(K_{>U}).$$

Now we prove part (ii). Since $D(K, G(\mathcal{Q}))$ is a model for $E_{\mathcal{F}}G$, it is in particular simply connected, and thus by Theorem 3.8 we get that G is isomorphic to the fundamental group of $G(\mathcal{Q})$. Since $D(K, G(\mathcal{Q}))$ and $D(B, G(\mathcal{Q}))$ are G–homotopy equivalent, we conclude that $D(B, G(\mathcal{Q}))$ is a model for $E_{\mathcal{F}}G$ as well. Clearly $D(B, G(\mathcal{Q}))$ is cocompact.

The formula for the dimension of $D(B, G(\mathcal{Q}))$ and formula (1-3) for $\mathrm{cd}_{\mathcal{F}} G$ follow now easily from combining Propositions 3.21 and 4.2. □

Remark 6.2 Theorem 6.1 holds true if we replace the complex K by any other panel complex over $\mathcal{Q}$ whose all panels are contractible (cf Remark 5.2). In particular, one can use the Bestvina complex B. Unlike in Theorem 1.6, here the dimension of the resulting Basic Construction $D(B, G(\mathcal{Q}), \psi)$ may not be optimal; nonetheless, since Bestvina complex in general has a smaller cell structure than the complex K, its use may simplify cohomological computations.

7 Deformation retractions and actions on trees

In this section we show that if the Bestvina complex for $G(\mathcal{Q})$ is a tree then it can be realised as an equivariant deformation retract of the standard development. This can be seen as a generalisation of results of Davis [9, Proposition 8.5.5] and the authors [26] to the case of infinite local groups. The key ingredient in the proof is the cohomological formula of Theorem 1.6. We remark that our approach relies neither on Dunwoody's accessibility theory [16] nor on Dicks and Dunwoody's almost stability theorem [14, III.8.5].

Theorem 7.1 *Let $G(\mathcal{Q})$ be a strictly developable thin simple complex of groups over a poset $\mathcal{Q}$ with fundamental group G and let $\mathcal{F}$ be the family generated by local groups. Suppose that $D(K, G(\mathcal{Q}))$ is a model for $E_{\mathcal{F}}G$. Then $\mathrm{cd}_{\mathcal{F}} G \leqslant 1$ if and only if $D(B, G(\mathcal{Q}))$ is a tree and an equivariant deformation retract of $D(K, G(\mathcal{Q}))$.*

Proof The proof is a verbatim translation of the proof of Theorem 4.8 of [26], which treats the case of finite local groups. The only place where that proof uses the fact that local groups are finite is the use of [26, Proposition 3.6], which gives a formula for the cohomological dimension of G for the family of finite subgroups. In Theorem 1.6 we prove that the same formula holds for a family $\mathcal{F}$ generated by arbitrary local groups,

$$\mathrm{cd}_{\mathcal{F}} G = \max\{n \in \mathbb{N} \mid \tilde{H}^{n-1}(K_{>J}) \neq 0 \text{ for some } J \in \mathcal{Q}\}.$$

Note that $\mathrm{cd}_{\mathcal{F}} G \leqslant 1$ implies that for any $J \in \mathcal{Q}$ we have $\tilde{H}^n(K_{>J}) = 0$ for all $n > 0$, and thus any $K_{>J}$ is a disjoint union of contractible spaces. This is the crucial piece of geometric information which is used in [26, Theorem 4.8] to build the Bestvina complex as an equivariant deformation retract of the standard development. □

In some cases the condition ensuring that $\mathrm{cd}_{\mathcal{F}} G \leqslant 1$ can be read from the global structure of the poset $\mathcal{Q}$.

Example 7.2 Suppose Q is a poset of simplices of a finite flag simplicial complex L. Then lcd $Q \leqslant 1$ if and only if the one skeleton $L^{(1)}$ of L is a *chordal graph*, ie for any cycle in $L^{(1)}$ of length at least four there is an edge connecting two nonconsecutive vertices of the cycle (a *chord*).

8 Applications and examples

8.1 Bredon cohomological dimension for finite subgroups

Proposition 8.1 *Let* $G(Q)$ *be a strictly developable simple complex of groups with collection of local groups* $\{P_J\}_{J \in Q}$ *and fundamental group* G. *Let* $\mathcal{R}$ *be the associated block poset. Suppose* $D(K, G(Q))$ *is a model for* $E_{\mathcal{F}}G$ *where* $\mathcal{F}$ *is the family generated by local groups and assume that* $\mathcal{F}$ *contains all finite subgroups of* G. *Then*

$$\underline{\mathrm{cd}}\, G \leqslant \mathrm{lcd}\, \mathcal{R} + \max\{\underline{\mathrm{cd}}\, P_J \mid J \in Q\}.$$

In particular, if $G(Q)$ *is thin then*

$$\underline{\mathrm{cd}}\, G \leqslant \mathrm{lcd}\, Q + \max\{\underline{\mathrm{cd}}\, P_J \mid J \in Q\}.$$

If G *is virtually torsion-free then both inequalities remain true if one replaces "*$\underline{\mathrm{cd}}$*" by "*vcd*".*

Proof For any discrete group G and for any family of subgroups $\mathcal{F}$ which contains all finite subgroups of G we have $\underline{\mathrm{cd}}\, G \leqslant \mathrm{cd}_{\mathcal{F}}\, G + \max\{\underline{\mathrm{cd}}\, F \mid F \in \mathcal{F}\}$ [13, Corollary 4.2]. Since every subgroup in $\mathcal{F}$ is subconjugate to a subgroup in $\{P_J\}_{J \in Q}$, we get that $\max\{\underline{\mathrm{cd}}\, F \mid F \in \mathcal{F}\} = \max\{\underline{\mathrm{cd}}\, P_J \mid P_J \in Q\}$. Both claims now follow from Lemma 3.22.

For the virtually torsion-free case, one first replaces G with a torsion-free finite-index subgroup G' and then one performs the same argument as above applied to ordinary cohomological dimension instead of the proper cohomological dimension. $\square$

8.2 Cohomology of buildings and their automorphisms

Groups acting chamber transitively on buildings form a large class of examples of actions on nonpositively curved complexes with a strict fundamental domain.

We recall some terminology. Let (W, S) be a Coxeter system with the set S finite. A subset $J \subseteq S$ is called *spherical* if the elements of S generate a finite subgroup of W (we assume that the empty set $\varnothing \subseteq S$ generates the trivial subgroup and thus it is spherical). Let Q be the poset of spherical subsets of S ordered by inclusion.

Now suppose that Δ is a building of type (W, S) and that a group G acts chamber transitively on Δ (see [8, Section I.3]). Such an action gives rise to a strictly developable simple complex of groups $G(Q)$ with fundamental group isomorphic to G. The standard *geometric realisation* of Δ is by definition the Basic Construction $D(K, G(Q))$ (by replacing K with another panel complex over Q one obtains a variety of

geometric realisations of Δ). By [8, Theorem 11.1] there is a complete CAT(0) metric on $D(K, G(\mathcal{Q}))$ such that G acts by isometries. Thus $D(K, G(\mathcal{Q}))$ is a cocompact model for $E_{\mathcal{F}}G$, where $\mathcal{F}$ is the family generated by the local groups. Let P_J denote the local group (ie the special parabolic subgroup) at the element $J \in Q$.

Lemma 8.2 *In the above setting, if $J \leqslant T$ then $P_J \leqslant P_T$ is a proper inclusion. In particular, the complex of groups $G(\mathcal{Q})$ is thin.*

Proof The proof is verbatim the proof of [22, Lemma 5.1], since the assumption that G acts properly on Δ was not used there.   □

We remark that in the case where $G = W$, the standard geometric realisation $D(K, G(\mathcal{Q}))$ of Δ is by definition the Davis complex of the system (W, S) and it is denoted by Σ_W.

We are now ready to prove the main result of this section, which is Corollary 1.7

Proof of Corollary 1.7 By definition $D(B, G(\mathcal{Q}))$ is a realisation of Δ. Since by Lemma 8.2 the complex $G(\mathcal{Q})$ is thin, Proposition 3.19(1) implies that $D(B, G(\mathcal{Q}))$ and $D(K, G(\mathcal{Q}))$ are G–homotopy equivalent. Thus $D(B, G(\mathcal{Q}))$ is a model for $E_{\mathcal{F}}G$, since $D(K, G(\mathcal{Q}))$ is a model. Since $D(B, G(\mathcal{Q}))$ is clearly cocompact, this establishes the first claim of the theorem.

The remaining claims follow directly from Theorem 1.6 as the formula for vcd W (see [15, Theorem 2] or [12, Theorem 5.4]) is identical to formula (1-3) for $\mathrm{cd}_{\mathcal{F}} G$.   □

Remark 8.3 $D(B, G(\mathcal{Q}))$ can also be constructed by first constructing $D(B, W(\mathcal{Q}))$ for the corresponding Coxeter group W and then realising the building with apartments modelled on $D(B, W(\mathcal{Q}))$.

We obtain the following corollary, first proven in [19, Theorem 4.1(ii)].

Corollary 8.4 *Let G be a virtually torsion-free group acting chamber transitively on a building of type (W, S). Then*

$$\mathrm{vcd}\, G \leqslant \mathrm{vcd}\, W + \max\{\mathrm{vcd}\, P \mid P \text{ is a special parabolic subgroup of } G\}.$$

Proof The corollary follows easily from combining Corollary 1.7 with Proposition 8.1, and the facts that $\mathrm{lcd}\, \mathcal{Q} = \mathrm{vcd}\, W$ and that local groups of $G(\mathcal{Q})$ are precisely the special parabolic subgroups of G. □

8.3 Graph products of groups

An example of a group acting chamber transitively on a building is a *graph product* of groups, such as for example the right-angled Artin group or the right-angled Coxeter group.

Definition 8.5 Consider a finite flag simplicial complex L on the vertex set S with groups P_s for every $s \in S$. The *graph product* G_L is defined as the quotient of the free product of groups P_s for $s \in S$ by the relations

$$\{[P_s, P_t] \mid [s, t] \text{ is an edge of } L\}.$$

In other words, elements of subgroups P_s and P_t commute if and only if there is an edge $[s, t]$ in L.

If we set $P_s \cong \mathbb{Z}/2$ for every $s \in S$, the corresponding graph product is called the *right-angled Coxeter group* and it is denoted by W_L.

If we set $P_s \cong \mathbb{Z}$ for every $s \in S$, the corresponding graph product is called the *right-angled Artin group* and it is denoted by A_L.

Theorem 8.6 [8, Theorem 5.1] *The group G_L acts chamber transitively on a building of type (W_L, S), where W_L is the right-angled Coxeter group corresponding to L.*

Thus G_L is the fundamental group of a simple complex of groups $G(\mathcal{Q})$, where $\mathcal{Q}$ is the poset of spherical subsets of S. Note that $\mathcal{Q}$ can be identified with the poset of simplices of L ordered by inclusion, together with the smallest element corresponding to the empty set. Consequently, the geometric realisation of $\mathcal{Q}$ is isomorphic to the cone over the barycentric subdivision of L. Moreover, the local group at simplex σ of L is the direct product $\prod_{s \in \sigma} P_s$ and the local group at $\varnothing$ is the trivial group.

Theorem 8.6 implies that Corollaries 1.7, 1.8, 1.9 and 1.10 apply to G_L.

8.4 Examples

Example 8.7 (barycentric subdivision and thinning) The first example shows that the thinning procedure may be intuitively seen as an inverse to the barycentric subdivision.

Let X be a G–simplicial complex with a strict fundamental domain Y, let $G(\mathcal{Q})$ be the associated complex of groups and let $\mathcal{F}$ be the family generated by the stabiliser subgroups. Thus $\mathcal{Q}$ is the poset of simplices of Y (ordered by the reverse inclusion). Assume that $G(\mathcal{Q})$ is thin.

Now let X' denote the barycentric subdivision of X, and consider the induced action of G on X'. The fundamental domain for this action is clearly Y'. Let $G(\mathcal{Q}')$ be the associated simple complex of groups, where $\mathcal{Q}'$ is the poset of simplices of Y'. Observe that $G(\mathcal{Q}')$ is not thin.

One easily sees that the fundamental groups of $G(\mathcal{Q})$ and $G(\mathcal{Q}')$ are isomorphic, and so are the families generated by local groups. However,

$$\text{lcd } \mathcal{Q}' = \dim(X') = \dim(X),$$

while in general lcd $\mathcal{Q}$ is strictly less than $\dim(X)$.

Proposition 8.8 *Let $G(\mathcal{Q})$ and $G(\mathcal{Q}')$ be as above. Let $\mathcal{R}$ denote the block poset associated to $G(\mathcal{Q}')$ and let $G(\mathcal{R})$ be the thinning of $G(\mathcal{Q}')$. Then $\mathcal{Q}$ and $\mathcal{R}$ are isomorphic, and simple complexes of groups $G(\mathcal{Q})$ and $G(\mathcal{R})$ are simply isomorphic.*

Proof Given a simplex $\sigma \subset Y$, all simplices of Y' of the form $\{\sigma_0 \subset \sigma_1 \subset \cdots \subset \sigma\}$ have the same local group equal to P_σ, where P_σ is the local group of $G(\mathcal{Q})$ at σ. Thus blocks of $\mathcal{Q}'$ are of the form $C_\sigma = \bigcup_k \{\sigma_0 \subset \sigma_1 \subset \cdots \subset \sigma_k \mid \sigma_k = \sigma\}$ and one can define a morphism $\mathcal{Q} \to \mathcal{R}$ by $\sigma \mapsto C_\sigma$. It is straightforward to check that it is an isomorphism and that so is the induced morphism $G(\mathcal{Q}) \to G(\mathcal{R})$. □

9 Reflection-like actions

In this section we introduce *reflection-like actions*, which generalise the actions of reflection groups on Euclidean spaces. Our main application is the construction of new counterexamples to the strong form of Brown's conjecture regarding the equality between vcd G and $\underline{\mathrm{gd}}\, G$ (see [6, Chapter 2] or [7, VIII.11]):

Brown's conjecture *Let G be a virtually torsion-free group with vcd $G < \infty$.*

  (i) **Weak form** *There is a contractible proper G–CW–complex of dimension vcd G.*

 (ii) **Strong form** $\underline{\mathrm{gd}}\, G = \mathrm{vcd}\, G$.

Our counterexamples are similar to those of [22], where the desired group G is a semidirect product of W_L and F, where W_L is a right-angled Coxeter group associated to a flag complex L and F is a finite group acting on L. However our method of producing these counterexamples is different. In our case, we require the action of F on L to be reflection-like and rely on an application of Theorem 6.1.

To the best of our knowledge, the only known example of a reflection-like action that serves as a counterexample to the strong form of Brown's conjecture is the action of A_5 on the 2–skeleton of the Poincaré homology sphere (see [22, Example 1]). In Example 9.13 we generalise this example. The reader may also look at the treatment of this example in [26], where the action is implicitly proven to be reflection-like.

Definition 9.1 (reflection-like action) Let F be a group acting admissibly on a connected, flag simplicial complex L of dimension $n \geqslant 1$, and let $Y \subseteq L$ be a strict fundamental domain for this action. We say that such an F–action is *reflection-like* if

  (i) the fundamental domain Y is homeomorphic to the ball B^n;

 (ii) every interior point of Y has the same stabiliser, which we denote by F_0;

(iii) F_0 is a proper subgroup of the stabiliser of any point in ∂B^n.

Note that, in particular, part (iii) implies that both the group F and its action on L are nontrivial.

Remark 9.2 In the above definition, the assumptions on the action and on the complex L are not very restrictive. Indeed, given an action of F on a polyhedral complex L, by taking barycentric subdivision of L one obtains an admissible action on a flag simplicial complex.

Lemma 9.3 *Consider a reflection-like action of F on L with a strict fundamental domain $Y \subset L$. Let $\mathcal{Q}$ denote the poset of simplices of Y ordered by reverse inclusion and let $F(\mathcal{Q})$ be the associated simple complex of groups* (see Theorem 3.8). *Then, the poset $\mathcal{Q}$ contains a block C with local group F_0 such that*

- (1) $K_C = K \cong Y \cong B^n$,
- (2) $K_{>C} \cong \partial(Y) \cong S^{n-1}$.

Proof The statement follows directly from the definition of a reflection-like action. Indeed, by Definition 9.1(ii) the local group at any (open) simplex which does not lie on the boundary of $Y \cong B^n$, is necessarily equal to F_0. On the other hand, by Definition 9.1(iii) the local group at any simplex on the boundary strictly contains F_0. □

Definition 9.4 Let F be a finite group with a reflection-like action on a connected, compact, n–dimensional flag simplicial complex L with a strict fundamental domain $Y \subseteq L$. Let W_L be the right-angled Coxeter group associated to L. Then the F–action of L induces an F–action on W_L. Let $G = W_L \rtimes F$ be the associated semidirect product.

In what follows, unless stated otherwise, let F, L, Y and G be as in Definition 9.4.

Proposition 9.5 *The group G acts on Davis complex Σ_{W_L} with strict fundamental domain and this action is proper and reflection-like.*

Proof The group G acts properly on the Davis complex Σ_{W_L} with a strict fundamental domain [22, Lemma 3.5]. One easily verifies that the fundamental domain is equal to $C(Y')$, the cone over the barycentric subdivision of Y. Since $Y \cong B^n$, we get that $C(Y') \cong B^{n+1}$ and thus part (i) of Definition 9.1 is satisfied. For parts (ii) and (iii) we need to identify G–stabilisers of the points in $C(Y')$. Recall that

$$C(Y') = Y' \times [0, 1] \big/ (x, 1) \sim (x', 1)$$

and let $[x, t]$ denote the equivalence class of a point $(x, t) \in Y' \times [0, 1]$.

- (1) For the points in the interior of $C(Y')$, ie points $[x, t]$ where $x \in \text{int}(Y')$ and $t \in (0, 1)$, we have $\text{Stab}_G[x, t] = F_0$ (where F_0 is the stabiliser of points in $\text{int}(Y)$ with respect to the F–action on L). This establishes part (ii) of Definition 9.1.
- (2) We have three types of points on the boundary of $C(Y')$:
  - (a) For the points $[x, 0]$ where $x \in Y'$, the stabiliser $\text{Stab}_G[x, t]$ is the Cartesian product of at least one generator of W_L and the stabiliser of $x \in Y$ with respect to the F–action on L.

(b) For the points (x, t) where $x \in \partial(Y')$ and $t \in (0, 1)$, the stabiliser $\mathrm{Stab}_G[x, t]$ is equal to the stabiliser $x \in \partial(Y)$ with respect to the F–action on L.

(c) The stabiliser of the point $[x, 1]$ is equal to the entire F.

Note that in each of the above cases, the stabiliser of $[x, t]$ strictly contains F_0. In case (a) this follows from the fact that there is at least one generator of W_L in the stabiliser, and in cases (b) and (c) this follows from the definition of a reflection-like action. Thus part (iii) of Definition 9.1 is satisfied, and therefore the G–action on Σ_{W_L} is reflection-like. $\qquad\square$

Lemma 9.6 *Let $G(\mathcal{Q})$ be a simple complex of groups associated to the G–action on Σ_{W_L}. Then G is isomorphic to the fundamental group of $G(\mathcal{Q})$ and*

$$\dim D(B, G(\mathcal{Q})) = \dim D(K, G(\mathcal{Q})) = \underline{\mathrm{gd}}\, G = \underline{\mathrm{cd}}\, G = n + 1.$$

Proof Since Σ_{W_L} is simply connected, by Theorem 3.8 we conclude that G is isomorphic to the fundamental group of $G(\mathcal{Q})$. The G–action on Σ_{W_L} is proper and cocompact, and since Σ_{W_L} is CAT(0), it follows that Σ_{W_L} is a cocompact G–CW–model for $\underline{E}G$. Note that $G(\mathcal{Q})$ is rigid, since all of its local groups are finite.

Thus the assumptions of Theorem 6.1 are satisfied and we can use it to compute the Bredon dimension of G. First note that since $\dim(\Sigma_{W_L}) = n + 1$, we get that $\underline{\mathrm{cd}}\, G \leqslant n + 1$. Thus it suffices to show that $\underline{\mathrm{cd}}\, G \geqslant n + 1$. By Proposition 9.5 the G–action on Σ_{W_L} is reflection-like and thus by Lemma 9.3 the poset Q contains a block C such that

(1) $K_C \cong C(Y') \cong B^{n+1}$,

(2) $K_{>C} \cong \partial(C(Y')) \cong S^n$.

Since $H^{n+1}(B^{n+1}, S^n) \cong \mathbb{Z} \neq 0$, by Theorem 6.1 we have that $\underline{\mathrm{cd}}\, G \geqslant n + 1$. $\qquad\square$

Lemma 9.7 *If $H^n(L) = 0$ then $\mathrm{vcd}\, G \leqslant n$.*

Proof Since G is a finite extension of W_L, we have that $\mathrm{vcd}\, G = \mathrm{vcd}\, W_L$. To prove that $\mathrm{vcd}\, W_L \leqslant n$, by [15, Theorem 2] it suffices to show that $H^n(\mathrm{Lk}(\sigma, L)) = 0$ for every simplex σ of L. For any nonempty simplex σ, the link $\mathrm{Lk}(\sigma, L)$ is at most $(n-1)$–dimensional, and thus $H^n(\mathrm{Lk}(\sigma, L)) = 0$. If σ is empty, $\mathrm{Lk}(\sigma, L) \cong L$ and by the assumption we have $H^n(L) = 0$. $\qquad\square$

The following theorem can be used to construct new cocompact counterexamples to the strong form of Brown's conjecture.

Theorem 9.8 *Let F be a finite group admitting a reflection-like action on a compact, connected, flag simplicial complex L of dimension $n \geqslant 1$. Let W_L be the right-angled Coxeter group associated to L and $G = W_L \rtimes F$ be the associated semidirect product. Suppose that $H^n(L) = 0$. Then*

$$\mathrm{vcd}\, G \leqslant n \quad \text{and} \quad \underline{\mathrm{cd}}\, G = n + 1.$$

Proof The statement follows immediately from combining Lemmas 9.6 and 9.7. $\qquad\square$

9.1 Examples of reflection-like actions

It remains to produce examples of groups satisfying the assumptions of Theorem 9.8. In every example discussed below, the underlying space admits an invariant polyhedral structure, which we will not specify (cf Remark 9.2).

We begin with the following two preparatory lemmas.

Lemma 9.9 *Suppose we have reflection-like actions of F_1 on an m–dimensional complex L_1 and of F_2 on an n–dimensional complex L_2. Then:*

(1) *The induced action of $F_1 \times F_2$ on $L_1 \times L_2$ is reflection-like. The fundamental domain is equal to the product of the respective fundamental domains and it is homeomorphic to B^{m+n}.*

(2) *The induced action of $F_1 \times F_2$ on the join $L_1 * L_2$ is reflection-like. The fundamental domain is equal to the join of the respective fundamental domains and it is homeomorphic to B^{m+n+1}.*

The proof is straightforward and follows at once from the definition of a reflection-like action.

Lemma 9.10 *Let L_1 be an m–dimensional finite complex and L_2 be an n–dimensional finite complex. Assume that either*

(1) $H^m(L_1) = 0$, *or*

(2) $H_m(L_1) = 0$, $H_n(L_2) = 0$ *and* $\mathrm{Tor}(H_{m-1}(L_1), H_{n-1}(L_2)) = 0$.

Then $H^{m+n}(L_1 \times L_2) = 0$ and $H^{m+n+1}(L_1 * L_2) = 0$.

Note that the assumption $\mathrm{Tor}(H_{m-1}(L_1), H_{n-1}(L_2)) = 0$ is equivalent to torsion subgroups of $H_{m-1}(L_1)$ and $H_{n-1}(L_2)$ having coprime orders.

Proof The claim follows easily from the Künneth formula, the universal coefficients theorem and the Mayer–Vietoris sequence for the join and the product. □

Note that Lemma 9.9 gives an easy way of producing new examples of reflection-like actions out of old ones, and Lemma 9.10 can be used to ensure that top-dimensional cohomology of the product/join will vanish. In order to construct genuinely new examples with vanishing top-dimensional cohomology, we first construct examples that do have nonzero top-dimensional cohomology, and then combine them into products or joins and use Lemma 9.10 to ensure that the top-dimensional cohomology vanishes.

The summary of the constructed examples is presented in Table 1.

Example 9.11 (finite reflection group) Let $F \leqslant O(n)$ be a finite subgroup generated by orthogonal reflections across hyperplanes in $\mathbb{R}^n$ (see [9, Chapter 6]). Then the induced action of F on the unit sphere $S^{n-1} \subset \mathbb{R}^n$ is reflection-like.

example	F	L	$\dim(L)$	$H^{\dim(L)}(L) = 0$?
9.11	$F \leqslant O(n+1)$, F finite	S^n	n	no
9.12	$(\mathbb{Z}/2)^n$	$\mathbb{R}P^n$	n	no
9.13	$\mathrm{PGL}_2(q)$, $q = 2^a$, $a \geqslant 2$	L_q	2	no, unless $q = 4$
9.14	D_k	M_k	2	no
9.15	$(\mathbb{Z}/2)^n \times \mathrm{PGL}_2(q)$, n even	$\mathbb{R}P^n \times L_q$	$n+2$	yes
		$\mathbb{R}P^n * L_q$	$n+3$	yes
9.16	$D_k \times D_l$, k and l coprime	$M_k \times M_l$	4	yes
		$M_k * M_l$	5	yes
9.17	$(\mathbb{Z}/2)^n \times D_k$, n even, k odd	$\mathbb{R}P^n \times M_k$	$n+2$	yes
		$\mathbb{R}P^n * M_k$	$n+3$	yes

Table 1: Examples of reflection-like actions, together with an indication whether they satisfy the assumptions of Theorem 9.8.

Example 9.12 Consider the action of $\mathbb{Z}/2$ on $\mathbb{R}$ given by $x \mapsto -x$ and consider the product action of $(\mathbb{Z}/2)^n$ on $\mathbb{R}^n$. Factoring out the action of the antipodal map $\iota \in (\mathbb{Z}/2)^n$, we obtain an action of $(\mathbb{Z}/2)^n / \langle \iota \rangle \cong (\mathbb{Z}/2)^{n-1}$ on the real projective space $\mathbb{R}P^{n-1}$. One easily verifies that this action is reflection-like, with the quotient being an $(n-1)$–simplex.

The above example is a special case of the so-called *small cover* of Davis and Januszkiewicz [10], which is an n–dimensional manifold together with a reflection-like action of $(\mathbb{Z}/2)^n$ whose quotient is isomorphic to an n–dimensional simple polytope.

Example 9.13 (Aschbacher–Segev) We outline a construction of a reflection-like action of the group $F = \mathrm{PGL}_2(q)$ for $q = 2^a$ with $a \geqslant 2$ on a flag 2–complex $L = L_q$ in order to illustrate the underlying simple complex of finite groups $F(\mathcal{Q})$. For more details we refer to [1, Section 9].

For the 1–skeleton $L_q^{(1)}$ take the barycentric subdivision of the complete graph on the projective line of $q + 1$ points $v_1, \ldots, v_{q+1}$ with the natural action of F. Fix a single conjugacy class $\mathcal{C}$ of cycles of order $q + 1$ in F. Every cycle of order $q + 1$ is conjugate to its inverse. Therefore, there are $\frac{1}{2}q(q-1)$ pairs of $(q+1)$–cycles $(\sigma_i, \sigma_i^{-1})$ in $\mathcal{C}$. Define L_q by attaching that many $(q+1)$–gons using the cycles σ_i to describe the attaching maps. Each 2–cell becomes a cone on its subdivided $(q+1)$–gonal boundary where σ_i acts by fixing the cone point. The 2–simplices of L_q are $q(q^2 - 1)$ right-angled triangles on which F acts simply transitively. Each one is a strict fundamental domain. Let Y be such a fundamental domain that contains a vertex v_j whose stabiliser is the Borel subgroup B of upper triangular matrices in F.

Figure 2, left, shows the fundamental domain Y together with local groups at cells. Figure 2, right, shows the fundamental domain $C(Y')$ for the associated action of $W_L \rtimes F$ on Σ_{W_L} together with local groups

$$Y \cong F/L \qquad\qquad C(Y') \cong \Sigma_{W_L}/W_L \rtimes F$$
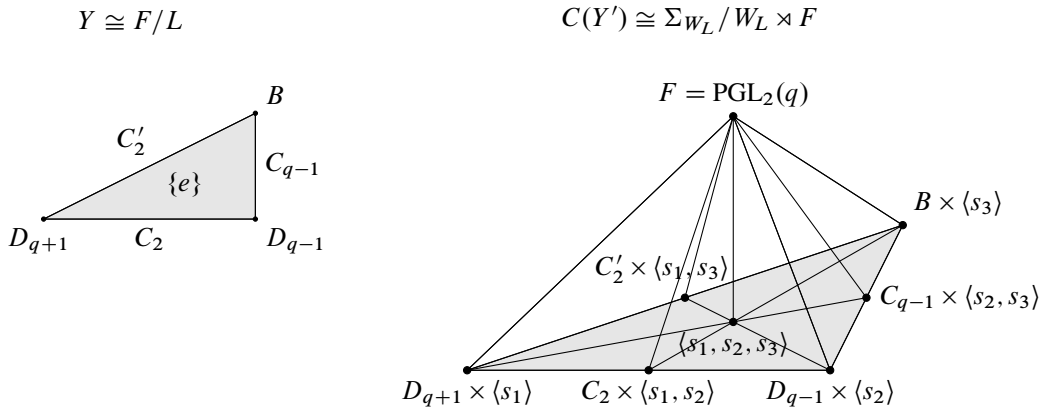
Figure 2: Fundamental domains Y, left, and $C(Y')$, right, together with stabilisers of cells and vertices respectively.

at vertices. Local groups at cells are given by the respective intersections of local groups at vertices. The generators of W_L corresponding to vertices of Y are denoted by s_1, s_2 and s_3.

($*$) For small values of q, the complex L_q is known to be $\mathbb{Q}$–acyclic, with first homology either trivial or elementary abelian of order r^{q-1}, where r is an odd prime. For $q = 4$, the complex L_q is homeomorphic to the Poincaré dodecahedron, and hence it is acyclic.

Example 9.14 (dihedral group acting on a Moore space) For a natural number $k \geqslant 2$, let M_k denote the Moore space $M(\mathbb{Z}/k, 1)$, ie a space obtained by attaching a disk to a circle along the map of degree k. Thus we have $\widetilde{H}_1(M_k) \cong \mathbb{Z}/k$ and $\widetilde{H}_i(M_k) = 0$ for all $i \neq 1$. We will describe a reflection-like action of the dihedral group D_k on M_k. Recall that D_k is generated by two reflections s and t and their product st is a rotation of order k.

Consider the standard action of D_k on a k–gon and the reflection action of $D_k/\langle st \rangle \cong \mathbb{Z}/2$ on a circle, both shown in Figure 3, left, (note that both actions reverse the orientation of the edges). The attaching map of the boundary of the k–gon is equivariant with respect to the homomorphism $D_k \to D_k/\langle st \rangle \cong \mathbb{Z}/2$, and thus we get a well-defined action of D_k on M_k. One easily checks that this action has a strict fundamental domain, which is a triangle. The fundamental domain together with its cell stabilisers is shown in Figure 3, right. By analysing the stabilisers, we conclude that the action of D_k on M_k is reflection-like.

We remark that in this setting M_k is homeomorphic to the Basic Construction $D(|\mathcal{Q}|, G(\mathcal{Q}), \psi)$, where $G(\mathcal{Q})$ is a simple complex of groups associated to the D_k–action on M_k, and $\psi : G(\mathcal{Q}) \to D_k$ is a simple morphism induced by sending all three vertex groups D_k into D_k via the identity map.

Finally, observe that for $k = 2$ in the above construction, $D_2 \cong \mathbb{Z}/2 \times \mathbb{Z}/2$ is an isometry group of a 2–gon and M_k is equivariantly homeomorphic to the real projective plane $\mathbb{R}P^2$ appearing in Example 9.12.
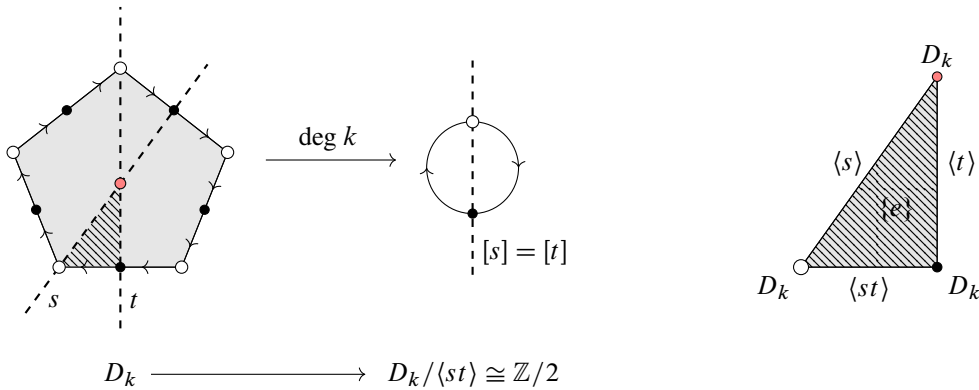
Figure 3: Left: reflection like action of D_k on a Moore space M_k. Right: the fundamental domain together with local groups at cells.

We are ready now to construct new counterexamples to the strong form of Brown's conjecture.

Example 9.15 Let L_q be a complex as in Example 9.13 satisfying ($*$). For an even integer n, consider the induced reflection-like actions of the product $(\mathbb{Z}/2)^n \times \mathrm{PGL}_2(q)$ on the product $\mathbb{R}P^n \times L_q$ and on the join $\mathbb{R}P^n * L_q$.

Since $H_n(\mathbb{R}P^n) = 0$, $H_{n-1}(\mathbb{R}P^n) = \mathbb{Z}/2$, $H_2(L_q) = 0$ and $H_1(L_q)$ is either trivial or elementary abelian of order being a power of an odd prime, by Lemma 9.10 we conclude that $H^{n+2}(\mathbb{R}P^n \times L_q) = 0$ and $H^{n+3}(\mathbb{R}P^n * L_q) = 0$.

Example 9.16 Consider M_k and M_l such that k and l are coprime. By Lemma 9.10 we get that $H^4(M_k \times M_l) = 0$ and $H^5(M_k * M_l) = 0$ (in fact $M_k * M_l$ is contractible).

Example 9.17 For an even integer n and an odd integer k consider the action of $(\mathbb{Z}/2)^n$ on the real projective space $\mathbb{R}P^n$, and the action of D_k on the Moore space M_k. By Lemma 9.10 we have $H^{n+2}(\mathbb{R}P^n \times M_k) = 0$ and $H^{n+3}(\mathbb{R}P^n * M_k) = 0$.

Remark 9.18 In contrast to Example 9.13 (and Example 9.15), Examples 9.16 and 9.17 are particularly simple in terms of algebraic structure of groups and cellular structure of complexes. The smallest group appearing in these examples is the product $D_2 \times D_3 \cong (\mathbb{Z}/2)^2 \times S_3$.

10 Final remarks and open questions

Let X be a G–CW–complex. We say that a G–CW–subcomplex Y is a *spine* of X if it is an equivariant deformation retract of X. When X is a model for $E_{\mathcal{F}}G$, then so is Y and $\dim(Y) \geqslant \mathrm{gd}_{\mathcal{F}} G$. Spines of minimal dimension (so equal to $\mathrm{gd}_{\mathcal{F}} G$) have been constructed, for example, for the actions of certain

arithmetic groups such as $\mathrm{SL}(n, \mathbb{Z})$ on the symmetric space [2], the actions of the outer automorphism groups of free groups on the outer space [27], and the actions of the mapping class groups of punctured surfaces on the Teichmüller space [18].

Question 10.1 *Suppose a group G acts on a $\mathrm{CAT}(0)$ polyhedral complex X with a strict fundamental domain. Denote by $\mathcal{F}$ the family generated by the stabilisers. Suppose the associated complex of groups $G(\mathcal{Q})$ is thin. Can $D(B, G(\mathcal{Q}))$ be constructed as a spine of X of the lowest possible dimension equal to $\mathrm{gd}_{\mathcal{F}} G$?*

Theorem 7.1 tells us that the answer is yes if $\mathrm{cd}_{\mathcal{F}} G \leqslant 1$. Also by Theorem 1.6, we know that $\dim D(B, G(\mathcal{Q})) = \mathrm{gd}_{\mathcal{F}} G$ and $D(B, G(\mathcal{Q}))$ is G–homotopy equivalent to X. The question whether $D(B, G(\mathcal{Q}))$ can be constructed as an equivariant deformation retract of X is open in general. In [26], we isolate a condition on a finite polyhedra which we call *subconical*. It is open whether every finite polyhedron is subconical. If this is the case, then a generalisation of [26, Proposition 4.7] to thin simple complexes of groups gives an affirmative answer to this question.

Question 10.2 *Does $D(B, G(\mathcal{Q}))$ attain the $\mathrm{CAT}(0)$ dimension of the group G?*

In many cases of interest, such as Coxeter groups or groups acting on buildings, the associated standard development $D(K, G(\mathcal{Q}))$ supports a G–invariant $\mathrm{CAT}(0)$ metric. Therefore it is natural to ask whether the Bestvina complex supports such a metric as well, or whether the dimension of Bestvina complex is equal to the $\mathrm{CAT}(0)$ *dimension of the group for the family $\mathcal{F}$*. The latter is defined as the minimal dimension of a model for $E_{\mathcal{F}} G$ that supports a G–invariant $\mathrm{CAT}(0)$ metric.

There are simple complexes of groups where the corresponding Bestvina complex does not admit any G–invariant piecewise linear $\mathrm{CAT}(0)$ metric (this will be shown in a forthcoming work of the second author). Moreover, we suspect that these examples also have $\mathrm{CAT}(0)$ dimension strictly larger than the Bredon cohomological dimension. The above examples are the right-angled Coxeter groups (or graph products) associated to certain 2–dimensional contractible but noncollapsible complexes. Consequently, the methods used to show the lack of $\mathrm{CAT}(0)$ metric do not carry through to higher dimensions, and to the best of our knowledge the question is open in all dimensions greater than 2.

The question is especially interesting when $\mathcal{F}$ is the family of all finite subgroups. In this case, the metric structure of $\underline{E}G$ can be used to study numerous features of G, eg by considering the visual boundary of $\underline{E}G$. Note that the positive answer to that question, combined with Example 9.16 (or 9.17), would result in a group of $\mathrm{CAT}(0)$ dimension four, whose finite-index overgroup has $\mathrm{CAT}(0)$ dimension equal to five.

Question 10.3 *Are the groups G constructed in Examples 9.16 or 9.17 also counterexamples to the weak form of Brown's conjecture?*

The weak form of Brown's conjecture is open in all dimensions except when vcd $G = 2$ [22]. A natural place to look for counterexamples are the groups that disprove the strong form of Brown's conjecture. Yet, most such groups G are known to act properly on a contractible complex of dimension vcd G. Take for example $G = W_L \rtimes F$. If L is contractible (see [22, Section 5] for examples), then there is a contractible subcomplex Y of Σ_{W_L} of dimension vcd G on which G acts properly. The subcomplex Y can be obtained by applying the Basic Construction to L' instead of CL'. Similarly, the finite extensions of Bestvina–Brady groups constructed in [21] or [24, 3.6] cannot be counterexamples to the weak form of the conjecture, because they act properly on the level sets of the Morse function which in these examples are contractible.

References

[1] **M Aschbacher**, **Y Segev**, *A fixed point theorem for groups acting on finite 2–dimensional acyclic simplicial complexes*, Proc. Lond. Math. Soc. 67 (1993) 329–354 MR Zbl

[2] **A Ash**, *Deformation retracts with lowest possible dimension of arithmetic quotients of self-adjoint homogeneous cones*, Math. Ann. 225 (1977) 69–76 MR Zbl

[3] **P Baum**, **A Connes**, **N Higson**, *Classifying space for proper actions and K–theory of group C^*–algebras*, from "C^*–algebras: 1943–1993" (R S Doran, editor), Contemp. Math. 167, Amer. Math. Soc., Providence, RI (1994) 240–291 MR Zbl

[4] **M Bestvina**, *The virtual cohomological dimension of Coxeter groups*, from "Geometric group theory, I" (G A Niblo, M A Roller, editors), Lond. Math. Soc. Lect. Note Ser. 181, Cambridge Univ. Press (1993) 19–23 MR Zbl

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

[6] **K S Brown**, *Groups of virtually finite dimension*, from "Homological group theory" (C T C Wall, editor), Lond. Math. Soc. Lect. Note Ser. 36, Cambridge Univ. Press (1979) 27–70 MR Zbl

[7] **K S Brown**, *Cohomology of groups*, Graduate Texts in Math. 87, Springer (1982) MR Zbl

[8] **M W Davis**, *Buildings are* CAT(0), from "Geometry and cohomology in group theory" (P H Kropholler, G A Niblo, R Stöhr, editors), Lond. Math. Soc. Lect. Note Ser. 252, Cambridge Univ. Press (1998) 108–123 MR Zbl

[9] **M W Davis**, *The geometry and topology of Coxeter groups*, Lond. Math. Soc. Monogr. Ser. 32, Princeton Univ. Press (2008) MR Zbl

[10] **M W Davis**, **T Januszkiewicz**, *Convex polytopes, Coxeter orbifolds and torus actions*, Duke Math. J. 62 (1991) 417–451 MR Zbl

[11] **D Degrijse**, *A cohomological characterization of locally virtually cyclic groups*, Adv. Math. 305 (2017) 935–952 MR Zbl

[12] **D Degrijse**, **C Martínez-Pérez**, *Dimension invariants for groups admitting a cocompact model for proper actions*, J. Reine Angew. Math. 721 (2016) 233–249 MR Zbl

[13] **F Dembegioti**, **N Petrosyan**, **O Talelli**, *Intermediaries in Bredon (co)homology and classifying spaces*, Publ. Mat. 56 (2012) 393–412 MR Zbl

[14] **W Dicks**, **M J Dunwoody**, *Groups acting on graphs*, Cambridge Stud. Adv. Math. 17, Cambridge Univ. Press (1989) MR Zbl

[15] **A N Dranishnikov**, *On the virtual cohomological dimensions of Coxeter groups*, Proc. Amer. Math. Soc. 125 (1997) 1885–1891 MR Zbl

[16] **M J Dunwoody**, *Accessibility and groups of cohomological dimension one*, Proc. Lond. Math. Soc. 38 (1979) 193–215 MR Zbl

[17] **M Fuentes**, *The equivariant K– and KO–theory of certain classifying spaces via an equivariant Atiyah–Hirzebruch spectral sequence*, preprint (2019) arXiv 1905.02972

[18] **J L Harer**, *The virtual cohomological dimension of the mapping class group of an orientable surface*, Invent. Math. 84 (1986) 157–176 MR Zbl

[19] **J Harlander**, *On the dimension of groups acting on buildings*, from "Groups St Andrews 1997 in Bath, I" (C M Campbell, E F Robertson, N Ruskuc, G C Smith, editors), Lond. Math. Soc. Lect. Note Ser. 260, Cambridge Univ. Press (1999) 318–328 MR Zbl

[20] **J Harlander**, **H Meinert**, *Higher generation subgroup sets and the virtual cohomological dimension of graph products of finite groups*, J. Lond. Math. Soc. 53 (1996) 99–117 MR Zbl

[21] **I J Leary**, **B E A Nucinkis**, *Some groups of type VF*, Invent. Math. 151 (2003) 135–165 MR Zbl

[22] **I J Leary**, **N Petrosyan**, *On dimensions of groups with cocompact classifying spaces for proper actions*, Adv. Math. 311 (2017) 730–747 MR Zbl

[23] **W Lück**, *Survey on classifying spaces for families of subgroups*, from "Infinite groups: geometric, combinatorial and dynamical aspects", Progr. Math. 248, Birkhäuser, Basel (2005) 269–322 MR Zbl

[24] **C Martínez-Pérez**, *Euler classes and Bredon cohomology for groups with restricted families of finite subgroups*, Math. Z. 275 (2013) 761–780 MR Zbl

[25] **C Martínez-Pérez**, **B E A Nucinkis**, *Bredon cohomological finiteness conditions for generalisations of Thompson groups*, Groups Geom. Dyn. 7 (2013) 931–959 MR Zbl

[26] **N Petrosyan**, **T Prytuła**, *Bestvina complex for group actions with a strict fundamental domain*, Groups Geom. Dyn. 14 (2020) 1277–1307 MR Zbl

[27] **K Vogtmann**, *Automorphisms of free groups and outer space*, Geom. Dedicata 94 (2002) 1–31 MR Zbl

School of Mathematical Sciences, University of Southampton
Southampton, United Kingdom

Department Of Applied Mathematics And Computer Science, Technical University of Denmark
Lyngby, Denmark

n.petrosyan@soton.ac.uk,   tomasz.prytula@alexandra.dk

On the decategorification of some higher actions in Heegaard Floer homology

ANDREW MANION

We decategorify the higher actions on bordered Heegaard Floer strands algebras from recent work of Rouquier and the author, and identify the decategorifications with certain actions on exterior powers of homology groups of surfaces. We also suggest an interpretation for these actions in the language of open-closed TQFT, and we prove a corresponding gluing formula.

1 Introduction

In [15], Raphaël Rouquier and the author define a tensor product operation for higher representations of the dg monoidal category of Khovanov [11], which we call $\mathcal{U}$, and use it to reformulate aspects of cornered Heegaard Floer homology; see Douglas, Lipshitz and Manolescu [3; 4]. Part of this work involves defining 2–actions of $\mathcal{U}$ on the dg algebras $\mathcal{A}(\mathcal{Z})$ that bordered Heegaard Floer homology assigns to combinatorial representations $\mathcal{Z}$ of surfaces.

Ignoring gradings and thus working with decategorifications over $\mathbb{F}_2$, one can view $\mathcal{U}$ as a categorification of the algebra $\mathbb{F}_2[E]/(E^2)$ (an $\mathbb{F}_2$ analogue of $U(\mathfrak{gl}(1|1)^+)$), while if $\mathcal{Z}$ is a representation of a surface F, then $\mathcal{A}(\mathcal{Z})$ categorifies the vector space $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ where S_+ is a distinguished subset of the boundary of F. Thus, the 2–actions from [15] should categorify actions of $\mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$; the goal of this paper is to identify these actions explicitly using certain topological operations and to give an interpretation of these actions in the setting of open-closed TQFT.

To make things more precise, we recall that following Zarev [23] (but generalizing his definition slightly), a sutured surface is (F, S_+, S_-, Λ) where F is a compact oriented surface and Λ is a finite set of points in ∂F dividing ∂F into alternating subsets S_+ and S_-. We impose no topological restrictions, but note that the sutured surfaces representable by arc diagrams $\mathcal{Z}$ are those such that in each connected component of F (not of ∂F), both S_+ and S_- are nonempty (unlike Zarev [23], we allow arc diagrams to have circle components as well as interval components, and we do not impose nondegeneracy). In particular, no closed surface can be represented by an arc diagram.

For an arc diagram $\mathscr{Z}$ representing a sutured surface (F, S_+, S_-, Λ), and each interval component I of S_+, the constructions of [15] define a 2–action of $\mathscr{U}$ on $\mathscr{A}(\mathscr{Z})$. On the other hand, there is a map ϕ_I from $H_1(F, S_+; \mathbb{F}_2)$ to $\mathbb{F}_2$ taking an element of $H_1(F, S_+; \mathbb{F}_2)$ to its boundary in $H_0(S_+; \mathbb{F}_2)$ and then pairing with the cohomology class of I. By summing ϕ_I over tensor factors, for $k \geq 1$ we get a map from $T^k H_1(F, S_+; \mathbb{F}_2)$ to $T^{k-1} H_1(F, S_+; \mathbb{F}_2)$ which induces a map Φ_I from $\wedge^k H_1(F, S_+; \mathbb{F}_2)$ to $\wedge^{k-1} H_1(F, S_+; \mathbb{F}_2)$.

Theorem 1.1 *The 2–action of $\mathscr{U}$ on $\mathscr{A}(\mathscr{Z})$ corresponding to I categorifies the action of $\mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ in which E acts by Φ_I.*

See Theorem 3.5 below for a more detailed statement of Theorem 1.1.

A TQFT interpretation

It is natural to ask whether the actions of $\mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ fit into a TQFT framework, with associated gluing results. Indeed, [15] reformulates and strengthens Douglas–Manolescu's gluing theorem for the algebras $\mathscr{A}(\mathscr{Z})$, which applies for certain decompositions of surfaces along 1–manifolds (given by certain decompositions of the arc diagram $\mathscr{Z}$). One could hope that such gluing theorems exist in even greater generality for the decategorified surface invariants $\wedge^* H_1(F, S_+; \mathbb{F}_2)$, yielding a TQFT-like construction for 1– and 2–manifolds.

Remark 1.2 Heegaard Floer homology is, in some nonaxiomatic sense, a 4–dimensional TQFT (space-times are 4–dimensional); accordingly, its decategorification should be a type of 3–dimensional TQFT involving the vector spaces $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ (and, for example, the Alexander polynomials of knots). The constructions under consideration for 1– and 2–manifolds should be part of a (loosely defined) extended-TQFT structure for decategorified Heegaard Floer homology.

A first observation is that a sutured surface (F, S_+, S_-, Λ) is nearly the same data as a morphism in the 2–dimensional open-closed cobordism category. As described by Lauda and Pfeiffer in [12], the objects of this category are finite disjoint unions of oriented intervals and circles. For two such objects X and Y, a morphism from X to Y is a compact oriented surface with its boundary decomposed into black regions (identified with $X \sqcup Y$) and colored regions. If (F, S_+, S_-, Λ) is a sutured surface and we label each component of S_+ as "incoming" or "outgoing", we get a morphism from S_+^{in} to S_+^{out} in this cobordism category. The black part of the boundary is S_+ and the colored part is S_-.

The actions of $\mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ suggest that one could try to assign the category of finite-dimensional $\mathbb{F}_2[E]/(E^2)$–modules to an interval. A sutured surface, with its S_+ boundary components labeled as incoming or outgoing, would be assigned a bimodule over tensor powers of $\mathbb{F}_2[E]/(E^2)$. For simplicity, we will restrict our attention here to sutured surfaces with no circular S_+ boundary components (all components of S_+ are intervals).
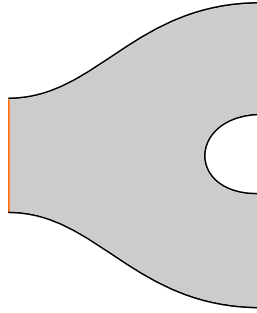
Figure 1: The open pair of pants; the S_+ boundary is shown in orange and the S_- boundary is shown in black (loosely following the visual conventions of [23]). Specifically, the input S_+ boundary is on the right while the output S_+ boundary is on the left.

For a surface F_1 with m intervals in its outgoing boundary and another surface F_2 with m intervals in its incoming boundary, let $F = F_2 \cup_{[0,1]^m} F_1$. We would want the bimodule of F to be a tensor product over $(\mathbb{F}_2[E]/(E^2))^{\otimes m}$ of the bimodules assigned to F_1 and F_2. The next theorem says this is true up to isomorphism; let $\mathbf{Alg}_{\mathbb{F}_2}$ denote the category whose objects are $\mathbb{F}_2$–algebras and whose morphisms are isomorphism classes of bimodules, with composition given by tensor product.

Theorem 1.3 *For F_1, F_2, and F as above, suppose that F_1 has m_{in} intervals in its incoming boundary and F_2 has m_{out} intervals in its outgoing boundary. We have a noncanonical isomorphism*

$$\wedge^* H_1(F, S_+; \mathbb{F}_2) \cong \wedge^* H_1(F_2, S_+; \mathbb{F}_2) \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes m}} \wedge^* H_1(F_1, S_+; \mathbb{F}_2)$$

as bimodules over $((\mathbb{F}_2[E]/(E^2))^{\otimes m_{\mathrm{out}}}, (\mathbb{F}_2[E]/(E^2))^{\otimes m_{\mathrm{in}}})$. Thus, the exterior algebra vector spaces $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ give a functor from the "open sector" of the open-closed cobordism category into $\mathbf{Alg}_{\mathbb{F}_2}$.

In fact, a slightly more general version of Theorem 1.3 holds in which F_1 and F_2 can have S_+ circles in their boundaries as long as we are not gluing along them; see Theorem 4.2 below.

The tensor product case

As a special case of Theorem 1.3, we can glue interval S_+ components of two surfaces F' and F'' to the two input intervals of the "open pair of pants" cobordism shown in Figure 1. Let $P = F_1$ be the open pair of pants, let $F_2 = F' \sqcup F''$, and let F be the glued surface. We can identify $\wedge^* H_1(P, S_+; \mathbb{F}_2)$ with $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$, with right action of $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ given by multiplication and left action of $\mathbb{F}_2[E]/(E^2)$ given by the coproduct

$$\Delta(E) = E \otimes 1 + 1 \otimes E$$

(in fact, $\mathbb{F}_2[E]/(E^2)$ is a Hopf algebra with this coproduct together with counit $\varepsilon(E) = 0$ and antipode $S(E) = E$).

Corollary 1.4 *We have*

$$\wedge^* H_1(F, S_+; \mathbb{F}_2) \cong \wedge^* H_1(F', S_+; \mathbb{F}_2) \otimes \wedge^* H_1(F'', S_+; \mathbb{F}_2),$$

where the tensor product $\otimes$ is taken in the tensor category of finite-dimensional modules over the Hopf algebra $\mathbb{F}_2[E]/(E^2)$.

We can view Corollary 1.4 as a decategorification of the gluing result from [15] based on the higher tensor product operation $\textcircled{\otimes}$. Thus, Theorem 1.3 suggests (at least at the decategorified level) a more general TQFT framework for the $\textcircled{\otimes}$–based gluing results of [15].

Relationship to other work

Probably the closest analogue to the structures considered here can be found in Honda, Kazez and Matić's paper [7]. The data of a sutured surface (F, S_+, S_-, Λ) as discussed here is equivalent to the data (Σ, F) considered in [7, Section 7.1] (our F is the Σ of Honda, Kazez and Matić and our Λ is their F). The vector space $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ is isomorphic to an $\mathbb{F}_2$ version of Honda, Kazez and Matić's $V(\Sigma, F)$ which was subsequently studied by Mathews [16; 17; 18; 19] and Mathews and Schoenfeld [20]. In our notation, Honda, Kazez and Matić view this vector space as the sutured Floer homology of $F \times S^1$ with sutures given by $\Lambda \times S^1$, rather than as a Grothendieck group associated to $\mathcal{A}(\mathcal{Z})$. In other words, their surface invariants come from "trace decategorification" of 3–dimensional Heegaard Floer invariants rather than from Grothendieck-group-based decategorification of 2–dimensional Heegaard Floer invariants; these notions often agree, as they do here. See Cooper [1] for related work in the contact setting that discusses vector spaces similar to $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ in relation to Grothendieck groups of formal contact categories.

We can think of the gluings in Theorem 1.3 as successive self-gluings of two S_+ intervals in a sutured surface. These gluings can be interpreted as special cases of Honda, Kazez and Matić's gluings, where their gluing subsets γ and γ' cover our gluing S_+ intervals and extend a small bit past them on both sides. However, Honda, Kazez and Matić only assert the existence of a gluing map from the vector space of the original surface to the vector space of the glued surface (satisfying certain properties). Theorem 1.3 goes farther for the special gluings under consideration in that it shows how the vector space of the larger surface is recovered up to isomorphism as a tensor product.

Integral versions of the vector spaces $\wedge^*(F, S_+; \mathbb{F}_2)$, especially for closed F, or F with one boundary component (and implicitly $|\Lambda| = 2$), have also been studied in the context of TQFT invariants for 3–manifolds starting with Frohman and Nicas in [5]; see also Donaldson [2] and Kerler [10]. Building on work of Petkova [21], Hom, Lidman and Watson show in [6] that bordered Heegaard Floer homology (in the original formulation of Lipshitz, Ozsváth and Thurston [14] where F is closed) can be viewed as categorifying the $2+1$ TQFT described in [2] in which a surface F is assigned $\wedge^* H_1(F)$. Our perspective here differs in that we follow Zarev [23] rather than [14] and in that instead of $2+1$ TQFT structure we are (loosely) looking at the lower two levels of a $1+1+1$ TQFT.

Finally, the fact that the topological gluing considered in [15] can be viewed as the above open-pair-of-pants gluing was already noted in [15, Section 7.2.5], which also contains speculations about the connection to open-closed TQFT and extended TQFT.

Future directions

It would be desirable to treat 1–, 2–, and 3–manifolds at the same time, integrating the gluing results for surfaces here with the 3–manifold invariants mentioned above in something like a $1 + 1 + 1$ TQFT. One obstacle to doing this appears to be that while the isomorphism in the statement of Theorem 1.3 seems like something that could conceivably be proved using Mayer–Vietoris sequences, we were not able to find such a proof; the isomorphism we construct is not canonical and depends on suitable choices of bases. Geometrically, the issue seems to be that given arbitrary elements of $\wedge^* H_1(F_1, S_+; \mathbb{F}_2)$ and $\wedge^* H_1(F_2, S_+; \mathbb{F}_2)$, it is not clear how to pair them to get an element of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ in a canonical way (the endpoints of arcs don't necessarily match up in any nice way at the gluing interface).

It would also be desirable to categorify Theorem 1.3, such that the $\otimes$–based gluing results of [15] are recovered by gluing with an open pair of pants as in Corollary 1.4. Just as the proof of Theorem 1.3 depends on a choice of basis, it seems likely that a categorification of this theorem will depend on the arc diagrams $\mathscr{Z}$ chosen to represent the surfaces. For general arc diagrams $\mathscr{Z}_1$ and $\mathscr{Z}_2$ representing the surfaces F_1 and F_2 of Theorem 1.3, it is not even clear how one should glue these diagrams to get an arc diagram for the glued surface F (speculatively, something like [8, Figure 5(b)] followed by an "unzip" operation may be relevant).

Finally, preliminary computations indicate that close relatives of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ should arise in a TQFT with better structural properties than the "open" TQFT considered here, specifically one that is extended down to points and defined at least for all 0–, 1–, and 2–manifolds, with appropriate gluing theorems (including for gluing along circles). In work in progress, we study this extended TQFT as well as its relationship to the constructions of this paper.

Organization

In Sections 2.1 through 2.3 we review $\mathcal{U}$, the algebras $\mathcal{A}(\mathscr{Z})$, and the higher actions from [15]. Section 2.4 discusses decategorification for $\mathcal{U}$ and $\mathcal{A}(\mathscr{Z})$, showing that in the sense considered here, $\mathcal{A}(\mathscr{Z})$ categorifies $\wedge^* H_1(F, S_+; \mathbb{F}_2)$. Section 3 decategorifies the 2–actions of $\mathcal{U}$ on $\mathcal{A}(\mathscr{Z})$ from [15] and proves Theorem 1.1. Section 4 proves a generalized version of Theorem 1.3, and Section 5 discusses Corollary 1.4 in more generality.

Acknowledgments

2 Decategorifying higher actions on strands algebras

2.1 The dg monoidal category $\mathcal{U}$

The following definition originated in [11] and was partly inspired by the strands dg algebras $\mathcal{A}(\mathcal{Z})$ in Heegaard Floer homology (we review these in Section 2.2). While Khovanov works over $\mathbb{Z}$, we work over $\mathbb{F}_2$ in order to interact properly with the $\mathbb{F}_2$–algebras $\mathcal{A}(\mathcal{Z})$.

Definition 2.1 Let $\mathcal{U}$ denote the strict $\mathbb{F}_2$–linear dg monoidal category freely generated (under $\otimes$ and composition) by an object e and an endomorphism τ of $e \otimes e$ modulo the relations $\tau^2 = 0$ and

$$(\mathrm{id}_e \otimes \tau) \circ (\tau \otimes \mathrm{id}_e) \circ (\mathrm{id}_e \otimes \tau) = (\tau \otimes \mathrm{id}_e) \circ (\mathrm{id}_e \otimes \tau) \circ (\tau \otimes \mathrm{id}_e).$$

We set $d(\tau) = 1$, and we let τ have degree -1 (we use the convention that differentials increase degree by 1).

The endomorphism algebra of $e^{\otimes n} \in \mathcal{U}$ is the dg algebra referred to as H_n^- in [11] (tensored with $\mathbb{F}_2$); in the language used in [15] it is a nil-Hecke algebra with a differential, and in the language used in [4] it is a nil-Coxeter algebra. We will use NC_n to denote the $\mathbb{F}_2$ version of this algebra. It has a graphical interpretation: $\mathbb{F}_2$–basis elements of NC_n are pictures like Figure 2, with n strands going from bottom to top (these pictures are in bijection with permutations on n letters). Multiplication is defined by vertical concatenation, with ab obtained by drawing a below b, except that if two strands cross and then uncross in the stacked picture (ie if the stacked picture has a double crossing) then the product is defined to be zero. The differential is defined by summing over all ways to resolve a crossing (see Figure 3), except that if a crossing resolution produces a double crossing between two strands then it contributes zero to the differential (see Figure 4). The endomorphism τ of $e \otimes e$ is represented by a single crossing between two strands.
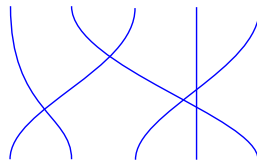
Figure 2: A basis element of NC_n for $n = 5$.
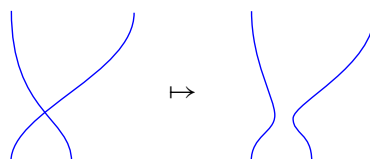
Figure 3: Resolving a crossing.
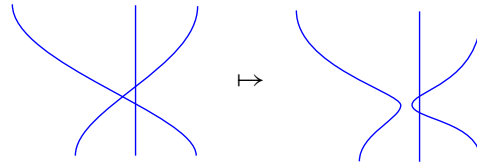
Figure 4: A resolution that produces a double crossing and thus does not contribute to the differential on NC_n.

2.2 Strands algebras

Let $\mathscr{Z}$ be an arc diagram as in [23, Definition 2.1.1], except that we allow (oriented) circles as well as intervals in $\mathbf{Z}$, and we do not impose any nondegeneracy condition. Thus, $\mathscr{Z}$ consists of:

- a finite collection $\mathbf{Z} = \{Z_1, \ldots, Z_l\}$ of oriented intervals and circles;
- a finite set of points $\mathbf{a}$ (with $|\mathbf{a}|$ even) in the interiors of the Z_i for $1 \leq i \leq l$;
- a two-to-one matching M of the points in $\mathbf{a}$.

An example is shown in Figure 5.

The definition of the dg strands algebra $\mathcal{A}(\mathscr{Z})$ over $\mathbb{F}_2$, from [23, Definition 2.2.2], generalizes in a straightforward way to this setting and is a special case of the general strands algebras treated in detail in [15]. One can view $\mathcal{A}(\mathscr{Z})$ as being defined by specifying an $\mathbb{F}_2$ basis consisting of certain pictures, along with rules for multiplying and differentiating basis elements.

Definition 2.2 A *strands picture* is a collection of strands drawn in $[0, 1] \times \mathbf{Z}$, each with its left endpoint in $\{0\} \times \mathbf{a}$ and its right endpoint in $\{1\} \times \mathbf{a}$. The strands can be either solid or dotted and are considered only up to homotopy relative to the endpoints; by convention, strands are drawn "taut", sometimes with a bit of curvature for visual effect (see Figure 6). They must satisfy the following rules:

- Strands cannot move against the orientation of $\mathbf{Z}$ when moving from left to right (from 0 to 1 in $[0, 1]$).
- No solid strands are horizontal, while all dotted strands are horizontal.
- If a solid strand has its left endpoint at $a \in \mathbf{a}$, and a is matched to $a' \in \mathbf{a}$ under M, then no strand can have its left endpoint at a', and similarly for right endpoints.
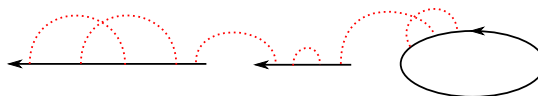
Figure 5: An arc diagram $\mathscr{Z} = (\mathbf{Z}, \mathbf{a}, M)$; $\mathbf{Z}$ consists of two intervals and a circle, $\mathbf{a}$ is the set of endpoints of the dotted (red) arcs, and M matches the two endpoints of each dotted arc.
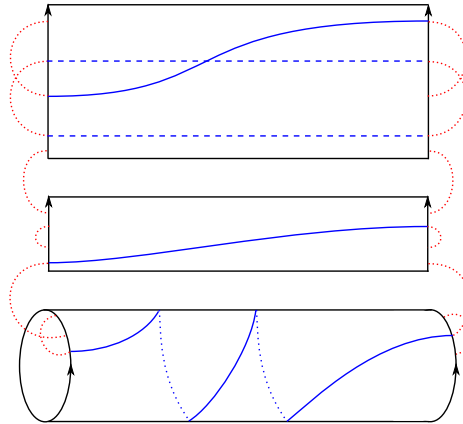
Figure 6: A strands picture (basis element for $\mathcal{A}(\mathcal{Z})$).

- If a dotted strand has its left (and thus right) endpoint at $a \in \boldsymbol{a}$, and a is matched to $a' \in \boldsymbol{a}$ under M, then there must be another dotted strand with its left (and thus right) endpoint at a' (we say this dotted strand is matched with the first one).

Definition 2.3   As an $\mathbb{F}_2$–vector space, $\mathcal{A}(\mathcal{Z})$ is defined to be the formal span of such strands pictures, so that strands pictures form an $\mathbb{F}_2$ basis for $\mathcal{A}(\mathcal{Z})$. The product of two basis elements of $\mathcal{A}(\mathcal{Z})$ is defined by concatenation (see Figure 7), with the following subtleties:

- If some solid strand has no strand to concatenate with, or if in some matched pair of dotted strands $\{s, s'\}$, neither s nor s' has a strand to concatenate with, the product is zero.
- When concatenating a solid strand with a dotted strand, one erases the dotted strand matched to the one involved in the concatenation, and makes the concatenated strand solid.
- If a double crossing is formed upon concatenation, the product of the basis elements is defined to be zero.

The differential of a basis element of $\mathcal{A}(\mathcal{Z})$ is the sum of all strands pictures formed by resolving a crossing in the original strands picture (in the sense of Figure 3 above), with the following subtleties:

- When resolving a crossing between a solid strand and a dotted strand, one erases the dotted strand matched to the one involved in the crossing resolution, and makes both the resolved strands solid.
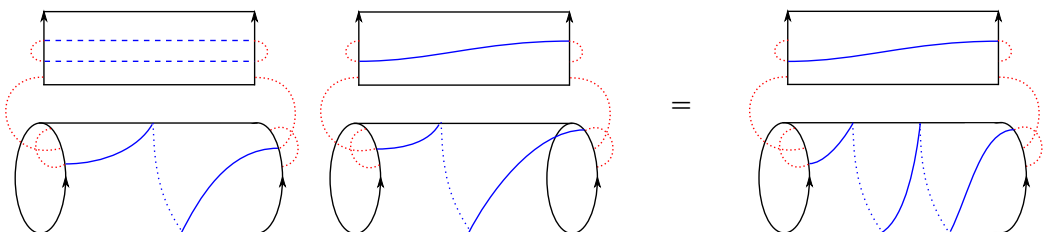
Figure 7: Example of a product in $\mathcal{A}(\mathcal{Z})$.

- If a double crossing is formed upon resolving a crossing (as in Figure 4 above), then this crossing resolution does not contribute a term to the differential.

Remark 2.4 Recall that a dg category over a field k is a category enriched in the symmetric monoidal category of chain complexes over k, ie graded $k[\partial]/(\partial^2)$–modules where ∂ has degree -1 or $+1$ depending on conventions, with the tensor product given as usual. Similarly, a differential category over k is a category enriched in the symmetric monoidal category of (ungraded) $k[\partial]/(\partial^2)$–modules (the symmetric monoidal structure is analogous to the graded case[1]).

While $\mathcal{U}$ is a dg category and not just a differential category, the grading on $\mathcal{A}(\mathcal{Z})$ is much more complicated: it is a grading by a nonabelian group $G(\mathcal{Z})$ rather than by $\mathcal{Z}$, and it depends on a choice of "grading refinement data". To avoid these complications, gradings were not fully treated in [15]; correspondingly, when decategorifying in this paper, we will work with Grothendieck groups defined over $\mathbb{F}_2$ rather than over $\mathbb{Z}$, and we will view $\mathcal{A}(\mathcal{Z})$ as a differential algebra.

Definition 2.5 We let $\mathcal{A}(\mathcal{Z}, k)$ be the $\mathbb{F}_2$–subspace of $\mathcal{A}(\mathcal{Z})$ spanned by strands pictures such that the number of solid strands plus half the number of dotted strands is k. In fact, $\mathcal{A}(\mathcal{Z}, k)$ is a dg subalgebra of $\mathcal{A}(\mathcal{Z})$ (ignoring unit), and if $|\boldsymbol{a}| = 2n$, we have $\mathcal{A}(\mathcal{Z}) = \bigoplus_{k=0}^{n} \mathcal{A}(\mathcal{Z}, k)$.

The basis elements of $\mathcal{A}(\mathcal{Z})$ with only dotted (horizontal) strands are idempotents of $\mathcal{A}(\mathcal{Z})$. Furthermore, for a general basis element a of $\mathcal{A}(\mathcal{Z})$, there is exactly one such idempotent (call it $\lambda(a)$) such that $\lambda(a)a = a$, and for all other such idempotents λ', we have $\lambda'a = 0$. We will refer to $\lambda(a)$ as the left idempotent of a; we can define a right idempotent $\rho(a)$ similarly.

Below we will identify $\mathcal{A}(\mathcal{Z})$ with the differential category whose objects are in bijection with the all-horizontal basis elements of $\mathcal{A}(\mathcal{Z})$, and whose morphism space from e to e' is $e'\mathcal{A}(\mathcal{Z})e$. Because each basis element of $\mathcal{A}(\mathcal{Z})$ has a unique left and right idempotent, we can view these elements as giving a basis for the morphism spaces of $\mathcal{A}(\mathcal{Z})$ as a category.

2.3 Higher actions on strands algebras

Let $\mathcal{Z} = (\boldsymbol{Z}, \boldsymbol{a}, M)$ be an arc diagram. As in [15, Section 7.2.4], we can view $\mathcal{Z}$ as a singular curve Z in the language of that paper, and $\mathcal{A}(\mathcal{Z})$ is the endomorphism algebra of a collection of objects in the strands category $\mathcal{S}(Z)$; see [15, Section 7.4.11]. For an interval I in $\boldsymbol{Z}$ (equivalently, a noncircular component of Z as in [15, Section 7.2.2]), the constructions of [15, Section 8.1.1] give us a differential bimodule E over $\mathcal{A}(\mathcal{Z})$.

Notation 2.6 We will call this bimodule $\mathcal{E}$ rather than E for notational clarity.

Closely related constructions appear in [4], although in that paper the relevant pictures were not explicitly organized into a bimodule over $\mathcal{A}(\mathcal{Z})$.

---

[1]And can be summarized by $\Delta(\partial) = \partial \otimes 1 + 1 \otimes \partial$, at least in characteristic 2, but our view is that in this paper "E" and "∂" are playing very different roles.
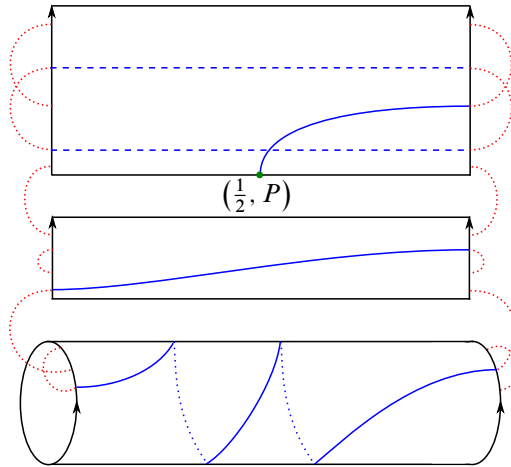
Figure 8: A strands picture for $\mathscr{E}$ (the distinguished interval I is the top interval).

As with the strands algebras, the bimodule $\mathscr{E}$ is defined by specifying an $\mathbb{F}_2$–basis of strands pictures, together with a differential and left and right actions of $\mathcal{A}(\mathcal{Z})$ in terms of basis elements. These strands pictures are almost the same as those described in Definition 2.2. To describe the difference, let P be the endpoint of the interval I such that in the orientation on $\mathbf{Z}$, I points from P to its other endpoint. Then, in a strands picture for $\mathscr{E}$, there should be one solid strand with its left endpoint at $\left(\frac{1}{2}, P\right) \in [0, 1] \times \mathbf{Z}$ and with its right endpoint in $\{1\} \times \mathbf{a}$. See Figure 8; all other rules in Definition 2.2 are unchanged.

Definition 2.7 As an $\mathbb{F}_2$–vector space, $\mathscr{E}$ is defined to be the formal span of the strands pictures described above, which form an $\mathbb{F}_2$–basis for $\mathscr{E}$. The left and right actions of $\mathcal{A}(\mathcal{Z})$ on $\mathscr{E}$, and the differential on $\mathscr{E}$, are defined by concatenation and resolution of crossings as in Definition 2.3. We let $\mathscr{E}(k)$ be the $\mathbb{F}_2$–subspace of $\mathscr{E}$ spanned by strands pictures such that the number of solid strands plus half the number of dotted strands is k; then $\mathscr{E}(k)$ is a differential subbimodule of $\mathscr{E}$, and if $|\mathbf{a}| = 2n$, we have $\mathscr{E} = \bigoplus_{k=1}^{n} \mathscr{E}(k)$. Furthermore, $\mathscr{E}(k)$ is a bimodule over $(\mathcal{A}(\mathcal{Z}, k-1), \mathcal{A}(\mathcal{Z}, k))$ with all other summands of $\mathcal{A}(\mathcal{Z})$ acting as zero on $\mathscr{E}(k)$.

As with the basis elements of $\mathcal{A}(\mathcal{Z})$, to each basis element x of $\mathscr{E}$ we can associate a left idempotent $\lambda(x)$ and a right idempotent $\rho(x)$. We have $x = \lambda(x) x \rho(x)$, while for any other purely horizontal basis elements $\lambda' \neq \lambda(x)$ and $\rho' \neq \rho(x)$ of $\mathcal{A}(\mathcal{Z})$, we have $\lambda' x = 0$ and $x\rho' = 0$.

By [15, Lemma 8.1.2], the bimodule $\mathscr{E} \otimes_{\mathcal{A}(\mathcal{Z})} \mathscr{E} \otimes_{\mathcal{A}(\mathcal{Z})} \cdots \otimes_{\mathcal{A}(\mathcal{Z})} \mathscr{E}$ (with m factors) is isomorphic to the bimodule defined analogously to $\mathscr{E}$, but having solid strands with left endpoints at

$$\left\{ \left(\frac{1}{m+1}, P\right), \left(\frac{2}{m+1}, P\right), \ldots, \left(\frac{m}{m+1}, P\right) \right\}.$$

This bimodule (which we will call $\mathscr{E}^{\otimes m}$) also appears in [4], and as in that paper it admits a left action of NC_m defined diagrammatically by sticking strands pictures for NC_m on the bottom of strands pictures
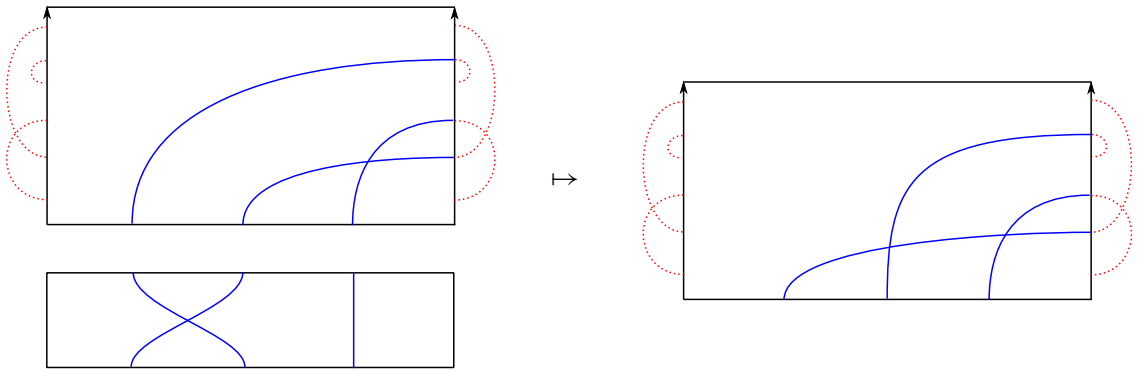
Figure 9: The action of an element of NC_3 on $\mathscr{E}^{\otimes 3}$.

for $\mathscr{E}^{\otimes m}$ (see Figure 9). These actions form a 2–action of $\mathscr{U}$ on $\mathscr{A}(\mathscr{Z})$ via differential bimodules and bimodule maps, which was defined in [15, Proposition 8.1.3]. In other words, they give a differential monoidal functor from $\mathscr{U}$ to the differential monoidal category of differential bimodules over $\mathscr{A}(\mathscr{Z})$ and chain complexes of bimodule maps between them.

2.4 Decategorification

2.4.1 Decategorifying $\mathscr{U}$

Definition 2.8 For a differential category A, we let $\bar{A}$ denote the smallest full differential subcategory of A–Mod (left differential modules over A) containing $\mathrm{Hom}(e, -)$ for all objects e of A and closed under mapping cones and isomorphisms. If A is a dg category, we let A–Mod be the category of left dg modules instead, and require that $\bar{A}$ be closed under degree shifts. We let $H(A)$ denote the homotopy category of A, and we let A^i denote the idempotent completion of A.

Remark 2.9 In the language of bordered Heegaard Floer homology [13; 14], $\bar{A}$ is essentially the same as the differential category of finitely generated bounded type D structures over A (in this setting it is typical to view A as a differential algebra with a distinguished set of idempotents rather than as a dg category).

It is a well-known result (see [9, Corollary 3.7]) that if A is a dg category, then $H(\bar{A})^i$ is equivalent to the full subcategory of the derived category $\mathscr{D}(A)$ (of left dg A–modules) on compact objects, ie the compact derived category of A.

We can view dg algebras such as NC_n as dg categories with one object. Khovanov shows in [11] that the Grothendieck group of the compact derived category of NC_n is zero for $n \geq 2$. For $n = 0$ and $n = 1$, NC_n is $\mathbb{F}_2$, so the Grothendieck group of its compact derived category is $\mathbb{Z}$ (Khovanov gets $\mathbb{Z}[q, q^{-1}]$ instead because he introduces an extra q–grading on NC_n which is identically zero, but we will not use this grading).

Corollary 2.10 *The Grothendieck group $K_0(H(\overline{NC}_n))$ is also $\mathbb{Z}$ for $n \in \{0, 1\}$ and is zero for $n \geq 2$, where $H(\overline{NC}_n)$ is the homotopy category of $\overline{NC}_n$.*

Proof The inclusion of the triangulated category $H(\overline{NC}_n)$ into its idempotent completion is a monomorphism by [22, Corollary 2.3]. In fact, by [22, Theorem 2.1], $H(\overline{NC}_n)$ is already idempotent complete. □

Since we will primarily work with Grothendieck groups over $\mathbb{F}_2$ here, we introduce the following definition.

Definition 2.11 Let $\mathscr{C}$ be a category equipped with a collection of distinguished triangles $X \to Y \to Z \rightsquigarrow$ as in a triangulated category (but we do not require $\mathscr{C}$ to be triangulated or even to have a shift functor; we place no requirements on the collection of distinguished triangles). We let $K_0^{\mathbb{F}_2}(\mathscr{C})$ be the $\mathbb{F}_2$–vector space with basis given by isomorphism classes of objects of $\mathscr{C}$ modulo relations $[X] + [Y] + [Z] = 0$ whenever there exists a distinguished triangle $X \to Y \to Z \rightsquigarrow$.

For a triangulated category $\mathscr{C}$, the above definition agrees with $K_0(\mathscr{C}) \otimes \mathbb{F}_2$. We see that $K_0^{\mathbb{F}_2}(H(\overline{NC}_n))$ is isomorphic to $\mathbb{F}_2$ for $n \in \{0, 1\}$ and is zero otherwise.

Now, since $\mathscr{U}$ is a direct sum of NC_n (as a one-object dg category) over all $n \geq 0$, $K_0^{\mathbb{F}_2}(H(\overline{\mathscr{U}})) \cong \mathbb{F}_2 \oplus \mathbb{F}_2$. For notational convenience, we let

$$K_0^{\mathbb{F}_2}(\mathscr{U}) := K_0^{\mathbb{F}_2}(H(\overline{\mathscr{U}})).$$

Taking the monoidal structure on $\mathscr{U}$ into account, we see that as an $\mathbb{F}_2$–algebra,

$$K_0^{\mathbb{F}_2}(\mathscr{U}) \cong \mathbb{F}_2[E]/(E^2)$$

(this is Khovanov's identification $K_0(H^-) \cong \mathbb{Z}[q, q^{-1}, E_1]/(E_1^2)$ from [11], adapted to our setting).

2.4.2 Decategorifying the strands algebras As mentioned above, we will view the strands algebras $\mathscr{A}(\mathscr{Z})$ as differential categories with multiple (but finitely many) objects in bijection with the set of purely horizontal strands pictures for $\mathscr{Z}$. The homotopy category $H(\overline{\mathscr{A}(\mathscr{Z})})$ has a collection of distinguished triangles, namely those isomorphic to the image in the homotopy category of $X \xrightarrow{f} Y \to \mathrm{Cone}(f) \rightsquigarrow$ for some closed morphism $f \colon X \to Y$ in $\overline{\mathscr{A}(\mathscr{Z})}$.

Recall that the construction of a sutured surface (F, S_+, S_-, Λ) from an arc diagram $\mathscr{Z} = (Z, a, \Lambda)$ starts by taking $Z \times [0, 1]$, a collection of rectangles and annuli, and gluing on some 2–dimensional 1–handles. For each pair of points $\{p, q\}$ of a matched by M, one glues on a 1–handle with attaching zero-sphere $\{(p, 1), (q, 1)\}$ compatibly with the orientation on Z. The result is F; one sets $S_+ := Z \times \{0\}$ and $\Lambda := (\partial Z) \times \{0\}$, with the rest of the boundary of F placed in S_-.

Proposition 2.12 [21] *For $\mathscr{Z} = (Z, a, M)$ with Z a single interval, $K_0(H(\overline{\mathscr{A}(\mathscr{Z})}))$ is isomorphic to $\wedge^* H_1(F; \mathbb{Z})$ where F is the surface represented by $\mathscr{Z}$. Specifically, for each k, $K_0(H(\overline{\mathscr{A}(\mathscr{Z}, k)}))$ is isomorphic to $\wedge^k H_1(F; \mathbb{Z})$.*

It follows that $K_0^{\mathbb{F}_2}(H(\overline{\mathcal{A}(\mathcal{Z})}))$ is isomorphic to $\wedge^* H_1(F; \mathbb{F}_2)$, and in the $\mathbb{F}_2$ setting we do not need to consider Petkova's absolute $\mathbb{Z}/2\mathbb{Z}$ homological grading on $\mathcal{A}(\mathcal{Z})$.

Remark 2.13 Petkova views the surface F associated to a one-interval arc diagram $\mathcal{Z}$ as being closed, while we view it as having S^1 boundary with one S_+ interval and one S_- interval. Letting $\overline{F}$ denote the closed surface and F denote the surface with boundary, we have natural identifications

$$H_1(\overline{F}) \cong H_1(F) \cong H_1(F, S_+)$$

(with either $\mathbb{Z}$ or $\mathbb{F}_2$ coefficients).

Petkova's arguments readily generalize to show that for general $\mathcal{Z}$ as defined above, $K_0^{\mathbb{F}_2}(H(\overline{\mathcal{A}(\mathcal{Z})}))$ has an $\mathbb{F}_2$–basis given by the set of objects of $\mathcal{A}(\mathcal{Z})$ as a dg category, ie by the purely horizontal strands pictures for $\mathcal{Z}$.

Proposition 2.14 *If (F, S_+, S_-, Λ) is the sutured surface represented by a general arc diagram $\mathcal{Z}$, then the vector space $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ has a basis in bijection with purely horizontal strands pictures for $\mathcal{Z}$.*

Proof It follows from the construction of (F, S_+, S_-, Λ) that F/S_+ is homotopy equivalent to a wedge product of circles, one for each pair of points of $\boldsymbol{a}$, and these circles form a basis for $H_1(F, S_+; \mathbb{F}_2)$. A basis for $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ is then given by all subsets of the set of these circles. For each such subset X, there is a corresponding purely horizontal strands picture for $\mathcal{Z}$; if a circle (corresponding to $\{p, q\}$ matched by M) is in X, one draws a pair of dotted horizontal strands at p and q in the strands picture. This correspondence is a bijection, proving the proposition. $\square$

Let $K_0^{\mathbb{F}_2}(\mathcal{A}(\mathcal{Z})) := K_0^{\mathbb{F}_2}(H(\overline{\mathcal{A}(\mathcal{Z})}))$ and $K_0^{\mathbb{F}_2}(\mathcal{A}(\mathcal{Z}, k)) := K_0^{\mathbb{F}_2}(H(\overline{\mathcal{A}(\mathcal{Z}, k)}))$.

Corollary 2.15 *We have natural identifications*

$$K_0^{\mathbb{F}_2}(\mathcal{A}(\mathcal{Z})) \cong \wedge^* H_1(F, S_+; \mathbb{F}_2) \quad \text{and} \quad K_0^{\mathbb{F}_2}(\mathcal{A}(\mathcal{Z}, k)) \cong \wedge^k H_1(F, S_+; \mathbb{F}_2).$$

3 Actions on exterior powers of homology

Let $\mathcal{Z} = (\boldsymbol{Z}, \boldsymbol{a}, M)$ be an arc diagram representing a sutured surface (F, S_+, S_-, Λ) as in Figure 10, and let I be an interval component of S_+ (equivalently, let I be an interval component of $\boldsymbol{Z}$). The endomorphism Φ_I of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ defined in the introduction squares to zero and thus gives us an action of $\mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ in which E acts by Φ_I. In this section we identify this action with the action of $K_0^{\mathbb{F}_2}(\mathcal{U})$ on $K_0^{\mathbb{F}_2}(\mathcal{A}(\mathcal{Z}))$ coming from the 2–action of $\mathcal{U}$ on $\mathcal{A}(\mathcal{Z})$ described in Section 2.3.
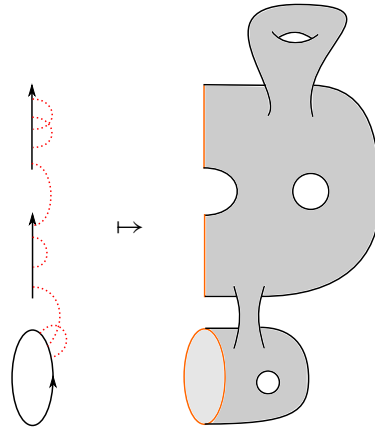
Figure 10: An arc diagram and the sutured surface it represents. The S_+ portion of the surface boundary is drawn in orange and the S_- portion is drawn in black.

Remark 3.1 For an element ω of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ that is a pure wedge product of arcs in F with boundary on S_+ and/or circles in F, we can depict ω by drawing all the arcs and circles of ω in a picture of F. See Figure 11 for an example. The element E of $\mathbb{F}_2[E]/(E^2)$ acts on this depiction of ω by summing over all ways of removing one arc incident with the component I of S_+; see Figure 12. An arc with both endpoints on I is "removed twice" which, in the sum with $\mathbb{F}_2$ coefficients, amounts to not being removed at all; indeed, such an arc represents the same homology class as a circle with no endpoints.

We first review an important structural property of the bimodule $\mathcal{E}$ from Section 2.3; the proposition below follows from [15, Section 8.1.4], but to keep this paper self-contained we include an independent proof.

Proposition 3.2 *As a left differential module over the differential category* $\mathcal{A}(\mathcal{Z})$, $\mathcal{E}$ *is an object of* $\overline{\mathcal{A}(\mathcal{Z})}$.
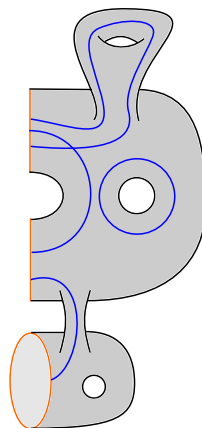
Figure 11: Depiction of a pure wedge-product element of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$.
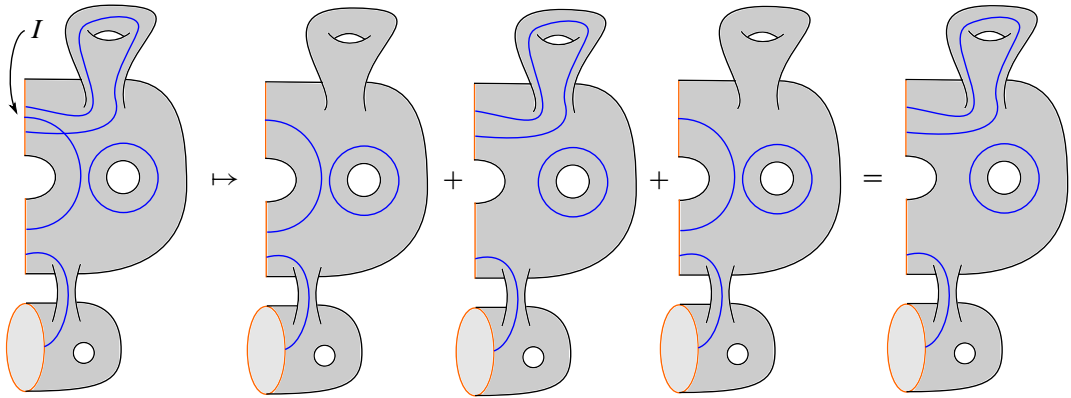
Figure 12: Action of $E \in \mathbb{F}_2[E]/(E^2)$ on $\omega \in \wedge^* H_1(F, S_+; \mathbb{F}_2)$ given a distinguished interval I of S_+.

Proof We first show that as a left module (disregarding the differential), $\mathscr{E}$ is isomorphic to a direct sum of modules of the form $\mathrm{Hom}(e, -)$ for objects e of $\mathcal{A}(\mathcal{Z})$. Indeed, consider the subset S of strands pictures for $\mathscr{E}$ (ie $\mathbb{F}_2$–basis elements of $\mathscr{E}$) such that the only moving strand is the one with left endpoint at $\left(\frac{1}{2}, P\right)$ in the language of Section 2.3. See Figure 13 for an example of an element of S. An arbitrary basis element x of $\mathscr{E}$ can be written as ay for unique basis elements $a \in \mathcal{A}(\mathcal{Z})$ and $y \in S$; indeed, after a homotopy relative to the endpoints, we can draw x such that all strands of x except the one with endpoint at $\left(\frac{1}{2}, P\right)$ only move on $\mathbf{Z} \times [0, \varepsilon]$ for some $\varepsilon < \frac{1}{2}$, and are horizontal on $\mathbf{Z} \times [\varepsilon, 1]$ (see Figure 14).

Cutting the diagram for x at $\mathbf{Z} \times \{\varepsilon\}$, we see a strands picture for a basis element $a \in \mathcal{A}(\mathcal{Z})$ on the left. On the right side of the cut, let y be the element of S obtained by making all the horizontal strands dotted and adding in their matching horizontal strands (according to the matching M). See Figure 15 for an example. We have $ay = x$; furthermore, for any $y \in S$ with left idempotent $\lambda(y)$, and any basis element
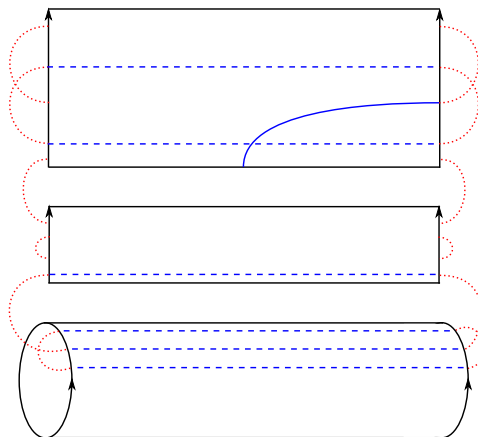
Figure 13: An element of the set S of special basis elements of $\mathscr{E}$.
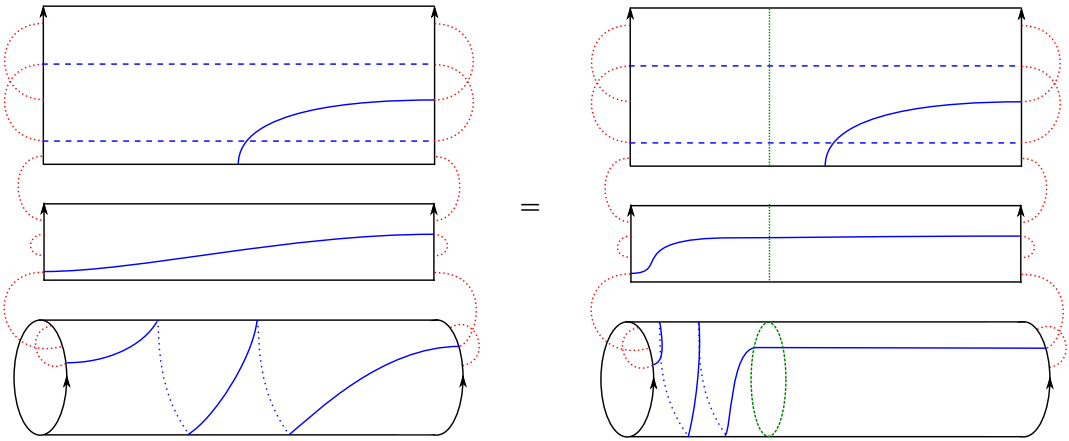
Figure 14: Stretching the basis element x of Figure 8 so that all "ordinary" moving strands only move on $\mathbf{Z} \times [0, \varepsilon]$; the green dashed lines on the right indicate where we will cut to factor x as ay with $a \in \mathcal{A}(\mathcal{Z})$ and $y \in S$.

a of $\mathrm{Hom}_{\mathcal{A}(\mathcal{Z})}(\lambda(y), -)$, we have that ay is a basis element for $\mathcal{E}$ and that a and y are recovered when splitting ay as above.

We have defined a bijection between our basis for $\mathcal{E}$ and the set of pairs (a, y) where y is an element of S with left idempotent $\lambda(y)$ and a is a basis element of $\mathrm{Hom}_{\mathcal{A}(\mathcal{Z})}(\lambda(y), -)$. Thus, we have an identification of $\mathcal{E}$ with $\bigoplus_{y \in S} \mathrm{Hom}_{\mathcal{A}(\mathcal{Z})}(\lambda(y), -)$ as vector spaces. This identification respects left multiplication by $\mathcal{A}(\mathcal{Z})$, so

$$\mathcal{E} \cong \bigoplus_{y \in S} \mathrm{Hom}_{\mathcal{A}(\mathcal{Z})}(\lambda(y), -)$$

as left modules over $\mathcal{A}(\mathcal{Z})$ (ignoring the differential).
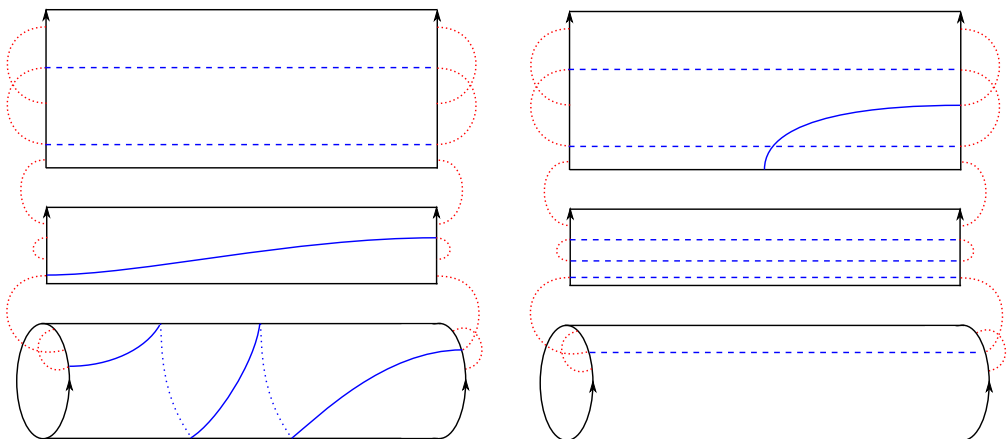
Figure 15: Factorizing the basis element x of Figure 8 as $a \in \mathcal{A}(\mathcal{Z})$ (left) times $y \in S$ (right).

Now, we can define a grading on the elements of S: say $y \in S$ has degree d if the moving strand σ of y with left endpoint $(\frac{1}{2}, P)$ encounters d points of $\boldsymbol{a}$ while traveling along a minimal path in $\boldsymbol{Z}$ from P to its right endpoint. Order the elements of S by increasing degree (choose any ordering of the elements of S in each given degree). Because the differential on $\mathscr{E}$, applied to $y \in S$, will only resolve crossings between the special strand σ of y and horizontal strands strictly below σ, the only nonzero terms of this differential will be of the form ay' for y' of degree strictly less than that of y (and thus y' that appear before y in the ordering on S). It follows that $\mathscr{E}$ is isomorphic to an iterated mapping cone built from $\mathrm{Hom}_{\mathscr{A}(\mathscr{Z})}(\lambda(y), -)$ for $y \in S$, so we have $\mathscr{E} \in \overline{\mathscr{A}(\mathscr{Z})}$. $\qquad \square$

Remark 3.3 In the language of bordered Heegaard Floer homology, Proposition 3.2 says that $\mathscr{E}$ is the differential bimodule associated to a finitely generated left bounded type DA bimodule over $\mathscr{A}(\mathscr{Z})$ with δ_i^1 zero for $i > 2$.

Proposition 3.2 gives us the following corollary.

Corollary 3.4 *We have a differential functor $\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} -$ from $\overline{\mathscr{A}(\mathscr{Z})}$ to itself, and thus a functor $\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} -$ from $H(\overline{\mathscr{A}(\mathscr{Z})})$ to itself.*

Proof Let $\mathscr{E} \cong \oplus_\alpha \mathscr{A}(\mathscr{Z}) \cdot e_\alpha$ (as a left module) and suppose we have $X \cong \oplus_\beta \mathscr{A}(\mathscr{Z}) \cdot x_\beta \in \overline{\mathscr{A}(\mathscr{Z})}$, where e_α and x_β are distinguished idempotents of $\mathscr{A}(\mathscr{Z})$, the sums over α and β are finite, for all (α, β) we have $e_\alpha \cdot' x_\beta \in \{e_\alpha, 0\}$ where $\cdot'$ denotes the right action of $\mathscr{A}(\mathscr{Z})$ on $\mathscr{E}$ (the proof of Proposition 3.2 implies this is possible), and there exist orderings of the α and β such that the differentials on $\mathscr{E}$ and X are strictly decreasing with respect to the order. Then

$$\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} X \cong \bigoplus_\beta \mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} (\mathscr{A}(\mathscr{Z}) \cdot x_\beta) \cong \bigoplus_\beta \mathscr{E} \cdot' x_\beta \cong \bigoplus_{\alpha, \beta} \mathscr{A}(\mathscr{Z}) \cdot (e_\alpha \cdot' x_\beta).$$

If we order the pairs (α, β) lexicographically such that the β coordinate dominates, then the differential on $\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} X$ is strictly decreasing with respect to the order. It follows that $\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} X \in \overline{\mathscr{A}(\mathscr{Z})}$; it is then a standard fact that $\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} -$ gives a differential endofunctor of $\overline{\mathscr{A}(\mathscr{Z})}$. $\qquad \square$

The differential functor $\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} -$ sends mapping cones to mapping cones, so the corresponding functor on homotopy categories sends distinguished triangles to distinguished triangles and thus induces an endomorphism $[\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} -]$ of $K_0^{\mathbb{F}_2}(\mathscr{A}(\mathscr{Z}))$.

Theorem 3.5 *Let $\mathscr{Z} = (\boldsymbol{Z}, \boldsymbol{a}, M)$ be an arc diagram and let (F, S_+, S_-, Λ) be the sutured surface represented by $\mathscr{Z}$. Let I be an interval component of S_+, or equivalently an interval component of $\boldsymbol{Z}$. Under the identification $K_0^{\mathbb{F}_2}(\mathscr{A}(\mathscr{Z})) \cong \wedge^* H_1(F, S_+; \mathbb{F}_2)$ from Corollary 2.15, the endomorphism $[\mathscr{E} \otimes_{\mathscr{A}(\mathscr{Z})} -]$ of $K_0^{\mathbb{F}_2}(\mathscr{A}(\mathscr{Z}))$ agrees with the endomorphism Φ_I of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ from the introduction. More specifically, the map $[\mathscr{E}(k) \otimes_{\mathscr{A}(\mathscr{Z}, k)} -]$ from $K_0^{\mathbb{F}_2}(\mathscr{A}(\mathscr{Z}), k)$ to $K_0^{\mathbb{F}_2}(\mathscr{A}(\mathscr{Z}), k-1)$ agrees with Φ_I as a map from $\wedge^k H_1(F, S_+; \mathbb{F}_2)$ to $\wedge^{k-1} H_1(F, S_+; \mathbb{F}_2)$.*

Proof Let e be an object of $\mathcal{A}(\mathcal{Z})$ (viewed as a differential category); we have a corresponding basis element $[\mathrm{Hom}(e, -)]$ of $K_0^{\mathbb{F}_2}(\mathcal{A}(\mathcal{Z}))$. Applying $[\mathcal{E} \otimes_{\mathcal{A}(\mathcal{Z})} -]$ to $[\mathrm{Hom}(e, -)]$, we get

$$\sum_{y \in S, \rho(y)=e} [\mathrm{Hom}(\lambda(y), -)].$$

Viewing e as a purely horizontal strands picture and defining S as in the proof of Proposition 3.2, there is one element $y_s \in S$ with $\rho(y_s) = e$ for each strand s of e with endpoints in the interval I, and these are all the elements $y \in S$ with $\rho(y) = e$. For each such strand s (say with endpoints at $Q \in I$), the element y_s has a moving strand between $(\frac{1}{2}, P)$ and $(1, Q)$, and has the same horizontal strands as e except for s and its partner s' under the matching. Thus, $\lambda(y_s)$ is e with the strands s and s' removed.

It follows that $[\mathcal{E} \otimes_{\mathcal{A}(\mathcal{Z})} -]([\mathrm{Hom}(e, -)])$ is the sum of $[\mathrm{Hom}(e', -)]$ over all e' obtained from e by choosing one strand s in $[0, 1] \times I$ and removing both s and its partner s'. In particular, for strands s in $[0, 1] \times I$ such that s' is also in $[0, 1] \times I$, the pair of strands (s, s') is removed from e twice, and since we are working over $\mathbb{F}_2$, removals of these strands contribute zero to $[\mathcal{E} \otimes_{\mathcal{A}(\mathcal{Z})} -]([\mathrm{Hom}(e, -)])$.

Now let ω be the element of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ corresponding to $[\mathrm{Hom}(e, -)]$ under the isomorphism of Corollary 2.15. Concretely, each pair of matched strands $\{s, s'\}$ of e gives a basis element of $H_1(F, S_+; \mathbb{F}_2)$, and ω is the wedge product of these elements over all such pairs $\{s, s'\}$. When we apply Φ_I to ω, we sum over all ways to remove a factor from this wedge product if the factor maps to $1 \in \mathbb{F}_2$ under the map ϕ_I from the introduction. Such factors are those corresponding to pairs of strands $\{s, s'\}$ of e in which one of $\{s, s'\}$, but not both, is in $[0, 1] \times I$. It follows that $\Phi_I(\omega)$ corresponds to $[\mathcal{E} \otimes_{\mathcal{A}(\mathcal{Z})} -]([\mathrm{Hom}(e, -)])$, as desired. $\qquad\square$

4 Gluing and TQFT

In this section, we prove (a slightly more general version of) Theorem 1.3 from the introduction. Let (F, S_+, S_-, Λ) be a sutured surface and suppose that $I_1 \neq I_2$ are interval components of S_+. Up to homeomorphism, there is a unique way to glue I_1 to I_2 and get an oriented surface $\overline{F}$. There are naturally defined subsets $\overline{S}_+$ and $\overline{S}_-$ of the boundary of $\overline{F}$, intersecting in a set of points $\overline{\Lambda}$ (which is Λ with the endpoints of I_1 and I_2 removed).

Lemma 4.1 *We have an isomorphism*

$$\wedge^* H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2) \cong (\wedge^* H_1(F, S_+; \mathbb{F}_2)) \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes 2}} \frac{\mathbb{F}_2[E]}{(E^2)},$$

where the action of $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ comes from the $\mathbb{F}_2[E]/(E^2)$ actions associated to I_1 and I_2, and the action of $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ on $\mathbb{F}_2[E]/(E^2)$ comes from multiplication. We can choose the isomorphism so that it intertwines the remaining actions of $\mathbb{F}_2[E]/(E^2)$ from S_+ intervals other than I_1 or I_2.
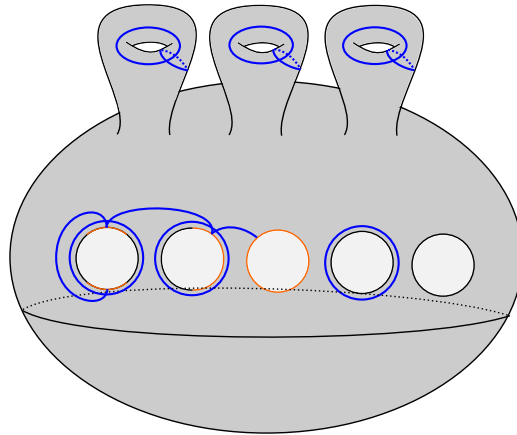
Figure 16: A standard model for a sutured surface, given by a sphere with some number of tori connect-summed on, as well as some number of disks removed and some even number of sutures on each boundary component. The S_+ boundary is drawn in orange and the S_- boundary is drawn in black. The set of blue arcs and circles gives a basis for $H_1(F, S_+; \mathbb{F}_2)$.

Proof Pick a homeomorphism between F and a finite disjoint union of standard sutured surfaces as shown in Figure 16 (spheres with some number of open disks removed and some even number of sutures on each boundary component, connect-summed with some number of tori). Figure 16 also indicates, with blue arcs and circles, a way to choose bases for $H_1(F, S_+; \mathbb{F}_2)$. One chooses

- for each torus that was connect-summed on, two circles giving a basis for the first homology of the torus;

- for all but one of the boundary components intersecting S_- nontrivially, a circle around the boundary component;

- a continuous map from a connected acyclic graph Γ_F to the surface F (an embedding on each edge of Γ_F) with one vertex on each component of S_+ — we will identify Γ_F with its image in F.

These circles, together with the edges of Γ_F, give a basis for $H_1(F, S_+; \mathbb{F}_2)$, so subsets of this set of arcs and circles give a basis for $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ consisting of wedge products of basis elements of $H_1(F, S_+; \mathbb{F}_2)$.

Now suppose I_1 and I_2 are intervals of S_+; we consider various cases.

Case 1 First, assume I_1 and I_2 live on distinct connected components of F. Choose Γ_F such that the vertices on I_1 and I_2 (say p_1 and p_2) are leaves of Γ_F, ie they have degree 1. When gluing F to get $\overline{F}$, we can ensure that p_1 and p_2 are glued to each other. If we let e_1 and e_2 denote the edges incident with p_1 and p_2, and modify Γ_F by removing p_1, p_2, e_1 and e_2 while adding the edge $e_1 \cup e_2$ as an embedded arc in $\overline{F}$, we get an acyclic graph $\Gamma_{\overline{F}}$ embedded in $\overline{F}$ with one vertex on each component of $\overline{S}_+$. See Figure 17.
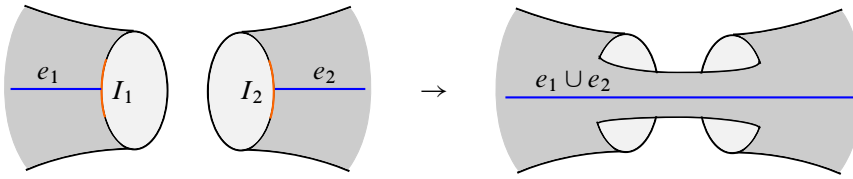
Figure 17: Left: arcs e_1 and e_2 in the surface F before gluing. Right: the arc $e_1 \cup e_2$ after gluing I_1 to I_2.

For an element ω of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$, obtained as a wedge product of basis elements of $H_1(F, S_+; \mathbb{F}_2)$, the I_1–action of $E \in \mathbb{F}_2[E]/(E^2)$ on ω is zero if e_1 is not a wedge factor of ω. Otherwise, write $\omega = e_1 \wedge \omega'$; we have $E \cdot \omega = \omega'$.

The I_2–action of $E \in \mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ is similar; informally, E acts by "removing e_2". It follows that $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ is a free module over $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ with an $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$–basis given by elements $e_1 \wedge e_2 \wedge \omega'$ for all wedge products ω' in the other basis elements (not e_1 or e_2) of $H_1(F, S_+; \mathbb{F}_2)$. Thus, a basis for

$$(\wedge^* H_1(F, S_+; \mathbb{F}_2)) \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes 2}} \frac{\mathbb{F}_2[E]}{(E^2)}$$

is given by the set of elements $e_1 \wedge e_2 \wedge \omega'$, together with the elements $e_1 \wedge \omega' = e_2 \wedge \omega'$ (in each case ω' is a wedge product of basis elements of $H_1(F, S_+; \mathbb{F}_2)$ that are not e_1 or e_2). Meanwhile, a basis for $\wedge^* H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2)$ is given by the set of elements $(e_1 \cup e_2) \wedge \omega'$ and ω' for the same set of ω'. We have a bijection between basis elements given by $e_1 \wedge e_2 \wedge \omega' \leftrightarrow (e_1 \cup e_2) \wedge \omega'$ and $(e_1 \wedge \omega' = e_2 \wedge \omega') \leftrightarrow \omega'$; this bijection is illustrated in Figure 18. Thus, we have an isomorphism of vector spaces as claimed in the statement of the theorem.

To see that this isomorphism intertwines the remaining actions of $\mathbb{F}_2[E]/(E^2)$ for S_+ intervals that are not I_1 or I_2, it suffices to consider the actions for the other two intervals (say I'_1 and I'_2) that intersect e_1 and e_2 respectively. We will consider the action for I'_1; the case of I'_2 is similar. In the terminology used above, there are four types of basis elements of $\wedge^* H_1(F, S_+; \mathbb{F}_2)$: those of the forms $e_1 \wedge e_2 \wedge \omega'$,
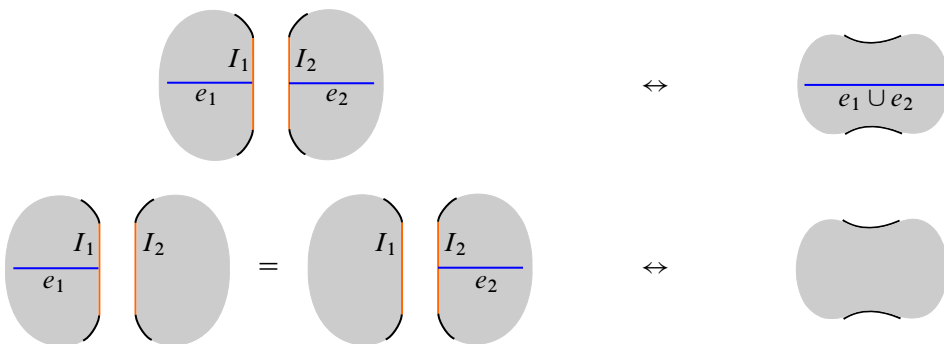
Figure 18: The bijection on basis elements in the first case of Lemma 4.1.
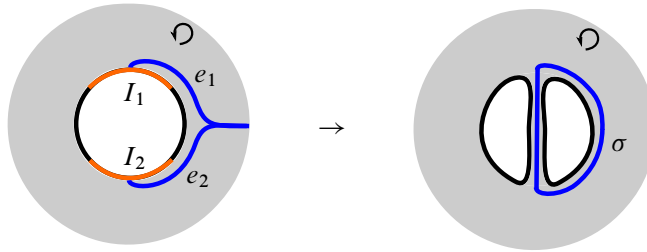
Figure 19: Left: local model near C for the arcs e_1 and e_2. Right: the circle σ after gluing I_1 to I_2. In both cases the curved arrow indicates the orientation on F; the induced boundary orientation on C is clockwise in this figure.

$e_1 \wedge \omega'$, $e_2 \wedge \omega'$, and ω'. The I_1'–action of $E \in \mathbb{F}_2[E]/(E^2)$ sums over all ways to remove one wedge factor corresponding to an arc with exactly one endpoint on I_1'; besides terms that modify ω', there is a "remove e_1" term that sends $e_1 \wedge e_2 \wedge \omega'$ to $e_2 \wedge \omega'$ and sends $e_1 \wedge \omega'$ to ω'. When we tensor over $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ with the identity map on $\mathbb{F}_2[E]/(E^2)$, the "remove e_1" term of the action of E sends $e_1 \wedge e_2 \wedge \omega'$ to $e_2 \wedge \omega' = e_1 \wedge \omega'$ and sends $e_1 \wedge \omega' = e_2 \wedge \omega'$ to zero. On the other hand, as above there are two types of basis elements of $\wedge^* H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2)$: those of the form $(e_1 \cup e_2) \wedge \omega'$ and those of the form ω'. The I_1'–action of E has terms modifying ω' in the same way as above, and it also has "remove $e_1 \cup e_2$" terms sending $(e_1 \cup e_2) \wedge \omega'$ to ω' and sending ω' to zero. It follows that our choice of isomorphism intertwines the I_1' action of $\mathbb{F}_2[E]/(E^2)$.

Case 2   Next, assume I_1 and I_2 live on the same connected component F' of F; without loss of generality we can assume F is connected so that $F' = F$. We consider two further cases: either I_1 and I_2 live on the same connected component of ∂F, or they live on different connected components of ∂F.

Case 2-1   First assume I_1 and I_2 live on the same component C of ∂F, so that gluing I_1 to I_2 increases the number of boundary components of F by one while keeping the genus the same. When choosing a basis for $H_1(F, S_+; \mathbb{F}_2)$ as above, we can choose C for the unique not-fully S_+ boundary component of F that does not get a circle around it. We can also ensure that in the acyclic graph Γ_F, the vertices p_1 on I_1 and p_2 on I_2 are leaves of Γ_F.

Case 2-1a   If there are any intervals of S_+ other than I_1 and I_2, or any fully S_+ circles, then p_1 and p_2 are incident with distinct edges $e_1 \neq e_2$ of Γ_F; we can furthermore choose Γ_F so that e_1 and e_2 share an endpoint q, and such that as embedded submanifolds of F, they look like the left side of Figure 19 in a small neighborhood of C and are identical outside this neighborhood (the picture should be appropriately modified if q lives on the circle C). As above, $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ is free over $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ and has four types of basis elements, namely $e_1 \wedge e_2 \wedge \omega'$, $e_1 \wedge \omega'$, $e_2 \wedge \omega'$, and ω'. A basis for

$$(\wedge^* H_1(F, S_+; \mathbb{F}_2)) \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes 2}} \frac{\mathbb{F}_2[E]}{(E^2)}$$

is given by the elements $e_1 \wedge e_2 \wedge \omega'$ along with the elements $e_1 \wedge \omega' = e_2 \wedge \omega'$. Meanwhile, we can take $\Gamma_{\overline{F}}$ to be Γ_F with the edges e_1 and e_2 removed, and when choosing circles around boundary components
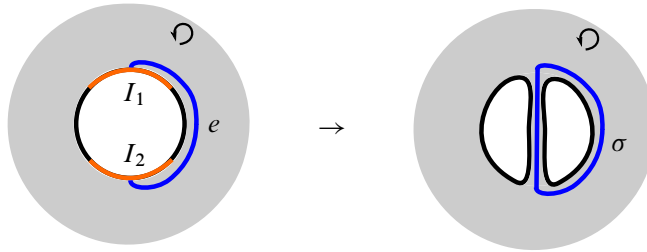
Figure 20: Left: local model near C for the arc e. Right: the circle σ after gluing I_1 to I_2.

to assemble a basis for $H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2)$, we can put a circle σ around the component of $\partial \overline{F}$ containing the segment of ∂F that goes from I_1 to I_2 when traversing the boundary in the oriented direction (see the right side of Figure 19). Then $\wedge^* H_1(\overline{F}, \overline{S}_+, \mathbb{F}_2)$ has basis elements of type $\sigma \wedge \omega'$ and ω'; we identify these with elements of type $e_1 \wedge e_2 \wedge \omega'$ and $e_1 \wedge \omega' = e_2 \wedge \omega'$ respectively. This bijection on basis elements gives us an isomorphism of vector spaces as in the statement of the theorem.

To see that this isomorphism intertwines the remaining actions of $\mathbb{F}_2[E]/(E^2)$ from S_+ intervals other than I_1 or I_2, it suffices to consider the interval I that contains the common endpoint q of e_1 and e_2. The I–action of $E \in \mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ has terms that modify ω' as well as "remove e_1" terms sending (for example) $e_1 \wedge e_2 \wedge \omega'$ to $e_2 \wedge \omega'$ and "remove e_2" terms sending (for example) $e_1 \wedge e_2 \wedge \omega'$ to $e_1 \wedge \omega'$. When we tensor over $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ with the identity map on $\mathbb{F}_2[E]/(E^2)$, both the "remove e_1" and the "remove e_2" terms send $e_1 \wedge e_2 \wedge \omega'$ to $e_1 \wedge \omega' = e_2 \wedge \omega'$, and they send $e_1 \wedge \omega' = e_2 \wedge \omega'$ to zero. Since the "remove e_1" and "remove e_2" terms act in the same way, their contribution to the overall action of E is zero, and only the "modify ω'" terms remain. On the other hand, the I–action of E on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ only modifies ω' in terms of type $\sigma \wedge \omega'$ or ω', since σ is closed. It follows that our choice of isomorphism intertwines the I–action of $\mathbb{F}_2[E]/(E^2)$.

Case 2-1b Now assume that I_1 and I_2 are the only intervals of S_+ (but they still live on the same component C of ∂F) and that there are no fully S_+ circles; it follows that Γ_F has a unique edge e and it connects p_1 to p_2. We can assume e lives in a small neighborhood of C, and that in this neighborhood it looks like the left side of Figure 20. The I_1–action and I_2–action of $E \in \mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ agree; they both send $e \wedge \omega'$ to ω' and send ω' to zero. Thus

$$(\wedge^* H_1(F, S_+; \mathbb{F}_2)) \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes 2}} \frac{\mathbb{F}_2[E]}{(E^2)}$$

is canonically isomorphic to $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ where no tensor operation is performed. Meanwhile, we can take $\Gamma_{\overline{F}}$ to be empty, but in assembling a basis for $H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2)$, we again put a circle σ around the component of $\partial \overline{F}$ containing the segment of ∂F that goes from I_1 to I_2 when traversing the boundary in the oriented direction (see the right side of Figure 20). The correspondences $e \wedge \omega' \leftrightarrow \sigma \wedge \omega'$ and $\omega' \leftrightarrow \omega'$ give an isomorphism of vector spaces as in the statement of the theorem. There are no remaining S_+ intervals, so we do not need to check that this isomorphism intertwines any actions.
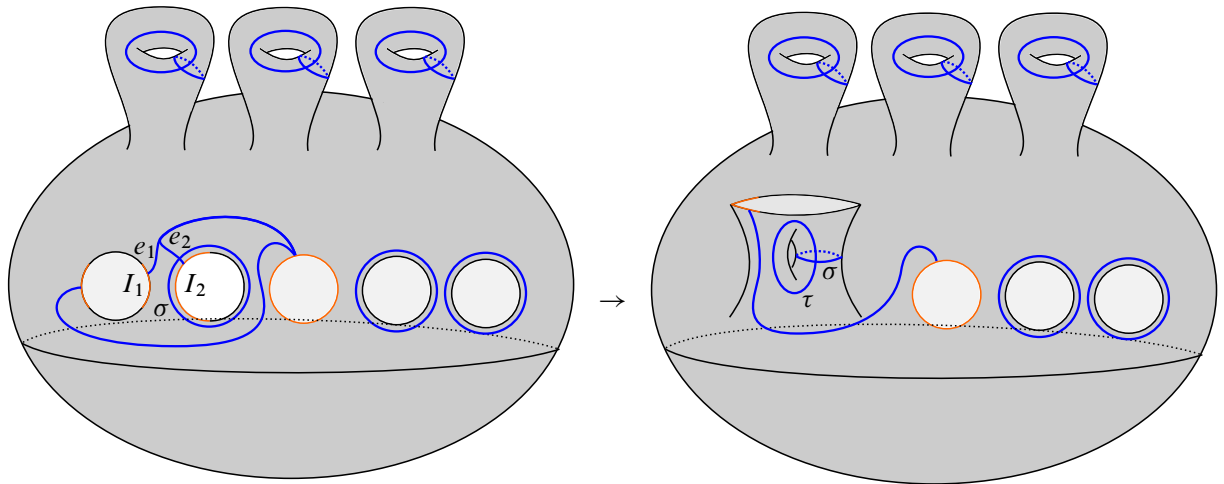
Figure 21: Left: F before gluing intervals I_1 and I_2 on the same component of F but different components of ∂F. Right: the glued surface $\overline{F}$.

Case 2-2 Next, assume that I_1 and I_2 live on different components C_1 and C_2 of ∂F; for visual simplicity, assume that in the model for F shown in Figure 16, C_1 and C_2 are next to each other. Gluing I_1 to I_2 decreases the number of boundary components of F by one and increases the genus of F by one.

Case 2-2a Also assume that there is either at least one S_+ interval that is not I_1 or I_2, or that there is at least one fully S_+ circle. As above, p_1 and p_2 are incident with distinct edges $e_1 \neq e_2$ of Γ_F, and we can choose Γ_F so that e_1 and e_2 share a vertex q and only diverge near C_1 and C_2. We also assume that C_1 is the unique not-fully S_+ boundary circle of F that does not get a circle around it as a basis element of $H_1(F, S_+; \mathbb{F}_2)$. Let σ be the circle around C_2; see the left side of Figure 21.

Basis elements for $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ can be of the form $e_1 \wedge e_2 \wedge \omega'$, $e_1 \wedge \omega'$, $e_2 \wedge \omega'$, or ω'; when we tensor with $\mathbb{F}_2[E]/(E^2)$ over $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$, we have a basis whose elements are of type $e_1 \wedge e_2 \wedge \omega'$ or $e_1 \wedge \omega' = e_2 \wedge \omega'$. Meanwhile, we choose a basis for $H_1(\overline{F}, \overline{S}_+, \mathbb{F}_2)$ by choosing a homeomorphism with the standard surface shown on the right side of Figure 21. The graph $\Gamma_{\overline{F}}$ can be understood as Γ_F with e_1 and e_2 removed; we also have basis elements σ and τ of $H_1(F, S_+; \mathbb{F}_2)$ where $\sigma \subset \overline{F}$ comes from $\sigma \subset F$ and τ comes from e_1 and e_2. Basis elements of $\wedge^* H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2)$ are of the form $\tau \wedge \omega'$ or ω', where ω' is a wedge product of basis elements for $H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2)$ that are not τ. The correspondence $e_1 \wedge e_2 \wedge \omega' \leftrightarrow \tau \wedge \omega'$ and $(e_1 \wedge \omega' = e_2 \wedge \omega') \leftrightarrow \omega'$ gives an isomorphism of vector spaces as in the statement of the theorem. The proof that this isomorphism intertwines the remaining actions of $\mathbb{F}_2[E]/(E^2)$ proceeds as above.

Case 2-2b Finally, assume that I_1 and I_2 are the only S_+ intervals and that there are no fully S_+ circles (while I_1 and I_2 still live on different components of ∂F). Letting e be the arc of Γ_F connecting $p_1 \in I_1$ to $p_2 \in I_2$, basis elements for $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ are of the form $e \wedge \omega'$ or ω'. Meanwhile, defining τ as in Figure 21, basis elements for $\wedge^* H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2)$ are of the form $\tau \wedge \omega'$ or ω'. The correspondence

$e \wedge \omega' \leftrightarrow \tau \wedge \omega'$ and $\omega' \wedge \omega'$ gives an isomorphism of vector spaces as in the statement of the theorem, and there are no remaining actions for this isomorphism to intertwine. □

Lemma 4.1 implies the following theorem.

Theorem 4.2 *Let* (F, S_+, S_-, Λ) *and* $(F', S'_+, S'_-, \Lambda')$ *be two sutured surfaces. For some* $m \geq 0$, *choose distinct intervals* $I_1, \ldots, I_m$ *of* S_+ *and distinct intervals* $I'_1, \ldots, I'_m$ *of* S'_+. *Use* $I_1, \ldots, I_m$ *to define an action of* $(\mathbb{F}_2[E]/(E^2))^{\otimes m}$ *on* $\wedge^* H_1(F, S_+; \mathbb{F}_2)$, *and similarly for* F'. *Let* $(\overline{F}, \overline{S}_+, \overline{S}_-, \overline{\Lambda})$ *be the sutured surface obtained by gluing* I_j *to* I'_j *for* $1 \leq j \leq m$ *(in such a way that the result is oriented). Then we have an isomorphism*

$$\wedge^* H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2) \cong \wedge^* H_1(F, S_+; \mathbb{F}_2) \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes m}} \wedge^* H_1(F', S'_+; \mathbb{F}_2)$$

that intertwines the remaining actions of $\mathbb{F}_2[E]/(E^2)$ *for intervals of* S_+ *and* S'_+ *that are not included in* $\{I_1, \ldots, I_m\}$ *or* $\{I'_1, \ldots, I'_m\}$.

Proof We can write $\wedge^* H_1(F, S_+; \mathbb{F}_2) \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes m}} \wedge^* H_1(F', S'_+; \mathbb{F}_2)$ as

$$\left((\wedge^* H_1(F \sqcup F', S_+ \sqcup S'_+; \mathbb{F}_2)) \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes 2}} \mathbb{F}_2[E]/(E^2) \right) \cdots \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes 2}} \mathbb{F}_2[E]/(E^2),$$

where there are m successive tensor products by $\mathbb{F}_2[E]/(E^2)$ over $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ (one for each pair (I_j, I'_j)). The result now follows from Lemma 4.1. □

Corollary 4.3 *There is a functor from the full subcategory of the* $(1+1)$*–dimensional oriented open-closed cobordism category on objects with no closed circles (the "open sector" of the open-closed cobordism category) to* $\mathbf{Alg}_{\mathbb{F}_2}$ *sending an object with* m *intervals to the algebra* $(\mathbb{F}_2[E]/(E^2))^{\otimes m}$ *and sending a morphism (viewed as a sutured surface* (F, S_+, S_-, Λ)*) to* $\wedge^* H_1(F, S_+; \mathbb{F}_2)$ *(viewed as a bimodule over tensor products of* $\mathbb{F}_2[E]/(E^2)$ *for the input and output intervals of the morphism).*

5 The tensor product case

Figure 22 shows the open pair of pants surface P with a sutured structure (P, S_+, S_-, Λ). Let e_1 and e_2 be the arcs shown in the figure and let I_1, I_2 and I_3 be the S_+ intervals shown in the figure. Since $\{e_1, e_2\}$ is a basis for $H_1(P, S_+; \mathbb{F}_2)$, we have a basis $\{1, e_1, e_2, e_1 \wedge e_2\}$ for $\wedge^* H_1(P, S_+; \mathbb{F}_2)$. The three actions of $\mathbb{F}_2[E]/(E^2)$ on $\wedge^* H_1(P, S_+; \mathbb{F}_2)$ can be described as follows:

- For the I_1–action, E sends $1 \mapsto 0$, $e_1 \mapsto 1$, $e_2 \mapsto 0$, and $e_1 \wedge e_2 \mapsto e_2$.
- For the I_2–action, E sends $1 \mapsto 0$, $e_1 \mapsto 0$, $e_2 \mapsto 1$, and $e_1 \wedge e_2 \mapsto e_1$.
- For the I_3–action, E sends $1 \mapsto 0$, $e_1 \mapsto 1$, $e_2 \mapsto 1$, and $e_1 \wedge e_2 \mapsto e_1 + e_2$.

Using the I_1 and I_2 actions to define an action of $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ on $\wedge^* H_1(P, S_+; \mathbb{F}_2)$, we see that $\wedge^* H_1(P, S_+; \mathbb{F}_2)$ is a free module of rank 1 over $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$ with an $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$–basis given
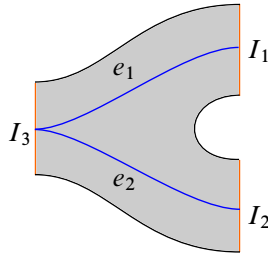
Figure 22: The open pair-of-pants surface P with sutured structure and basis $\{e_1, e_2\}$ for $H_1(P, S_+; \mathbb{F}_2)$.

by $\{e_1 \wedge e_2\}$. The I_3–action of $\mathbb{F}_2[E]/(E^2)$ is then given by applying the coproduct $\Delta(E) = E \otimes 1 + 1 \otimes E$, followed by multiplication in $(\mathbb{F}_2[E]/(E^2))^{\otimes 2}$.

Now, if we have sutured surfaces $(F', S'_+, S'_-, \Lambda')$ and $(F'', S''_+, S''_-, \Lambda'')$ with chosen intervals I' and I'' in S'_+ and S''_+ respectively, we can glue $F' \sqcup F''$ to P by gluing I' to I_1 and I'' to I_2. Applying Theorem 4.2 with $F_1 := F' \sqcup F''$ and $F_2 := P$, and letting $(\overline{F}, \overline{S}_+, \overline{S}_-, \overline{\Lambda})$ denote the glued surface,

$$\wedge^* H_1(\overline{F}, \overline{S}_+; \mathbb{F}_2) \cong (\mathbb{F}_2[E]/(E^2))^{\otimes 2} \otimes_{(\mathbb{F}_2[E]/(E^2))^{\otimes 2}} \wedge^* H_1(F' \sqcup F'', S'_+ \sqcup S''_+; \mathbb{F}_2)$$

$$\cong \wedge^* H_1(F' \sqcup F'', S'_+ \sqcup S''_+; \mathbb{F}_2)$$

$$\cong \wedge^* H_1(F', S'_+; \mathbb{F}_2) \otimes \wedge^* H_1(F'', S''_+; \mathbb{F}_2)$$

with I_3–action of E given by taking $\Delta(E) = E \otimes 1 + 1 \otimes E$ and then acting on the tensor product $\wedge^* H_1(F', S'_+; \mathbb{F}_2) \otimes \wedge^* H_1(F'', S''_+; \mathbb{F}_2)$. Corollary 1.4 follows from this computation.

References

[1] **B Cooper**, *Formal contact categories*, preprint (2015) arXiv 1511.04765 To appear in Algebr. Geom. Topol.

[2] **S K Donaldson**, *Topological field theories and formulae of Casson and Meng–Taubes*, from "Proceedings of the Kirbyfest" (J Hass, M Scharlemann, editors), Geom. Topol. Monogr. 2, Geom. Topol. Publ., Coventry (1999) 87–102 MR Zbl

[3] **C L Douglas**, **R Lipshitz**, **C Manolescu**, *Cornered Heegaard Floer homology*, Mem. Amer. Math. Soc. 1266, Amer. Math. Soc., Providence, RI (2019) MR Zbl

[4] **C L Douglas**, **C Manolescu**, *On the algebra of cornered Floer homology*, J. Topol. 7 (2014) 1–68 MR Zbl

[5] **C Frohman**, **A Nicas**, *The Alexander polynomial via topological quantum field theory*, from "Differential geometry, global analysis, and topology" (A Nicas, W F Shadwick, editors), CMS Conf. Proc. 12, Amer. Math. Soc., Providence, RI (1991) 27–40 MR Zbl

[6] **J Hom**, **T Lidman**, **L Watson**, *The Alexander module, Seifert forms, and categorification*, J. Topol. 10 (2017) 22–100 MR Zbl

[7] **K Honda**, **W H Kazez**, **G Matić**, *Contact structures, sutured Floer homology and TQFT*, preprint (2008) arXiv 0807.2431

[8]   **R M Kaufmann**, **R C Penner**, *Closed/open string diagrammatics*, Nuclear Phys. B 748 (2006) 335–379 MR Zbl

[9]   **B Keller**, *On differential graded categories*, from "International Congress of Mathematicians, II" (M Sanz-Solé, J Soria, J L Varona, J Verdera, editors), Eur. Math. Soc., Zürich (2006) 151–190 MR Zbl

[10] **T Kerler**, *Homology TQFT's and the Alexander–Reidemeister invariant of 3–manifolds via Hopf algebras and skein theory*, Canad. J. Math. 55 (2003) 766–821 MR Zbl

[11] **M Khovanov**, *How to categorify one-half of quantum $\mathfrak{gl}(1|2)$*, from "Knots in Poland 3, III" (J H Przytycki, P Traczyk, editors), Banach Center Publ. 103, Polish Acad. Sci. Inst. Math., Warsaw (2014) 211–232 MR Zbl

[12] **A D Lauda**, **H Pfeiffer**, *Open-closed strings: two-dimensional extended TQFTs and Frobenius algebras*, Topology Appl. 155 (2008) 623–666 MR Zbl

[13] **R Lipshitz**, **P S Ozsváth**, **D P Thurston**, *Bimodules in bordered Heegaard Floer homology*, Geom. Topol. 19 (2015) 525–724 MR Zbl

[14] **R Lipshitz**, **P S Ozsváth**, **D P Thurston**, *Bordered Heegaard Floer homology*, Mem. Amer. Math. Soc. 1216, Amer. Math. Soc., Providence, RI (2018) MR Zbl

[15] **A Manion**, **R Rouquier**, *Higher representations and cornered Heegaard Floer homology*, preprint (2020) arXiv 2009.09627

[16] **D Mathews**, *Chord diagrams, contact-topological quantum field theory and contact categories*, Algebr. Geom. Topol. 10 (2010) 2091–2189 MR Zbl

[17] **D V Mathews**, *Sutured Floer homology, sutured TQFT and noncommutative QFT*, Algebr. Geom. Topol. 11 (2011) 2681–2739 MR Zbl

[18] **D V Mathews**, *Sutured TQFT, torsion and tori*, Int. J. Math. 24 (2013) art. id. 1350039 MR Zbl

[19] **D V Mathews**, *Itsy bitsy topological field theory*, Ann. Henri Poincaré 15 (2014) 1801–1865 MR Zbl

[20] **D V Mathews**, **E Schoenfeld**, *Dimensionally reduced sutured Floer homology as a string homology*, Algebr. Geom. Topol. 15 (2015) 691–731 MR Zbl

[21] **I Petkova**, *The decategorification of bordered Heegaard Floer homology*, J. Symplectic Geom. 16 (2018) 227–277 MR Zbl

[22] **R W Thomason**, *The classification of triangulated subcategories*, Compos. Math. 105 (1997) 1–27 MR Zbl

[23] **R Zarev**, *Bordered sutured Floer homology*, PhD thesis, Columbia University (2011) Available at https://www.proquest.com/docview/868276640

Department of Mathematics, North Carolina State University
Raleigh, NC, United States

ajmanion@ncsu.edu

https://math.sciences.ncsu.edu/people/ajmanion/

A simplicial version of the 2–dimensional Fulton–MacPherson operad

NATHANIEL BOTTMAN

We define an operad in Top, called FM_2^W. The spaces in FM_2^W come with CW decompositions such that the operad compositions are cellular. In fact, each space in FM_2^W is the realization of a simplicial set. We expect, but do not prove here, that FM_2^W is isomorphic to the 2–dimensional Fulton–MacPherson operad FM_2. Our construction is connected to the author's work on the symplectic $(A_\infty, 2)$–category, and suggests a strategy toward equipping the symplectic cochain complex with the structure of a homotopy Batalin–Vilkovisky algebra.

1 Introduction

Getzler and Jones [1994] introduced the Fulton–MacPherson operad

$$\mathrm{FM}_2 = (\mathrm{FM}_2(k))_{k \geq 1}, \tag{1}$$

where $\mathrm{FM}_2(k)$ is the compactification à la Fulton and MacPherson [1994] of the configuration space of k distinct labeled points in $\mathbb{R}^2$, modulo translations and dilations. Getzler and Jones proposed in the same paper a collection of cellular decompositions of the spaces in FM_2, such that these decompositions are compatible with the operad maps $\circ_i \colon \mathrm{FM}_2(k) \times \mathrm{FM}_2(l) \to \mathrm{FM}_2(k + l - 1)$. These decompositions formed the basis for a significant amount of work related to the Deligne conjecture, including a proof in [Getzler and Jones 1994] of that conjecture.

Unfortunately, Tamarkin found an error in Getzler and Jones' decomposition. In particular, in the 9–dimensional space $\mathrm{FM}_2(6)$, there are two disjoint open 6–cells C_1 and C_2 with the property that $\overline{C}_1 \cap C_2$ is nonempty, as described in [Voronov 2000, Section 1.2.2]. Salvatore [2022] used meromorphic differentials to construct cellular decompositions of the spaces in FM. His approach is completely different from Getzler and Jones'.

We construct an operad of CW complexes, which we conjecture to be isomorphic in Top to FM_2. Under this expected isomorphism, our decompositions are refinements of Getzler and Jones' attempted decompositions. The context for the current paper is the author's program (as developed in [Bottman 2015; 2019a; 2019b; 2020; Bottman and Carmeli 2021; Bottman and Oblomkov 2019; Bottman and Wehrheim 2018]) to construct Symp, the symplectic $(A_\infty, 2)$–category. Specifically, the author plans to use the decompositions of FM that we construct here to understand the axioms for identity 1–morphisms
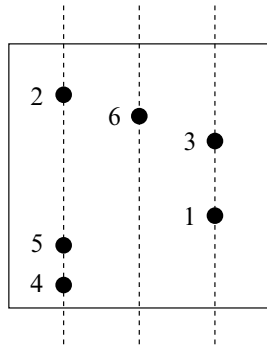
Figure 1

in an $(A_\infty, 2)$–category. In the context of Symp, this suggests a strategy toward endowing symplectic cohomology with a chain-level homotopy Gerstenhaber (and eventually, homotopy BV) algebra structure that is finite in each arity, thus answering Conjecture 2.6.1 of [Abouzaid 2015]. We note that our approach is compatible with the operations in Symp, unlike Salvatore's; in addition, we expect our approach to generalize to the Fulton–MacPherson operad of any dimension.

1.1 Getzler and Jones' attempted decomposition

Getzer and Jones' attempted decomposition is an adaptation to the case of FM_2 of Fox and Neuwirth's decomposition [1962] of the one-point compactification of the configuration space $(\mathbb{R}^2)^k \setminus \Delta$ of k points in $\mathbb{R}^2$, where Δ is the fat diagonal. A Fox–Neuwirth cell corresponds to a choice of which subsets of the points $p_1, \ldots, p_k$ should be vertically aligned, the left-to-right order in which these subsets of points should appear, and the top-to-bottom order in which each subset of the points should appear. For instance, Figure 1 is a real-codimension-3 cell in $((\mathbb{R}^2)^6 \setminus \Delta)^*$. Getzler and Jones observed that the Fox–Neuwirth cells are invariant under translations and dilations, and moreover that one can define a similar type of cell for the boundary locus. The elements in the boundary of $FM_2(k)$ are trees of "screens", and these "boundary cells" are defined by partitioning and ordering the points on each of the screen in the same way as with Fox–Neuwirth cells.

1.2 Tamarkin's counterexample

As described in [Voronov 2000], Tamarkin observed a way in which Getzler and Jones' supposed decomposition fails. Consider $FM_2(6)$, the open locus of which parametrizes configurations of six distinct points in $\mathbb{R}^2$, up to translations and dilations. Next, we consider the two 6–cells C_1 and C_2 in Figure 2 (we omit the numberings). The jth bubble in C_2 (for $j = 1, 2$) carries a modulus λ_j defined in the following way: by translating and dilating, we can move the left and right lines to $x = 0$ and $x = 1$, respectively; we then denote by λ_j the position of the middle line. The intersection $\overline{C}_1 \cap C_2$ is the codimension-1 locus in C_2 in which $\lambda_1 = \lambda_2$. What Getzler and Jones proposed is therefore not a cellular decomposition, because the intersection of the closures of two distinct n cells should be contained in the $(n-1)$–skeleton.
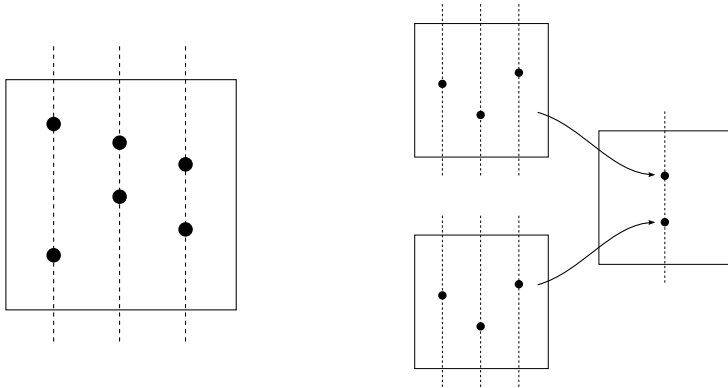
Figure 2

In our construction, C_1, C_2, and $\overline{C}_1 \cap C_2$ will each be a union of cells.

1.3 An overview of our construction

We construct a collection of CW complexes $\mathrm{FM}_2^W(k)$ and maps

$$(2) \qquad \circ_i : \mathrm{FM}_2^W(k) \times \mathrm{FM}_2^W(l) \to \mathrm{FM}_2^W(k+l-1) \quad \text{for } 1 \le i \le k.$$

Here is our main result:

Main Theorem *The spaces* $(\mathrm{FM}_2^W(k))_{k \ge 1}$ *together with the composition operations* $\circ_i$ *form a non-Σ operad, and the composition maps*

$$(3) \qquad \circ_i : \mathrm{FM}_2^W(k) \times \mathrm{FM}_2^W(l) \to \mathrm{FM}_2^W(k+l-1)$$

are cellular.

We will now give a brief overview of the definition of $\mathrm{FM}_2^W(k)$.

(i) First, we define a "W–version" W_n^W of the 2–associahedra by the analogy

$$(4) \qquad K_r : W(\mathrm{Ass}) \quad :: \quad W_n : W_n^W.$$

Here K_r is the $(r-2)$–dimensional associahedron, and $W(\mathrm{Ass})$ is the Boardman–Vogt W–construction applied to the associative operad, which is defined in terms of metric stable trees and yields an operad of CW complexes that is isomorphic to the associahedral operad K in Top. W_n is an $(|n|+r-3)$–dimensional 2–associahedron, and W_n^W is a CW complex that we define in Section 2 in terms of metric stable tree-pairs and which we expect to be homeomorphic to W_n. We then refine the CW structure on W_n^W to a simplicial decomposition.

(ii) Toward our construction of $\mathrm{FM}_2^W(k)$, we decompose $\mathrm{FM}_2(k)$ into Getzler–Jones cells, then identify each open Getzler–Jones cell with a product of open 2–associahedra. We then replace each such product by the corresponding product of interiors of the spaces W_n^W described in the previous step. This product

comes with a decomposition into products of simplices, and we refine this to a simplicial structure. Finally, we attach these decomposed Getzler–Jones cells together to produce $FM_2^W(k)$. This part of the construction appears in Section 3.

The essential property of $FM_2^W(k)$ that we must verify is that our CW decomposition is valid. It is clear that our putative open cells disjointly decompose our space, and that they are homeomorphic to open balls. The only nontrivial check we need to make is that the n–cells are attached to the $(n-1)$–skeleton. This is where Getzler and Jones' attempted decomposition fails: the 6–cell C_1 that we described in Section 1.2 is not attached to the 5–skeleton. Our decomposition satisfies this property by construction: we attach a given n–cell by taking a closed n–simplex, then attaching it to the existing skeleton via quotient maps from the boundary $(n-1)$–simplices to the $(n-1)$–skeleton. In fact, the boundary of an n–cell is a union of cells of dimension at most $n-1$.

1.4 The relationship between our construction and Symp

The genesis of the construction of FM_2^W was a connection between the symplectic $(A_\infty, 2)$–category Symp and E_2 suggested by Jacob Lurie in 2016. (The construction of Symp is a long-term project of the author, building on work of Ma'u, Wehrheim, and Woodward; see [Bottman 2015; 2019a; 2019b; 2020; Bottman and Carmeli 2021; Bottman and Wehrheim 2018; Ma'u et al. 2018].) We can express this connection concretely, via a collection of maps

$$(5) \qquad\qquad f_\sigma^W : W_{\boldsymbol{n}}^W \to FM_2^W(|\boldsymbol{n}|),$$

where σ is a 2–*permutation*, as defined in Section 3.2. The idea of this map is very simple. The map f_σ forgets the data of the lines, then labels the points according to the 2–permutation σ. Then f_σ extends continuously to the boundary of $W_{\boldsymbol{n}}$; it is an embedding on the interior of its domain, but contracts some boundary cells.

Example 1.1 In Figure 3, we depict W_{111} and its image under an appropriate map f_σ. More precisely, we depict their nets — to "assemble" both CW complexes, one would cut them out, then glue together like-numbered edges. As is evident, most of the 2–cells of W_{111} are contracted by f_σ.

While it would take us too far afield to explain the relationship between FM_2 and Symp (and their W–counterparts) in detail, let us indicate the basic idea. Symp, being an $(A_\infty, 2)$–category, assigns to a chain in a 2–associahedron $W_{\boldsymbol{n}}$ an operation on 2–morphisms. (For instance, the objects of Symp are symplectic manifolds, and given two objects M_0 and M_1, the 1–morphism category is $\text{Fuk}(M_0^- \times M_1)$; 2–associahedra W_n, where n is a single positive integer, act on this Fukaya category by the usual A_∞–operations.) The current definition of an $(A_\infty, 2)$–category, appearing in [Bottman and Carmeli 2021], does not equip identity 1–morphisms with all the possible structure. Indeed, when defining operations on 2–morphisms in the situation where some of the 1–morphisms are identities, those 1–morphisms should be allowed to be "moved past" the other 1–morphisms. To make this precise, one exactly needs to understand the maps f_σ, and to equip their targets with a CW structure so that f_σ is cellular. One way to
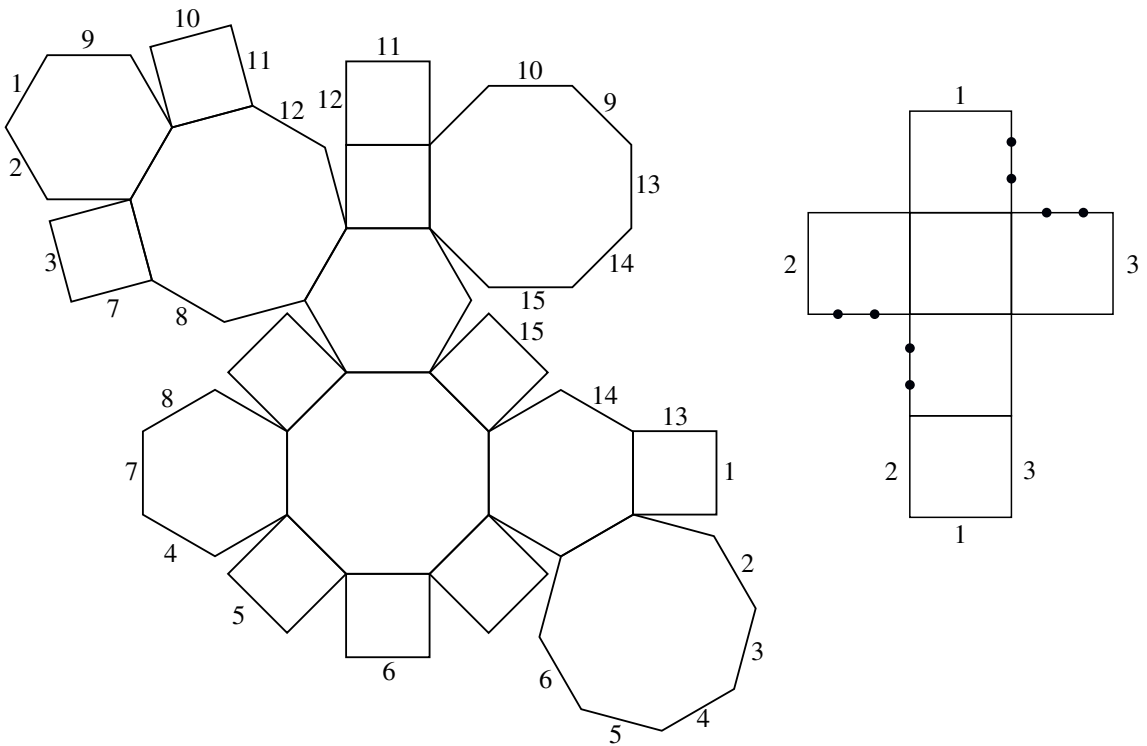
Figure 3

proceed toward this goal is to first decompose FM_2^W so that f_σ^W is cellular, and next construct coherent homeomorphisms $W_{\boldsymbol{n}} \cong W_{\boldsymbol{n}}^W$ and $\mathrm{FM}_2(k) \cong \mathrm{FM}_2^W(k)$.

The following result therefore shows the way toward a connection between the symplectic $(A_\infty, 2)$–category and FM_2^W. It is an immediate consequence of our construction of $W_{\boldsymbol{n}}^W$ and $\mathrm{FM}_2^W(k)$, and it forms the content of [Remark 3.14](#).

Proposition *Fix* $r \geq 1$, $\boldsymbol{n} \in \mathbb{Z}_{\geq 0}^r \setminus \{\boldsymbol{0}\}$, *and a 2–permutation* σ *of type* $\boldsymbol{n}$. *Then the associated map*

$$\tag{6} f_\sigma^W : W_{\boldsymbol{n}}^W \to \mathrm{FM}_2^W(|\boldsymbol{n}|)$$

is cellular.

1.5 Future directions

The author plans to develop several aspects of the current paper. In particular:

- With several collaborators, the author plans to extend this work to produce cellular decompositions of FM_k^W for all $k \geq 1$, and to show that FM_k^W is isomorphic to FM_k in Top.

- This paper can be construed as a way of incorporating identity 1–morphisms into the symplectic $(A_\infty, 2)$–category. The author plans to formalize this in future work on the algebra of $(A_\infty, 2)$–categories.

- We plan to upgrade this work to give a cellular model for the framed analogue of the Fulton–MacPherson operad. This suggests a way of endowing symplectic cohomology with a chain-level BV algebra structure, which is the subject of Conjecture 2.6.1 of [Abouzaid 2015].

Acknowledgments

2 A "W–version" of the 2–associahedra

In this section, we construct a "W–version" of the 2–associahedra. (The 2–associahedra were originally defined in [Bottman 2019a].) This is an essential ingredient in our definition of $FM_2^W(k)$, which will appear in Section 3.

2.1 A warm-up: K^W, ie $W(\mathrm{Ass})$, ie a W–version of the associahedra

In this subsection, we recall a certain operad, which we will denote by $K^W = (K_r^W)_{r \geq 1}$. This is simply the Boardman–Vogt W–construction applied to the associative operad Ass. We construct only K^W rather than recalling the general definition of the W–construction, because this one-off construction will be a useful warm-up to our construction of W^W later in this section. As noted in [Barber 2013], K^W is isomorphic in Top to the associahedral operad K.

The following proposition summarizes what we will prove about K^W:

Proposition 2.1 *The spaces $(K_r^W)_{r \geq 1}$ form a non-Σ operad of CW complexes, and the composition maps*

(7) $$\circ_i \colon K_r^W \times K_s^W \to K_{r+s-1}^W$$

defined in Definition 2.11 are cellular.

We will prove Proposition 2.1 at the end of the current subsection.

We begin with a definition of rooted ribbon trees. Stable rooted ribbon trees with r leaves index the strata of the associahedron K_r, and they will be an integral part of the definition of K_r^W.

Definition 2.2 [Bottman 2019a, Definition 2.2] A *rooted ribbon tree* (RRT) is a tree T with a choice of a root $\alpha_{\text{root}} \in T$ and a cyclic ordering of the edges incident to each vertex; we orient such a tree toward the root. We say that a vertex α of an RRT T is *interior* if the set $\text{in}(\alpha)$ of its incoming neighbors is nonempty, and we denote the set of interior vertices of T by T_{int}. An RRT T is *stable* if every interior vertex has at least two incoming edges. We define K_r^{tree} to be the set of all isomorphism classes of stable rooted ribbon trees with r leaves.

We denote the ith leaf of an RRT T by λ_i^T. For any $\alpha, \beta \in T$, $T_{\alpha\beta}$ denotes those vertices γ such that the path $[\alpha, \gamma]$ from α to γ passes through β. We define $T_\alpha := T_{\alpha_{\text{root}}\alpha}$.

Remark 2.3 Ribbon trees (resp. rooted ribbon trees) are often referred to as planar trees (resp. planted trees).

Next, we define a version of RRTs with internal edge lengths:

Definition 2.4 A *metric RRT* $(T, (\ell_e))$ is the data of

- an RRT T, and
- for every edge e of T not incident to a leaf (but possibly incident to the root), a length $\ell_e \in [0, 1]$.

We call this a *metric RRT of type T*.

Now we will define a "dimension" function d on stable RRTs:

Definition 2.5 [Bottman 2019a, Definition 2.4] For T a stable RRT in K_r^{tree}, we define its *dimension* $d(T) \in [0, r-2]$ like so:

$$(8) \qquad\qquad d(T) := r - \#T_{\text{int}} - 1.$$

Definition 2.6 Given a stable tree T, the *cell associated to T* is denoted by C_T and is defined to consist of all metric RRTs of type T.

Note that we can canonically identify C_T with the closed cube of dimension equal to the number of internal edges of T. That is:

$$(9) \qquad\qquad C_T \cong [0, 1]^{\#T_{\text{int}}-1} = [0, 1]^{r-2-d(T)}.$$

As we will see, K_r^W is $(r-2)$–dimensional; it follows that $d(T)$ is the codimension of C_T in K_r^W. (The unfortunate clash of terminology between "dimension" and "codimension" is due to the fact that, in K_r, the cell indexed by T has dimension $d(T)$.)

We now define K_r^W by taking the union of the cells C_T for T any stable RRT with r leaves, then collapsing edges of length 0.
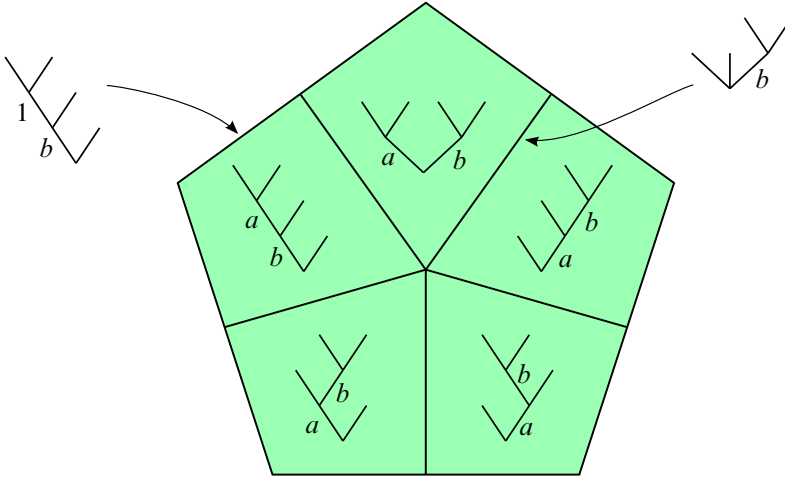
Figure 4

Definition 2.7 Given $r \geq 1$, we define K_r^W to be the following quotient:

$$K_r^W := \left(\bigsqcup_{T \in K_r^{\text{tree}}} C_T \right) \Big/ \sim. \tag{10}$$

Here $\sim$ identifies $(T, (\ell_e))$ and $(T', (\ell_e'))$ if, after collapsing all edges e of T with $\ell_e = 0$ and all edges e of T' with $\ell_e' = 0$, both metric RRTs reduce to the same metric RRT $(T'', (\ell_e''))$.

Example 2.8 In Figure 4, we depict the CW complex K_4^W. Note that this is a refinement of K_4, which (as a CW complex) is a pentagon. We have labeled the open top cells by the metric stable RRTs that they parametrize, where each a and b is allowed to vary in $[0, 1]$. The closed top cells are glued together along the cells where some of the edge lengths are 0 — for instance, we have indicated how the top and top-right cubes are joined along the internal edge of the pentagon where the edge length b in both cells becomes 0. The boundary of K_r^W is the union of cells where at least one edge length is 1.

Finally, we define a simplicial refinement of the CW structure on K_r^W. To approach this, we note that if P is the poset $\{0, 1\}^k$, where $\sigma_1 < \sigma_2$ if σ_2 can be gotten by changing some of the 0s of σ_1 to 1s, then the nerve of P is a simplicial decomposition of the cube $[0, 1]^k$. More concretely, the top simplices are the sets of the form

$$\{(x_1, \ldots, x_k) \in [0, 1]^k \mid 0 < x_{\sigma(1)} < \cdots < x_{\sigma(k)} < 1\}, \tag{11}$$

where σ is a permutation on k letters. The remaining simplices are the result of replacing some of these inequalities by equalities.

Definition 2.9 We refine the CW structure on K_r^W by decomposing each cell C_T in K_r^W like so: we make the identification $C_T \cong [0, 1]^{r-2-d(T)}$, then perform the simplicial decomposition described in the previous paragraph. This refinement equips K_r^W with a simplicial decomposition.
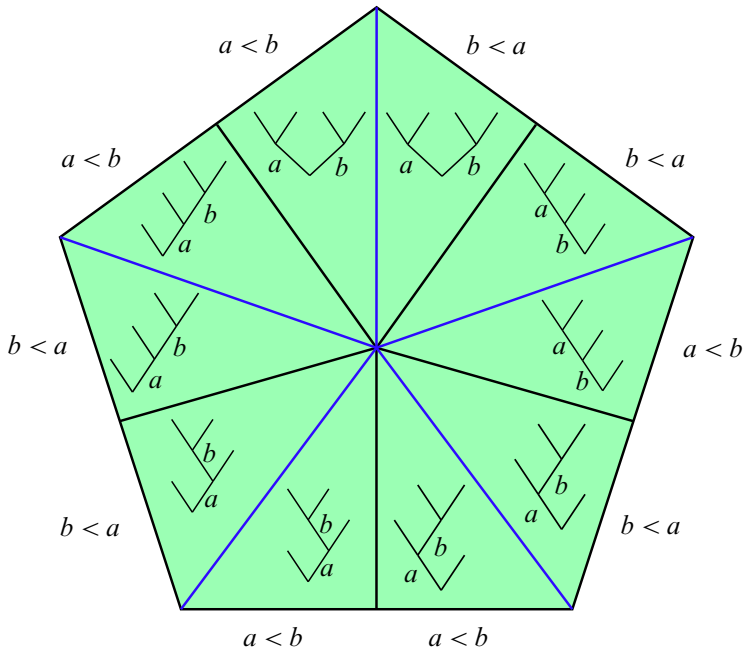
Figure 5

Example 2.10 In Figure 5, we depict the simplicial complex K_4^W. This is the refinement of our initial cubical CW decomposition of K_r^W gotten by subdividing each of the five squares into two triangles. We indicate the new edges by coloring them blue.

Now that we have constructed the spaces K_r^W, we can prove Proposition 2.1, which states that (K_r^W) is a non-Σ operad and that the operad maps are cellular.

Definition 2.11 Fix r, s, and $i \in [1, r]$. We wish to define the composition map

$$\circ_i \colon K_r^W \times K_s^W \to K_{r+s-1}^W. \tag{12}$$

We do so cell by cell. That is, fix cells $C_T \subset K_r^W$ and $C_{T'} \subset K_s^W$. Define T'' to be the result of grafting T' to the ith leaf of T. Then we define $\circ_i$ on $C_T \times C_{T'}$ like so: given collections of edge lengths on T and T', combine them to produce a collection of edge lengths on T'', where we assign to the single newly formed interior edge the length 1.

Proof of Proposition 2.1 Fix r, s, and $i \in [1, r]$, and consider the composition map

$$\circ_i \colon K_r^W \times K_s^W \to K_{r+s-1}^W. \tag{13}$$

To show that $\circ_i$ is cellular, let's consider the restriction of $\circ_i$ to a product $C_T \times C_{T'}$ of closed cubes, for $T \in K_r$ and $T' \in K_s$. Denote by T'' the tree obtained by grafting the root of T' to the ith leaf of T. Then $\circ_i$ includes $C_T \times C_{T'}$ into $C_{T''}$ as the face gotten by requiring the outgoing edge of the root of T' to have length 1. The CW structure of this face of $C_{T''}$ is finer than that of $C_T \times C_{T''}$, so $\circ_i$ is indeed cellular. $\square$

2.2 Metric tree-pairs and the definition of W_n^W

Just as we defined K_r^W to be the parameter space of metric stable RRTs, we will define W_n^W to parametrize metric stable tree-pairs. The definition of metric stable tree-pairs is somewhat involved, so we devote the current subsection to this definition.

Before defining metric stable tree-pairs, we recall the definition of stable tree-pairs:

Definition 2.12 [Bottman 2019a, Definition 3.1] A *stable tree-pair of type n* is a datum $2T = T_b \xrightarrow{f} T_s$, with T_b, T_s, and f described below:

- The *bubble tree* T_b is an RRT whose edges are either solid or dashed, which must satisfy these properties:

    – The vertices of T_b are partitioned as $V(T_b) = V_{\text{comp}} \sqcup V_{\text{seam}} \sqcup V_{\text{mark}}$, where
        * every $\alpha \in V_{\text{comp}}$ has at least 1 solid incoming edge, no dashed incoming edges, and either a dashed or no outgoing edge;
        * every $\alpha \in V_{\text{seam}}$ has zero or more dashed incoming edges, no solid incoming edges, and a solid outgoing edge; and
        * every $\alpha \in V_{\text{mark}}$ has no incoming edges and either a dashed or no outgoing edge.
      We partition $V_{\text{comp}} =: V_{\text{comp}}^1 \sqcup V_{\text{comp}}^{\geq 2}$ according to the number of incoming edges of a given vertex.

    – **Stability** If α is a vertex in V_{comp}^1 and β is its incoming neighbor, then $\#\text{in}(\beta) \geq 2$; if α is a vertex in $V_{\text{comp}}^{\geq 2}$ and $\beta_1, \ldots, \beta_l$ are its incoming neighbors, then there exists j with $\#\text{in}(\beta_j) \geq 1$.

- The *seam tree* T_s is an element of K_r^{tree}.

- The *coherence map* is a map $f: T_b \to T_s$ of sets having these properties:

    – f sends root to root, and if $\beta \in \text{in}(\alpha)$ in T_b, then either $f(\beta) \in \text{in}(f(\alpha))$ or $f(\alpha) = f(\beta)$.

    – f contracts all dashed edges, and every solid edge whose terminal vertex is in V_{comp}^1.

    – For any $\alpha \in V_{\text{comp}}^{\geq 2}$, f maps the incoming edges of α bijectively onto the incoming edges of $f(\alpha)$, compatibly with $<_\alpha$ and $<_{f(\alpha)}$.

    – f sends every element of V_{mark} to a leaf of T_s, and if $\lambda_i^{T_s}$ is the ith leaf of T_s, then $f^{-1}\{\lambda_i^{T_s}\}$ contains n_i elements of V_{mark}, which we denote by $\mu_{i1}^{T_b}, \ldots, \mu_{in_i}^{T_b}$.

We denote by W_n^{tree} the set of isomorphism classes of stable tree-pairs of type n. Here an isomorphism from $T_b \xrightarrow{f} T_s$ to $T_b' \xrightarrow{f'} T_s'$ is a pair of maps $\varphi_b: T_b \to T_b'$ and $\varphi_s: T_s \to T_s'$ that fit into a commutative square in the obvious way and that respect all the structure of the bubble trees and seam trees.

Next, we define metric stable tree-pairs. This notion is more subtle than that of metric stable RRTs, because we must impose conditions on the edge-lengths. (This should be compared to Bottman and Oblomkov's similar constraints [2019, Section 3], imposed in order to define local charts on a complexified version of W_n.)

Definition 2.13 A *metric stable tree-pair* $(2T, (L_e), (\ell_e))$ is the following data:

- $2T$ is a stable tree-pair.

- We have, for every interior dashed edge e of T_b, a length $L_e \in [0, 1]$, and, for every interior edge e of T_s, a length $\ell_e \in [0, 1]$, subject to the following coherence conditions (where for convenience we set $L_\alpha := L_e$ for $\alpha \in V_{\text{comp}}(T_b) \setminus \{\alpha_{\text{root}}\}$ and e the outgoing edge of α, and similarly for the edge-lengths in T_s):

  - For every $\alpha_1, \alpha_2 \in V_{\text{comp}}^{\geq 2}(T_b)$ and $\beta \in V_{\text{comp}}^1(T_b)$ with $f(\alpha_1) = f(\alpha_2) = f(\beta)$, we require

(14)
$$\max_{\gamma \in [\alpha_1, \beta)} L_\gamma = \max_{\gamma \in [\alpha_2, \beta)} L_\gamma.$$

  - For every $\rho \in V_{\text{int}}(T_s) \setminus \{\rho_{\text{root}}\}$ and $\alpha \in V_{\text{comp}}^{\geq 2}(T_b)$ with $f(\alpha) = \rho$, we require

(15)
$$\ell_\rho = \max_{\gamma \in [\alpha, \beta_\alpha)} L_\gamma,$$

where we define β_α to be the first element of $V_{\text{comp}}^{\geq 2}(T_b)$ that the path from α to α_{root} passes through.

Finally, we recall the *dimension* of a stable tree-pair. Similarly to the dimension of a stable RRT, this will be the codimension in W_n^W of the cell corresponding to the stable tree-pair in question.

Definition 2.14 [Bottman 2019a, Definition 3.3] For $2T$ a stable tree-pair, we define the *dimension* $d(2T) \in [0, |\boldsymbol{n}| + r - 3]$ like so:

(16)
$$d(2T) := |\boldsymbol{n}| + r - \#V_{\text{comp}}^1(T_b) - \#(T_s)_{\text{int}} - 2.$$

We are now prepared to define W_n^W, the "W–version" of the 2–associahedron. We will define W_n^W by attaching together the cells C_{2T}, which consist of metric stable tree-pairs.

Definition 2.15 Given a stable tree-pair $2T$, the *cell associated to* $2T$ is the collection of all metric stable tree-pairs of type $2T$. We denote this cell by C_{2T}.

Note that we can identify C_{2T} with the subset of the cube $[0, 1]^k$ defined by the equalities (14) and (15), where k is the number of interior dashed edges of T_b plus the number of interior edges of T_s.

Definition 2.16 Fix $r \geq 1$ and $\boldsymbol{n} \in \mathbb{Z}_{\geq 0}^r \setminus \{\boldsymbol{0}\}$. We define W_n^W similarly to how we defined K_r^W in Definition 2.7:

$$W_n^W := \left(\bigsqcup_{2T \in W_n^{\text{tree}}} C_{2T} \right) \Big/ \sim.$$

The quotient here is somewhat subtler than the quotient that appeared in Definition 2.7, specifically when it comes to T_b. In T_s, we simply contract any edges of length 0. We indicate in Figure 6 how to perform the necessary contractions in T_b when some edge-lengths are 0. The reader should think of the left contraction as undoing a type-1 move (as in [Bottman 2019a, Section 3.1]), whereas the right contraction undoes either a type-2 or a type-3 move. Note that we are using the coherences enforced in
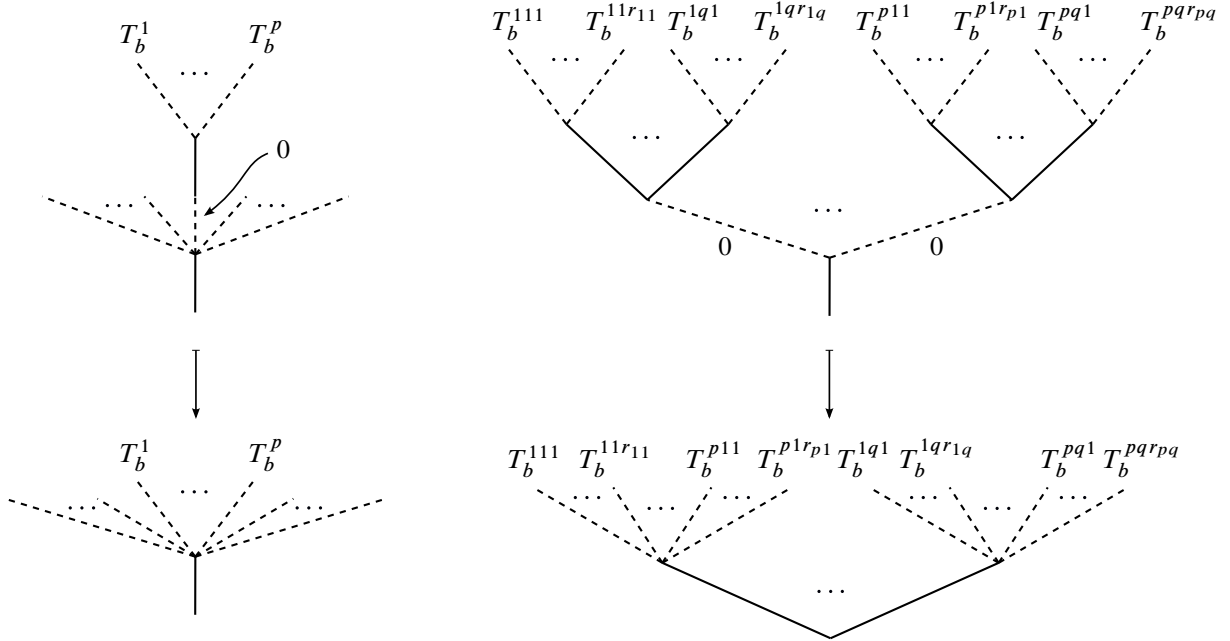
Figure 6

Definition 2.13 — for instance, these mean that we do not have to consider a situation as in the right-hand side of the above figure, but where only some of the edge-lengths in this portion of T_b are 0.

Example 2.17 In Figure 7, we depict the CW complex W_{21}^W. Each of the parameters a and b lie in $[0, 1]$; they do not have the same meaning across different cells. The eight interior edges (resp. sixteen boundary edges) correspond to the loci in the top cells where a parameter goes to 0 (resp. to 1).

Finally, we refine the CW structure on W_n^W to a simplicial decomposition.

Lemma 2.18 *Fix a stable tree-pair $2T$. For every simplex S in the standard simplicial decomposition of $[0, 1]^k \supset C_{2T}$, S is either contained in C_{2T} or disjoint from it. The collection of such simplices that are contained in C_{2T} form a simplicial decomposition of C_{2T}.*

Proof Fix a simplex S. S is defined by a collection of equalities and inequalities of the form

(17) $$0 * x_{\sigma(1)} * \cdots * x_{\sigma(k)} * 1,$$

where each "$*$" is either a "$<$" or an "$=$" and where σ is a permutation on k letters. After imposing these (in)equalities, the left- and right-hand sides of the equalities (14) and (15) become single variables. This collection of equalities will either be always satisfied or never satisfied, depending on the constraints in (17). Depending on which of these is the case, S is either contained in C_{2T} or disjoint from it.

It follows immediately that the collection of simplices that are contained in C_{2T} form a simplicial decomposition of C_{2T}. $\qquad\square$
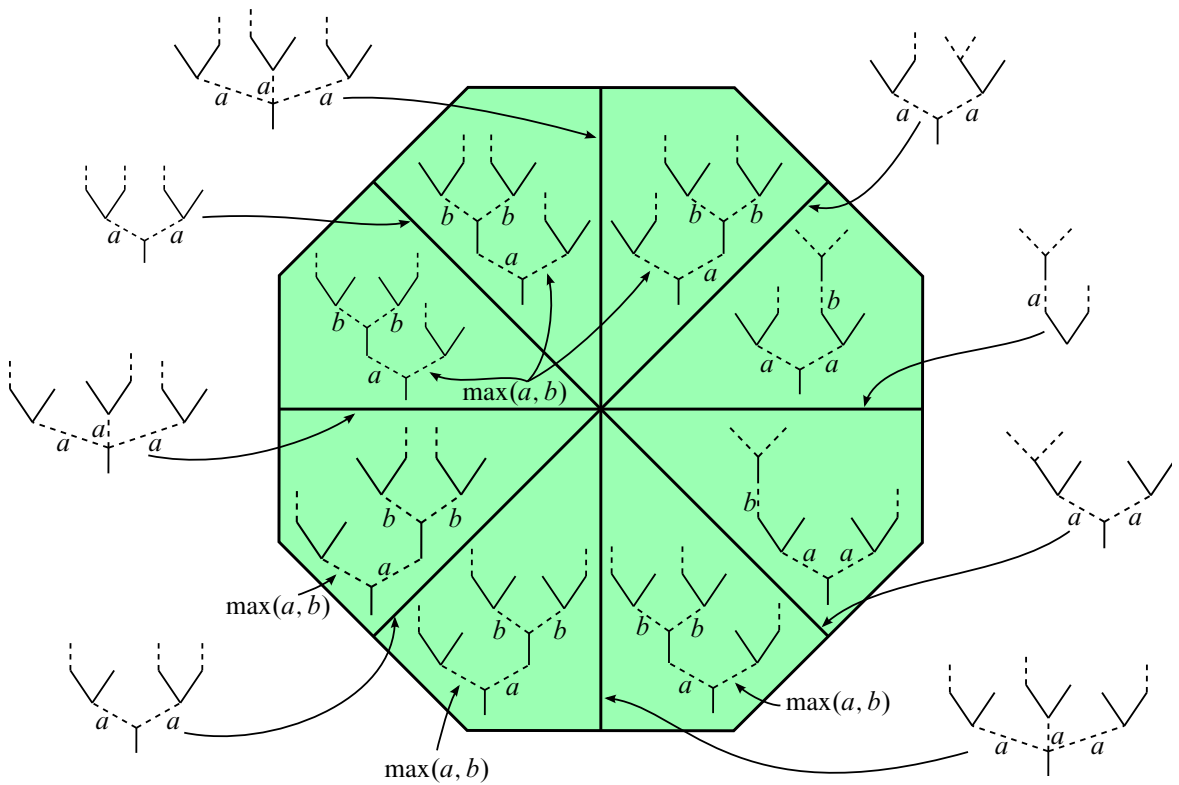
Figure 7

Example 2.19 In Figure 8, we illustrate the closed cell in W_{40}^W associated to the underlying tree-pair of the (top-dimensional) metric tree-pair shown on the right. The restriction on the lengths $a, b, c, d \in [0, 1]$ is that they must satisfy $\max(a, b) = \max(c, d)$; as a result, this cell has the CW type of a square pyramid.

We indicate the simplicial refinement of this cell: the square pyramid is subdivided into eight 3–simplices, which are defined by imposing inequalities and equalities as shown in this figure.

$a < c < b = d$

$a < d < b = c$

$d < a < b = c$

$c < a < b = d$

$d < b < a = c$

$c < b < a = d$

$b < c < a = d$

$b < d < a = c$
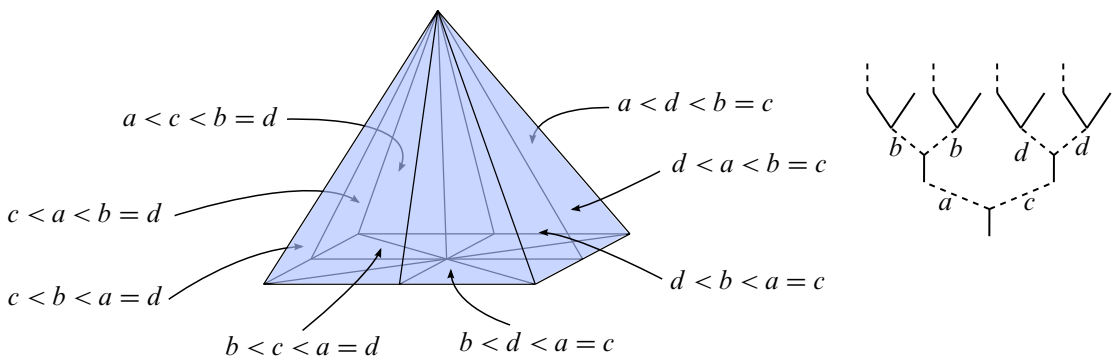
Figure 8

3 The construction of FM_2^W

In this final section, we will construct a collection of CW complexes $(\mathrm{FM}_2^W(k))_{k \geq 1}$ and a collection of operations

$$\tag{18} \circ_i \colon \mathrm{FM}_2^W(k) \times \mathrm{FM}_2^W(l) \to \mathrm{FM}_2^W(k+l-1)$$

such that these data form an operad.

We will now give an overview of our construction of $\mathrm{FM}_2^W(k)$. This is an expansion of step (ii) in the overview we gave in Section 1.3, and we label the parts accordingly:

(iia) Each open Getzler–Jones cell in $\mathrm{FM}_2(k)$ can be identified with a product of open 2–associahedra, ie a product of the form $\mathring{W}_{\boldsymbol{m}^1} \times \cdots \times \mathring{W}_{\boldsymbol{m}^a}$ (where "$\mathring{X}$" is our notation for the interior of a space X). For each such open cell, we replace these 2–associahedra by their W–construction equivalents thusly: $\mathring{W}_{\boldsymbol{m}^1}^W \times \cdots \times \mathring{W}_{\boldsymbol{m}^a}^W$. This product comes with the product CW structure, and we refine this in a way that endows $\mathring{W}_{\boldsymbol{m}^1}^W \times \cdots \times \mathring{W}_{\boldsymbol{m}^a}^W$ with the structure of a simplicial complex.

(iib) While an open Getzler–Jones cell can be identified with a product $\mathring{W}_{\boldsymbol{m}^1} \times \cdots \times \mathring{W}_{\boldsymbol{m}^a}$ of 2–associahedra, their compactifications (in $\mathrm{FM}_2(k)$ and $W_{\boldsymbol{m}^1} \times \cdots \times W_{\boldsymbol{m}^a}$, respectively) are different: the compactification of the former is smaller than the compactification of the latter. This is reflected in how we glue our products $\mathring{W}_{\boldsymbol{m}^1}^W \times \cdots \times \mathring{W}_{\boldsymbol{m}^a}^W$ together. Specifically, we perform this gluing by applying a quotient map to each simplex in the boundary of $W_{\boldsymbol{m}^1}^W \times \cdots \times W_{\boldsymbol{m}^a}^W$. This quotient map is closely related to the maps $f_\sigma \colon W_{\boldsymbol{n}} \to \mathrm{FM}_2(k)$ that we described in Section 1.4: they reflect the fact that the compactification used to define $W_{\boldsymbol{n}}$ allows lines with no marked points, whereas the compactification of a Getzler–Jones cell does not allow this.

The following is the main result of this section, which we stated in the introduction and record again here:

Main Theorem *The spaces $(\mathrm{FM}_2^W(k))_{k \geq 1}$ together with the composition operations $\circ_i$ defined in Definition 3.11 form a non-Σ operad, and the composition maps*

$$\tag{19} \circ_i \colon \mathrm{FM}_2^W(k) \times \mathrm{FM}_2^W(l) \to \mathrm{FM}_2^W(k+l-1)$$

are cellular.

Proof Combine Lemmata 3.12 and 3.13 below. □

3.1 Quotient maps on 2–associahedra

Before we can define the quotient involved in (24), we will define for every cell F in $\partial W_{\boldsymbol{n}}^W$ a map q_F from F to a certain product of 2–associahedra, where this target will vary for difference choices of F. We begin with two preliminary definitions:

Definition 3.1 Fix $r \geq 1$ and $\boldsymbol{n} \in \mathbb{Z}_{\geq 0}^r \setminus \{\boldsymbol{0}\}$, and fix $i \in [1, r]$ such that $n_i = 0$. Define $\tilde{\boldsymbol{n}} := (n_1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_r)$. We then define a map of posets $\pi_i^{\mathrm{tree}} \colon W_{\boldsymbol{n}}^{\mathrm{tree}} \to W_{\tilde{\boldsymbol{n}}}^{\mathrm{tree}}$ by applying the following procedure to $2T = T_b \xrightarrow{f} T_s \in W_{\boldsymbol{n}}^{\mathrm{tree}}$:

(i) Denote by e_0 the edge in T_s incident to the ith leaf $\lambda_i^{T_s}$. If e is a solid edge in T_b that is mapped identically under f to e_0, then we delete e. Next, we delete e_0. We modify f in the obvious way.

(ii) After performing these deletions, our tree-pair may no longer be stable. We rectify this in T_b (resp. T_s) by performing the contractions indicated on the left (resp. right):
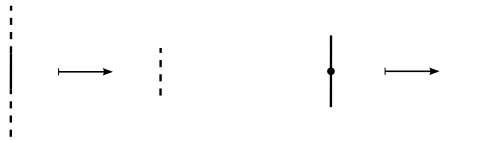
Specifically, we perform these contractions as many times as necessary for the tree-pair to be stable. Denoting the end result of this procedure by $\widetilde{2T}$, we define $\pi_i^{\text{tree}}(2T) := \widetilde{2T}$.

Next, we define another map of posets. Fix $r \geq 1$ and $\boldsymbol{n} \in \mathbb{Z}_{\geq 0}^r \setminus \{\boldsymbol{0}\}$. Denote by $\tilde{\boldsymbol{n}}$ the result of deleting all the zeroes from $\boldsymbol{n}$, and set $\tilde{r}$ to be the length of $\tilde{\boldsymbol{n}}$. We define $\pi^{\text{tree}} \colon W_{\boldsymbol{n}}^{\text{tree}} \to W_{\tilde{\boldsymbol{n}}}^{\text{tree}}$ by applying the map π_i^{tree} once for each i with $n_i = 0$.

It is not hard to check that the choices implicit in this definition do not matter, and that the resulting maps are indeed maps of posets.

Definition 3.2 Fix $r \geq 1$ and $\boldsymbol{n} \in \mathbb{Z}_{\geq 0}^r \setminus \{\boldsymbol{0}\}$. We define a map $\pi^W \colon W_{\boldsymbol{n}}^W \to W_{\tilde{\boldsymbol{n}}}^W$ in the same fashion as π^{tree}, with the provision that when we contract adjacent edges of lengths ℓ_1 and ℓ_2 (whether in T_b or T_s) we equip the resulting edge with length $\max(\ell_1, \ell_2)$.

Next, we recall a W–version analogue of two properties of the 2–associahedra:

W–version analogue of the forgetful property of [Bottman 2019a, Theorem 4.1] Fix $r \geq 1$ and $\boldsymbol{n} \in \mathbb{Z}_{\geq 0}^r \setminus \{\boldsymbol{0}\}$. There is a surjection $W_{\boldsymbol{n}}^W \to K_r^W$ which sends a metric stable tree-pair $(T_b \xrightarrow{f} T_s, (L_e), (\ell_e))$ to the metric stable RRT $(T_s, (\ell_e))$.

W–version analogue of the recursive property of [Bottman 2019a, Theorem 4.1] Fix a stable tree-pair $2T = T_b \xrightarrow{f} T_s \in W_{\boldsymbol{n}}^{\text{tree}}$. There is an inclusion of CW complexes

$$(20) \qquad \Gamma_{2T} \colon \prod_{\substack{\alpha \in V_{\text{comp}}^1(T_b) \\ \text{in}(\alpha) = (\beta)}} W_{\#\text{in}(\beta)}^W \times \prod_{\rho \in V_{\text{int}}(T_s)} \prod_{\substack{\alpha \in V_{\text{comp}}^{\geq 2}(T_b) \cap f^{-1}\{\rho\} \\ \text{in}(\alpha) = (\beta_1, \ldots, \beta_{\#\text{in}(\rho)})}}^{K_{\#\text{in}(\rho)}^W} W_{\#\text{in}(\beta_1), \ldots, \#\text{in}(\beta_{\#\text{in}(\alpha)})}^W \hookrightarrow W_{\boldsymbol{n}}^W,$$

where the superscript on one of the product symbols indicates that it is a fiber product with respect to the maps in the description of the forgetful property above.

The map Γ_{2T} defined in [Bottman 2019a], which is defined for the posets $W_{\boldsymbol{n}}^{\text{tree}}$, is defined by attaching stable tree-pairs together in a way specified by the stable tree-pair $2T$. This map is similar, but we are attaching together *metric* stable tree-pairs. We assign the length 1 to the edges along which we attach the trees. (The image of Γ_{2T} is a union of cells in $\partial W_{\boldsymbol{n}}^W$.)

We can now define the quotient maps q_F on W_n^W:

Definition 3.3 Fix $r \geq 1$, $\boldsymbol{n} \in \mathbb{Z}_{\geq 0}^r \setminus \{\boldsymbol{0}\}$, a stable type-$\boldsymbol{n}$ tree-pair $\widetilde{2T}$, and a face F of the associated cell $C_{\widetilde{2T}}$ in W_n^W with the property that F lies in ∂W_n^W. (Equivalently, the metric tree-pairs in F have at least one length that is identically equal to 1.) The *quotient map associated to F* is a map q_F from F to a product of 2–associahedra. Given a metric stable tree-pair $(2T, (L_e), (\ell_e))$, we define its image under π in the following fashion:

(i) Break up T_b and T_s along the edges that are identically 1 in F. Equivalently, choose $2T$ of minimal dimension with the property that F lies in the image of Γ_{2T}, then identify F as a top cell in a product of fiber products of the following form:

$$
(21) \qquad \prod_{\substack{\alpha \in V_{\mathrm{comp}}^1(T_b) \\ \mathrm{in}(\alpha)=(\beta)}} W_{\#\mathrm{in}(\beta)}^W \times \prod_{\rho \in V_{\mathrm{int}}(T_s)} \prod_{\substack{\alpha \in V_{\mathrm{comp}}^{\geq 2}(T_b) \cap f^{-1}\{\rho\} \\ \mathrm{in}(\alpha)=(\beta_1,\ldots,\beta_{\#\mathrm{in}(\rho)})}}^{K_{\#\mathrm{in}(\rho)}^W} W_{\#\mathrm{in}(\beta_1),\ldots,\#\mathrm{in}(\beta_{\#\mathrm{in}(\alpha)})}^W .
$$

As a result, we obtain a list of metric stable tree-pairs, which we can regard as lying inside a product $W_{\boldsymbol{m}^1}^W \times \cdots \times W_{\boldsymbol{m}^a}^W$.

(ii) We then apply the map π^W to each of the factors in the product just recorded, hence producing an element of $W_{\widetilde{\boldsymbol{m}}^1}^W \times \cdots \times W_{\widetilde{\boldsymbol{m}}^a}^W$. (As in Definitions 3.1 and 3.2, $\widetilde{\boldsymbol{m}}^i$ denotes the result of removing the 0s from $\boldsymbol{m}^i$.)

Note that for two cells F_1 and F_2 in the boundary of W_n^W, the targets of q_{F_1} and q_{F_2} are typically different.

Example 3.4 In Figure 9, we illustrate several things about W_{21}^W. Initially, W_{21}^W is an octagon, decomposed into eight squares; this is indicated by the black lines. The simplicial refinement divides each square into two 2–simplices. We have indicated the metric tree-pairs that correspond to each of the eight squares, as well as those corresponding to the sixteen 1–simplices that comprise ∂W_{21}^W. (Some dashed edges are not labeled; these should be interpreted as having length $\max(a, b)$.)

Finally, we have indicated the behavior of the quotient maps on W_{21}^W. These maps are the identity on every edge except for those indicated in red. Each pair of red edges is contracted to a point. One reflection of this is that in Example 1.1, the octagons in W_{111} are taken to the (cellular) hexagons in the Getzler–Jones cell indicted on the right.

3.2 The construction of $\mathrm{FM}_2^W(k)$

In this subsection, we tackle the construction of $\mathrm{FM}_2^W(k)$. First, we will describe our version of the Getzler–Jones cells. Next, we will explain how to glue these spaces together.

To define the Getzler–Jones cells, we must introduce 2–*permutations*, which will allow us to enforce the alignment and ordering of special points on screens as in Figure 1.
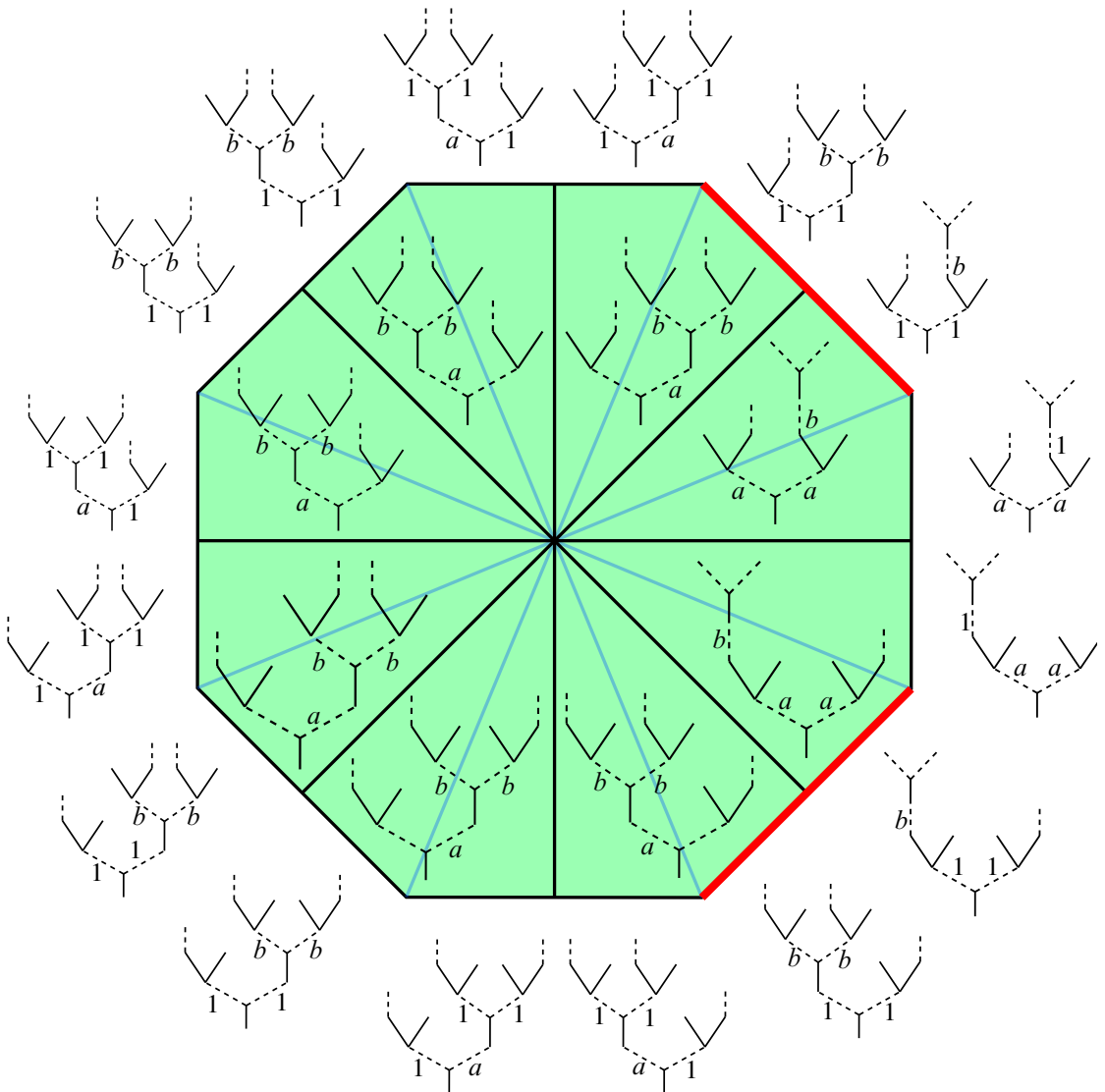
Figure 9

Definition 3.5 Fix a finite set A. A 2–*permutation* σ *on* A is the data

- an ordered decomposition

(22)
$$A = A_1 \sqcup \cdots \sqcup A_r,$$

where A_r is allowed to be empty, and

- for each i, a linear order on A_i.

We define the *type* of σ to be the vector $\boldsymbol{n} := (|A_1|, \ldots, |A_r|)$. If σ is a 2–permutation whose type $\boldsymbol{n}$ has no zero entries, then we say that σ *has no empty part*.

Remark 3.6 A type-$(\overbrace{1, \ldots, 1}^{r})$ 2–permutation is exactly the data of a permutation on r letters. The same is true of a type-(n) 2–permutation.

Next, we define a *Getzler–Jones datum*, the set of which indexes the Getzler–Jones cells in $\mathrm{FM}_2^W(k)$.

Definition 3.7 Fix $k \geq 2$. A *Getzler–Jones datum* consists of

- a stable rooted tree T with k leaves, together with a numbering of its leaves from 1 through k, and
- for every interior vertex $v \in T_{\mathrm{int}}$, a 2–permutation σ on its incoming vertices $V_{\mathrm{in}}(T)$ such that σ has no empty part.

We denote the type of the 2–permutation associated to v by $\boldsymbol{n}(v)$. We will abuse notation and denote the entire Getzler–Jones datum by T.

Finally, we can define the *Getzler–Jones cells of type k*:

Definition 3.8 Fix $k \geq 2$ and a Getzler–Jones datum T. Then we define

$$(23) \qquad \mathrm{GJ}_T := \prod_{v \in T_{\mathrm{int}}} \mathring{W}_{\boldsymbol{n}(v)}^W \quad \text{and} \quad \widetilde{\mathrm{GJ}}_T := \prod_{v \in T_{\mathrm{int}}} W_{\boldsymbol{n}(v)}^W.$$

We call GJ_T the *Getzler–Jones cell GJ_T associated to T*, and refer to GJ_T as a *type-k Getzler–Jones cell*.

In Lemma 2.18 we equipped $W_{\boldsymbol{n}}^W$ with the structure of a simplicial complex, which induces a CW structure on GJ_T and $\widetilde{\mathrm{GJ}}_T$. We refine these to equip GJ_T and $\widetilde{\mathrm{GJ}}_T$ with simplicial decompositions, in the fashion of Lemma 2.18.

Remark 3.9 The reason why we do not refer to $\widetilde{\mathrm{GJ}}_T$ as a "closed Getzler–Jones cell" is because it is *not* the closure in $\mathrm{FM}_2^W(k)$ of GJ_T. In fact, it is larger than this closure. Our reason for making this second definition is that $\widetilde{\mathrm{GJ}}_T$ will be an integral part of our definition of $\mathrm{FM}_2^W(k)$.

We will define $\mathrm{FM}_2^W(k)$ as a quotient of the following form, where T varies over type-k Getzler–Jones data:

$$(24) \qquad \mathrm{FM}_2^W(k) := \left(\coprod_T \widetilde{\mathrm{GJ}}_T \right) \Big/ \sim.$$

The remaining ingredient is the collection of maps that we will use to attach these spaces. As a consequence of the definition of these maps, $\mathrm{FM}_2^W(k)$ will decompose as a set into the union of all type-k Getzler–Jones cells.

Finally, we come to the definition of $\mathrm{FM}_2^W(k)$:

Definition 3.10 Fix $k \geq 2$. We construct $\mathrm{FM}_2^W(k)$ like so:

(i) Begin with the following disjoint union, where T varies over type-k Getzler–Jones data:

$$(25) \qquad \coprod_T \widetilde{\mathrm{GJ}}_T.$$

(ii) Fix a type-k Getzler–Jones datum T, and fix a cell F in the boundary of $\widetilde{\mathrm{GJ}}_T = \prod_{v \in T_{\mathrm{int}}} W^W_{\boldsymbol{n}(v)}$. F lies inside a product of cells in the 2–associahedra that comprise $\widetilde{\mathrm{GJ}}_T$ — that is, we may write $F \subset \prod_{v \in T_{\mathrm{int}}} F_v \subset \prod_{v \in T_{\mathrm{int}}} W^W_{\boldsymbol{n}(v)}$, where F_v is a cell in $W^W_{\boldsymbol{n}(v)}$. For every v, we have a map q_v from $W^W_{\boldsymbol{n}(v)}$ to a product of 2–associahedra; by combining these, we obtain a map from F to a product of 2–associahedra. In fact, we can regard the target of this map as a Getzler–Jones cell.

(iii) We take the quotient of the disjoint union in (25) by attaching the constituent spaces together via the maps we defined in the last step.

We define $\mathrm{FM}_2^W(1)$ to be a point.

It is a consequence of the simplicial structure of the $\widetilde{\mathrm{GJ}}_T$ that each $\mathrm{FM}_2^W(k)$ has the structure of a CW complex. As noted above, a result of our definition is that $\mathrm{FM}_2^W(k)$ decomposes as a union of Getzler–Jones cells, over all Getzler–Jones data of type k.

3.3 The operad structure on FM_2^W

Definition 3.11 Fix k, l, and $i \in [1, k]$. We wish to define the map

$$\text{(26)} \qquad \circ_i : \mathrm{FM}_2^W(k) \times \mathrm{FM}_2^W(l) \to \mathrm{FM}_2^W(k + l - 1).$$

To do so, fix Getzler–Jones data T and T' of types k and l, respectively, and fix cells $F \subset \mathrm{GJ}_T$ and $F' \subset \mathrm{GJ}_{T'}$. We will define $\circ_i$ on

$$\text{(27)} \qquad \mathrm{GJ}_T \times \mathrm{GJ}_{T'} = \prod_{v \in T_{\mathrm{int}} \sqcup T'_{\mathrm{int}}} W^W_{\boldsymbol{n}(v)}.$$

Define T'' to be the result of grafting T' to the ith leaf of T, and completing it to a Getzler–Jones datum in the obvious way. We define $\circ_i$ on $\mathrm{GJ}_T \times \mathrm{GJ}_{T'}$ to be the identification of $\mathrm{GJ}_T \times \mathrm{GJ}_{T'}$ with $\mathrm{GJ}_{T''}$.

Lemma 3.12 *Taken together, the spaces* $(\mathrm{FM}_2^W(k))_{k \geq 1}$ *together with the composition operations* $\circ_i$ *form a non-Σ operad.*

Proof This is immediate from the definition. □

Lemma 3.13 *The composition maps*

$$\text{(28)} \qquad \circ_i : \mathrm{FM}_2^W(k) \times \mathrm{FM}_2^W(l) \to \mathrm{FM}_2^W(k + l - 1)$$

are cellular.

Proof This is similar to the proof of Proposition 2.1. □

Remark 3.14 Fix $r \geq 1$, $\boldsymbol{n} \in \mathbb{Z}^r_{\geq 0} \setminus \{\boldsymbol{0}\}$, and a 2–permutation σ of type $\boldsymbol{n}$. Then the associated forgetful map

$$\text{(29)} \qquad f_\sigma^W : W^W_{\boldsymbol{n}} \to \mathrm{FM}_2^W(|\boldsymbol{n}|)$$

is cellular. This map is defined in the obvious way: we first identify $W_{\boldsymbol{n}}^W$ with the corresponding $\widetilde{\mathrm{GJ}}_T$, where T is a Getzler–Jones datum whose associated tree T is a corolla with $|\boldsymbol{n}|$ leaves. Then, we include $\widetilde{\mathrm{GJ}}_T$ into the disjoint union $\bigsqcup_T \widetilde{\mathrm{GJ}}_T$, and finally take the quotient to land in $\mathrm{FM}_2^W(|\boldsymbol{n}|)$.

References

[Abouzaid 2015] **M Abouzaid**, *Symplectic cohomology and Viterbo's theorem*, from "Free loop spaces in geometry and topology" (J Latschev, A Oancea, editors), IRMA Lect. Math. Theor. Phys. 24, Eur. Math. Soc., Zürich (2015) 271–485 MR Zbl

[Barber 2013] **D A Barber**, *A comparison of models for the Fulton–MacPherson operads*, PhD thesis, University of Sheffield (2013) Available at https://core.ac.uk/download/pdf/131322152.pdf

[Bottman 2015] **N S Bottman**, *Pseudoholomorphic quilts with figure eight singularity*, PhD thesis, Massachusetts Institute of Technology (2015) Available at http://hdl.handle.net/1721.1/101823

[Bottman 2019a] **N Bottman**, 2–*Associahedra*, Algebr. Geom. Topol. 19 (2019) 743–806 MR Zbl

[Bottman 2019b] **N Bottman**, *Moduli spaces of witch curves topologically realize the 2–associahedra*, J. Symplectic Geom. 17 (2019) 1649–1682 MR Zbl

[Bottman 2020] **N Bottman**, *Pseudoholomorphic quilts with figure eight singularity*, J. Symplectic Geom. 18 (2020) 1–55 MR Zbl

[Bottman and Carmeli 2021] **N Bottman**, **S Carmeli**, $(A_\infty, 2)$–*categories and relative 2–operads*, High. Struct. 5 (2021) 401–421 MR Zbl

[Bottman and Oblomkov 2019] **N Bottman**, **A Oblomkov**, *A compactification of the moduli space of marked vertical lines in* $\mathbb{C}^2$, preprint (2019) arXiv 1910.02037

[Bottman and Wehrheim 2018] **N Bottman**, **K Wehrheim**, *Gromov compactness for squiggly strip shrinking in pseudoholomorphic quilts*, Selecta Math. 24 (2018) 3381–3443 MR Zbl

[Fox and Neuwirth 1962] **R Fox**, **L Neuwirth**, *The braid groups*, Math. Scand. 10 (1962) 119–126 MR Zbl

[Fulton and MacPherson 1994] **W Fulton**, **R MacPherson**, *A compactification of configuration spaces*, Ann. of Math. 139 (1994) 183–225 MR Zbl

[Getzler and Jones 1994] **E Getzler**, **J D S Jones**, *Operads, homotopy algebra and iterated integrals for double loop spaces*, preprint (1994) arXiv hep-th/9403055

[Ma'u et al. 2018] **S Ma'u**, **K Wehrheim**, **C Woodward**, A_∞ *functors for Lagrangian correspondences*, Selecta Math. 24 (2018) 1913–2002 MR Zbl

[Salvatore 2022] **P Salvatore**, *A cell decomposition of the Fulton MacPherson operad*, J. Topol. 15 (2022) 443–504 MR Zbl

[Voronov 2000] **A A Voronov**, *Homotopy Gerstenhaber algebras*, from "Conférence Moshé Flato 1999, II: Quantization, deformations, and symmetries" (G Dito, D Sternheimer, editors), Math. Phys. Stud. 22, Kluwer, Dordrecht (2000) 307–331 MR Zbl

Max Planck Institute for Mathematics
Bonn, Germany

natebottman@gmail.com

msp

Intrinsically knotted graphs with linklessly embeddable simple minors

THOMAS W MATTMAN

RAMIN NAIMI

ANDREI PAVELESCU

ELENA PAVELESCU

It has been an open question whether the deletion or contraction of an edge in an intrinsically knotted graph always yields an intrinsically linked graph. We present a new intrinsically knotted graph that shows the answer to both questions is no.

05C10; 57M15, 57K10

1 Introduction

A graph is *intrinsically knotted* (resp. *intrinsically linked*) if every embedding of it in S^3 contains a nontrivial knot (resp. 2–component link). We abbreviate intrinsically knotted (resp. linked) as IK (resp. IL), and not intrinsically knotted (resp. linked) as nIK (resp. nIL). Robertson, Seymour, and Thomas [12] showed that every IK graph is IL. It is also known that coning one vertex over an IL graph yields an IK graph. (This is shown by combining [12] and the work of Foisy [4] and Sachs [13].) However, it has been difficult to make the relationship between IK and IL graphs stronger. For example, Adams [1] asked if deleting a vertex from an IK graph always yields an IL graph, but Foisy [5] provided a counterexample. Deleting a vertex from a graph also deletes all edges incident to that vertex, so it might seem more likely that deleting, or contracting, a single edge of an IK graph should leave it IL. Naimi, Pavelescu, and Schwartz [10] tried to show that this is the case when the edge belongs to a 3–cycle, but their proof contained an error (which we will describe in Section 6). They also asked if deleting or contracting an edge in an IK graph always yields an IL graph. We verify (using a computer program) that the answer to this question is yes for graphs of order at most 9, but we show that in general the answer is no. We present an IK graph $G_{11,35}$ of order 11 and size 35 with edges e and f such that neither $G_{11,35}-e$ (edge deletion) nor $G_{11,35}/f$ (edge contraction) is IL. We argue that $G_{11,35}$ is a minimal-order example of an IK graph that yields a nIL graph by deleting one edge, and that ten is the smallest order for an IK graph that yields a nIL graph by contracting one edge. The graph $G_{11,35}$ is also a counterexample to the main result of [10].

Graphs that are IK but yield a nIL graph by deleting one vertex or edge or by contracting one edge are intriguing from the perspective of Colin de Verdière's graph invariant μ. This is an integer-valued graph
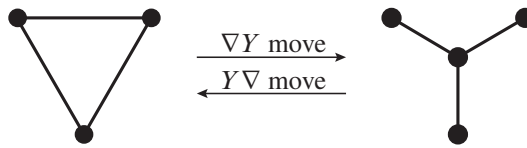
Figure 1: ∇Y and $Y\nabla$ moves.

invariant that is difficult to compute in general; its value is known only for certain classes of graphs with "nice" topological properties. For example, for any graph G, $\mu(G) \leq 3$ if and only if G is planar (see Colin de Verdière [2]), and $\mu(G) \leq 4$ if and only if G is nIL; see van der Holst, Lovász, and Schrijver [7].

An important open question is how to characterize graphs G with $\mu(G) \leq 5$. Even though many known minor-minimal IK (MMIK) graphs have μ–invariant 6, intrinsic knottedness is not the answer. A *minor* of a graph G is a graph obtained by contracting zero or more edges in a subgraph of G. We'll say an *edge deletion minor* (resp. *edge contraction minor*) of G is a graph obtained by deleting (resp. contracting) exactly one edge of G. Both are called *simple minors* of G. As we explain in Section 5, if an IK graph G has a nIL simple minor then $\mu(G) = 5$. Thus, our graph $G_{11,35}$, together with other IK graphs obtained from it (as described in Section 5), join Foisy's graph as new examples of IK graphs with μ–invariant 5. These examples show that $\mu(G) \leq 5$ is not equivalent to G being nIK.

In the next section we describe the graph $G_{11,35}$ and we show it is IK and minor-minimal for that property in Sections 3 and 4, respectively. In Section 5 we make some observations about the Colin de Verdière invariant and prove that 10 is the least order for an IK graph with an edge-contraction minor that is IL. Section 6 goes over the error in [10], and we conclude with an appendix that provides edge lists for three graphs we discuss.

To complete this introduction, we provide several definitions. A graph G is *n–apex* if one can delete n vertices from G to obtain a planar graph; G is *apex* if it is 1–apex, and 0–apex is a synonym for planar. A graph G is *minor minimal* with respect to a property if G has that property but no minor of it has that property. The complete graph on n vertices is denoted by K_n. $V(G)$ and $E(G)$ denote the vertex set and the edge set of G, respectively. A graph G is the *clique sum* of two subgraphs G_1 and G_2 over K_n if $V(G) = V(G_1) \cup V(G_2)$, $E(G) = E(G_1) \cup E(G_2)$, and the subgraphs induced in G_1 and G_2 by $V(G_1) \cap V(G_2)$ are both isomorphic to K_n. We use the notation $G = G_1 \oplus_{K_n} G_2$. The *∇Y–move* and *$Y\nabla$–move* are defined as shown in Figure 1. The *family* of a graph G is the set of all graphs obtained from G by doing zero or more ∇Y and $Y\nabla$ moves. The Petersen family of graphs is the family of the Petersen graph (which is also the family of K_6).

2 The graph $G_{11,35}$

We describe a sequence of graphs and graph operations used to construct $G_{11,35}$. Let H denote the graph in Figure 2, left. Deleting the vertex labeled 4, one obtains the maximal planar graph H', depicted in
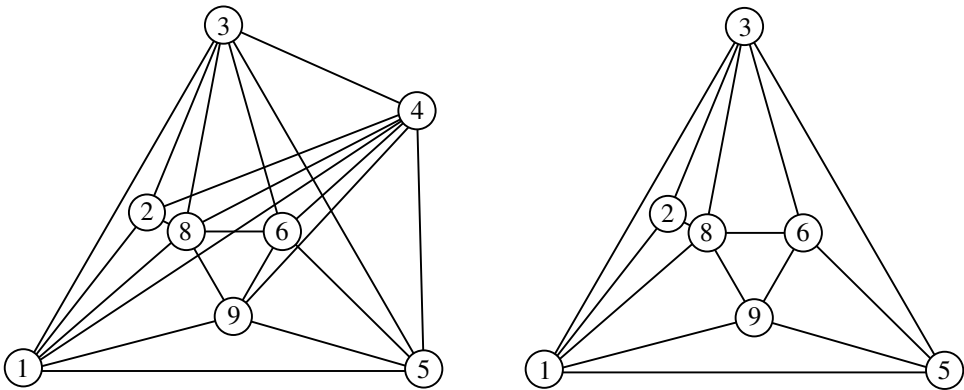
Figure 2: Left: H is apex. Right: H' is maximal planar.

Figure 2, right. This implies that H is an apex graph; thus it is nIL by [13]. Similarly, the graph K shown in Figure 3, left, is nIL since deleting vertex 5 from K yields a maximal planar graph, as in Figure 3, right.

Notice that deleting the vertices 3, 4, 5, and 6 from both H and K produces connected subgraphs. So, by [9, Lemma 14], the clique sum of H and K over the K_4 induced by $\{3, 4, 5, 6\}$ is a nIL graph, denoted by M and depicted in Figure 4.

The graph $G_{11,35}$ is obtained by adding the edge $(2, 11)$ to the nIL graph M (see Figure 5). We prove in Section 3 that $G_{11,35}$ is IK. We have thus obtained an IK graph that has a nIL edge deletion minor. Further, since the edge $(2, 11)$ is in a 3–cycle in $G_{11,35}$, this also gives a counterexample to the main result of [10]. Notice that contracting the edge $(2, 3)$ in $G_{11,35}$ yields a graph that is a minor of M, and therefore nIL. Hence, $G_{11,35}$ also has a nIL edge contraction minor. The edge list of $G_{11,35}$ is given in the appendix.

Remark The edge $(2, 3)$ in $G_{11,35}$ is triangular (ie it belongs to one or more triangles), so contracting it results in the deletion of parallel edges. One can ask whether contracting a nontriangular edge in an IK
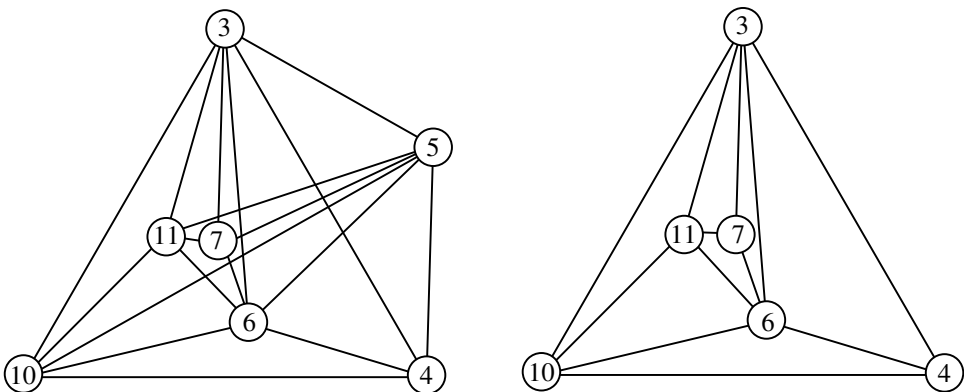
Figure 3: Left: K is apex. Right: K' is maximal planar.
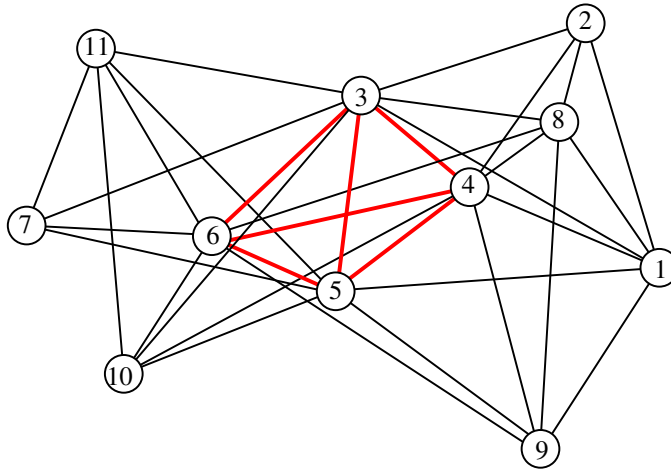
Figure 4: $M \simeq H \oplus_{K_4} K$.

graph can result in a nIL graph. The answer is yes: In $G_{11,35}$, if we do a ∇Y move on the triangle with vertices $2, 3$, and 11, we obtain a new IK graph G' with a new vertex, denoted by x. Contracting the edge $(x, 3)$ — which is nontriangular — in G' yields a graph isomorphic to $G_{11,35} - (2, 11)$, which is nIL.

Remark The graph $G_{11,35}$ is a minimal-order IK graph with a nIL edge deletion minor. To verify this, we took every maxnIL graph of order 10 (there are 107 of them [11]), and checked (with computer assistance) that adding one edge to it never yields an IK graph. However, 11 is not the smallest order of an IK graph that has a nIL edge contraction minor. The graph $G_{10,30}$, depicted in Figure 6, is a minor-minimal IK graph of order 10. Contracting the edge $(2, 6)$ gives the nIL minor in Figure 7, left. This graph is
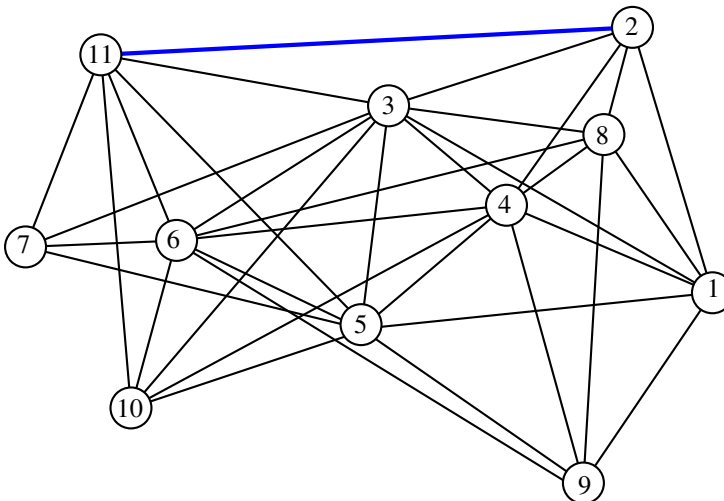
Figure 5: The graph $G_{11,35}$.
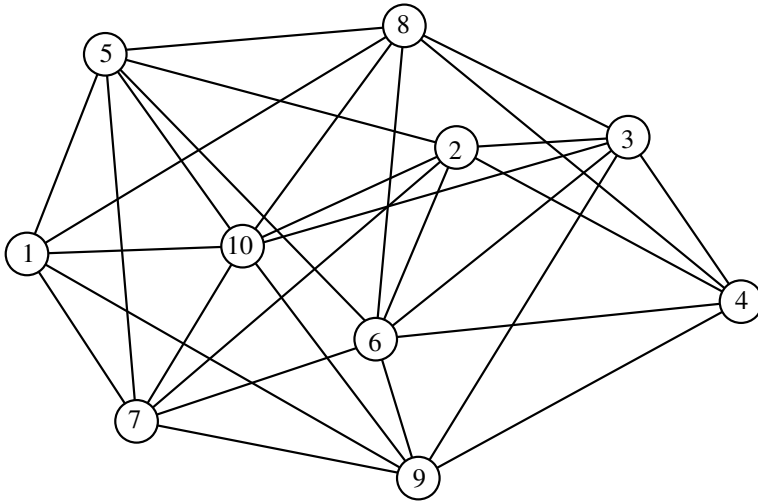
Figure 6: The graph $G_{10,30}$.
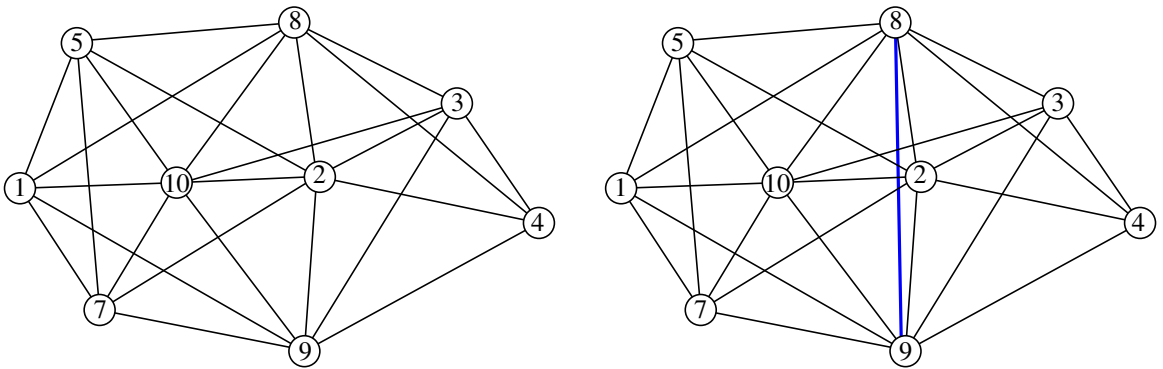
Figure 7: Left: the contraction minor of $G_{10,30}$. Right: $H \oplus_{K_4} K_5$.

nIL since adding the edge $(8, 9)$ produces a graph isomorphic to the clique sum, over the K_4 subgraph induced by $\{2, 3, 8, 9\}$, of K_5 and a subgraph isomorphic to H, introduced in Figure 2. By the following proposition, $G_{10,30}$ is a minimal-order IK graph with a nIL edge contraction minor. In Section 5, we show that $G_{10,30}$ has μ–invariant 5. Furthermore, according to our computer program, this graph is MMIK.

Proposition 2.1 *Ten is the smallest order for an IK graph which admits a nIL edge contraction minor.*

We defer the proof to Section 5.

3 $G_{11,35}$ is IK

We prove $G_{11,35}$ is IK by showing that the graph $G_{10,26}$ in Figure 8 is an IK minor of $G_{11,35}$. (In fact, $G_{10,26}$ is MMIK; we show this in the next section.)
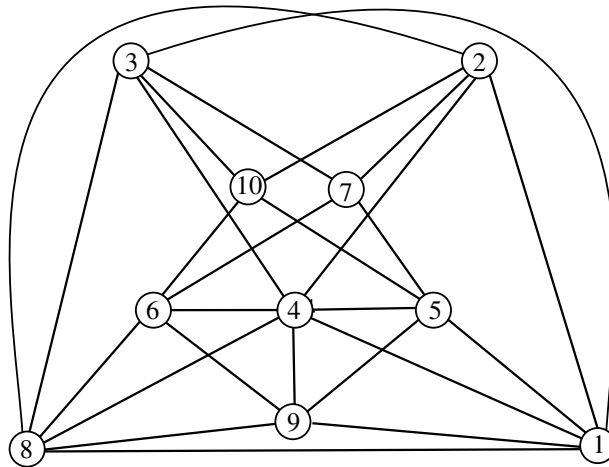
Figure 8: $G_{10,26}$.

The graph $G_{10,26}$ is obtained from $G_{11,35}$ by contracting the edge $(2,11)$ and deleting the edges $(2,3)$, $(2,5)$, $(2,6)$, $(3,5)$, $(3,6)$, $(4,10)$, and $(5,6)$.

To prove $G_{10,26}$ is IK, we use the technique developed by Foisy in [4], which we explain below. The D_4 graph is the (multi)graph shown in Figure 9. A *double-linked* D_4 is a D_4 graph embedded in S^3 so that each pair of opposite 2–cycles ($C_1 \cup C_3$ and $C_2 \cup C_4$) has odd linking number. The following lemma was proved by Foisy [4]; a more general version was proved independently by Taniyama and Yasuhara [14].

Lemma 3.1 *Every double-linked D_4 contains a nontrivial knot.*

We will also use the following (well known and easy to prove) lemma.

Lemma 3.2 *Suppose α, β_1, and β_2 are simple closed curves in S^3 such that $\beta_1 \cap \beta_2$ is an arc and α has odd linking number with $(\beta_1 \cup \beta_2) \setminus \text{interior}(\beta_1 \cap \beta_2)$. Then α has odd linking number with β_1 or β_2.*

Theorem 3.3 *The graph $G_{10,26}$ in Figure 8 is IK.*
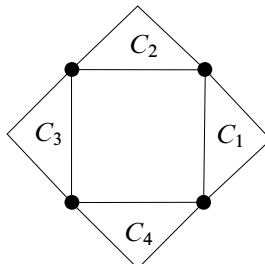
Figure 9: The D_4 graph.
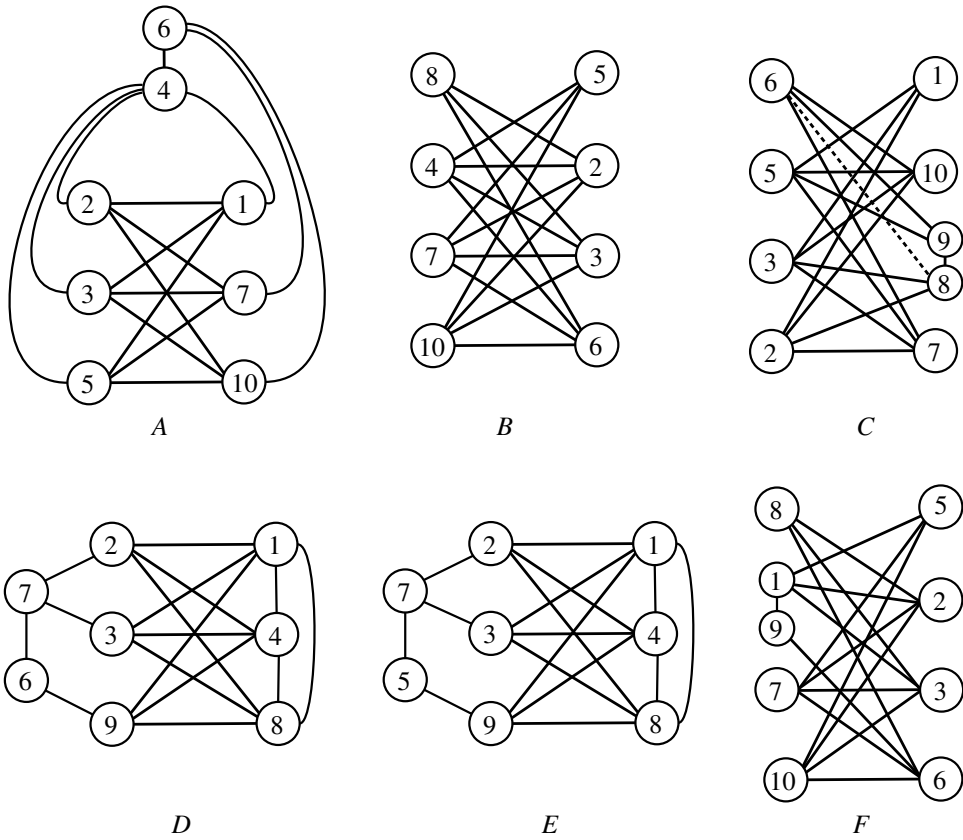
Figure 10: Selected subgraphs of $G_{10,26}$.

Proof We shall prove that every embedding of $G_{10,26}$ has a double-linked D_4 minor. It then follows from Lemma 3.1 that $G_{10,26}$ is IK. For the remainder of this proof, we will say two disjoint simple closed curves α and β in S^3 are *linked*, or α *links* β, if $\alpha \cup \beta$ has odd linking number.

In $G_{10,26}$ we select the subgraphs A, B, C, D, E, and F shown in Figure 10 (these are not induced subgraphs). All these subgraphs are either in the Petersen family of graphs or have minors in this family, and are therefore intrinsically linked: A contains a $K_{3,3,1}$ minor obtained by contracting the edge $(4,6)$; B is isomorphic to $K_{4,4}^-$; C and F contain $K_{4,4}^-$ minors obtained by contracting the edges $(8,9)$ and $(1,9)$, respectively; D and E contain G_7 minors obtained by contracting the edges $(6,7)$ and $(5,7)$, respectively.

We organize the proof into several cases and subcases, according to which two cycles of each subgraph are linked. We start with the subgraph A. The vertices of $G_{10,26}$ can be partitioned into six equivalence classes up to symmetry: $\{1,8\}$, $\{2,3\}$, $\{4\}$, $\{5,6\}$, $\{7,10\}$, and $\{9\}$. All of these except vertex 9 are in A. This gives, up to symmetry, four different pairs of cycles in A:

$$(A1) \quad (4,1,5) \cup (2,7,3,10), \qquad (A3) \quad (4,6,7,5) \cup (2,1,3,10),$$
$$(A2) \quad (4,1,2) \cup (3,7,5,10), \qquad (A4) \quad (4,6,7,2) \cup (3,1,5,10).$$
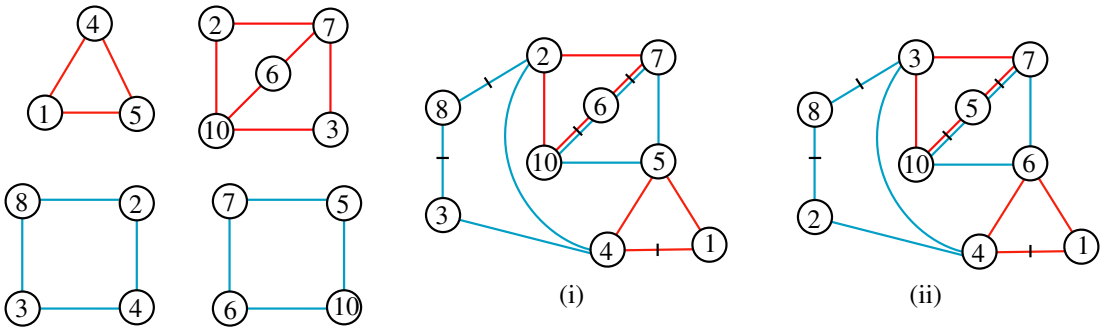
Figure 11: Diagrams for the subcase (A1)-(B1).

Since A is intrinsically linked, given any embedding of $G_{10,26}$, we can relabel (if necessary) the vertices of $G_{10,26}$ within each equivalence class so that at least one of these four pairs of cycles is linked. We subdivide each of the four cases (A1)–(A4): (A1) is split into subcases according to which two cycles of B are linked, (A2) according to C, (A3) according to D, and (A4) according to B. For each subcase a diagram is drawn with the nontrivial link in A drawn in red. The two cycles in each of the subgraphs B through F are drawn in blue. Each diagram contains some marked edges; contracting these marked edges in $G_{10,26}$ gives a double-linked D_4 minor.

Case (A1) Assume $(4, 1, 5) \cup (2, 7, 3, 10)$ is a nontrivial link of A. We identify a nontrivial link in B and show the existence of a double-linked D_4 in every subcase. Based on the symmetries of $G_{10,26}$, B has four different types of pairs of cycles. We match the link in (A1) with each of the four types of links in B:

$$\text{(B1)} \quad (8, 2, 4, 3) \cup (7, 5, 10, 6), \qquad \text{(B3)} \quad (8, 2, 7, 6) \cup (4, 5, 10, 3),$$
$$\text{(B2)} \quad (8, 2, 7, 3) \cup (4, 5, 10, 6), \qquad \text{(B4)} \quad (8, 2, 4, 6) \cup (7, 5, 10, 3).$$

Subcase (A1)-(B1) From this point forward, we abbreviate "the cycles X and Y are linked" as just "$X \cup Y$". Assume $(8, 2, 4, 3) \cup (7, 5, 10, 6)$. Since $(4, 1, 5) \cup (2, 7, 3, 10)$, by Lemma 3.2 we have either (i) $(4, 1, 5) \cup (2, 7, 6, 10)$ or (ii) $(4, 1, 5) \cup (3, 7, 6, 10)$. See Figure 11.

Subcase (A1)-(B2) Assume $(8, 2, 7, 3) \cup (4, 5, 10, 6)$. See Figure 12, left.
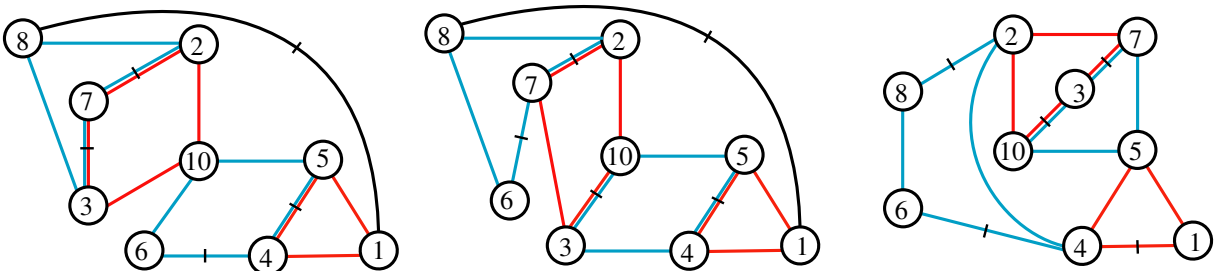
Figure 12: Diagrams for subcases. Left: (A1)-(B2). Center: (A1)-(B3). Right: (A1)-(B4).
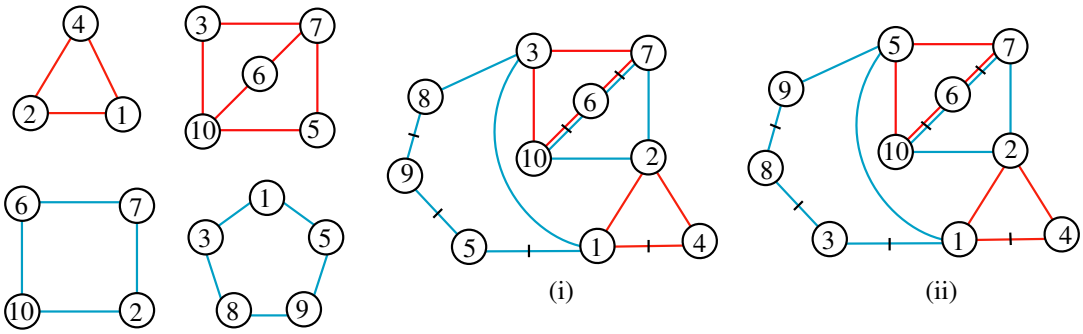
Figure 13: Diagrams for the subcase (A2)-(C1).

Subcase (A1)-(B3) Assume $(8, 2, 7, 6) \cup (4, 5, 10, 3)$. See Figure 12, center.

Subcase (A1)-(B4) Assume $(8, 2, 4, 6) \cup (7, 5, 10, 3)$. See Figure 12, right.

Case (A2) Assume $(4, 1, 2) \cup (3, 7, 5, 10)$ is a nontrivial link of A. We identify a nontrivial link in C and show the existence of a double-linked D_4. We note that vertices 8 and 9 and the edge between them act as one vertex of the $K_{4,4}^-$. Based on the symmetries of G, C has four different types of pairs of cycles. Since in the (A2) link of A vertices 2 and 3 are distinguished, they need also be distinguished within the linked cycles of C. We match the link in (A2) with each link of C:

> (C1) $(6, 7, 2, 10) \cup (1, 5, 9, 8, 3),$ (C4) $(6, 7, 2, 8, 9) \cup (1, 3, 10, 5),$
>
> (C2) $(6, 7, 3, 10) \cup (1, 5, 9, 8, 2),$ (C5) $(6, 7, 3, 8, 9) \cup (1, 2, 10, 5),$
>
> (C3) $(6, 7, 5, 10) \cup (1, 2, 8, 3),$ (C6) $(6, 7, 5, 9) \cup (1, 2, 10, 3).$

Subcase (A2)-(C1) Assume $(6, 7, 2, 10) \cup (1, 5, 9, 8, 3)$. Since $(4, 1, 2) \cup (3, 7, 5, 10)$, by Lemma 3.2 we have either (i) $(4, 1, 2) \cup (3, 7, 6, 10)$ or (ii) $(4, 1, 2) \cup (5, 7, 6, 10)$. See Figure 13.

Subcase (A2)-(C2) Assume $(6, 7, 3, 10) \cup (1, 5, 9, 8, 2)$. See Figure 14, left.

Subcase (A2)-(C3) Assume $(6, 7, 5, 10) \cup (1, 2, 8, 3)$. See Figure 14, center.

Subcase (A2)-(C4) Assume $(6, 7, 2, 8, 9) \cup (1, 3, 10, 5)$. See Figure 14, right.
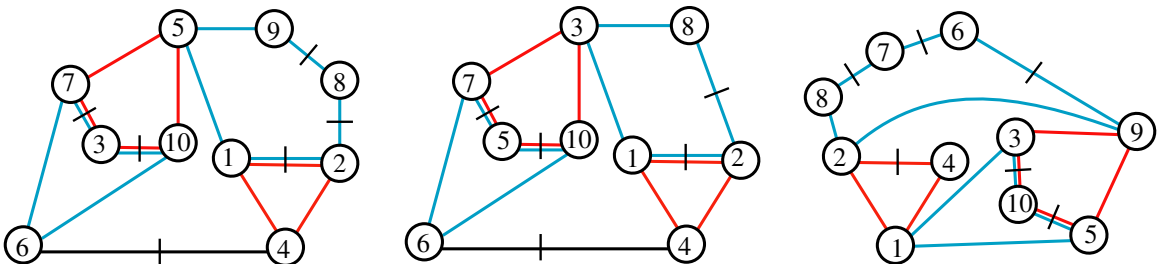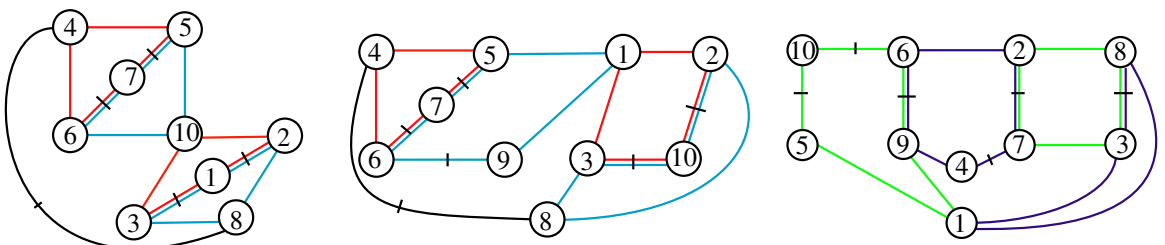

Figure 14: Diagrams for subcases. Left: (A2)-(C2). Center: (A2)-(C3). Right: (A2)-(C4).
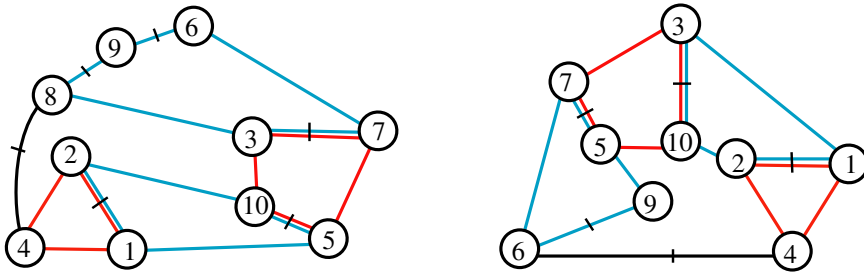
Figure 15: Diagrams for subcases. Left: (A2)-(C5). Right: (A2)-(C6).

Subcase (A2)-(C5) Assume $(6, 7, 3, 8, 9) \cup (1, 2, 10, 5)$. See Figure 15, left.

Subcase (A2)-(C6) Assume $(6, 7, 3, 8, 9) \cup (1, 2, 10, 5)$. See Figure 15, right.

Case (A3) Assume $(4, 6, 7, 5) \cup (2, 1, 3, 10)$ is a nontrivial link of A. We identify a nontrivial link in D and show the existence of a double-linked D_4 for all cases except one. We then identify a nontrivial link in F and show the existence of a double-linked D_4 for all cases except one. If both exceptional cases occur at the same time, the existence of a double-linked D_4 is shown.

We note that if the edge $(6, 7)$ is contracted in the graph D, a G_7 graph is obtained. Based on the symmetries of G, D has four different types of pairs of cycles. Since the (A3) link of A contains vertex 1 but does not contain vertex 8, vertices 1 and 8 need also be distinguished within the linked cycles of D. We match the link in (A3) with each link type of D:

$$(D1) \quad (7, 2, 4, 3) \cup (1, 8, 9), \qquad (D4) \quad (7, 2, 1, 9, 6) \cup (4, 3, 8),$$
$$(D2) \quad (7, 2, 1, 3) \cup (4, 8, 9), \qquad (D5) \quad (7, 2, 8, 9, 6) \cup (4, 3, 1),$$
$$(D3) \quad (7, 2, 8, 3) \cup (4, 1, 9), \qquad (D6)* \quad (7, 2, 4, 9, 6) \cup (1, 3, 8).$$

Subcase (A3)-(D1) Assume $(7, 2, 4, 3) \cup (1, 8, 9)$. Then (i) $(7, 6, 4, 2) \cup (1, 8, 9)$ or (ii) $(7, 6, 4, 3) \cup (1, 8, 9)$. See Figure 16.
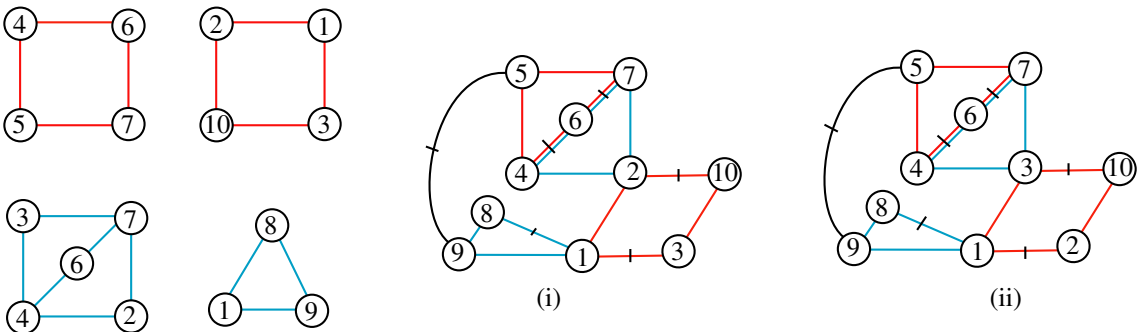
Figure 16: Diagrams for the subcase (A3)-(D1).
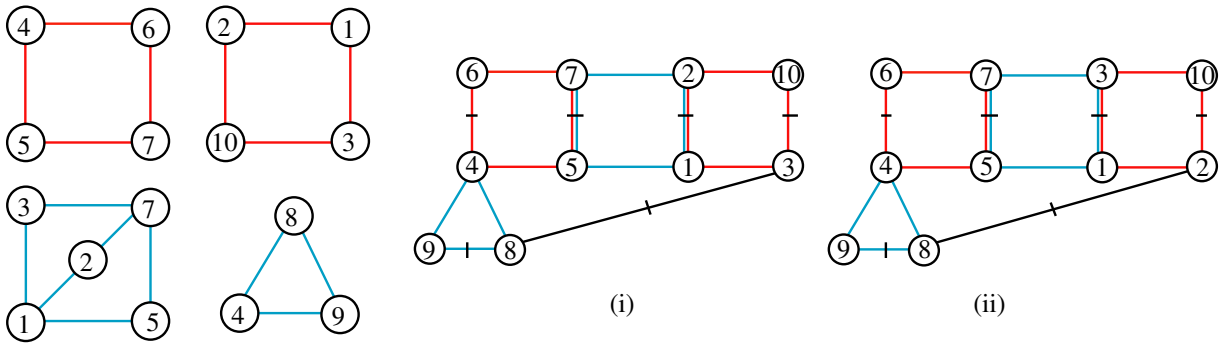
Figure 17: Diagrams for the subcase (A3)-(D2).
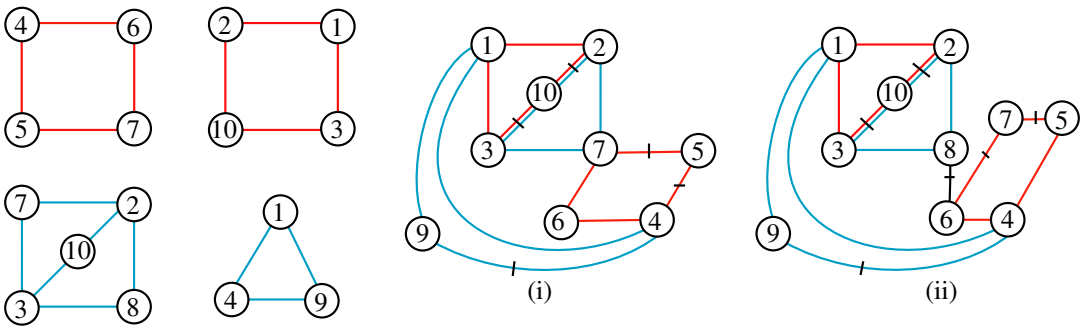
Figure 18: Diagrams for the subcase (A3)-(D3).

Subcase (A3)-(D2) Assume $(7, 2, 1, 3) \cup (4, 8, 9)$. Then (i) $(7, 2, 1, 5) \cup (4, 8, 9)$ or (ii) $(7, 3, 1, 5) \cup (4, 8, 9)$. See Figure 17.

Subcase (A3)-(D3) Assume $(7, 2, 8, 3) \cup (4, 1, 9)$. Then (i) $(7, 2, 10, 3) \cup (4, 1, 9)$ or (ii) $(8, 2, 10, 3) \cup (4, 1, 9)$. See Figure 18.

Subcase (A3)-(D4) Assume $(7, 2, 1, 9, 6) \cup (4, 3, 8)$. See Figure 19, left.

Subcase (A3)-(D5) Assume $(7, 2, 8, 9, 6) \cup (4, 3, 1)$. See Figure 19, right.

If none of the five D-subcases above occurs, then there exists a nontrivial link (D6) $(7, 2, 4, 9, 6) \cup (1, 3, 8)$.
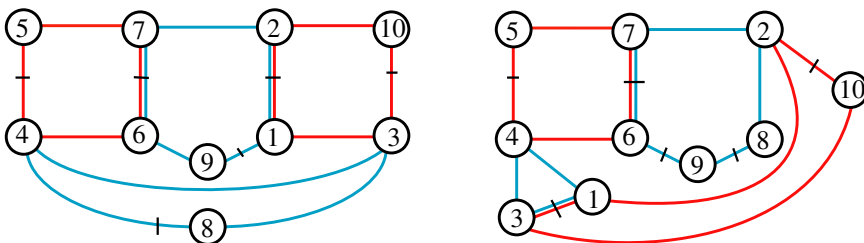
Figure 19: Diagrams for subcases. Left: (A3)-(D4). Right: (A3)-(D5).
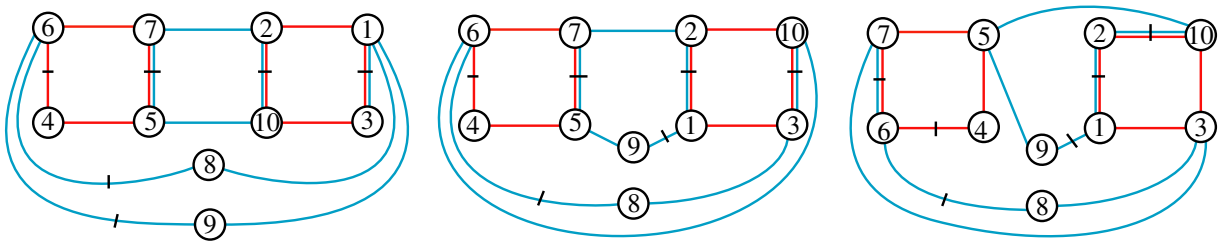
Figure 20: Diagrams for subcases. Left: (A3)-(F1). Center: (A3)-(F2). Right: (A3)-(F3).

We now match the link in (A3) with each link type of F:

(F1)	$(5, 7, 2, 10) \cup (3, 1, 9, 6, 8)$,	(F4)	$(5, 7, 6, 10) \cup (2, 1, 3, 8)$,
(F2)	$(5, 7, 2, 1) \cup (3, 10, 6, 8)$,	(F5)	$(5, 7, 6, 9, 1) \cup (2, 10, 3, 8)$,
(F3)	$(5, 10, 2, 19) \cup (3, 7, 6, 8)$,	(F6)*	$(5, 10, 6, 9, 1) \cup (2, 7, 3, 8)$.

Subcase (A3)-(F1) Assume $(5, 7, 2, 10) \cup (3, 1, 9, 6, 8)$. See Figure 20, left.

Subcase (A3)-(F2) Assume $(5, 7, 2, 1) \cup (3, 10, 6, 8)$. See Figure 20, center.

Subcase (A3)-(F3) Assume $(5, 10, 2, 19) \cup (3, 7, 6, 8)$. See Figure 20, right.

Subcase (A3)-(F4) Assume $(5, 7, 6, 10) \cup (2, 1, 3, 8)$. See Figure 21, left.

Subcase (A3)-(F5) Assume $(5, 7, 6, 9, 1) \cup (2, 10, 3, 8)$. See Figure 21, center.

If none of the five F-subcases solved above occurs, then we have (F6) $(5, 10, 6, 9, 1) \cup (2, 7, 3, 8)$. This coupled with the remaining (D6) subcase gives:

Subcase (D6)-(F6) Assume $(7, 2, 4, 9, 6) \cup (1, 3, 8)$ and $(5, 10, 6, 9, 1) \cup (2, 7, 3, 8)$. See Figure 21, right.

Case (A4) Assume $(4, 6, 7, 2) \cup (3, 1, 5, 10)$ is a nontrivial link. We look at possible nontrivial links in the graph B. Based on the symmetries of $G_{10,26}$, B has four different types of pairs of cycles. Since vertices 2 and 3 and vertices 7 and 10, respectively, are distinguished in the link A4, they need to be distinguished within the cycles of B. We match the link in (A4) with each link in B. There is one
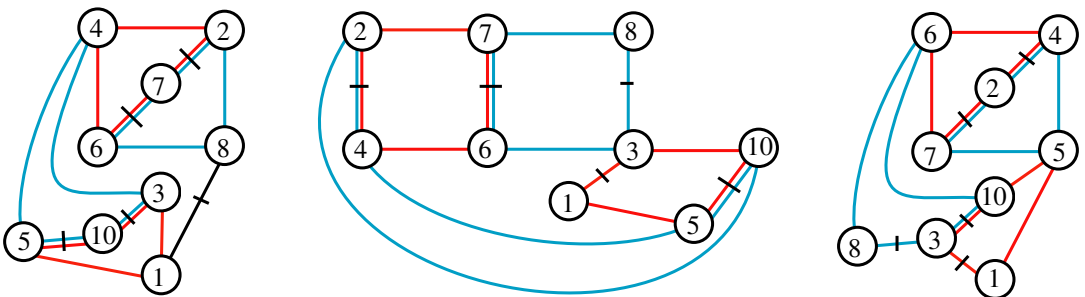

Figure 21: Diagrams for subcases. Left: (A3)-(F4). Center: (A3)-(F5). Right: (D6)-(F6).
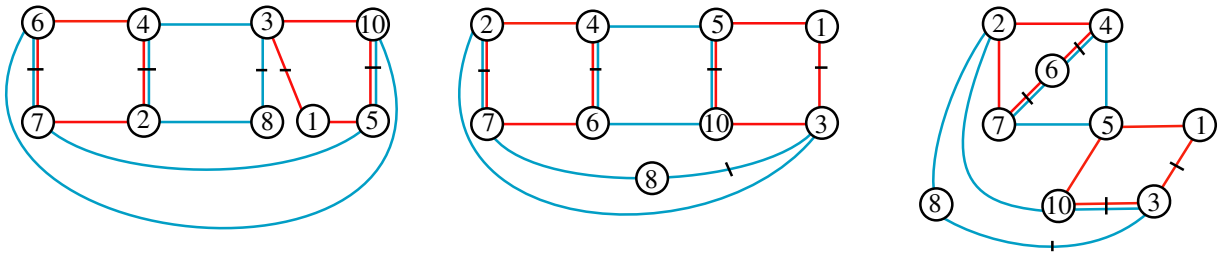
Figure 22: Diagrams for subcases. Left: (A4)-(B1). Center: (A4)-(B2). Right: (A4)-(B3).

exceptional case which cannot be solved this way. Then we look at possible nontrivial links in the graph E and we match the link in (A4) with each link in E. There are two exceptional cases which cannot be solved this way. We match the two pairs of exceptional cases to complete the proof.

(B1)	$(8, 2, 4, 3) \cup (7, 5, 10, 6),$	(B6)	$(8, 3, 10, 6) \cup (4, 5, 7, 2),$
(B2)	$(8, 2, 7, 3) \cup (4, 5, 10, 6),$	(B7)$*$	$(8, 2, 10, 6) \cup (4, 5, 7, 3),$
(B3)	$(8, 2, 10, 3) \cup (4, 5, 7, 6),$	(B8)	$(8, 2, 4, 6) \cup (7, 5, 10, 3),$
(B4)	$(8, 2, 7, 6) \cup (4, 5, 10, 3),$	(B9)	$(8, 3, 4, 6) \cup (7, 5, 10, 2),$
(B5)	$(8, 3, 7, 6) \cup (4, 5, 10, 2).$		

Subcase (A4)-(B1) Assume $(8, 2, 4, 3) \cup (7, 5, 10, 6)$. See Figure 22, left.

Subcase (A4)-(B2) Assume $(8, 2, 7, 3) \cup (4, 5, 10, 6)$. See Figure 22, center.

Subcase (A4)-(B3) Assume $(8, 2, 10, 3) \cup (4, 5, 7, 6)$. See Figure 22, right.

Subcase (A4)-(B4) Assume $(8, 2, 7, 6) \cup (4, 5, 10, 3)$. See Figure 23, left.

Subcase (A4)-(B5) Assume $(8, 3, 7, 6) \cup (4, 5, 10, 2)$. See Figure 23, center.

Subcase (A4)-(B6) Assume $(8, 3, 10, 6) \cup (4, 5, 7, 2)$. See Figure 23, right.

Subcase (A4)-(B8) Assume $(8, 2, 4, 6) \cup (7, 5, 10, 3)$. See Figure 24, left.

Subcase (A4)-(B9) Assume $(8, 3, 4, 6) \cup (7, 5, 10, 2)$. See Figure 24, center.

Figure 23: Diagrams for subcases. Left: (A4)-(B4). Center: (A4)-(B5). Right: (A4)-(B6).
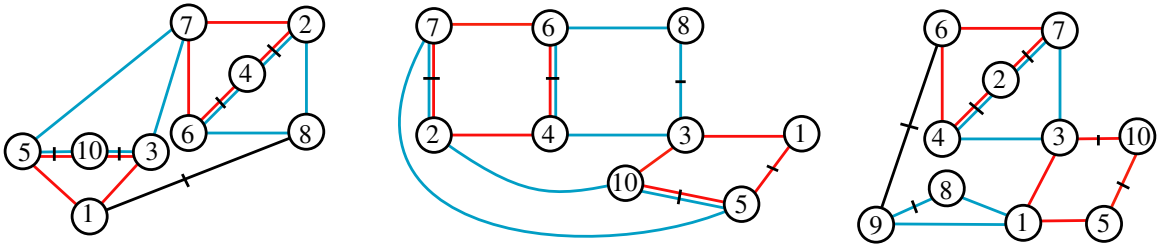
Figure 24: Diagrams for subcases. Left: (A4)-(B8). Center: (A4)-(B9). Right: (A4)-(E1).

We look at possible nontrivial links in the graph E and we match the link in (A4) with each link in E:

(E1)	$(7, 2, 4, 3) \cup (1, 9, 8)$,	(E6)	$(7, 2, 1, 9, 5) \cup (4, 3, 8)$,
(E2)	$(7, 2, 1, 3) \cup (4, 9, 8)$,	(E7)	$(7, 2, 8, 9, 5) \cup (4, 3, 1)$,
(E3)	$(7, 2, 8, 3) \cup (4, 9, 1)$,	(E8)∗	$(7, 3, 8, 9, 5) \cup (4, 2, 1)$,
(E4)	$(7, 2, 4, 9, 5) \cup (3, 1, 8)$,	(E9)∗	$(7, 3, 1, 9, 5) \cup (4, 2, 8)$,
(E5)	$(7, 3, 4, 9, 5) \cup (2, 1, 8)$.		

Subcase (A4)-(E1) Assume $(7, 2, 4, 3) \cup (1, 9, 8)$. See Figure 24, right.

Subcase (A4)-(E2) Assume $(7, 2, 1, 3) \cup (4, 9, 8)$. See Figure 25, left.

Subcase (A4)-(E3) Assume $(7, 2, 8, 3) \cup (4, 9, 1)$. See Figure 25, center.

Subcase (A4)-(E4) Assume $(7, 2, 4, 9, 5) \cup (3, 1, 8)$. See Figure 25, right.

Subcase (A4)-(E5) Assume $(7, 3, 4, 9, 5) \cup (2, 1, 8)$. Then (i) $(7, 5, 10, 3) \cup (2, 1, 8)$ or (ii) $(5, 10, 3, 4, 9) \cup (2, 1, 8)$. See Figure 26.

Subcase (A4)-(E6) Assume $(7, 2, 1, 9, 5) \cup (4, 3, 8)$. Then (i) $(5, 7, 6, 9) \cup (4, 3, 8)$ or (ii) $(7, 6, 9, 1, 2) \cup (4, 3, 8)$. See Figure 27.

Subcase (A4)-(E7) Assume $(7, 2, 8, 9, 5) \cup (4, 3, 1)$. See Figure 28, left.

Subcase (B7)-(E8) Assume $(8, 2, 10, 6) \cup (4, 5, 7, 3)$ and $(7, 3, 8, 9, 5) \cup (4, 2, 1)$. See Figure 28, center.

Subcase (B7)-(E9) Assume $(8, 2, 10, 6) \cup (4, 5, 7, 3)$ and $(7, 3, 1, 9, 5) \cup (4, 2, 8)$. See Figure 28, right. □
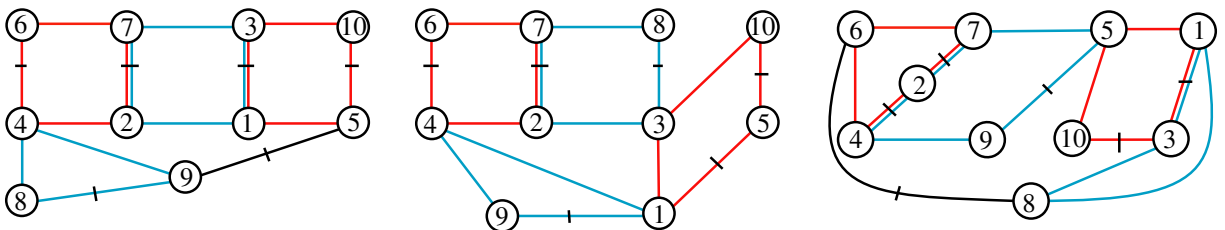
Figure 25: Diagrams for subcases. Left: (A4)-(E2). Center: (A4)-(E3). Right: (A4)-(E4).
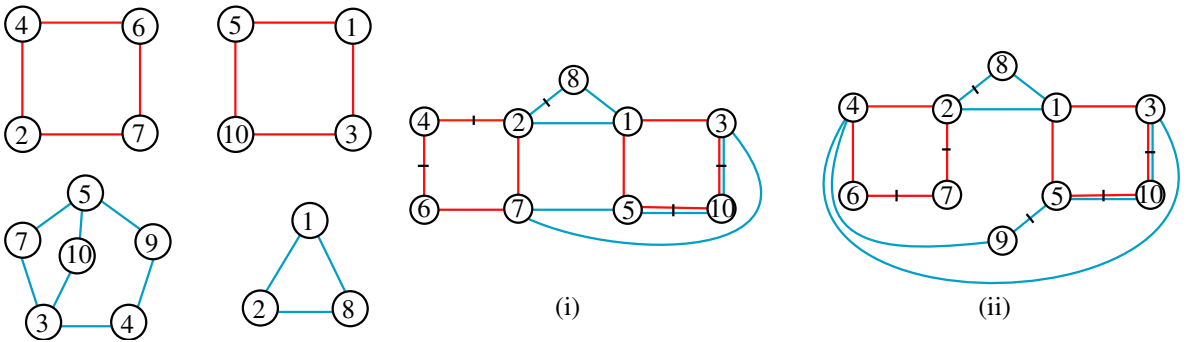
Figure 26: Diagrams for the subcase (A4)-(E5).
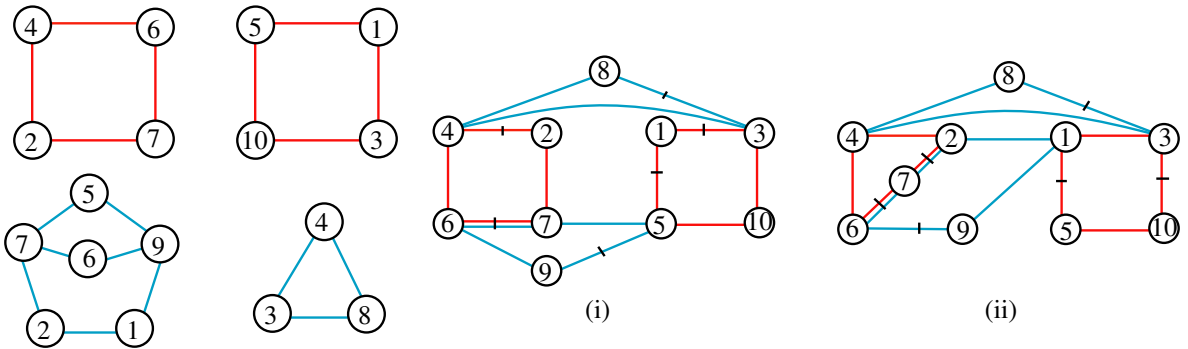
Figure 27: Diagrams for the subcase (A4)-(E6).
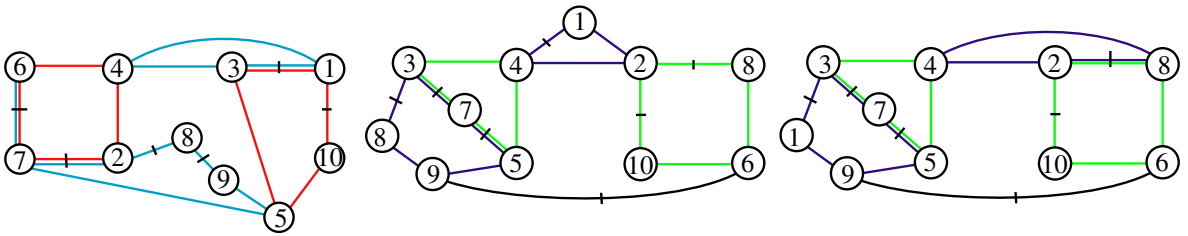
Figure 28: Diagrams for subcases. Left: (A4)-(E7). Center: (B7)-(E8). Right: (B7)-(E9).

4 $G_{10,26}$ is MMIK

In this section we prove $G_{10,26}$ is MMIK by showing that each of its simple minors is nIK. The graph $G_{10,26}$ has ten vertices, labeled $1, 2, \ldots, 10$. Due to the symmetries of the graph, the vertices can be partitioned into six equivalence classes: $\{1, 8\}$, $\{2, 3\}$, $\{4\}$, $\{5, 6\}$, $\{7, 10\}$, and $\{9\}$. Up to symmetry, $G_{10,26}$ has eleven types of edges. Representatives for each possible type of edge are listed in the first column of Table 1. For each such edge type, we constructed two graphs, one by deleting the edge and one by contracting the edge. The graph obtained by deleting the edge is 2–apex, since the removal of the

edge	deletion	contraction
$(1, 2)$	$4, 7$	$1, 3$
$(1, 4)$	$2, 6$	$1, 7$
$(1, 5)$	$2, 3$	$1, 2$
$(1, 8)$	$2, 3$	$1, 4$
$(1, 9)$	$2, 5$	$2, 3$
$(2, 4)$	$5, 6$	$2, 3$
$(2, 7)$	$3, 4$	$2, 4$
$(4, 5)$	$*$	$2, 4$
$(4, 9)$	$2, 3$	$4, 7$
$(5, 7)$	$2, 4$	$2, 4$
$(5, 9)$	$2, 6$	$2, 5$

Table 1: The graph obtained by deleting the edge in the first column becomes planar when deleting the two vertices in the second column. The graph obtained by contracting the edge in the first column becomes planar when deleting the two vertices in the second column.

two vertices listed in the second column gives a planar graph. There is one exception: the graph obtained by deleting the edge $(4, 5)$ is not 2–apex. This graph is shown to be nIK in the next paragraph. The graph obtained by contracting the edge listed in the first column is 2–apex, since the removal of the two vertices listed in the third column gives a planar graph. When contracting an edge e, the new vertex inherits the smaller label among the endpoints of e, and all vertices not incident to e maintain their labels.

The graph G' obtained from $G_{10,26}$ by deleting the edge $(4, 5)$ is not 2–apex. We show it is nIK. Denote by G'' the graph obtained from G' through a ∇Y–move on the triangle $(1, 5, 9)$. Call the new vertex 11;
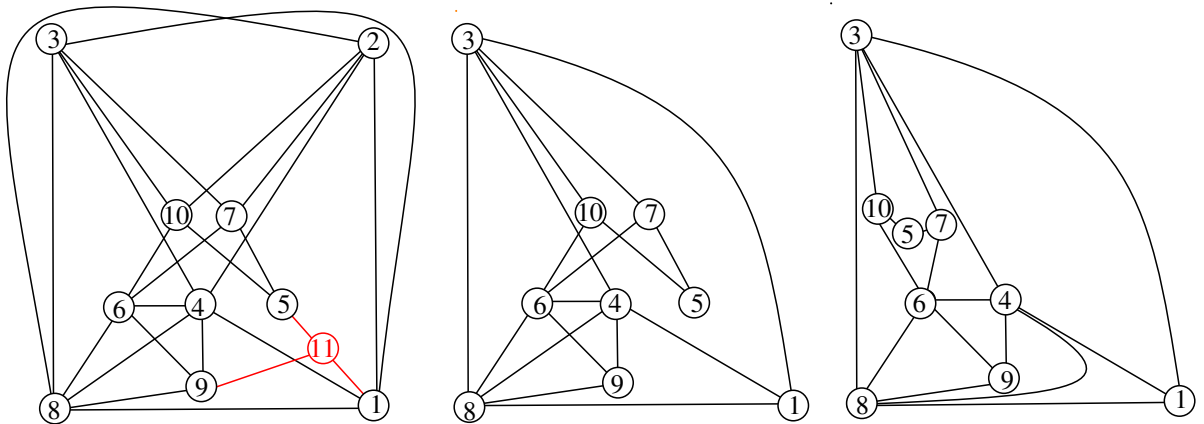
Figure 29: Left: the graph G'' obtained from $G_{10,26}$ by removing the edge $(4, 5)$ followed by a ∇Y–move on the triangle $(1, 5, 9)$. Center: the graph G''' obtained from G'' by deleting vertices 2 and 11. Right: the planar embedding of G'''.

see Figure 29. Delete vertices 2 and 11 of G'' to obtain a planar graph. This proves G'' is 2–apex, and thus nIK. Sachs [13] showed that the ∇Y–move preserves intrinsic linking. Essentially the same argument shows that the ∇Y–move also preserves intrinsic knotting. So the graph G' is nIK.

5 $\mu = 5$ IK graphs

In this section we describe what is known about graphs G with Colin de Verdière invariant 5. We begin with some basic observations. Let $K_1 * G$ denote the graph obtained by coning a vertex over G, ie we add a vertex a to G along with edge av for every $v \in V(G)$.

Lemma 5.1 [7] *Let G be a graph.*

  (1) *If G has at least one edge, then $\mu(K_1 * G) = \mu(G) + 1$.*
  (2) *If G' is a minor of G, then $\mu(G') \leq \mu(G)$.*

Lemma 5.2 [2; 7]   (1) $\mu(G) \leq 3$ *if and only if G is planar.*

  (2) $\mu(G) \leq 4$ *if and only if G is nIL.*

Lemma 5.3 [7] *If $\mu(G) \geq 4$ and a ∇Y move on G produces G', then $\mu(G) = \mu(G')$.*

For $v \in V(G)$, let $G - v$ denote the graph that results after deleting v and all its edges.

Lemma 5.4 *If G is n–apex for $n \geq 0$, then $\mu(G) \leq n + 3$.*

Proof We use induction on n. If $n = 0$, the result follows from Lemma 5.2. Suppose G is $(n+1)$–apex and $v \in V$ is such that $G - v$ is n–apex. Then G is a subgraph of $K_1 * (G - v)$, and, by Lemma 5.1, $\mu(G) \leq \mu(G - v) + 1 \leq (n + 1) + 3$. □

Lemma 5.5 *If G is IK and there is a vertex v such that $G - v$ is nIL, then $\mu(G) = 5$.*

Proof Robertson, Seymour, and Thomas [12] established that G being IK implies G is IL. By Lemma 5.2, $\mu(G) \geq 5$ and $\mu(G - v) \leq 4$. Since G is a subgraph of $K_1 * (G - v)$, using Lemma 5.1, $\mu(G) \leq 5$. □

For $e \in E(G)$, let $G - e$ denote the edge deletion minor and G/e the edge contraction minor of G.

Lemma 5.6 *If G is IK and has a nIL simple minor, then $\mu(G) = 5$.*

Proof The proof is similar to that of the previous lemma. In particular $\mu(G) \geq 5$. By definition, there is an edge e such that $G - e$ or G/e is nIL. Suppose first that $G - e$ is nIL. By Lemma 5.2, $\mu(G - e) \leq 4$. We can form a graph G' homeomorphic to G by adding a degree-two vertex between a and b, the vertices of e. Then G' is a subgraph of $K_1 * (G - e)$, and, using Lemma 5.1, $\mu(G') \leq 5$. Since G is a minor of G', by Lemma 5.1, $\mu(G) \leq 5$.

Next, suppose G/e is nIL, so that $\mu(G/e) \leq 4$. We can again recognize G as a subgraph of $K_1 * (G/e)$, which implies $\mu(G) \leq 5$. □
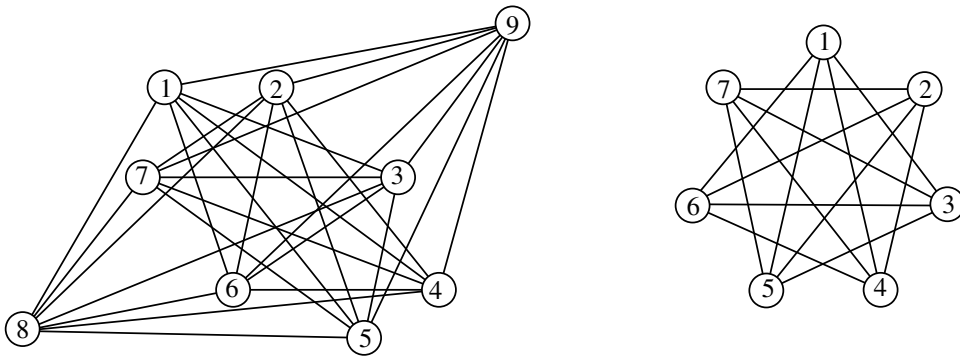
Figure 30: Left: the graph $G_{9,28}$. Right: the complement of a 7–cycle.

We remark that many of the known MMIK graphs have $\mu = 6$. In [3], the authors provide a listing of 264 MMIK graphs, of which 105 are in the families of K_7, $K_{3,3,1,1}$, and $E_9 + e$. We will now verify that each of these three graphs has $\mu = 6$. By Lemma 5.3, all 105 graphs have μ–invariant 6. As shown in [2], $\mu(K_n) = n - 1$ when $n > 1$, so $\mu(K_7) = 6$. The graph $K_{3,3,1,1}$ is $K_1 * K_{3,3,1}$. Since $K_{3,3,1}$ is an obstruction for intrinsic linking [12], by Lemma 5.2, $\mu(K_{3,3,1,1}) = \mu(K_{3,3,1}) + 1 \geq 6$. On the other hand, $K_{3,3,1,1}$ is 3–apex, which, by Lemma 5.4, shows $\mu(K_{3,3,1,1}) \leq 6$. Since E_9 is in the K_7 family, by Lemma 5.3, $\mu(E_9) = \mu(K_7) = 6$. By Lemma 5.1, $\mu(E_9 + e) \geq \mu(E_9) = 6$. On the other hand, $E_9 + e$ is 3–apex, so, by Lemma 5.4, $\mu(E_9 + e) \leq 6$. By Lemma 5.3, all 110 graphs in the $E_9 + e$ family have $\mu = 6$ (not just the 33 that are MMIK). Note that these 110 graphs are all IK [6].

In contrast, here we have introduced several new examples of IK graphs with $\mu = 5$. Such examples were known previously. For example, Foisy [5] provided an example of an MMIK graph F that becomes nIL on deletion of a vertex. By Lemma 5.5, $\mu(F) = 5$. By Lemma 5.6, $\mu(G_{11,35}) = 5$ as it is IK with both a nIL edge deletion minor as well a nIL edge contraction minor. Similarly, $\mu(G_{10,30}) = 5$ since it is IK with a nIL edge contraction minor. Finally, we argue that $\mu(G_{10,26}) = 5$. Since $G_{10,26}$ is a minor of $G_{11,35}$, we have $\mu(G_{10,26}) \leq \mu(G_{11,35}) = 5$. On the other hand, as we proved in Section 3, $G_{10,26}$ is IK, hence IL [12], and $\mu(G_{10,26}) \geq 5$ by Lemma 5.2. By Lemma 5.3, graphs in the families of $G_{10,26}$, $G_{10,30}$, and $G_{11,35}$ also have $\mu = 5$. Using computers, the $G_{10,26}$ family alone provides more than 600 new examples of IK graphs with Colin de Verdière invariant 5.

Proof of Proposition 2.1   Assume there exists an IK graph G of order less than 10 which admits a nIL edge contraction minor. As such, by Lemma 5.6, $\mu(G) = 5$. Since μ is minor monotone (Lemma 5.1), any MMIK minor of G must have $\mu = 5$. By work of Goldberg, Mattman, and Naimi [6], and Mattman, Morris, and Ryker [8], the MMIK graphs of order at most 9 are known. With the exception of $G_{9,28}$, depicted in Figure 30, left, all the others are either in the K_7 family, the $K_{3,3,1,1}$ family, or the E_9+e family, and thus have $\mu = 6$. It follows that G must have order 9 and that $G_{9,28}$ is a subgraph of G. If contracting an edge e of G produces a nIL minor, then deleting either endpoint of e must also produce

a nIL minor (subgraph). Since $G_{9,28}$ is a subgraph of G, deleting the same vertex must produce a nIL subgraph of $G_{9,28}$. The graph $G_{9,28}$ is highly symmetric, having a rich automorphism group, and it is structured as two nonadjacent cones over the complement of a 7–cycle (the graph depicted in Figure 30, right). Up to isomorphism, there are only two induced subgraphs of order 8 inside $G_{9,28}$: the graph obtained by deleting the vertex labeled 9, and the graph obtained by deleting the vertex labeled 7. Neither of these are nIL, since they both have a K_6 minor. For the first graph, contracting the edges $(4, 7)$ and $(2, 6)$ produces a complete minor on the 6 vertices. For the second graph, contracting the edges $(4, 9)$ and $(2, 6)$ also produces a complete minor on the 6 vertices. □

6 Erratum

In this section we discuss an error in the proof of [10, Proposition 2]. The proposition asserts that if a graph G has a paneled embedding, and an edge is added to G between two vertices a and b that have a common adjacent vertex v, then $G + ab$ has a knotless embedding.

In the proof of Proposition 2, it is first shown that one can assume there is a path $P_{ab} \subset G$ from a to b disjoint from v. Next, the proof claims that, in any paneled embedding Γ of G, if D is a panel for the cycle $P_{ab} \cup av \cup vb$ in Γ, then embedding the new edge ab in D yields a knotless embedding Γ' of $G + ab$. Figure 31 shows a counterexample to this claim, and will be used to explain where the error in the proof of Proposition 2 lies.

It is not difficult to see that in Figure 31, left, every cycle in Γ is paneled. In particular, the cycle $acdbva$ bounds a panel D such that vc and vd lie below and above D, respectively, in the figure. If we embed the edge ab in D as in Figure 31, right, we see that the cycle $abcvda$ is a trefoil, and hence Γ' isn't a knotless embedding as claimed.

The error is specifically in the last few sentences of the penultimate paragraph in the proof, where it mentions a type 1 Reidemeister move on $P_1 \cup \{e\}$. The proof overlooks the possibility that P_{bv} may prevent this Reidemeister move, as is the case in Figure 31, right (for reference, the paths $acdb$, adv, and bcv in Figure 31 represent the paths P_{ab}, P_{av}, and P_{bv}, respectively, in the proof of Proposition 2).
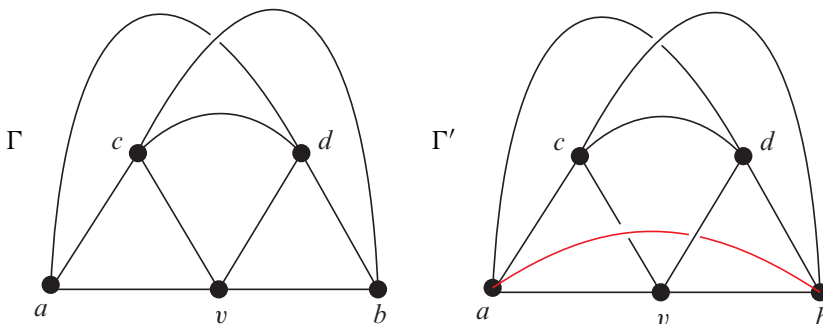

Figure 31: Left: every cycle in Γ is paneled. Right: Γ' contains a trefoil.

Appendix

We give edge lists for the graphs $G_{11,35}$, $G_{10,30}$, and $G_{10,26}$:

$E(G_{11,35})$

$= \{(1,2), (1,3), (1,4), (1,5), (1,8), (1,9), (2,3), (2,4), (2,8), (3,4), (3,5), (3,6),$
$(3,7), (3,8), (3,10), (3,11), (4,5), (4,6), (4,8), (4,9), (4,10), (5,6), (5,7), (5,9),$
$(5,10), (5,11), (6,7), (6,8), (6,9), (6,10), (6,11), (7,11), (8,9), (10,11), (2,11)\}$

$E(G_{10,30})$

$= \{(1,5), (1,7), (1,8), (1,9), (1,10), (2,3), (2,4), (2,5), (2,6), (2,7), (2,10), (3,4), (3,6), (3,8), (3,9),$
$(3,10), (4,6), (4,8), (4,9), (5,6), (5,7), (5,8), (5,10), (6,7), (6,8), (6,9), (7,9), (7,10), (8,10), (9,10)\}$

$E(G_{10,26})$

$= \{(1,2), (1,3), (1,4), (1,5), (1,8), (1,9), (2,4), (2,7), (2,8), (2,10), (3,4), (3,7), (3,8),$
$(3,10), (4,5), (4,6), (4,8), (4,9), (5,7), (5,9), (5,10), (6,7), (6,8), (6,9), (6,10), (8,9)\}$

References

[1] **C C Adams**, *The knot book*: *an elementary introduction to the mathematical theory of knots*, Freeman, New York (1994) MR Zbl

[2] **Y Colin de Verdière**, *Sur un nouvel invariant des graphes et un critère de planarité*, J. Combin. Theory Ser. B 50 (1990) 11–21 MR Zbl

[3] **E Flapan**, **T W Mattman**, **B Mellor**, **R Naimi**, **R Nikkuni**, *Recent developments in spatial graph theory*, from "Knots, links, spatial graphs, and algebraic invariants" (E Flapan, A Henrich, A Kaestner, S Nelson, editors), Contemp. Math. 689, Amer. Math. Soc., Providence, RI (2017) 81–102 MR Zbl

[4] **J Foisy**, *Intrinsically knotted graphs*, J. Graph Theory 39 (2002) 178–187 MR Zbl

[5] **J Foisy**, *A newly recognized intrinsically knotted graph*, J. Graph Theory 43 (2003) 199–209 MR Zbl

[6] **N Goldberg**, **T W Mattman**, **R Naimi**, *Many, many more intrinsically knotted graphs*, Algebr. Geom. Topol. 14 (2014) 1801–1823 MR Zbl

[7] **H van der Holst**, **L Lovász**, **A Schrijver**, *The Colin de Verdière graph parameter*, from "Graph theory and combinatorial biology" (L Lovász, A Gyárfás, G Katona, A Recski, L Székely, editors), Bolyai Soc. Math. Stud. 7, Bolyai Math. Soc., Budapest (1999) 29–85 MR Zbl

[8] **T W Mattman**, **C Morris**, **J Ryker**, *Order nine MMIK graphs*, from "Knots, links, spatial graphs, and algebraic invariants" (E Flapan, A Henrich, A Kaestner, S Nelson, editors), Contemp. Math. 689, Amer. Math. Soc., Providence, RI (2017) 103–124 MR Zbl

[9] **R Naimi**, **A Pavelescu**, **E Pavelescu**, *New bounds on maximal linkless graphs*, Algebr. Geom. Topol. 23 (2023) 2545–2559 MR Zbl

[10] **R Naimi**, **E Pavelescu**, **H Schwartz**, *Deleting an edge of a 3–cycle in an intrinsically knotted graph gives an intrinsically linked graph*, J. Knot Theory Ramifications 23 (2014) art. id. 1450075 MR Zbl

[11] **R Odeneal**, **R Naimi**, **A Pavelescu**, **E Pavelescu**, *The complement problem for linklessly embeddable graphs*, J. Knot Theory Ramifications 31 (2022) art. id. 2250075 MR Zbl

[12] **N Robertson**, **P D Seymour**, **R Thomas**, *Linkless embeddings of graphs in 3–space*, Bull. Amer. Math. Soc. 28 (1993) 84–89 MR Zbl

[13] **H Sachs**, *On spatial representations of finite graphs*, from "Finite and infinite sets, II" (A Hajnal, L Lovász, V T Sós, editors), Colloq. Math. Soc. János Bolyai 37, North-Holland, Amsterdam (1984) 649–662 MR Zbl

[14] **K Taniyama**, **A Yasuhara**, *Realization of knots and links in a spatial graph*, Topology Appl. 112 (2001) 87–109 MR Zbl

Department of Mathematics and Statistics, California State University at Chico
Chico, CA, United States

Department of Mathematics, Occidental College
Los Angeles, CA, United States

Mathematics and Statistics Department, University of South Alabama
Mobile, AL, United States

Mathematics and Statistics Department, University of South Alabama
Mobile, AL, United States

tmattman@csuchico.edu,   rnaimi@oxy.edu,   andreipavelescu@southalabama.edu,
elenapavelescu@southalabama.edu

http://tmattman.yourweb.csuchico.edu,   http://faculty.oxy.edu/rnaimi/,
http://apaveles.wixsite.com/scientist-site

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.